The Book of Statistical Proofs

DOI: 10.5281/zenodo.4305949

https://statproofbook.github.io/

StatProofBook@gmail.com

2024-07-18, 10:50

Contents

Chapter I

General Theorems

1 Probability theory

1.1 Random experiments

1.1.1 Random experiment

Definition: A random experiment is any repeatable procedure that results in one (→ I/1.2.2) out of a well-defined set of possible outcomes.

- The set of possible outcomes is called sample space (→ I/1.1.2).
- A set of zero or more outcomes is called a random event (→ I/1.2.1).
- A function that maps from events to probabilities is called a probability function (→ I/1.5.1).

Together, sample space (→ I/1.1.2), event space (→ I/1.1.3) and probability function (→ I/1.1.4) characterize a random experiment.

Sources:

- Wikipedia (2020): "Experiment (probability theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: https://en.wikipedia.org/wiki/Experiment_(probability_theory).

1.1.2 Sample space

Definition: Given a random experiment (→ I/1.1.1), the set of all possible outcomes from this experiment is called the sample space of the experiment. A sample space is usually denoted as Ω and specified using set notation.

Sources:

- Wikipedia (2021): "Sample space"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: https://en.wikipedia.org/wiki/Sample_space.

1.1.3 Event space

Definition: Given a random experiment (→ I/1.1.1), an event space $\mathcal{E}$ is any set of events, where an event (→ I/1.2.1) is any set of zero or more elements from the sample space (→ I/1.1.2) Ω of this experiment.

Sources:

- Wikipedia (2021): "Event (probability theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: https://en.wikipedia.org/wiki/Event_(probability_theory).

1.1.4 Probability space

Definition: Given a random experiment (→ I/1.1.1), a probability space $(\Omega, \mathcal{E}, P)$ is a triple consisting of

- the sample space (→ I/1.1.2) Ω, i.e. the set of all possible outcomes from this experiment;
- an event space (→ I/1.1.3) $\mathcal{E} \subseteq 2^\Omega$, i.e. a set of subsets from the sample space, called events (→ I/1.2.1);
- a probability measure $P: \; \mathcal{E} \rightarrow [0,1]$, i.e. a function mapping from the event space (→ I/1.1.3) to the real numbers, observing the axioms of probability (→ I/1.4.1).

Sources:
- Wikipedia (2021): "Probability space"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: https://en.wikipedia.org/wiki/Probability_space#Definition.

1.2 Random variables

1.2.1 Random event

Definition: A random event E is the outcome of a random experiment ($\rightarrow$ I/1.1.1) which can be described by a statement that is either true or false.
- If the statement is true, the event is said to take place, denoted as E.
- If the statement is false, the complement of E occurs, denoted as $\overline{E}$.

In other words, a random event is a random variable ($\rightarrow$ I/1.2.2) with two possible values (true and false, or 1 and 0). A random experiment ($\rightarrow$ I/1.1.1) with two possible outcomes is called a Bernoulli trial ($\rightarrow$ II/1.2.1).

Sources:
- Wikipedia (2020): "Event (probability theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: https://en.wikipedia.org/wiki/Event_(probability_theory).

1.2.2 Random variable

Definition: A random variable may be understood
- informally, as a real number $X \in \mathbb{R}$ whose value is the outcome of a random experiment ($\rightarrow$ I/1.1.1);
- formally, as a measurable function X defined on a probability space ($\rightarrow$ I/1.1.4) $(\Omega, \mathcal{E}, P)$ that maps from a sample space ($\rightarrow$ I/1.1.2) Ω to the real numbers $\mathbb{R}$ using an event space ($\rightarrow$ I/1.1.3) $\mathcal{E}$ and a probability function ($\rightarrow$ I/1.5.1) P;
- more broadly, as any random quantity X such as a random event ($\rightarrow$ I/1.2.1), a random scalar ($\rightarrow$ I/1.2.2), a random vector ($\rightarrow$ I/1.2.3) or a random matrix ($\rightarrow$ I/1.2.4).

Sources:
- Wikipedia (2020): "Random variable"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Random_variable#Definition.

1.2.3 Random vector

Definition: A random vector, also called "multivariate random variable", is an n-dimensional column vector $X \in \mathbb{R}^{n \times 1}$ whose entries are random variables ($\rightarrow$ I/1.2.2).

Sources:
- Wikipedia (2020): "Multivariate random variable"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Multivariate_random_variable.

1.2.4 Random matrix

Definition: A random matrix, also called "matrix-valued random variable", is an $n \times p$ matrix $X \in \mathbb{R}^{n \times p}$ whose entries are random variables (→ I/1.2.2). Equivalently, a random matrix is an $n \times p$ matrix whose columns are n-dimensional random vectors (→ I/1.2.3).

Sources:
- Wikipedia (2020): "Random matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Random_matrix.

1.2.5 Constant

Definition: A constant is a quantity which does not change and thus always has the same value. From a statistical perspective, a constant is a random variable (→ I/1.2.2) which is equal to its expected value (→ I/1.10.1)

$$X = \mathrm{E}(X) \tag{1}$$

or equivalently, whose variance (→ I/1.11.1) is zero

$$\mathrm{Var}(X) = 0 \; . \tag{2}$$

Sources:
- ProofWiki (2020): "Definition: Constant"; in: *ProofWiki*, retrieved on 2020-09-09; URL: https://proofwiki.org/wiki/Definition:Constant#Definition.

1.2.6 Discrete vs. continuous

Definition: Let X be a random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$. Then,
- X is called a discrete random variable, if $\mathcal{X}$ is either a finite set or a countably infinite set; in this case, X can be described by a probability mass function (→ I/1.6.1);
- X is called a continuous random variable, if $\mathcal{X}$ is an uncountably infinite set; if it is absolutely continuous, X can be described by a probability density function (→ I/1.7.1).

Sources:
- Wikipedia (2020): "Random variable"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-29; URL: https://en.wikipedia.org/wiki/Random_variable#Standard_case.

1.2.7 Univariate vs. multivariate

Definition: Let X be a random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$. Then,
- X is called a two-valued random variable or random event (→ I/1.2.1), if $\mathcal{X}$ has exactly two elements, e.g. $\mathcal{X} = \left\{E, \overline{E}\right\}$ or $\mathcal{X} = \{\mathrm{true}, \mathrm{false}\}$ or $\mathcal{X} = \{1, 0\}$;
- X is called a univariate random variable or random scalar (→ I/1.2.2), if $\mathcal{X}$ is one-dimensional, i.e. (a subset of) the real numbers $\mathbb{R}$;
- X is called a multivariate random variable or random vector (→ I/1.2.3), if $\mathcal{X}$ is multi-dimensional, e.g. (a subset of) the n-dimensional Euclidean space $\mathbb{R}^n$;

- X is called a matrix-valued random variable or random matrix (→ I/1.2.4), if $\mathcal{X}$ is (a subset of) the set of $n \times p$ real matrices $\mathbb{R}^{n \times p}$.

Sources:
- Wikipedia (2020): "Multivariate random variable"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-06; URL: https://en.wikipedia.org/wiki/Multivariate_random_variable.

1.3 Probability

1.3.1 Probability

Definition: Let E be a statement about an arbitrary event such as the outcome of a random experiment (→ I/1.1.1). Then, $p(E)$ is called the probability of E and may be interpreted as
- (objectivist interpretation of probability:) some physical state of affairs, e.g. the relative frequency of occurrence of E, when repeating the experiment ("Frequentist probability"); or
- (subjectivist interpretation of probability:) a degree of belief in E, e.g. the price at which someone would buy or sell a bet that pays 1 unit of utility if E and 0 if not E ("Bayesian probability").

Sources:
- Wikipedia (2020): "Probability"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Probability#Interpretations.

1.3.2 Joint probability

Definition: Let A and B be two arbitrary statements about random variables (→ I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then, $p(A, B)$ is called the joint probability of A and B and is defined as the probability (→ I/1.3.1) that A and B are both true.

Sources:
- Wikipedia (2020): "Joint probability distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Joint_probability_distribution.
- Jason Browlee (2019): "A Gentle Introduction to Joint, Marginal, and Conditional Probability"; in: *Machine Learning Mastery*, retrieved on 2021-08-01; URL: https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/.

1.3.3 Marginal probability

Definition: (law of marginal probability, also called "sum rule") Let A and X be two arbitrary statements about random variables (→ I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Furthermore, assume a joint probability (→ I/1.3.2) distribution $p(A, X)$. Then, $p(A)$ is called the marginal probability of A and,
1) if X is a discrete (→ I/1.2.6) random variable (→ I/1.2.2) with domain $\mathcal{X}$, is given by

$$p(A) = \sum_{x \in \mathcal{X}} p(A, x) \; ; \tag{1}$$

2) if X is a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with domain $\mathcal{X}$, is given by

$$p(A) = \int_{\mathcal{X}} p(A,x) \, \mathrm{d}x \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Marginal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Marginal_distribution#Definition.
- Jason Browlee (2019): "A Gentle Introduction to Joint, Marginal, and Conditional Probability"; in: *Machine Learning Mastery*, retrieved on 2021-08-01; URL: https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/.

1.3.4 Conditional probability

Definition: (law of conditional probability, also called "product rule") Let A and B be two arbitrary statements about random variables (→ I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Furthermore, assume a joint probability (→ I/1.3.2) distribution $p(A,B)$. Then, $p(A|B)$ is called the conditional probability that A is true, given that B is true, and is given by

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{1}$$

where $p(B)$ is the marginal probability (→ I/1.3.3) of B.

Sources:
- Wikipedia (2020): "Conditional probability"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Conditional_probability#Definition.
- Jason Browlee (2019): "A Gentle Introduction to Joint, Marginal, and Conditional Probability"; in: *Machine Learning Mastery*, retrieved on 2021-08-01; URL: https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/.

1.3.5 Exceedance probability

Definition: Let $X = \{X_1, \ldots, X_n\}$ be a set of n random variables (→ I/1.2.2) which the joint probability distribution (→ I/1.5.2) $p(X) = p(X_1, \ldots, X_n)$. Then, the exceedance probability for random variable X_i is the probability (→ I/1.3.1) that X_i is larger than all other random variables $X_j, \; j \neq i$:

$$\begin{split} \varphi(X_i) &= \Pr\left(\forall j \in \{1, \ldots, n | j \neq i\} : \; X_i > X_j\right) \\ &= \Pr\left(\bigwedge_{j \neq i} X_i > X_j\right) \\ &= \Pr\left(X_i = \max(\{X_1, \ldots, X_n\})\right) \\ &= \int_{X_i = \max(X)} p(X) \, \mathrm{d}X \; . \end{split} \tag{1}$$

Sources:

- Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009): "Bayesian model selection for group studies"; in: *NeuroImage*, vol. 46, pp. 1004–1017, eq. 16; URL: https://www.sciencedirect.com/science/article/abs/pii/S1053811909002638; DOI: 10.1016/j.neuroimage.2009.03.025.
- Soch J, Allefeld C (2016): "Exceedance Probabilities for the Dirichlet Distribution"; in: *arXiv stat.AP*, 1611.01439; URL: https://arxiv.org/abs/1611.01439.

1.3.6 Statistical independence

Definition: Generally speaking, random variables (→ I/1.2.2) are statistically independent, if their joint probability (→ I/1.3.2) can be expressed in terms of their marginal probabilities (→ I/1.3.3).

1) A set of discrete random variables (→ I/1.2.2) $X_1, \ldots, X_n$ with possible values $\mathcal{X}_1, \ldots, \mathcal{X}_n$ is called statistically independent, if

$$p(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \quad \text{for all } x_i \in \mathcal{X}_i, \; i = 1, \ldots, n \tag{1}$$

where $p(x_1, \ldots, x_n)$ are the joint probabilities (→ I/1.3.2) of $X_1, \ldots, X_n$ and $p(x_i)$ are the marginal probabilities (→ I/1.3.3) of X_i.

2) A set of continuous random variables (→ I/1.2.2) $X_1, \ldots, X_n$ defined on the domains $\mathcal{X}_1, \ldots, \mathcal{X}_n$ is called statistically independent, if

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i, \; i = 1, \ldots, n \tag{2}$$

or equivalently, if the probability densities (→ I/1.7.1) exist, if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i, \; i = 1, \ldots, n \tag{3}$$

where F are the joint (→ I/1.5.2) or marginal (→ I/1.5.3) cumulative distribution functions (→ I/1.8.1) and f are the respective probability density functions (→ I/1.7.1).

Sources:

- Wikipedia (2020): "Independence (probability theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Independence_(probability_theory)#Definition.

1.3.7 Conditional independence

Definition: Generally speaking, random variables (→ I/1.2.2) are conditionally independent given another random variable, if they are statistically independent (→ I/1.3.6) in their conditional probability distributions (→ I/1.5.4) given this random variable.

1) A set of discrete random variables (→ I/1.2.6) $X_1, \ldots, X_n$ with possible values $\mathcal{X}_1, \ldots, \mathcal{X}_n$ is called conditionally independent given the random variable Y with possible values $\mathcal{Y}$, if

$$p(X_1 = x_1, \ldots, X_n = x_n|Y = y) = \prod_{i=1}^{n} p(X_i = x_i|Y = y) \quad \text{for all } x_i \in \mathcal{X}_i \quad \text{and all } y \in \mathcal{Y} \tag{1}$$

where $p(x_1, \ldots, x_n|y)$ are the joint (conditional) probabilities (→ I/1.3.2) of $X_1, \ldots, X_n$ given Y and $p(x_i|y)$ are the marginal (conditional) probabilities (→ I/1.3.3) of X_i given Y.

2) A set of continuous random variables (→ I/1.2.6) $X_1, \ldots, X_n$ with possible values $\mathcal{X}_1, \ldots, \mathcal{X}_n$ is called conditionally independent given the random variable Y with possible values $\mathcal{Y}$, if

$$F_{X_1,\ldots,X_n|Y=y}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i|Y=y}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i \quad \text{and all } y \in \mathcal{Y} \tag{2}$$

or equivalently, if the probability densities (→ I/1.7.1) exist, if

$$f_{X_1,\ldots,X_n|Y=y}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i|Y=y}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i \quad \text{and all } y \in \mathcal{Y} \tag{3}$$

where F are the joint (conditional) (→ I/1.5.2) or marginal (conditional) (→ I/1.5.3) cumulative distribution functions (→ I/1.8.1) and f are the respective probability density functions (→ I/1.7.1).

Sources:
- Wikipedia (2020): "Conditional independence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: https://en.wikipedia.org/wiki/Conditional_independence#Conditional_independence_of_random_variables.

1.3.8 Probability under independence

Theorem: Let A and B be two statements about random variables (→ I/1.2.2). Then, if A and B are independent (→ I/1.3.6), marginal (→ I/1.3.3) and conditional (→ I/1.3.4) probabilities are equal:

$$\begin{aligned} p(A) &= p(A|B) \\ p(B) &= p(B|A) \; . \end{aligned} \tag{1}$$

Proof: If A and B are independent (→ I/1.3.6), then the joint probability (→ I/1.3.2) is equal to the product of the marginal probabilities (→ I/1.3.3):

$$p(A, B) = p(A) \cdot p(B) \; . \tag{2}$$

The law of conditional probability (→ I/1.3.4) states that

$$p(A|B) = \frac{p(A, B)}{p(B)} \; . \tag{3}$$

Combining (2) and (3), we have:

$$p(A|B) = \frac{p(A) \cdot p(B)}{p(B)} = p(A) \; . \tag{4}$$

Equivalently, we can write:

$$p(B|A) \overset{(3)}{=} \frac{p(A,B)}{p(A)} \overset{(2)}{=} \frac{p(A) \cdot p(B)}{p(A)} = p(B) \; . \tag{5}$$

■

Sources:
- Wikipedia (2021): "Independence (probability theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: https://en.wikipedia.org/wiki/Independence_(probability_theory)#Definition.

1.3.9 Mutual exclusivity

Definition: Generally speaking, random events (→ I/1.2.1) are mutually exclusive, if they cannot occur together, such that their intersection is equal to the empty set (→ I/1.4.3).

More precisely, a set of statements $A_1, \ldots, A_n$ is called mutually exclusive, if

$$p(A_1, \ldots, A_n) = 0 \tag{1}$$

where $p(A_1, \ldots, A_n)$ is the joint probability (→ I/1.3.2) of the statements $A_1, \ldots, A_n$.

Sources:
- Wikipedia (2021): "Mutual exclusivity"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: https://en.wikipedia.org/wiki/Mutual_exclusivity#Probability.

1.3.10 Probability under exclusivity

Theorem: Let A and B be two statements about random variables (→ I/1.2.2). Then, if A and B are mutually exclusive (→ I/1.3.9), the probability (→ I/1.3.1) of their disjunction is equal to the sum of the marginal probabilities (→ I/1.3.3):

$$p(A \vee B) = p(A) + p(B) \; . \tag{1}$$

Proof: If A and B are mutually exclusive (→ I/1.3.9), then their joint probability (→ I/1.3.2) is zero:

$$p(A,B) = 0 \; . \tag{2}$$

The addition law of probability (→ I/1.3.3) states that

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \tag{3}$$

which, in logical rather than set-theoretic expression, becomes

$$p(A \vee B) = p(A) + p(B) - p(A,B) \; . \tag{4}$$

Because the union of mutually exclusive events is the empty set (→ I/1.3.9) and the probability of the empty set is zero (→ I/1.4.3), the joint probability (→ I/1.3.2) term cancels out:

$$p(A \vee B) = p(A) + p(B) - p(A,B) \overset{(2)}{=} p(A) + p(B) \; . \tag{5}$$

■

Sources:
- Wikipedia (2021): "Mutual exclusivity"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: https://en.wikipedia.org/wiki/Mutual_exclusivity#Probability.

1.4 Probability axioms

1.4.1 Axioms of probability

Definition: Let there be a sample space (→ I/1.1.2) Ω, an event space (→ I/1.1.3) $\mathcal{E}$ and a probability measure P, such that $P(E)$ is the probability (→ I/1.3.1) of some event (→ I/1.2.1) $E \in \mathcal{E}$. Then, we introduce three axioms of probability:
- First axiom: The probability of an event is a non-negative real number:

$$P(E) \in \mathbb{R}, \; P(E) \geq 0, \; \text{for all } E \in \mathcal{E} \; . \tag{1}$$

- Second axiom: The probability that at least one elementary event in the sample space will occur is one:

$$P(\Omega) = 1 \; . \tag{2}$$

- Third axiom: The probability of any countable sequence of disjoint (i.e. mutually exclusive (→ I/1.3.9)) events $E_1, E_2, E_3, \ldots$ is equal to the sum of the probabilities of the individual events:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \; . \tag{3}$$

Sources:
- A.N. Kolmogorov (1950): "Elementary Theory of Probability"; in: *Foundations of the Theory of Probability*, p. 2; URL: https://archive.org/details/foundationsofthe00kolm/page/2/mode/2up.
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eqs. 8.2-8.4; URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549.
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: https://en.wikipedia.org/wiki/Probability_axioms#Axioms.

1.4.2 Monotonicity of probability

Theorem: Probability (→ I/1.3.1) is monotonic, i.e. if A is a subset of or equal to B, then the probability of A is smaller than or equal to B:

$$A \subseteq B \quad \Rightarrow \quad P(A) \leq P(B) \; . \tag{1}$$

Proof: Set $E_1 = A$, $E_2 = B \setminus A$ and $E_i = \emptyset$ for $i \geq 3$. Then, the sets E_i are pairwise disjoint and $E_1 \cup E_2 \cup \ldots = B$, because $A \subseteq B$. Thus, from the third axiom of probability (→ I/1.4.1), we have:

$$P(B) = P(A) + P(B \setminus A) + \sum_{i=3}^{\infty} P(E_i) \; . \tag{2}$$

Since, by the first axiom of probability (→ I/1.4.1), the right-hand side is a series of non-negative numbers converging to $P(B)$ on the left-hand side, it follows that

$$P(A) \leq P(B) \; . \tag{3}$$

■

Sources:

- A.N. Kolmogorov (1950): "Elementary Theory of Probability"; in: *Foundations of the Theory of Probability*, p. 6; URL: https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up.
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: https://en.wikipedia.org/wiki/Probability_axioms#Monotonicity.

1.4.3 Probability of the empty set

Theorem: The probability (→ I/1.3.1) of the empty set is zero:

$$P(\emptyset) = 0 \; . \tag{1}$$

Proof: Let A and B be two events fulfilling $A \subseteq B$. Set $E_1 = A$, $E_2 = B \setminus A$ and $E_i = \emptyset$ for $i \geq 3$. Then, the sets E_i are pairwise disjoint and $E_1 \cup E_2 \cup \ldots = B$. Thus, from the third axiom of probability (→ I/1.4.1), we have:

$$P(B) = P(A) + P(B \setminus A) + \sum_{i=3}^{\infty} P(E_i) \; . \tag{2}$$

Assume that the probability of the empty set is not zero, i.e. $P(\emptyset) > 0$. Then, the right-hand side of (2) would be infinite. However, by the first axiom of probability (→ I/1.4.1), the left-hand side must be finite. This is a contradiction. Therefore, $P(\emptyset) = 0$.

■

Sources:

- A.N. Kolmogorov (1950): "Elementary Theory of Probability"; in: *Foundations of the Theory of Probability*, p. 6, eq. 3; URL: https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up.
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eq. (b); URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549.
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: https://en.wikipedia.org/wiki/Probability_axioms#The_probability_of_the_empty_set.

1.4.4 Probability of the complement

Theorem: The probability (→ I/1.3.1) of a complement of a set is one minus the probability of this set:

$$P(A^c) = 1 - P(A) \tag{1}$$

where $A^c = \Omega \setminus A$ and Ω is the sample space (→ I/1.1.2).

Proof: Since A and A^c are mutually exclusive (→ I/1.3.9) and $A \cup A^c = \Omega$, the third axiom of probability (→ I/1.4.1) implies:

$$\begin{split} P(A \cup A^c) &= P(A) + P(A^c) \\ P(\Omega) &= P(A) + P(A^c) \\ P(A^c) &= P(\Omega) - P(A) \; . \end{split} \tag{2}$$

The second axiom of probability (→ I/1.4.1) states that $P(\Omega) = 1$, such that we obtain:

$$P(A^c) = 1 - P(A) \; . \tag{3}$$

■

Sources:
- A.N. Kolmogorov (1950): "Elementary Theory of Probability"; in: *Foundations of the Theory of Probability*, p. 6, eq. 2; URL: https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up.
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eq. (c); URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549.
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: https://en.wikipedia.org/wiki/Probability_axioms#The_complement_rule.

1.4.5 Range of probability

Theorem: The probability (→ I/1.3.1) of an event is bounded between 0 and 1:

$$0 \leq P(E) \leq 1 \; . \tag{1}$$

Proof: From the first axiom of probability (→ I/1.4.1), we have:

$$P(E) \geq 0 \; . \tag{2}$$

By combining the first axiom of probability (→ I/1.4.1) and the probability of the complement (→ I/1.4.4), we obtain:

$$\begin{split} 1 - P(E) = P(E^c) &\geq 0 \\ 1 - P(E) &\geq 0 \\ P(E) &\leq 1 \; . \end{split} \tag{3}$$

Together, (2) and (3) imply that

$$0 \leq P(E) \leq 1 \; . \tag{4}$$

■

Sources:
- A.N. Kolmogorov (1950): "Elementary Theory of Probability"; in: *Foundations of the Theory of Probability*, p. 6; URL: https://archive.org/details/foundationsofthe00kolm/page/6/mode/2up.
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: https://en.wikipedia.org/wiki/Probability_axioms#The_numeric_bound.

1.4.6 Addition law of probability

Theorem: The probability (→ I/1.3.1) of the union of A and B is the sum of the probabilities of A and B minus the probability of the intersection of A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \; . \tag{1}$$

Proof: Let $E_1 = A$ and $E_2 = B \setminus A$, such that $E_1 \cup E_2 = A \cup B$. Then, by the third axiom of probability (→ I/1.4.1), we have:

$$\begin{split} P(A \cup B) &= P(A) + P(B \setminus A) \\ P(A \cup B) &= P(A) + P(B \setminus [A \cap B]) \; . \end{split} \tag{2}$$

Then, let $E_1 = B \setminus [A \cap B]$ and $E_2 = A \cap B$, such that $E_1 \cup E_2 = B$. Again, from the third axiom of probability (→ I/1.4.1), we obtain:

$$\begin{split} P(B) &= P(B \setminus [A \cap B]) + P(A \cap B) \\ P(B \setminus [A \cap B]) &= P(B) - P(A \cap B) \; . \end{split} \tag{3}$$

Plugging (3) into (2), we finally get:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \; . \tag{4}$$

■

Sources:
- A.N. Kolmogorov (1950): "Elementary Theory of Probability"; in: *Foundations of the Theory of Probability*, p. 2; URL: https://archive.org/details/foundationsofthe00kolm/page/2/mode/2up.
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, ch. 8.6, p. 288, eq. (a); URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549.
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-30; URL: https://en.wikipedia.org/wiki/Probability_axioms#Further_consequences.

1.4.7 Law of total probability

Theorem: Let A be a subset of sample space (→ I/1.1.2) Ω and let $B_1, \ldots, B_n$ be finite or countably infinite partition of Ω, such that $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\cup_i B_i = \Omega$. Then, the probability (→ I/1.3.1) of the event A is

$$P(A) = \sum_i P(A \cap B_i) \; . \tag{1}$$

Proof: Because all B_i are disjoint, sets $(A \cap B_i)$ are also disjoint:

$$B_i \cap B_j = \emptyset \quad \Rightarrow \quad (A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = A \cap \emptyset = \emptyset \; . \tag{2}$$

Because the B_i are exhaustive, the sets $(A \cap B_i)$ are also exhaustive:

$$\cup_i B_i = \Omega \quad \Rightarrow \quad \cup_i (A \cap B_i) = A \cap (\cup_i B_i) = A \cap \Omega = A \; . \tag{3}$$

Thus, the third axiom of probability (→ I/1.4.1) implies that

$$P(A) = \sum_i P(A \cap B_i) \; . \tag{4}$$

■

Sources:

- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, p. 288, eq. (d); p. 289, eq. 8.7; URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9780470669549.
- Wikipedia (2021): "Law of total probability"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-08; URL: https://en.wikipedia.org/wiki/Law_of_total_probability#Statement.

1.4.8 Probability of exhaustive events

Theorem: Let $B_1, \ldots, B_n$ be mutually exclusive (→ I/1.3.9) and collectively exhaustive subsets of a sample space (→ I/1.1.2) Ω. Then, their total probability (→ I/1.4.7) is one:

$$\sum_i P(B_i) = 1 \; . \tag{1}$$

Proof: Because all B_i are mutually exclusive, we have:

$$B_i \cap B_j = \emptyset \quad \text{for all} \quad i \neq j \; . \tag{2}$$

Because the B_i are collectively exhaustive, we have:

$$\cup_i B_i = \Omega \; . \tag{3}$$

Thus, the third axiom of probability (→ I/1.4.1) implies that

$$\sum_i P(B_i) = P(\Omega) \; . \tag{4}$$

and the second axiom of probability (→ I/1.4.1) implies that

$$\sum_i P(B_i) = 1 \; . \tag{5}$$

■

Sources:
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9
- Wikipedia (2021): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-08; URL: https://en.wikipedia.org/wiki/Probability_axioms#Axioms.

1.4.9 Probability of exhaustive events

Theorem: Let $B_1, \ldots, B_n$ be mutually exclusive (→ I/1.3.9) and collectively exhaustive subsets of a sample space (→ I/1.1.2) Ω. Then, their total probability (→ I/1.4.7) is one:

$$\sum_i P(B_i) = 1 \; . \tag{1}$$

Proof: The addition law of probability (→ I/1.4.6) states that for two events (→ I/1.2.1) A and B, the probability (→ I/1.3.1) of at least one of them occurring is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \; . \tag{2}$$

Recursively applying this law to the events $B_1, \ldots, B_n$, we have:

$$\begin{aligned}
P(B_1 \cup \ldots \cup B_n) &= P(B_1) + P(B_2 \cup \ldots \cup B_n) - P(B_1 \cap [B_2 \cup \ldots \cup B_n]) \\
&= P(B_1) + P(B_2) + P(B_3 \cup \ldots \cup B_n) - P(B_2 \cap [B_3 \cup \ldots \cup B_n]) - P(B_1 \cap [B_2 \cup \ldots \cup B_n]) \\
&\;\vdots \\
&= P(B_1) + \ldots + P(B_n) - P(B_1 \cap [B_2 \cup \ldots \cup B_n]) - \ldots - P(B_{n-1} \cap B_n) \\
P(\cup_i^n B_i) &= \sum_i^n P(B_i) - \sum_i^{n-1} P(B_i \cap [\cup_{j=i+1}^n B_j]) \\
&= \sum_i^n P(B_i) - \sum_i^{n-1} P(\cup_{j=i+1}^n [B_i \cap B_j]) \; .
\end{aligned} \tag{3}$$

Because all B_i are mutually exclusive, we have:

$$B_i \cap B_j = \emptyset \quad \text{for all} \quad i \neq j \; . \tag{4}$$

Since the probability of the empty set is zero (→ I/1.4.3), this means that the second sum on the right-hand side of (3) disappears:

$$P(\cup_i^n B_i) = \sum_i^n P(B_i) \; . \tag{5}$$

Because the B_i are collectively exhaustive, we have:

$$\cup_i B_i = \Omega \; . \tag{6}$$

Since the probability of the sample space is one (→ I/1.4.1), this means that the left-hand side of (5) becomes equal to one:

$$1 = \sum_i^n P(B_i) \; . \tag{7}$$

This proofs the statement in (1).

■

Sources:
- Alan Stuart & J. Keith Ord (1994): "Probability and Statistical Inference"; in: *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory*, pp. 288-289; URL: https://www.wiley.com/en-us/Kendall%27s+Advanced+Theory+of+Statistics%2C+3+Volumes%2C+Set%2C+6th+Edition-p-9
- Wikipedia (2022): "Probability axioms"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-27; URL: https://en.wikipedia.org/wiki/Probability_axioms#Consequences.

1.5 Probability distributions

1.5.1 Probability distribution

Definition: Let X be a random variable (→ I/1.2.2) with the set of possible outcomes $\mathcal{X}$. Then, a probability distribution of X is a mathematical function that gives the probabilities (→ I/1.3.1) of occurrence of all possible outcomes $x \in \mathcal{X}$ of this random variable.

Sources:
- Wikipedia (2020): "Probability distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Probability_distribution.

1.5.2 Joint distribution

Definition: Let X and Y be random variables (→ I/1.2.2) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, a joint distribution of X and Y is a probability distribution (→ I/1.5.1) that specifies the probability of the event that $X = x$ and $Y = y$ for each possible combination of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- The joint distribution of two scalar random variables (→ I/1.2.2) is called a bivariate distribution.
- The joint distribution of a random vector (→ I/1.2.3) is called a multivariate distribution.
- The joint distribution of a random matrix (→ I/1.2.4) is called a matrix-variate distribution.

Sources:
- Wikipedia (2020): "Joint probability distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Joint_probability_distribution.

1.5.3 Marginal distribution

Definition: Let X and Y be random variables (→ I/1.2.2) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, the marginal distribution of X is a probability distribution (→ I/1.5.1) that specifies the probability of the event that $X = x$ irrespective of the value of Y for each possible value $x \in \mathcal{X}$. The marginal distribution can be obtained from the joint distribution (→ I/1.5.2) of X and Y using the law of marginal probability (→ I/1.3.3).

Sources:
- Wikipedia (2020): "Marginal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Marginal_distribution.

1.5.4 Conditional distribution

Definition: Let X and Y be random variables (→ I/1.2.2) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, the conditional distribution of X given that Y is a probability distribution (→ I/1.5.1) that specifies the probability of the event that $X = x$ given that $Y = y$ for each possible combination of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The conditional distribution of X can be obtained from the joint distribution (→ I/1.5.2) of X and Y and the marginal distribution (→ I/1.5.3) of Y using the law of conditional probability (→ I/1.3.4).

Sources:
- Wikipedia (2020): "Conditional probability distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-17; URL: https://en.wikipedia.org/wiki/Conditional_probability_distribution.

1.5.5 Sampling distribution

Definition: Let there be a random sample with finite sample size. Then, the probability distribution (→ I/1.5.1) of a given statistic computed from this sample, e.g. a test statistic (→ I/4.3.5), is called a sampling distribution.

Sources:
- Wikipedia (2021): "Sampling distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: https://en.wikipedia.org/wiki/Sampling_distribution.

1.6 Probability mass function

1.6.1 Definition

Definition: Let X be a discrete (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$. Then, $f_X(x) : \mathbb{R} \to [0, 1]$ is the probability mass function (PMF) of X, if

$$f_X(x) = 0 \tag{1}$$

for all $x \notin \mathcal{X}$,

$$\Pr(X = x) = f_X(x) \tag{2}$$

for all $x \in \mathcal{X}$ and

$$\sum_{x \in \mathcal{X}} f_X(x) = 1 \; . \tag{3}$$

Sources:
- Wikipedia (2020): "Probability mass function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Probability_mass_function.

1.6.2 Probability mass function of sum of independents

Theorem: Let X and Y be two independent (→ I/1.3.6) discrete (→ I/1.2.6) random variables (→ I/1.2.2) with possible values $\mathcal{X}$ and $\mathcal{Y}$ and let $Z = X + Y$. Then, the probability mass function (→ I/1.6.1) of Z is given by

$$\begin{split} f_Z(z) &= \sum_{y \in \mathcal{Y}} f_X(z-y) f_Y(y) \\ \text{or} \quad f_Z(z) &= \sum_{x \in \mathcal{X}} f_Y(z-x) f_X(x) \end{split} \tag{1}$$

where $f_X(x)$, $f_Y(y)$ and $f_Z(z)$ are the probability mass functions (→ I/1.6.1) of X, Y and Z.

Proof: Using the definition of the probability mass function (→ I/1.6.1) and the expected value (→ I/1.10.1), the first equation can be derived as follows:

$$\begin{split} f_Z(z) &= \mathrm{Pr}(Z = z) \\ &= \mathrm{Pr}(X + Y = z) \\ &= \mathrm{Pr}(X = z - Y) \\ &= \mathrm{E}\left[\mathrm{Pr}(X = z - Y \vert Y = y) \right] \; . \end{split} \tag{2}$$

By construction, X and Y are independent (→ I/1.3.6), such that conditional probabilities are equal to marginal probabilities (→ I/1.3.8), i.e. $\mathrm{Pr}(X = z - Y \vert Y = y) = \mathrm{Pr}(X = z - Y)$ and we have:

$$\begin{split} f_Z(z) &= \mathrm{E}\left[\mathrm{Pr}(X = z - Y) \right] \\ &= \mathrm{E}\left[f_X(z - Y) \right] \\ &= \sum_{y \in \mathcal{Y}} f_X(z-y) f_Y(y) \; . \end{split} \tag{3}$$

The second equation can be derived by switching X and Y.

■

Sources:
- Taboga, Marco (2017): "Sums of independent random variables"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: https://www.statlect.com/fundamentals-of-probabi sums-of-independent-random-variables.

1.6.3 Probability mass function of strictly increasing function

Theorem: Let X be a discrete (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly increasing function on the support of X. Then, the probability mass function (→ I/1.6.1) of $Y = g(X)$ is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \{ y = g(x) : x \in \mathcal{X} \} \; . \tag{2}$$

Proof: Because a strictly increasing function is invertible, the probability mass function (→ I/1.6.1) of Y can be derived as follows:

$$\begin{aligned} f_Y(y) &= \mathrm{Pr}(Y = y) \\ &= \mathrm{Pr}(g(X) = y) \\ &= \mathrm{Pr}(X = g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \; . \end{aligned} \tag{3}$$

■

Sources:

- Taboga, Marco (2017): "Functions of random variables and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-10-29; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid3.

1.6.4 Probability mass function of strictly decreasing function

Theorem: Let X be a discrete (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly decreasing function on the support of X. Then, the probability mass function (→ I/1.6.1) of $Y = g(X)$ is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \{ y = g(x) : x \in \mathcal{X} \} \; . \tag{2}$$

Proof: Because a strictly decreasing function is invertible, the probability mass function (→ I/1.6.1) of Y can be derived as follows:

$$\begin{aligned} f_Y(y) &= \mathrm{Pr}(Y = y) \\ &= \mathrm{Pr}(g(X) = y) \\ &= \mathrm{Pr}(X = g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \; . \end{aligned} \tag{3}$$

■

Sources:
- Taboga, Marco (2017): "Functions of random variables and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-11-06; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid6.

1.6.5 Probability mass function of invertible function

Theorem: Let X be an $n \times 1$ random vector (→ I/1.2.3) of discrete random variables (→ I/1.2.6) with possible outcomes $\mathcal{X}$ and let $g:\ \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an invertible function on the support of X. Then, the probability mass function (→ I/1.6.1) of $Y = g(X)$ is given by

$$f_Y(y) = \left\{ \begin{array}{rl} f_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{array} \right. \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \left\{ y = g(x) : x \in \mathcal{X} \right\} \; . \tag{2}$$

Proof: Because an invertible function is a one-to-one mapping, the probability mass function (→ I/1.6.1) of Y can be derived as follows:

$$\begin{split} f_Y(y) &= \mathrm{Pr}(Y = y) \\ &= \mathrm{Pr}(g(X) = y) \\ &= \mathrm{Pr}(X = g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \; . \end{split} \tag{3}$$

■

Sources:
- Taboga, Marco (2017): "Functions of random vectors and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors.

1.7 Probability density function

1.7.1 Definition

Definition: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$. Then, $f_X(x) : \mathbb{R} \rightarrow \mathbb{R}$ is the probability density function (PDF) of X, if

$$f_X(x) \geq 0 \tag{1}$$

for all $x \in \mathbb{R}$,

$$\mathrm{Pr}(X \in A) = \int_A f_X(x) \, \mathrm{d}x \tag{2}$$

for any $A \subset \mathcal{X}$ and

$$\int_{\mathcal{X}} f_X(x) \, \mathrm{d}x = 1 \; . \tag{3}$$

Sources:
- Wikipedia (2020): "Probability density function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Probability_density_function.

1.7.2 Probability density function of sum of independents

Theorem: Let X and Y be two independent (→ I/1.3.6) continuous (→ I/1.2.6) random variables (→ I/1.2.2) with possible values $\mathcal{X}$ and $\mathcal{Y}$ and let $Z = X + Y$. Then, the probability density function (→ I/1.7.1) of Z is given by

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) \, \mathrm{d}y \\ \text{or} \quad f_Z(z) &= \int_{-\infty}^{+\infty} f_Y(z-x) f_X(x) \, \mathrm{d}x \end{aligned} \tag{1}$$

where $f_X(x)$, $f_Y(y)$ and $f_Z(z)$ are the probability density functions (→ I/1.7.1) of X, Y and Z.

Proof: The cumulative distribution function of a sum of independent random variables (→ I/1.8.2) is

$$F_Z(z) = \mathrm{E}\left[F_X(z-Y)\right] \; . \tag{2}$$

The probability density function is the first derivative of the cumulative distribution function (→ I/1.7.7), such that

$$\begin{aligned} f_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) \\ &= \frac{\mathrm{d}}{\mathrm{d}z} \mathrm{E}\left[F_X(z-Y)\right] \\ &= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}z} F_X(z-Y)\right] \\ &= \mathrm{E}\left[f_X(z-Y)\right] \\ &= \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) \, \mathrm{d}y \; . \end{aligned} \tag{3}$$

The second equation can be derived by switching X and Y.

■

Sources:
- Taboga, Marco (2017): "Sums of independent random variables"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: https://www.statlect.com/fundamentals-of-probabi sums-of-independent-random-variables.

1.7.3 Probability density function of strictly increasing function

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly increasing function on the support of X. Then, the probability density function (→ I/1.7.1) of $Y = g(X)$ is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} \; . \tag{2}$$

Proof: The cumulative distribution function of a strictly increasing function (→ I/1.8.3) is

$$F_Y(y) = \begin{cases} 0 \; , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 1 \; , & \text{if } y > \max(\mathcal{Y}) \end{cases} \tag{3}$$

Because the probability density function is the first derivative of the cumulative distribution function (→ I/1.7.7)

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \; , \tag{4}$$

the probability density function (→ I/1.7.1) of Y can be derived as follows:
1) If y does not belong to the support of Y, $F_Y(y)$ is constant, such that

$$f_Y(y) = 0, \quad \text{if} \quad y \notin \mathcal{Y} \; . \tag{5}$$

2) If y belongs to the support of Y, then $f_Y(y)$ can be derived using the chain rule:

$$\begin{aligned} f_Y(y) &\overset{(4)}{=} \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &\overset{(3)}{=} \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; . \end{aligned} \tag{6}$$

Taking together (5) and (6), eventually proves (1).

■

Sources:

- Taboga, Marco (2017): "Functions of random variables and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-10-29; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid4.

1.7.4 Probability density function of strictly decreasing function

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly decreasing function on the support of X. Then, the probability density function (→ I/1.7.1) of $Y = g(X)$ is given by

$$f_Y(y) = \begin{cases} -f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \{ y = g(x) : x \in \mathcal{X} \} \; . \tag{2}$$

Proof: The cumulative distribution function of a strictly decreasing function (→ I/1.8.4) is

$$F_Y(y) = \begin{cases} 1 \; , & \text{if } y > \max(\mathcal{Y}) \\ 1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y < \min(\mathcal{Y}) \end{cases} \tag{3}$$

Note that for continuous random variables, the probability (→ I/1.7.1) of point events is

$$\Pr(X = a) = \int_a^a f_X(x) \, \mathrm{d}x = 0 \; . \tag{4}$$

Because the probability density function is the first derivative of the cumulative distribution function (→ I/1.7.7)

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \; , \tag{5}$$

the probability density function (→ I/1.7.1) of Y can be derived as follows:
1) If y does not belong to the support of Y, $F_Y(y)$ is constant, such that

$$f_Y(y) = 0, \quad \text{if} \quad y \notin \mathcal{Y} \; . \tag{6}$$

2) If y belongs to the support of Y, then $f_Y(y)$ can be derived using the chain rule:

$$\begin{aligned} f_Y(y) &\overset{(5)}{=} \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &\overset{(3)}{=} \frac{\mathrm{d}}{\mathrm{d}y} \left[1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)) \right] \\ &\overset{(4)}{=} \frac{\mathrm{d}}{\mathrm{d}y} \left[1 - F_X(g^{-1}(y)) \right] \\ &= -\frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) \\ &= -f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; . \end{aligned} \tag{7}$$

Taking together (6) and (7), eventually proves (1).

■

Sources:

- Taboga, Marco (2017): "Functions of random variables and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-11-06; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid7.

1.7.5 Probability density function of invertible function

Theorem: Let X be an $n \times 1$ random vector (→ I/1.2.3) of continuous random variables (→ I/1.2.6) with possible outcomes $\mathcal{X} \subseteq \mathbb{R}^n$ and let $g:\ \mathbb{R}^n \to \mathbb{R}^n$ be an invertible and differentiable function on the support of X. Then, the probability density function (→ I/1.7.1) of $Y = g(X)$ is given by

$$f_Y(y) = \left\{ \begin{array}{rl} f_X(g^{-1}(y)) \, \left| J_{g^{-1}}(y) \right| \; , & \text{if} \; y \in \mathcal{Y} \\ 0 \; , & \text{if} \; y \notin \mathcal{Y} \end{array} \right. \; , \tag{1}$$

if the Jacobian determinant satisfies

$$\left| J_{g^{-1}}(y) \right| \neq 0 \quad \text{for all} \quad y \in \mathcal{Y} \tag{2}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$, $J_{g^{-1}}(y)$ is the Jacobian matrix of $g^{-1}(y)$

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}y_1} & \ldots & \frac{\mathrm{d}x_1}{\mathrm{d}y_n} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}x_n}{\mathrm{d}y_1} & \ldots & \frac{\mathrm{d}x_n}{\mathrm{d}y_n} \end{bmatrix} \; , \tag{3}$$

$|J|$ is the determinant of J and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \left\{ y = g(x) : x \in \mathcal{X} \right\} \; . \tag{4}$$

Proof:
1) First, we obtain the cumulative distribution function (→ I/1.8.1) of $Y = g(X)$. The joint CDF (→ I/1.8.10) is given by

$$\begin{split} F_Y(y) &= \mathrm{Pr}(Y_1 \leq y_1, \ldots, Y_n \leq y_n) \\ &= \mathrm{Pr}(g_1(X) \leq y_1, \ldots, g_n(X) \leq y_n) \\ &= \int_{A(y)} f_X(x) \, \mathrm{d}x \end{split} \tag{5}$$

where $A(y)$ is the following subset of the n-dimensional Euclidean space:

$$A(y) = \left\{ x \in \mathbb{R}^n : g_j(x) \leq y_j \; \text{for all} \; j = 1, \ldots, n \right\} \tag{6}$$

and $g_j(X)$ is the function which returns the j-th element of Y, given a vector X.

2) Next, we substitute $x = g^{-1}(y)$ into the integral which gives us

$$
\begin{aligned}
F_Y(z) &= \int_{B(z)} f_X(g^{-1}(y)) \, \mathrm{d}g^{-1}(y) \\
&= \int_{-\infty}^{z_n} \ldots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) \, \mathrm{d}g^{-1}(y) \; .
\end{aligned} \tag{7}
$$

where we have the modified the integration regime $B(z)$ which reads

$$
B(z) = \{y \in \mathbb{R}^n : y \leq z_j \text{ for all } j = 1, \ldots, n\} \; . \tag{8}
$$

3) The formula for change of variables in multivariable calculus states that

$$
y = f(x) \quad \Rightarrow \quad \mathrm{d}y = |J_f(x)| \; \mathrm{d}x \; . \tag{9}
$$

Applied to equation (7), this yields

$$
\begin{aligned}
F_Y(z) &= \int_{-\infty}^{z_n} \ldots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) \; |J_{g^{-1}}(y)| \; \mathrm{d}y \\
&= \int_{-\infty}^{z_n} \ldots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) \; |J_{g^{-1}}(y)| \; \mathrm{d}y_1 \ldots \mathrm{d}y_n \; .
\end{aligned} \tag{10}
$$

4) Finally, we obtain the probability density function (→ I/1.7.1) of $Y = g(X)$. Because the PDF is the derivative of the CDF (→ I/1.7.7), we can differentiate the joint CDF to get

$$
\begin{aligned}
f_Y(z) &= \frac{\mathrm{d}^n}{\mathrm{d}z_1 \ldots \mathrm{d}z_n} F_Y(z) \\
&= \frac{\mathrm{d}^n}{\mathrm{d}z_1 \ldots \mathrm{d}z_n} \int_{-\infty}^{z_n} \ldots \int_{-\infty}^{z_1} f_X(g^{-1}(y)) \; |J_{g^{-1}}(y)| \; \mathrm{d}y_1 \ldots \mathrm{d}y_n \\
&= f_X(g^{-1}(z)) \; |J_{g^{-1}}(z)|
\end{aligned} \tag{11}
$$

which can also be written as

$$
f_Y(y) = f_X(g^{-1}(y)) \; |J_{g^{-1}}(y)| \; . \tag{12}
$$

■

Sources:

- Taboga, Marco (2017): "Functions of random vectors and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors.
- Lebanon, Guy (2017): "Functions of a Random Vector"; in: *Probability: The Analysis of Data, Vol. 1*, retrieved on 2021-08-30; URL: http://theanalysisofdata.com/probability/4_4.html.
- Poirier, Dale J. (1995): "Distributions of Functions of Random Variables"; in: *Intermediate Statistics and Econometrics: A Comparative Approach*, ch. 4, pp. 149ff.; URL: https://books.google.de/books?id=K52_YvD1YNwC&hl=de&source=gbs_navlinks_s.
- Devore, Jay L.; Berk, Kennth N. (2011): "Conditional Distributions"; in: *Modern Mathematical Statistics with Applications*, ch. 5.2, pp. 253ff.; URL: https://books.google.de/books?id=5PRLUho-YYgC&hl=de&source=gbs_navlinks_s.

- peek-a-boo (2019): "How to come up with the Jacobian in the change of variables formula"; in: *StackExchange Mathematics*, retrieved on 2021-08-30; URL: https://math.stackexchange.com/a/3239222.
- Bazett, Trefor (2019): "Change of Variables & The Jacobian | Multi-variable Integration"; in: *YouTube*, retrieved on 2021-08-30; URL: https://www.youtube.com/watch?v=wUF-lyyWpUc.

1.7.6 Probability density function of linear transformation

Theorem: Let X be an $n \times 1$ random vector (→ I/1.2.3) of continuous random variables (→ I/1.2.6) with possible outcomes $\mathcal{X} \subseteq \mathbb{R}^n$ and let $Y = \Sigma X + \mu$ be a linear transformation of this random variable with constant (→ I/1.2.5) $n \times 1$ vector μ and constant (→ I/1.2.5) $n \times n$ matrix Σ. Then, the probability density function (→ I/1.7.1) of Y is

$$f_Y(y) = \begin{cases} \frac{1}{|\Sigma|} f_X(\Sigma^{-1}(y-\mu)) \; , & \text{if} \; y \in \mathcal{Y} \\ 0 \; , & \text{if} \; y \notin \mathcal{Y} \end{cases} \tag{1}$$

where $|\Sigma|$ is the determinant of Σ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \{ y = \Sigma x + \mu : x \in \mathcal{X} \} \; . \tag{2}$$

Proof: Because the linear function $g(X) = \Sigma X + \mu$ is invertible and differentiable, we can determine the probability density function of an invertible function of a continuous random vector (→ I/1.7.5) using the relation

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \, |J_{g^{-1}}(y)| \; , & \text{if} \; y \in \mathcal{Y} \\ 0 \; , & \text{if} \; y \notin \mathcal{Y} \end{cases} \; . \tag{3}$$

The inverse function is

$$X = g^{-1}(Y) = \Sigma^{-1}(Y - \mu) = \Sigma^{-1} Y - \Sigma^{-1} \mu \tag{4}$$

and the Jacobian matrix is

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}y_1} & \ldots & \frac{\mathrm{d}x_1}{\mathrm{d}y_n} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}x_n}{\mathrm{d}y_1} & \ldots & \frac{\mathrm{d}x_n}{\mathrm{d}y_n} \end{bmatrix} = \Sigma^{-1} \; . \tag{5}$$

Plugging (4) and (5) into (3) and applying the determinant property $|A^{-1}| = |A|^{-1}$, we obtain

$$f_Y(y) = \frac{1}{|\Sigma|} f_X(\Sigma^{-1}(y-\mu)) \; . \tag{6}$$

■

Sources:

- Taboga, Marco (2017): "Functions of random vectors and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors.

1.7.7 Probability density function in terms of cumulative distribution function

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2). Then, the probability distribution function (→ I/1.7.1) of X is the first derivative of the cumulative distribution function (→ I/1.8.1) of X:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \; . \tag{1}$$

Proof: The cumulative distribution function in terms of the probability density function of a continuous random variable (→ I/1.8.6) is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, \mathrm{d}t, \; x \in \mathbb{R} \; . \tag{2}$$

Taking the derivative with respect to x, we have:

$$\frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x} \int_{-\infty}^{x} f_X(t) \, \mathrm{d}t \; . \tag{3}$$

The fundamental theorem of calculus states that, if $f(x)$ is a continuous real-valued function defined on the interval $[a, b]$, then it holds that

$$F(x) = \int_{a}^{x} f(t) \, \mathrm{d}t \quad \Rightarrow \quad F'(x) = f(x) \quad \text{for all} \quad x \in (a, b) \; . \tag{4}$$

Applying (4) to (2), it follows that

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, \mathrm{d}t \quad \Rightarrow \quad \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = f_X(x) \quad \text{for all} \quad x \in \mathbb{R} \; . \tag{5}$$

■

Sources:
- Wikipedia (2020): "Fundamental theorem of calculus"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: https://en.wikipedia.org/wiki/Fundamental_theorem_of_calculus#Formal_statements.

1.8 Cumulative distribution function

1.8.1 Definition

Definition: The cumulative distribution function (CDF) of a random variable (→ I/1.2.2) X at a given value x is defined as the probability (→ I/1.3.1) that X is smaller than x:

$$F_X(x) = \Pr(X \leq x) \; . \tag{1}$$

1) If X is a discrete (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and the probability mass function (→ I/1.6.1) $f_X(x)$, then the cumulative distribution function is the function (→ I/1.8.5) $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ with

$$F_X(x) = \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) \; . \tag{2}$$

2) If X is a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and the probability density function (→ I/1.7.1) $f_X(x)$, then the cumulative distribution function is the function (→ I/1.8.6) $F_X(x) : \mathbb{R} \to [0, 1]$ with

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, \mathrm{d}t \; . \tag{3}$$

Sources:

- Wikipedia (2020): "Cumulative distribution function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Cumulative_distribution_function#Definition.

1.8.2 Cumulative distribution function of sum of independents

Theorem: Let X and Y be two independent (→ I/1.3.6) random variables (→ I/1.2.2) and let $Z = X + Y$. Then, the cumulative distribution function (→ I/1.8.1) of Z is given by

$$\begin{aligned} & F_Z(z) = \mathrm{E}\left[F_X(z - Y)\right] \\ \text{or} \quad & F_Z(z) = \mathrm{E}\left[F_Y(z - X)\right] \end{aligned} \tag{1}$$

where $F_X(x)$, $F_Y(y)$ and $F_Z(z)$ are the cumulative distribution functions (→ I/1.8.1) of X, Y and Z and $\mathrm{E}\left[\cdot\right]$ denotes the expected value (→ I/1.10.1).

Proof: Using the definition of the cumulative distribution function (→ I/1.8.1), the first equation can be derived as follows:

$$\begin{aligned} F_Z(z) &= \Pr(Z \leq z) \\ &= \Pr(X + Y \leq z) \\ &= \Pr(X \leq z - Y) \\ &= \mathrm{E}\left[\Pr(X \leq z - Y | Y = y)\right] \\ &= \mathrm{E}\left[\Pr(X \leq z - Y)\right] \\ &= \mathrm{E}\left[F_X(z - Y)\right] \; . \end{aligned} \tag{2}$$

Note that the second-last transition is justified by the fact that X and Y are independent (→ I/1.3.6), such that conditional probabilities are equal to marginal probabilities (→ I/1.3.8). The second equation can be derived by switching X and Y.

■

Sources:

- Taboga, Marco (2017): "Sums of independent random variables"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-08-30; URL: https://www.statlect.com/fundamentals-of-probabi sums-of-independent-random-variables.

1.8.3 Cumulative distribution function of strictly increasing function

Theorem: Let X be a random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly increasing function on the support of X. Then, the cumulative distribution function (→ I/1.8.1) of $Y = g(X)$ is given by

$$F_Y(y) = \begin{cases} 0 \; , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 1 \; , & \text{if } y > \max(\mathcal{Y}) \end{cases} \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \left\{ y = g(x) : x \in \mathcal{X} \right\} \; . \tag{2}$$

Proof: The support of Y is determined by $g(x)$ and by the set of possible outcomes of X. Moreover, if $g(x)$ is strictly increasing, then $g^{-1}(y)$ is also strictly increasing. Therefore, the cumulative distribution function (→ I/1.8.1) of Y can be derived as follows:
1) If y is lower than the lowest value (→ I/1.17.1) Y can take, then $\mathrm{Pr}(Y \leq y) = 0$, so

$$F_Y(y) = 0, \quad \text{if} \quad y < \min(\mathcal{Y}) \; . \tag{3}$$

2) If y belongs to the support of Y, then $F_Y(y)$ can be derived as follows:

$$\begin{split} F_Y(y) &= \mathrm{Pr}(Y \leq y) \\ &= \mathrm{Pr}(g(X) \leq y) \\ &= \mathrm{Pr}(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \; . \end{split} \tag{4}$$

3) If y is higher than the highest value (→ I/1.17.2) Y can take, then $\mathrm{Pr}(Y \leq y) = 1$, so

$$F_Y(y) = 1, \quad \text{if} \quad y > \max(\mathcal{Y}) \; . \tag{5}$$

Taking together (3), (4), (5), eventually proves (1).

■

Sources:

- Taboga, Marco (2017): "Functions of random variables and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-10-29; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid2.

1.8.4 Cumulative distribution function of strictly decreasing function

Theorem: Let X be a random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly decreasing function on the support of X. Then, the cumulative distribution function (→ I/1.8.1) of $Y = g(X)$ is given by

$$F_Y(y) = \begin{cases} 1 \; , & \text{if } y > \max(\mathcal{Y}) \\ 1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y < \min(\mathcal{Y}) \end{cases} \tag{1}$$

where $g^{-1}(y)$ is the inverse function of $g(x)$ and $\mathcal{Y}$ is the set of possible outcomes of Y:

$$\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\} \; . \tag{2}$$

Proof: The support of Y is determined by $g(x)$ and by the set of possible outcomes of X. Moreover, if $g(x)$ is strictly decreasing, then $g^{-1}(y)$ is also strictly decreasing. Therefore, the cumulative distribution function (→ I/1.8.1) of Y can be derived as follows:
1) If y is higher than the highest value (→ I/1.17.2) Y can take, then $\Pr(Y \leq y) = 1$, so

$$F_Y(y) = 1, \quad \text{if} \quad y > \max(\mathcal{Y}) \; . \tag{3}$$

2) If y belongs to the support of Y, then $F_Y(y)$ can be derived as follows:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= 1 - \Pr(Y > y) \\ &= 1 - \Pr(g(X) > y) \\ &= 1 - \Pr(X < g^{-1}(y)) \\ &= 1 - \Pr(X < g^{-1}(y)) - \Pr(X = g^{-1}(y)) + \Pr(X = g^{-1}(y)) \\ &= 1 - \left[\Pr(X < g^{-1}(y)) + \Pr(X = g^{-1}(y))\right] + \Pr(X = g^{-1}(y)) \\ &= 1 - \Pr(X \leq g^{-1}(y)) + \Pr(X = g^{-1}(y)) \\ &= 1 - F_X(g^{-1}(y)) + \Pr(X = g^{-1}(y)) \; . \end{aligned} \tag{4}$$

3) If y is lower than the lowest value (→ I/1.17.1) Y can take, then $\Pr(Y \leq y) = 0$, so

$$F_Y(y) = 0, \quad \text{if} \quad y < \min(\mathcal{Y}) \; . \tag{5}$$

Taking together (3), (4), (5), eventually proves (1).

■

Sources:

- Taboga, Marco (2017): "Functions of random variables and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2020-11-06; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-variables-and-their-distribution#hid5.

1.8.5 Cumulative distribution function of discrete random variable

Theorem: Let X be a discrete (→ I/1.2.6) random variable (→ I/1.2.2) with possible values $\mathcal{X}$ and probability mass function (→ I/1.6.1) $f_X(x)$. Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) \; . \tag{1}$$

Proof: The cumulative distribution function (→ I/1.8.1) of a random variable (→ I/1.2.2) X is defined as the probability that X is smaller than x:

$$F_X(x) = \Pr(X \leq x) \; . \tag{2}$$

The probability mass function (→ I/1.6.1) of a discrete (→ I/1.2.6) random variable (→ I/1.2.2) X returns the probability that X takes a particular value x:

$$f_X(x) = \Pr(X = x) \; . \tag{3}$$

Taking these two definitions together, we have:

$$\begin{aligned} F_X(x) &\overset{(2)}{=} \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} \Pr(X = t) \\ &\overset{(3)}{=} \sum_{\substack{t \in \mathcal{X} \\ t \leq x}} f_X(t) \; . \end{aligned} \tag{4}$$

■

1.8.6 Cumulative distribution function of continuous random variable

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with possible values $\mathcal{X}$ and probability density function (→ I/1.7.1) $f_X(x)$. Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, \mathrm{d}t \; . \tag{1}$$

Proof: The cumulative distribution function (→ I/1.8.1) of a random variable (→ I/1.2.2) X is defined as the probability that X is smaller than x:

$$F_X(x) = \Pr(X \leq x) \; . \tag{2}$$

The probability density function (→ I/1.7.1) of a continuous (→ I/1.2.6) random variable (→ I/1.2.2) X can be used to calculate the probability that X falls into a particular interval A:

$$\Pr(X \in A) = \int_A f_X(x) \, \mathrm{d}x \; . \tag{3}$$

Taking these two definitions together, we have:

$$\begin{aligned} F_X(x) &\overset{(2)}{=} \Pr(X \in (-\infty, x]) \\ &\overset{(3)}{=} \int_{-\infty}^{x} f_X(t) \, \mathrm{d}t \; . \end{aligned} \tag{4}$$

■

1.8.7 Probability integral transform

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with invertible (→ I/1.9.1) cumulative distribution function (→ I/1.8.1) $F_X(x)$. Then, the random variable (→ I/1.2.2)

$$Y = F_X(X) \tag{1}$$

has a standard uniform distribution (→ II/3.1.2).

Proof: The cumulative distribution function (→ I/1.8.1) of $Y = F_X(X)$ can be derived as

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(F_X(X) \leq y) \\ &= \Pr(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned} \tag{2}$$

which is the cumulative distribution function of a continuous uniform distribution (→ II/3.1.4) with $a = 0$ and $b = 1$, i.e. the cumulative distribution function (→ I/1.8.1) of the standard uniform distribution (→ II/3.1.2) $\mathcal{U}(0, 1)$.

■

Sources:

- Wikipedia (2021): "Probability integral transform"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-07; URL: https://en.wikipedia.org/wiki/Probability_integral_transform#Proof.

1.8.8 Inverse transformation method

Theorem: Let U be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) having a standard uniform distribution (→ II/3.1.2). Then, the random variable (→ I/1.2.2)

$$X = F_X^{-1}(U) \tag{1}$$

has a probability distribution (→ I/1.5.1) characterized by the invertible (→ I/1.9.1) cumulative distribution function (→ I/1.8.1) $F_X(x)$.

Proof: The cumulative distribution function (→ I/1.8.1) of the transformation $X = F_X^{-1}(U)$ can be derived as

$$\begin{aligned} &\Pr(X \leq x) \\ &= \Pr(F_X^{-1}(U) \leq x) \\ &= \Pr(U \leq F_X(x)) \\ &= F_X(x) \; , \end{aligned} \tag{2}$$

because the cumulative distribution function (→ I/1.8.1) of the standard uniform distribution (→ II/3.1.2) $\mathcal{U}(0, 1)$ is

$$U \sim \mathcal{U}(0,1) \quad \Rightarrow \quad F_U(u) = \Pr(U \leq u) = u \; . \tag{3}$$

■

Sources:
- Wikipedia (2021): "Inverse transform sampling"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-07; URL: https://en.wikipedia.org/wiki/Inverse_transform_sampling#Proof_of_correctness.

1.8.9 Distributional transformation

Theorem: Let X and Y be two continuous (→ I/1.2.6) random variables (→ I/1.2.2) with cumulative distribution function (→ I/1.8.1) $F_X(x)$ and invertible cumulative distribution function (→ I/1.8.1) $F_Y(y)$. Then, the random variable (→ I/1.2.2)

$$\tilde{X} = F_Y^{-1}(F_X(X)) \tag{1}$$

has the same probability distribution (→ I/1.5.1) as Y.

Proof: The cumulative distribution function (→ I/1.8.1) of the transformation $\tilde{X} = F_Y^{-1}(F_X(X))$ can be derived as

$$\begin{aligned} F_{\tilde{X}}(y) &= \mathrm{Pr}\left(\tilde{X} \le y\right) \\ &= \mathrm{Pr}\left(F_Y^{-1}(F_X(X)) \le y\right) \\ &= \mathrm{Pr}\left(F_X(X) \le F_Y(y)\right) \\ &= \mathrm{Pr}\left(X \le F_X^{-1}(F_Y(y))\right) \\ &= F_X\left(F_X^{-1}(F_Y(y))\right) \\ &= F_Y(y) \end{aligned} \tag{2}$$

which shows that $\tilde{X}$ and Y have the same cumulative distribution function (→ I/1.8.1) and are thus identically distributed (→ I/1.5.1).

■

Sources:
- Soch, Joram (2020): "Distributional Transformation Improves Decoding Accuracy When Predicting Chronological Age From Structural MRI"; in: *Frontiers in Psychiatry*, vol. 11, art. 604268; URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.604268/full; DOI: 10.3389/fpsyt.2020.60426

1.8.10 Joint cumulative distribution function

Definition: Let $X \in \mathbb{R}^{n \times 1}$ be an $n \times 1$ random vector (→ I/1.2.3). Then, the joint (→ I/1.5.2) cumulative distribution function (→ I/1.8.1) of X is defined as the probability (→ I/1.3.1) that each entry X_i is smaller than a specific value x_i for $i = 1, \ldots, n$:

$$F_X(x) = \mathrm{Pr}(X_1 \le x_1, \ldots, X_n \le x_n) \; . \tag{1}$$

Sources:
- Wikipedia (2021): "Cumulative distribution function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-07; URL: https://en.wikipedia.org/wiki/Cumulative_distribution_function#Definition_for_more_than_two_random_variables.

1.9 Other probability functions

1.9.1 Quantile function

Definition: Let X be a random variable (→ I/1.2.2) with the cumulative distribution function (→ I/1.8.1) (CDF) $F_X(x)$. Then, the function $Q_X(p) : [0,1] \to \mathbb{R}$ which is the inverse CDF is the quantile function (QF) of X. More precisely, the QF is the function that, for a given quantile $p \in [0,1]$, returns the smallest x for which $F_X(x) = p$:

$$Q_X(p) = \min \left\{ x \in \mathbb{R} \, | \, F_X(x) = p \right\} \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Probability density function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Quantile_function#Definition.

1.9.2 Quantile function in terms of cumulative distribution function

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2) with the cumulative distribution function (→ I/1.8.1) $F_X(x)$. If the cumulative distribution function is strictly monotonically increasing, then the quantile function (→ I/1.9.1) is identical to the inverse of $F_X(x)$:

$$Q_X(p) = F_X^{-1}(x) \; . \tag{1}$$

Proof: The quantile function (→ I/1.9.1) $Q_X(p)$ is defined as the function that, for a given quantile $p \in [0,1]$, returns the smallest x for which $F_X(x) = p$:

$$Q_X(p) = \min \left\{ x \in \mathbb{R} \, | \, F_X(x) = p \right\} \; . \tag{2}$$

If $F_X(x)$ is continuous and strictly monotonically increasing, then there is exactly one x for which $F_X(x) = p$ and $F_X(x)$ is an invertible function, such that

$$Q_X(p) = F_X^{-1}(x) \; . \tag{3}$$

■

Sources:
- Wikipedia (2020): "Quantile function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: https://en.wikipedia.org/wiki/Quantile_function#Definition.

1.9.3 Characteristic function

Definition:
1) The characteristic function of a random variable (→ I/1.2.2) $X \in \mathbb{R}$ is

$$\varphi_X(t) = \mathrm{E}\left[e^{itX} \right] \; , \quad t \in \mathbb{R} \; . \tag{1}$$

2) The characteristic function of a random vector (→ I/1.2.3) $X \in \mathbb{R}^n$ is

$$\varphi_X(t) = \mathrm{E}\left[e^{it^\mathrm{T} X} \right] \; , \quad t \in \mathbb{R}^n \; . \tag{2}$$

3) The characteristic function of a random matrix (→ I/1.2.4) $X \in \mathbb{R}^{n \times p}$ is

$$\varphi_X(t) = \mathrm{E}\left[e^{i\,\mathrm{tr}\left(t^\mathrm{T} X\right)}\right] , \quad t \in \mathbb{R}^{n \times p} \; . \tag{3}$$

Sources:
- Wikipedia (2021): "Characteristic function (probability theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-09-22; URL: https://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)#Definition.
- Taboga, Marco (2017): "Joint characteristic function"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-10-07; URL: https://www.statlect.com/fundamentals-of-probability/joint-characteristic-function.

1.9.4 Characteristic function of arbitrary function

Theorem: Let X be a random variable (→ I/1.2.2) with the expected value (→ I/1.10.1) function $\mathrm{E}_X[\cdot]$. Then, the characteristic function (→ I/1.9.3) of $Y = g(X)$ is equal to

$$\varphi_Y(t) = \mathrm{E}_X\left[\exp(it\, g(X))\right] \; . \tag{1}$$

Proof: The characteristic function (→ I/1.9.3) is defined as

$$\varphi_Y(t) = \mathrm{E}\left[\exp(it\, Y)\right] \; . \tag{2}$$

Due of the law of the unconscious statistician (→ I/1.10.12)

$$\begin{aligned} \mathrm{E}[g(X)] &= \sum_{x \in \mathcal{X}} g(x) f_X(x) \\ \mathrm{E}[g(X)] &= \int_{\mathcal{X}} g(x) f_X(x) \, \mathrm{d}x \; , \end{aligned} \tag{3}$$

$Y = g(X)$ can simply be substituted into (2) to give

$$\varphi_Y(t) = \mathrm{E}_X\left[\exp(it\, g(X))\right] \; . \tag{4}$$

■

Sources:
- Taboga, Marco (2017): "Functions of random vectors and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-09-22; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors.

1.9.5 Moment-generating function

Definition:
1) The moment-generating function of a random variable (→ I/1.2.2) $X \in \mathbb{R}$ is

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] , \quad t \in \mathbb{R} \; . \tag{1}$$

2) The moment-generating function of a random vector (→ I/1.2.3) $X \in \mathbb{R}^n$ is

$$M_X(t) = \mathrm{E}\left[e^{t^\mathrm{T} X}\right], \quad t \in \mathbb{R}^n \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Moment-generating function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-22; URL: https://en.wikipedia.org/wiki/Moment-generating_function#Definition.
- Taboga, Marco (2017): "Joint moment generating function"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-10-07; URL: https://www.statlect.com/fundamentals-of-probability/joint-moment-generating-function.

1.9.6 Moment-generating function of arbitrary function

Theorem: Let X be a random variable (→ I/1.2.2) with the expected value (→ I/1.10.1) function $\mathrm{E}_X[\cdot]$. Then, the moment-generating function (→ I/1.9.5) of $Y = g(X)$ is equal to

$$M_Y(t) = \mathrm{E}_X\left[\exp(t\, g(X))\right] \; . \tag{1}$$

Proof: The moment-generating function (→ I/1.9.5) is defined as

$$M_Y(t) = \mathrm{E}\left[\exp(t\, Y)\right] \; . \tag{2}$$

Due of the law of the unconscious statistician (→ I/1.10.12)

$$\begin{aligned} \mathrm{E}[g(X)] &= \sum_{x \in \mathcal{X}} g(x) f_X(x) \\ \mathrm{E}[g(X)] &= \int_{\mathcal{X}} g(x) f_X(x) \, \mathrm{d}x \; , \end{aligned} \tag{3}$$

$Y = g(X)$ can simply be substituted into (2) to give

$$M_Y(t) = \mathrm{E}_X\left[\exp(t\, g(X))\right] \; . \tag{4}$$

■

Sources:
- Taboga, Marco (2017): "Functions of random vectors and their distribution"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-09-22; URL: https://www.statlect.com/fundamentals-of-probability/functions-of-random-vectors.

1.9.7 Moment-generating function of linear transformation

Theorem: Let X be an $n \times 1$ random vector (→ I/1.2.3) with the moment-generating function (→ I/1.9.5) $M_X(t)$. Then, the moment-generating function of the linear transformation $Y = AX + b$ is given by

$$M_Y(t) = \exp\left[t^\mathrm{T} b\right] \cdot M_X(At) \tag{1}$$

where A is an $m \times n$ matrix and b is an $m \times 1$ vector.

Proof: The moment-generating function of a random vector (→ I/1.9.5) X is

$$M_X(t) = \mathrm{E}\left(\exp\left[t^\mathrm{T} X\right]\right) \tag{2}$$

and therefore the moment-generating function of the random vector (→ I/1.2.3) Y is given by

$$\begin{aligned} M_Y(t) &= \mathrm{E}\left(\exp\left[t^\mathrm{T}(AX+b)\right]\right) \\ &= \mathrm{E}\left(\exp\left[t^\mathrm{T} AX\right] \cdot \exp\left[t^\mathrm{T} b\right]\right) \\ &= \exp\left[t^\mathrm{T} b\right] \cdot \mathrm{E}\left(\exp\left[(At)^\mathrm{T} X\right]\right) \\ &= \exp\left[t^\mathrm{T} b\right] \cdot M_X(At) \; . \end{aligned} \tag{3}$$

■

Sources:
- ProofWiki (2020): "Moment Generating Function of Linear Transformation of Random Variable"; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Moment_Generating_Function_of_Linear_Transformation_of_Random_Variable.

1.9.8 Moment-generating function of linear combination

Theorem: Let $X_1, \ldots, X_n$ be n independent (→ I/1.3.6) random variables (→ I/1.2.2) with moment-generating functions (→ I/1.9.5) $M_{X_i}(t)$. Then, the moment-generating function of the linear combination $X = \sum_{i=1}^n a_i X_i$ is given by

$$M_X(t) = \prod_{i=1}^n M_{X_i}(a_i t) \tag{1}$$

where $a_1, \ldots, a_n$ are n real numbers.

Proof: The moment-generating function of a random variable (→ I/1.9.5) X_i is

$$M_{X_i}(t) = \mathrm{E}\left(\exp\left[t X_i\right]\right) \tag{2}$$

and therefore the moment-generating function of the linear combination X is given by

$$\begin{aligned} M_X(t) &= \mathrm{E}\left(\exp\left[tX\right]\right) \\ &= \mathrm{E}\left(\exp\left[t \sum_{i=1}^n a_i X_i\right]\right) \\ &= \mathrm{E}\left(\prod_{i=1}^n \exp\left[t\, a_i X_i\right]\right) \; . \end{aligned} \tag{3}$$

Because the expected value is multiplicative for independent random variables (→ I/1.10.7), we have

$$\begin{aligned} M_X(t) &= \prod_{i=1}^n \mathrm{E}\left(\exp\left[(a_i t) X_i\right]\right) \\ &= \prod_{i=1}^n M_{X_i}(a_i t) \; . \end{aligned} \tag{4}$$

■

Sources:

- ProofWiki (2020): "Moment Generating Function of Linear Combination of Independent Random Variables"; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Moment_Generating_Function_of_Linear_Combination_of_Independent_Random_Variables.

1.9.9 Probability-generating function

Definition:
1) If X is a discrete (→ I/1.2.6) random variable (→ I/1.2.2) taking values in the non-negative integers $\{0, 1, \ldots\}$, then the probability-generating function of X is defined as

$$G_X(z) = \sum_{x=0}^{\infty} p(x)\, z^x \tag{1}$$

where $z \in \mathbb{C}$ and $p(x)$ is the probability mass function (→ I/1.6.1) of X.
2) If X is a discrete (→ I/1.2.6) random vector (→ I/1.2.3) taking values in the n-dimensional integer lattice $x \in \{0, 1, \ldots\}^n$, then the probability-generating function of X is defined as

$$G_X(z) = \sum_{x_1=0}^{\infty} \cdots \sum_{x_n=0}^{\infty} p(x_1, \ldots, x_n)\, {z_1}^{x_1} \cdot \ldots \cdot {z_n}^{x_n} \tag{2}$$

where $z \in \mathbb{C}^n$ and $p(x)$ is the probability mass function (→ I/1.6.1) of X.

Sources:

- Wikipedia (2020): "Probability-generating function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Probability-generating_function#Definition.

1.9.10 Probability-generating function in terms of expected value

Theorem: Let X be a discrete (→ I/1.2.6) random variable (→ I/1.2.2) whose set of possible values $\mathcal{X}$ is a subset of the natural numbers $\mathbb{N}$. Then, the probability-generating function (→ I/1.9.9) of X can be expressed in terms of an expected value (→ I/1.10.1) of a function of X

$$G_X(z) = \mathrm{E}\left[z^X\right] \tag{1}$$

where $z \in \mathbb{C}$.

Proof: The law of the unconscious statistician (→ I/1.10.12) states that

$$\mathrm{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) \tag{2}$$

where $f_X(x)$ is the probability mass function (→ I/1.6.1) of X. Here, we have $g(X) = z^X$, such that

$$\mathrm{E}\left[z^X\right] = \sum_{x \in \mathcal{X}} z^x f_X(x) \; . \tag{3}$$

By the definition of X, this is equal to

$$\mathrm{E}\left[z^X\right] = \sum_{x=0}^{\infty} f_X(x)\, z^x \; . \tag{4}$$

The right-hand side is equal to the probability-generating function (→ I/1.9.9) of X:

$$\mathrm{E}\left[z^X\right] = G_X(z) \; . \tag{5}$$

■

Sources:
- ProofWiki (2022): "Probability Generating Function as Expectation"; in: *ProofWiki*, retrieved on 2022-10-11; URL: https://proofwiki.org/wiki/Probability_Generating_Function_as_Expectation.

1.9.11 Probability-generating function of zero

Theorem: Let X be a random variable (→ I/1.2.2) with probability-generating function (→ I/1.9.9) $G_X(z)$ and probability mass function (→ I/1.6.1) $f_X(x)$. Then, the value of the probability-generating function at zero is equal to the value of the probability mass function at zero:

$$G_X(0) = f_X(0) \; . \tag{1}$$

Proof: The probability-generating function (→ I/1.9.9) of X is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x)\, z^x \tag{2}$$

where $f_X(x)$ is the probability mass function (→ I/1.6.1) of X. Setting $z = 0$, we obtain:

$$\begin{split} G_X(0) &= \sum_{x=0}^{\infty} f_X(x) \cdot 0^x \\ &= f_X(0) + 0^1 \cdot f_X(1) + 0^2 \cdot f_X(2) + \ldots \\ &= f_X(0) + 0 + 0 + \ldots \\ &= f_X(0) \; . \end{split} \tag{3}$$

■

Sources:
- ProofWiki (2022): "Probability Generating Function of Zero"; in: *ProofWiki*, retrieved on 2022-10-11; URL: https://proofwiki.org/wiki/Probability_Generating_Function_of_Zero.

1.9.12 Probability-generating function of one

Theorem: Let X be a random variable (→ I/1.2.2) with probability-generating function (→ I/1.9.9) $G_X(z)$ and set of possible values $\mathcal{X}$. Then, the value of the probability-generating function at one is equal to one:

$$G_X(1) = 1 \; . \tag{1}$$

Proof: The probability-generating function (→ I/1.9.9) of X is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x)\, z^x \tag{2}$$

where $f_X(x)$ is the probability mass function (→ I/1.6.1) of X. Setting $z = 1$, we obtain:

$$\begin{split} G_X(1) &= \sum_{x=0}^{\infty} f_X(x) \cdot 1^x \\ &= \sum_{x=0}^{\infty} f_X(x) \cdot 1 \\ &= \sum_{x=0}^{\infty} f_X(x) \; . \end{split} \tag{3}$$

Because the probability mass function (→ I/1.6.1) sums up to one, this becomes:

$$\begin{split} G_X(1) &= \sum_{x \in \mathcal{X}} f_X(x) \\ &= 1 \; . \end{split} \tag{4}$$

■

Sources:
- ProofWiki (2022): "Probability Generating Function of One"; in: *ProofWiki*, retrieved on 2022-10-11; URL: https://proofwiki.org/wiki/Probability_Generating_Function_of_One.

1.9.13 Cumulant-generating function

Definition:
1) The cumulant-generating function of a random variable (→ I/1.2.2) $X \in \mathbb{R}$ is

$$K_X(t) = \log \mathrm{E}\left[e^{tX}\right], \quad t \in \mathbb{R} \; . \tag{1}$$

2) The cumulant-generating function of a random vector (→ I/1.2.3) $X \in \mathbb{R}^n$ is

$$K_X(t) = \log \mathrm{E}\left[e^{t^\mathrm{T} X}\right], \quad t \in \mathbb{R}^n \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Cumulant"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Cumulant#Definition.

1.10 Expected value

1.10.1 Definition

Definition:
1) The expected value (or, mean) of a discrete random variable (→ I/1.2.2) X with domain $\mathcal{X}$ is

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{1}$$

where $f_X(x)$ is the probability mass function (→ I/1.6.1) of X.

2) The expected value (or, mean) of a continuous random variable (→ I/1.2.2) X with domain $\mathcal{X}$ is

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{2}$$

where $f_X(x)$ is the probability density function (→ I/1.7.1) of X.

Sources:
- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Expected_value#Definition.

1.10.2 Sample mean

Definition: Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the sample mean of x is denoted as $\bar{x}$ and is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \; . \tag{1}$$

Sources:
- Wikipedia (2021): "Sample mean and covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-16; URL: https://en.wikipedia.org/wiki/Sample_mean_and_covariance#Definition_of_the_sample_mean.

1.10.3 Non-negative random variable

Theorem: Let X be a non-negative random variable (→ I/1.2.2). Then, the expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \int_0^\infty (1 - F_X(x)) \, \mathrm{d}x \tag{1}$$

where $F_X(x)$ is the cumulative distribution function (→ I/1.8.1) of X.

Proof: Because the cumulative distribution function gives the probability of a random variable being smaller than a given value (→ I/1.8.1),

$$F_X(x) = \Pr(X \leq x) \; , \tag{2}$$

we have

$$1 - F_X(x) = \Pr(X > x) \; , \tag{3}$$

such that

$$\int_0^\infty (1 - F_X(x)) \, \mathrm{d}x = \int_0^\infty \Pr(X > x) \, \mathrm{d}x \tag{4}$$

which, using the probability density function (→ I/1.7.1) of X, can be rewritten as

$$\begin{split} \int_0^\infty (1 - F_X(x)) \, \mathrm{d}x &= \int_0^\infty \int_x^\infty f_X(z) \, \mathrm{d}z \, \mathrm{d}x \\ &= \int_0^\infty \int_0^z f_X(z) \, \mathrm{d}x \, \mathrm{d}z \\ &= \int_0^\infty f_X(z) \int_0^z 1 \, \mathrm{d}x \, \mathrm{d}z \\ &= \int_0^\infty [x]_0^z \cdot f_X(z) \, \mathrm{d}z \\ &= \int_0^\infty z \cdot f_X(z) \, \mathrm{d}z \end{split} \tag{5}$$

and by applying the definition of the expected value (→ I/1.10.1), we see that

$$\int_0^\infty (1 - F_X(x)) \, \mathrm{d}x = \int_0^\infty z \cdot f_X(z) \, \mathrm{d}z = \mathrm{E}(X) \tag{6}$$

which proves the identity given above.

■

Sources:
- Kemp, Graham (2014): "Expected value of a non-negative random variable"; in: *StackExchange Mathematics*, retrieved on 2020-05-18; URL: https://math.stackexchange.com/questions/958472/expected-value-of-a-non-negative-random-variable.

1.10.4 Non-negativity

Theorem: If a random variable (→ I/1.2.2) is strictly non-negative, its expected value (→ I/1.10.1) is also non-negative, i.e.

$$\mathrm{E}(X) \geq 0, \quad \text{if} \quad X \geq 0 \; . \tag{1}$$

Proof:
1) If $X \geq 0$ is a discrete random variable, then, because the probability mass function (→ I/1.6.1) is always non-negative, all the addends in

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{2}$$

are non-negative, thus the entire sum must be non-negative.

2) If $X \geq 0$ is a continuous random variable, then, because the probability density function (→ I/1.7.1) is always non-negative, the integrand in

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{3}$$

is strictly non-negative, thus the term on the right-hand side is a Lebesgue integral, so that the result on the left-hand side must be non-negative.

■

Sources:
- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.

1.10.5 Linearity

Theorem: The expected value (→ I/1.10.1) is a linear operator, i.e.

$$\begin{aligned} \mathrm{E}(X+Y) &= \mathrm{E}(X) + \mathrm{E}(Y) \\ \mathrm{E}(a\,X) &= a\,\mathrm{E}(X) \end{aligned} \tag{1}$$

for random variables (→ I/1.2.2) X and Y and a constant (→ I/1.2.5) a.

Proof:
1) If X and Y are discrete random variables (→ I/1.2.6), the expected value (→ I/1.10.1) is

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \tag{2}$$

and the law of marginal probability (→ I/1.3.3) states that

$$p(x) = \sum_{y \in \mathcal{Y}} p(x,y) \; . \tag{3}$$

Applying this, we have

$$\begin{aligned} \mathrm{E}(X+Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x+y) \cdot f_{X,Y}(x,y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot f_{X,Y}(x,y) + \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y \cdot f_{X,Y}(x,y) \\ &= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} f_{X,Y}(x,y) + \sum_{y \in \mathcal{Y}} y \sum_{x \in \mathcal{X}} f_{X,Y}(x,y) \\ &\overset{(3)}{=} \sum_{x \in \mathcal{X}} x \cdot f_X(x) + \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) \\ &\overset{(2)}{=} \mathrm{E}(X) + \mathrm{E}(Y) \end{aligned} \tag{4}$$

as well as

$$
\begin{aligned}
\mathrm{E}(a\,X) &= \sum_{x \in \mathcal{X}} a\,x \cdot f_X(x) \\
&= a \sum_{x \in \mathcal{X}} x \cdot f_X(x) \\
&\overset{(2)}{=} a\,\mathrm{E}(X) \; .
\end{aligned} \tag{5}
$$

2) If X and Y are continuous random variables (→ I/1.2.6), the expected value (→ I/1.10.1) is

$$
\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{6}
$$

and the law of marginal probability (→ I/1.3.3) states that

$$
p(x) = \int_{\mathcal{Y}} p(x,y) \, \mathrm{d}y \; . \tag{7}
$$

Applying this, we have

$$
\begin{aligned}
\mathrm{E}(X+Y) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x+y) \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} x \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x + \int_{\mathcal{X}} \int_{\mathcal{Y}} y \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x + \int_{\mathcal{Y}} y \int_{\mathcal{X}} f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y \\
&\overset{(7)}{=} \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x + \int_{\mathcal{Y}} y \cdot f_Y(y) \, \mathrm{d}y \\
&\overset{(6)}{=} \mathrm{E}(X) + \mathrm{E}(Y)
\end{aligned} \tag{8}
$$

as well as

$$
\begin{aligned}
\mathrm{E}(a\,X) &= \int_{\mathcal{X}} a\,x \cdot f_X(x) \, \mathrm{d}x \\
&= a \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \\
&\overset{(6)}{=} a\,\mathrm{E}(X) \; .
\end{aligned} \tag{9}
$$

Collectively, this shows that both requirements for linearity are fulfilled for the expected value (→ I/1.10.1), for discrete (→ I/1.2.6) as well as for continuous (→ I/1.2.6) random variables.

■

Sources:

- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.
- Michael B, Kuldeep Guha Mazumder, Geoff Pilling et al. (2020): "Linearity of Expectation"; in: *brilliant.org*, retrieved on 2020-02-13; URL: https://brilliant.org/wiki/linearity-of-expectation/.

1.10.6 Monotonicity

Theorem: The expected value (→ I/1.10.1) is monotonic, i.e.

$$\mathrm{E}(X) \leq \mathrm{E}(Y), \quad \text{if} \quad X \leq Y \; . \tag{1}$$

Proof: Let $Z = Y - X$. Due to the linearity of the expected value (→ I/1.10.5), we have

$$\mathrm{E}(Z) = \mathrm{E}(Y - X) = \mathrm{E}(Y) - \mathrm{E}(X) \; . \tag{2}$$

With the non-negativity property of the expected value (→ I/1.10.4), it also holds that

$$Z \geq 0 \quad \Rightarrow \quad \mathrm{E}(Z) \geq 0 \; . \tag{3}$$

Together with (2), this yields

$$\mathrm{E}(Y) - \mathrm{E}(X) \geq 0 \; . \tag{4}$$

■

Sources:
- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.

1.10.7 (Non-)Multiplicativity

Theorem:
1) If two random variables (→ I/1.2.2) X and Y are independent (→ I/1.3.6), the expected value (→ I/1.10.1) is multiplicative, i.e.

$$\mathrm{E}(X\,Y) = \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{1}$$

2) If two random variables (→ I/1.2.2) X and Y are dependent (→ I/1.3.6), the expected value (→ I/1.10.1) is not necessarily multiplicative, i.e. there exist X and Y such that

$$\mathrm{E}(X\,Y) \neq \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{2}$$

Proof:
1) If X and Y are independent (→ I/1.3.6), it holds that

$$p(x,y) = p(x)\,p(y) \quad \text{for all} \quad x \in \mathcal{X}, y \in \mathcal{Y} \; . \tag{3}$$

Applying this to the expected value for discrete random variables (→ I/1.10.1), we have

$$
\begin{aligned}
\mathrm{E}(X\,Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x,y) \\
&\overset{(3)}{=} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \\
&= \sum_{x \in \mathcal{X}} x \cdot f_X(x) \, \sum_{y \in \mathcal{Y}} y \cdot f_Y(y) \\
&= \sum_{x \in \mathcal{X}} x \cdot f_X(x) \cdot \mathrm{E}(Y) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) \; .
\end{aligned} \tag{4}
$$

And applying it to the expected value for continuous random variables (→ I/1.10.1), we have

$$
\begin{aligned}
\mathrm{E}(X\,Y) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot f_{X,Y}(x,y) \, \mathrm{d}y \, \mathrm{d}x \\
&\overset{(3)}{=} \int_{\mathcal{X}} \int_{\mathcal{Y}} (x \cdot y) \cdot (f_X(x) \cdot f_Y(y)) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} x \cdot f_X(x) \, \int_{\mathcal{Y}} y \cdot f_Y(y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} x \cdot f_X(x) \cdot \mathrm{E}(Y) \, \mathrm{d}x \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) \; .
\end{aligned} \tag{5}
$$

2) Let X and Y be Bernoulli random variables (→ II/1.2.1) with the following joint probability (→ I/1.3.2) mass function (→ I/1.6.1)

$$
\begin{aligned}
p(X=0, Y=0) &= 1/2 \\
p(X=0, Y=1) &= 0 \\
p(X=1, Y=0) &= 0 \\
p(X=1, Y=1) &= 1/2
\end{aligned} \tag{6}
$$

and thus, the following marginal probabilities:

$$
\begin{aligned}
p(X=0) = p(X=1) = 1/2 \\
p(Y=0) = p(Y=1) = 1/2 \; .
\end{aligned} \tag{7}
$$

Then, X and Y are dependent, because

$$
p(X=0, Y=1) \overset{(6)}{=} 0 \neq \frac{1}{2} \cdot \frac{1}{2} \overset{(7)}{=} p(X=0)\,p(Y=1) \; , \tag{8}
$$

and the expected value of their product is

$$
\begin{aligned}
\mathrm{E}(X\,Y) &= \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} (x \cdot y) \cdot p(x,y) \\
&= (1 \cdot 1) \cdot p(X=1, Y=1) \\
&\overset{(6)}{=} \frac{1}{2}
\end{aligned} \tag{9}
$$

while the product of their expected values is

$$
\begin{aligned}
\mathrm{E}(X)\,\mathrm{E}(Y) &= \left(\sum_{x \in \{0,1\}} x \cdot p(x) \right) \cdot \left(\sum_{y \in \{0,1\}} y \cdot p(y) \right) \\
&= (1 \cdot p(X=1)) \cdot (1 \cdot p(Y=1)) \\
&\overset{(7)}{=} \frac{1}{4}
\end{aligned} \tag{10}
$$

and thus,

$$
\mathrm{E}(X\,Y) \neq \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{11}
$$

■

Sources:

- Wikipedia (2020): "Expected value"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-17; URL: https://en.wikipedia.org/wiki/Expected_value#Basic_properties.

1.10.8 Expectation of a trace

Theorem: Let A be an $n \times n$ random matrix (→ I/1.2.4). Then, the expectation (→ I/1.10.1) of the trace of A is equal to the trace of the expectation (→ I/1.10.1) of A:

$$
\mathrm{E}\left[\mathrm{tr}(A)\right] = \mathrm{tr}\left(\mathrm{E}[A]\right) \; . \tag{1}
$$

Proof: The trace of an $n \times n$ matrix A is defined as:

$$
\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii} \; . \tag{2}
$$

Using this definition of the trace, the linearity of the expected value (→ I/1.10.5) and the expected value of a random matrix (→ I/1.10.14), we have:

$$
\begin{aligned}
\mathrm{E}\left[\mathrm{tr}(A)\right] &= \mathrm{E}\left[\sum_{i=1}^{n} a_{ii}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[a_{ii}\right] \\
&= \mathrm{tr}\left(\begin{bmatrix} \mathrm{E}[a_{11}] & \ldots & \mathrm{E}[a_{1n}] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[a_{n1}] & \ldots & \mathrm{E}[a_{nn}] \end{bmatrix}\right) \\
&= \mathrm{tr}\left(\mathrm{E}[A]\right) \; .
\end{aligned} \tag{3}
$$

■

Sources:

- drerD (2018): "'Trace trick' for expectations of quadratic forms"; in: *StackExchange Mathematics*, retrieved on 2021-12-07; URL: https://math.stackexchange.com/a/3004034/480910.

1.10.9 Expectation of a quadratic form

Theorem: Let X be an $n \times 1$ random vector (→ I/1.2.3) with mean (→ I/1.10.1) μ and covariance (→ I/1.13.1) Σ and let A be a symmetric $n \times n$ matrix. Then, the expectation of the quadratic form $X^\mathrm{T} A X$ is

$$
\mathrm{E}\left[X^\mathrm{T} A X\right] = \mu^\mathrm{T} A \mu + \mathrm{tr}(A\Sigma) \; . \tag{1}
$$

Proof: Note that $X^\mathrm{T} A X$ is a 1×1 matrix. We can therefore write

$$
\mathrm{E}\left[X^\mathrm{T} A X\right] = \mathrm{E}\left[\mathrm{tr}\left(X^\mathrm{T} A X\right)\right] \; . \tag{2}
$$

Using the trace property $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, this becomes

$$
\mathrm{E}\left[X^\mathrm{T} A X\right] = \mathrm{E}\left[\mathrm{tr}\left(A X X^\mathrm{T}\right)\right] \; . \tag{3}
$$

Because mean and trace are linear operators (→ I/1.10.8), we have

$$
\mathrm{E}\left[X^\mathrm{T} A X\right] = \mathrm{tr}\left(A \, \mathrm{E}\left[X X^\mathrm{T}\right]\right) \; . \tag{4}
$$

Note that the covariance matrix can be partitioned into expected values (→ I/1.13.11)

$$
\mathrm{Cov}(X, X) = \mathrm{E}(X X^\mathrm{T}) - \mathrm{E}(X)\mathrm{E}(X)^\mathrm{T} \; , \tag{5}
$$

such that the expected value of the quadratic form becomes

$$
\mathrm{E}\left[X^\mathrm{T} A X\right] = \mathrm{tr}\left(A \left[\mathrm{Cov}(X, X) + \mathrm{E}(X)\mathrm{E}(X)^\mathrm{T}\right]\right) \; . \tag{6}
$$

Finally, applying mean and covariance of X, we have

$$
\begin{aligned}
\mathrm{E}\left[X^{\mathrm{T}} A X\right] &= \mathrm{tr}\left(A\left[\Sigma + \mu\mu^{\mathrm{T}}\right]\right) \\
&= \mathrm{tr}\left(A\Sigma + A\mu\mu^{\mathrm{T}}\right) \\
&= \mathrm{tr}(A\Sigma) + \mathrm{tr}(A\mu\mu^{\mathrm{T}}) \\
&= \mathrm{tr}(A\Sigma) + \mathrm{tr}(\mu^{\mathrm{T}} A\mu) \\
&= \mu^{\mathrm{T}} A\mu + \mathrm{tr}(A\Sigma) \; .
\end{aligned} \tag{7}
$$

■

Sources:

- Kendrick, David (1981): "Expectation of a quadratic form"; in: *Stochastic Control for Economic Models*, pp. 170-171.
- Wikipedia (2020): "Multivariate random variable"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-13; URL: https://en.wikipedia.org/wiki/Multivariate_random_variable#Expectation_of_a_quadratic_form.
- Halvorsen, Kjetil B. (2012): "Expected value and variance of trace function"; in: *StackExchange CrossValidated*, retrieved on 2020-07-13; URL: https://stats.stackexchange.com/questions/34477/expected-value-and-variance-of-trace-function.
- Sarwate, Dilip (2013): "Expected Value of Quadratic Form"; in: *StackExchange CrossValidated*, retrieved on 2020-07-13; URL: https://stats.stackexchange.com/questions/48066/expected-value-of-quadrat

1.10.10 Squared expectation of a product

Theorem: Let X and Y be two random variables (→ I/1.2.2) with expected values (→ I/1.10.1) $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ and let $\mathrm{E}(XY)$ exist and be finite. Then, the square of the expectation of the product of X and Y is less than or equal to the product of the expectation of the squares of X and Y:

$$
[\mathrm{E}(XY)]^2 \leq \mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) \; . \tag{1}
$$

Proof: Note that Y^2 is a non-negative random variable (→ I/1.2.2) whose expected value is also non-negative (→ I/1.10.4):

$$
\mathrm{E}\left(Y^2\right) \geq 0 \; . \tag{2}
$$

1) First, consider the case that $\mathrm{E}\left(Y^2\right) > 0$. Define a new random variable Z as

$$
Z = X - Y\,\frac{\mathrm{E}(XY)}{\mathrm{E}\left(Y^2\right)} \; . \tag{3}
$$

Once again, because Z^2 is always non-negative, we have the expected value:

$$
\mathrm{E}\left(Z^2\right) \geq 0 \; . \tag{4}
$$

Thus, using the linearity of the expected value (→ I/1.10.5), we have

$$
\begin{split}
0 &\leq \mathrm{E}\left(Z^2\right) \\
&\leq \mathrm{E}\left(\left(X - Y\,\frac{\mathrm{E}(XY)}{\mathrm{E}\left(Y^2\right)}\right)^2\right) \\
&\leq \mathrm{E}\left(X^2 - 2XY\,\frac{\mathrm{E}(XY)}{\mathrm{E}\left(Y^2\right)} + Y^2\,\frac{\left[\mathrm{E}(XY)\right]^2}{\left[\mathrm{E}\left(Y^2\right)\right]^2}\right) \\
&\leq \mathrm{E}\left(X^2\right) - 2\,\mathrm{E}(XY)\,\frac{\mathrm{E}(XY)}{\mathrm{E}\left(Y^2\right)} + \mathrm{E}\left(Y^2\right)\,\frac{\left[\mathrm{E}(XY)\right]^2}{\left[\mathrm{E}\left(Y^2\right)\right]^2} \\
&\leq \mathrm{E}\left(X^2\right) - 2\,\frac{\left[\mathrm{E}(XY)\right]^2}{\mathrm{E}\left(Y^2\right)} + \frac{\left[\mathrm{E}(XY)\right]^2}{\mathrm{E}\left(Y^2\right)} \\
&\leq \mathrm{E}\left(X^2\right) - \frac{\left[\mathrm{E}(XY)\right]^2}{\mathrm{E}\left(Y^2\right)} \; ,
\end{split}
\tag{5}
$$

giving

$$\left[\mathrm{E}(XY)\right]^2 \leq \mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) \tag{6}$$

as required.
2) Next, consider the case that $\mathrm{E}\left(Y^2\right) = 0$. In this case, Y must be a constant (→ I/1.2.5) with mean (→ I/1.10.1) $\mathrm{E}(Y) = 0$ and variance (→ I/1.11.1) $\mathrm{Var}(Y) = 0$, thus we have

$$\Pr(Y = 0) = 1 \; . \tag{7}$$

This implies

$$\Pr(XY = 0) = 1 \; , \tag{8}$$

such that

$$\mathrm{E}(XY) = 0 \; . \tag{9}$$

Thus, we can write

$$0 = \left[\mathrm{E}(XY)\right]^2 = \mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) = 0 \; , \tag{10}$$

giving

$$\left[\mathrm{E}(XY)\right]^2 \leq \mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) \tag{11}$$

as required.

■

Sources:

- ProofWiki (2022): "Square of Expectation of Product is Less Than or Equal to Product of Expectation of Squares"; in: *ProofWiki*, retrieved on 2022-10-11; URL: https://proofwiki.org/wiki/Square_of_Expectation_of_Product_is_Less_Than_or_Equal_to_Product_of_Expectation_of_Squares.

1.10.11 Law of total expectation

Theorem: (law of total expectation, also called "law of iterated expectations") Let X be a random variable (→ I/1.2.2) with expected value (→ I/1.10.1) $\mathrm{E}(X)$ and let Y be any random variable (→ I/1.11.1) defined on the same probability space (→ I/1.1.4). Then, the expected value (→ I/1.10.1) of the conditional expectation of X given Y is the same as the expected value (→ I/1.10.1) of X:

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|Y)] \; . \tag{1}$$

Proof: Let X and Y be discrete random variables (→ I/1.2.6) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, the expectation of the conditional expetectation can be rewritten as:

$$\begin{split}
\mathrm{E}[\mathrm{E}(X|Y)] &= \mathrm{E}\left[\sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x|Y) \right] \\
&= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x|Y = y) \right] \cdot \mathrm{Pr}(Y = y) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot \mathrm{Pr}(X = x|Y = y) \cdot \mathrm{Pr}(Y = y) \; .
\end{split} \tag{2}$$

Using the law of conditional probability (→ I/1.3.4), this becomes:

$$\begin{split}
\mathrm{E}[\mathrm{E}(X|Y)] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \cdot \mathrm{Pr}(X = x, Y = y) \\
&= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} \mathrm{Pr}(X = x, Y = y) \; .
\end{split} \tag{3}$$

Using the law of marginal probability (→ I/1.3.3), this becomes:

$$\begin{split}
\mathrm{E}[\mathrm{E}(X|Y)] &= \sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x) \\
&= \mathrm{E}(X) \; .
\end{split} \tag{4}$$

■

Sources:
- Wikipedia (2021): "Law of total expectation"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: https://en.wikipedia.org/wiki/Law_of_total_expectation#Proof_in_the_finite_and_countable_cases.

1.10.12 Law of the unconscious statistician

Theorem: Let X be a random variable (→ I/1.2.2) and let $Y = g(X)$ be a function of this random variable.
1) If X is a discrete random variable with possible outcomes $\mathcal{X}$ and probability mass function (→ I/1.6.1) $f_X(x)$, the expected value (→ I/1.10.1) of $g(X)$ is

$$\mathrm{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) \; . \tag{1}$$

2) If X is a continuous random variable with possible outcomes $\mathcal{X}$ and probability density function (→ I/1.7.1) $f_X(x)$, the expected value (→ I/1.10.1) of $g(X)$ is

$$\mathrm{E}[g(X)] = \int_{\mathcal{X}} g(x) f_X(x) \, \mathrm{d}x \; . \tag{2}$$

Proof: Suppose that g is differentiable and that its inverse g^{-1} is monotonic.
1) The expected value (→ I/1.10.1) of $Y = g(X)$ is defined as

$$\mathrm{E}[Y] = \sum_{y \in \mathcal{Y}} y \, f_Y(y) \; . \tag{3}$$

Writing the probability mass function $f_Y(y)$ in terms of $y = g(x)$, we have:

$$\begin{split}
\mathrm{E}[g(X)] &= \sum_{y \in \mathcal{Y}} y \, \mathrm{Pr}(g(x) = y) \\
&= \sum_{y \in \mathcal{Y}} y \, \mathrm{Pr}(x = g^{-1}(y)) \\
&= \sum_{y \in \mathcal{Y}} y \sum_{x = g^{-1}(y)} f_X(x) \\
&= \sum_{y \in \mathcal{Y}} \sum_{x = g^{-1}(y)} y f_X(x) \\
&= \sum_{y \in \mathcal{Y}} \sum_{x = g^{-1}(y)} g(x) f_X(x) \; .
\end{split} \tag{4}$$

Finally, noting that "for all y, then for all $x = g^{-1}(y)$" is equivalent to "for all x" if g^{-1} is a monotonic function, we can conclude that

$$\mathrm{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) \; . \tag{5}$$

2) Let $y = g(x)$. The derivative of an inverse function is

$$\frac{\mathrm{d}}{\mathrm{d}y}(g^{-1}(y)) = \frac{1}{g'(g^{-1}(y))} \tag{6}$$

Because $x = g^{-1}(y)$, this can be rearranged into

$$\mathrm{d}x = \frac{1}{g'(g^{-1}(y))} \, \mathrm{d}y \tag{7}$$

and subsitituing (7) into (2), we get

$$\int_{\mathcal{X}} g(x) f_X(x) \, \mathrm{d}x = \int_{\mathcal{Y}} y \, f_X(g^{-1}(y)) \, \frac{1}{g'(g^{-1}(y))} \, \mathrm{d}y \; . \tag{8}$$

Considering the cumulative distribution function (→ I/1.8.1) of Y, one can deduce:

$$
\begin{aligned}
F_Y(y) &= \Pr(Y \leq y) \\
&= \Pr(g(X) \leq y) \\
&= \Pr(X \leq g^{-1}(y)) \\
&= F_X(g^{-1}(y)) \; .
\end{aligned} \tag{9}
$$

Differentiating to get the probability density function (→ I/1.7.1) of Y, the result is:

$$
\begin{aligned}
f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\
&\overset{(9)}{=} \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) \\
&= f_X(g^{-1}(y)) \, \frac{\mathrm{d}}{\mathrm{d}y}(g^{-1}(y)) \\
&\overset{(6)}{=} f_X(g^{-1}(y)) \, \frac{1}{g'(g^{-1}(y))} \; .
\end{aligned} \tag{10}
$$

Finally, substituing (10) into (8), we have:

$$
\int_{\mathcal{X}} g(x) f_X(x) \, \mathrm{d}x = \int_{\mathcal{Y}} y \, f_Y(y) \, \mathrm{d}y = \mathrm{E}[Y] = \mathrm{E}[g(X)] \; . \tag{11}
$$

■

Sources:
- Wikipedia (2020): "Law of the unconscious statistician"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician#Proof.
- Taboga, Marco (2017): "Transformation theorem"; in: *Lectures on probability and mathematical statistics*, retrieved on 2021-09-22; URL: https://www.statlect.com/glossary/transformation-theorem.

1.10.13 Expected value of a random vector

Definition: Let X be an $n \times 1$ random vector (→ I/1.2.3). Then, the expected value (→ I/1.10.1) of X is an $n \times 1$ vector whose entries correspond to the expected values of the entries of the random vector:

$$
\mathrm{E}(X) = \mathrm{E}\left(\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \right) = \begin{bmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{bmatrix} \; . \tag{1}
$$

Sources:
- Taboga, Marco (2017): "Expected value"; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2021-07-08; URL: https://www.statlect.com/fundamentals-of-probability/expected-value#hid12.
- Wikipedia (2021): "Multivariate random variable"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-08; URL: https://en.wikipedia.org/wiki/Multivariate_random_variable#Expected_value.

1.10.14 Expected value of a random matrix

Definition: Let X be an $n \times p$ random matrix (→ I/1.2.4). Then, the expected value (→ I/1.10.1) of X is an $n \times p$ matrix whose entries correspond to the expected values of the entries of the random matrix:

$$\mathrm{E}(X) = \mathrm{E}\left(\begin{bmatrix} X_{11} & \ldots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{np} \end{bmatrix} \right) = \begin{bmatrix} \mathrm{E}(X_{11}) & \ldots & \mathrm{E}(X_{1p}) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(X_{n1}) & \ldots & \mathrm{E}(X_{np}) \end{bmatrix} . \tag{1}$$

Sources:

- Taboga, Marco (2017): "Expected value"; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2021-07-08; URL: https://www.statlect.com/fundamentals-of-probability/expected-value#hid13.

1.11 Variance

1.11.1 Definition

Definition: The variance of a random variable (→ I/1.2.2) X is defined as the expected value (→ I/1.10.1) of the squared deviation from its expected value (→ I/1.10.1):

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2 \right] . \tag{1}$$

Sources:

- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-13; URL: https://en.wikipedia.org/wiki/Variance#Definition.

1.11.2 Sample variance

Definition: Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the sample variance of x is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1}$$

and the unbiased sample variance of x is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

where $\bar{x}$ is the sample mean (→ I/1.10.2).

Sources:

- Wikipedia (2021): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-16; URL: https://en.wikipedia.org/wiki/Variance#Sample_variance.

1.11.3 Partition into expected values

Theorem: Let X be a random variable (→ I/1.2.2). Then, the variance (→ I/1.11.1) of X is equal to the mean (→ I/1.10.1) of the square of X minus the square of the mean (→ I/1.10.1) of X:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; . \tag{1}$$

Proof: The variance (→ I/1.11.1) of X is defined as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \tag{2}$$

which, due to the linearity of the expected value (→ I/1.10.5), can be rewritten as

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\
&= \mathrm{E}\left[X^2 - 2\,X\,\mathrm{E}(X) + \mathrm{E}(X)^2\right] \\
&= \mathrm{E}(X^2) - 2\,\mathrm{E}(X)\,\mathrm{E}(X) + \mathrm{E}(X)^2 \\
&= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; .
\end{aligned} \tag{3}$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-19; URL: https://en.wikipedia.org/wiki/Variance#Definition.

1.11.4 Non-negativity

Theorem: The variance (→ I/1.11.1) is always non-negative, i.e.

$$\mathrm{Var}(X) \geq 0 \; . \tag{1}$$

Proof: The variance (→ I/1.11.1) of a random variable (→ I/1.2.2) is defined as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \; . \tag{2}$$

1) If X is a discrete random variable (→ I/1.2.2), then, because squares and probabilities are strictly non-negative, all the addends in

$$\mathrm{Var}(X) = \sum_{x \in \mathcal{X}} (x - \mathrm{E}(X))^2 \cdot f_X(x) \tag{3}$$

are also non-negative, thus the entire sum must be non-negative.

2) If X is a continuous random variable (→ I/1.2.2), then, because squares and probability densities are strictly non-negative, the integrand in

$$\mathrm{Var}(X) = \int_{\mathcal{X}} (x - \mathrm{E}(X))^2 \cdot f_X(x) \, \mathrm{d}x \tag{4}$$

is always non-negative, thus the term on the right-hand side is a Lebesgue integral, so that the result on the left-hand side must be non-negative.

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.5 Variance of a constant

Theorem: The variance (→ I/1.11.1) of a constant (→ I/1.2.5) is zero

$$a = \text{const.} \quad \Rightarrow \quad \mathrm{Var}(a) = 0 \tag{1}$$

and if the variance (→ I/1.11.1) of X is zero, then X is a constant (→ I/1.2.5)

$$\mathrm{Var}(X) = 0 \quad \Rightarrow \quad X = \text{const.} \tag{2}$$

Proof:
1) A constant (→ I/1.2.5) is defined as a quantity that always has the same value. Thus, if understood as a random variable (→ I/1.2.2), the expected value (→ I/1.10.1) of a constant is equal to itself:

$$\mathrm{E}(a) = a \; . \tag{3}$$

Plugged into the formula of the variance (→ I/1.11.1), we have

$$\begin{split} \mathrm{Var}(a) &= \mathrm{E}\left[(a - \mathrm{E}(a))^2\right] \\ &= \mathrm{E}\left[(a - a)^2\right] \\ &= \mathrm{E}(0) \; . \end{split} \tag{4}$$

Applied to the formula of the expected value (→ I/1.10.1), this gives

$$\mathrm{E}(0) = \sum_{x=0} x \cdot f_X(x) = 0 \cdot 1 = 0 \; . \tag{5}$$

Together, (4) and (5) imply (1).

2) The variance (→ I/1.11.1) is defined as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \; . \tag{6}$$

Because $(X - \mathrm{E}(X))^2$ is strictly non-negative (→ I/1.10.4), the only way for the variance to become zero is, if the squared deviation is always zero:

$$(X - \mathrm{E}(X))^2 = 0 \; . \tag{7}$$

This, in turn, requires that X is equal to its expected value (→ I/1.10.1)

$$X = \mathrm{E}(X) \tag{8}$$

which can only be the case, if X always has the same value (→ I/1.2.5):

$$X = \text{const.} \tag{9}$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-27; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.6 Invariance under addition

Theorem: The variance (→ I/1.11.1) is invariant under addition of a constant (→ I/1.2.5):

$$\mathrm{Var}(X+a) = \mathrm{Var}(X) \tag{1}$$

Proof: The variance (→ I/1.11.1) is defined in terms of the expected value (→ I/1.10.1) as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \; . \tag{2}$$

Using this and the linearity of the expected value (→ I/1.10.5), we can derive (1) as follows:

$$\begin{split} \mathrm{Var}(X+a) &\overset{(2)}{=} \mathrm{E}\left[((X+a)-\mathrm{E}(X+a))^2\right] \\ &= \mathrm{E}\left[(X+a-\mathrm{E}(X)-a)^2\right] \\ &= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \\ &\overset{(2)}{=} \mathrm{Var}(X) \; . \end{split} \tag{3}$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.7 Scaling upon multiplication

Theorem: The variance (→ I/1.11.1) scales upon multiplication with a constant (→ I/1.2.5):

$$\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X) \tag{1}$$

Proof: The variance (→ I/1.11.1) is defined in terms of the expected value (→ I/1.10.1) as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \; . \tag{2}$$

Using this and the linearity of the expected value (→ I/1.10.5), we can derive (1) as follows:

$$\begin{split} \mathrm{Var}(aX) &\overset{(2)}{=} \mathrm{E}\left[((aX)-\mathrm{E}(aX))^2\right] \\ &= \mathrm{E}\left[(aX-a\mathrm{E}(X))^2\right] \\ &= \mathrm{E}\left[(a[X-\mathrm{E}(X)])^2\right] \\ &= \mathrm{E}\left[a^2(X-\mathrm{E}(X))^2\right] \\ &= a^2\,\mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \\ &\overset{(2)}{=} a^2\,\mathrm{Var}(X) \; . \end{split} \tag{3}$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.8 Variance of a sum

Theorem: The variance (→ I/1.11.1) of the sum of two random variables (→ I/1.2.2) equals the sum of the variances of those random variables, plus two times their covariance (→ I/1.13.1):

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) \; . \tag{1}$$

Proof: The variance (→ I/1.11.1) is defined in terms of the expected value (→ I/1.10.1) as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \; . \tag{2}$$

Using this and the linearity of the expected value (→ I/1.10.5), we can derive (1) as follows:

$$\begin{split}
\mathrm{Var}(X+Y) &\overset{(2)}{=} \mathrm{E}\left[((X+Y)-\mathrm{E}(X+Y))^2\right] \\
&= \mathrm{E}\left[([X-\mathrm{E}(X)]+[Y-\mathrm{E}(Y)])^2\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))^2 + (Y-\mathrm{E}(Y))^2 + 2\,(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] + \mathrm{E}\left[(Y-\mathrm{E}(Y))^2\right] + \mathrm{E}\left[2\,(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&\overset{(2)}{=} \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) \; .
\end{split} \tag{3}$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.9 Variance of linear combination

Theorem: The variance (→ I/1.11.1) of the linear combination of two random variables (→ I/1.2.2) is a function of the variances as well as the covariance (→ I/1.13.1) of those random variables:

$$\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y) \; . \tag{1}$$

Proof: The variance (→ I/1.11.1) is defined in terms of the expected value (→ I/1.10.1) as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] \; . \tag{2}$$

Using this and the linearity of the expected value (→ I/1.10.5), we can derive (1) as follows:

$$
\begin{aligned}
\mathrm{Var}(aX+bY) &\overset{(2)}{=} \mathrm{E}\left[((aX+bY)-\mathrm{E}(aX+bY))^2\right] \\
&= \mathrm{E}\left[(a[X-\mathrm{E}(X)]+b[Y-\mathrm{E}(Y)])^2\right] \\
&= \mathrm{E}\left[a^2\,(X-\mathrm{E}(X))^2+b^2\,(Y-\mathrm{E}(Y))^2+2ab\,(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&= \mathrm{E}\left[a^2\,(X-\mathrm{E}(X))^2\right]+\mathrm{E}\left[b^2\,(Y-\mathrm{E}(Y))^2\right]+\mathrm{E}\left[2ab\,(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right] \\
&\overset{(2)}{=} a^2\,\mathrm{Var}(X)+b^2\,\mathrm{Var}(Y)+2ab\,\mathrm{Cov}(X,Y)\;.
\end{aligned} \tag{3}
$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.10 Additivity under independence

Theorem: The variance (→ I/1.11.1) is additive for independent (→ I/1.3.6) random variables (→ I/1.2.2):

$$
p(X,Y)=p(X)\,p(Y) \quad \Rightarrow \quad \mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)\;. \tag{1}
$$

Proof: The variance of the sum of two random variables (→ I/1.11.8) is given by

$$
\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)\;. \tag{2}
$$

The covariance of independent random variables (→ I/1.13.6) is zero:

$$
p(X,Y)=p(X)\,p(Y) \quad \Rightarrow \quad \mathrm{Cov}(X,Y)=0\;. \tag{3}
$$

Combining (2) and (3), we have:

$$
\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)\;. \tag{4}
$$

■

Sources:
- Wikipedia (2020): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-07; URL: https://en.wikipedia.org/wiki/Variance#Basic_properties.

1.11.11 Law of total variance

Theorem: (law of total variance, also called "conditional variance formula") Let X and Y be random variables (→ I/1.2.2) defined on the same probability space (→ I/1.1.4) and assume that the variance (→ I/1.11.1) of Y is finite. Then, the sum of the expectation (→ I/1.10.1) of the conditional variance and the variance (→ I/1.11.1) of the conditional expectation of Y given X is equal to the variance (→ I/1.11.1) of Y:

$$
\mathrm{Var}(Y)=\mathrm{E}[\mathrm{Var}(Y|X)]+\mathrm{Var}[\mathrm{E}(Y|X)]\;. \tag{1}
$$

Proof: The variance can be decomposed into expected values (→ I/1.11.3) as follows:

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2 \; . \tag{2}$$

This can be rearranged into:

$$\mathrm{E}(Y^2) = \mathrm{Var}(Y) + \mathrm{E}(Y)^2 \; . \tag{3}$$

Applying the law of total expectation (→ I/1.10.11), we have:

$$\mathrm{E}(Y^2) = \mathrm{E}\left[\mathrm{Var}(Y|X) + \mathrm{E}(Y|X)^2\right] \; . \tag{4}$$

Now subtract the second term from (2):

$$\mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \mathrm{E}\left[\mathrm{Var}(Y|X) + \mathrm{E}(Y|X)^2\right] - \mathrm{E}(Y)^2 \; . \tag{5}$$

Again applying the law of total expectation (→ I/1.10.11), we have:

$$\mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \mathrm{E}\left[\mathrm{Var}(Y|X) + \mathrm{E}(Y|X)^2\right] - \mathrm{E}\left[\mathrm{E}(Y|X)\right]^2 \; . \tag{6}$$

With the linearity of the expected value (→ I/1.10.5), the terms can be regrouped to give:

$$\mathrm{E}(Y^2) - \mathrm{E}(Y)^2 = \mathrm{E}\left[\mathrm{Var}(Y|X)\right] + \left(\mathrm{E}\left[\mathrm{E}(Y|X)^2\right] - \mathrm{E}\left[\mathrm{E}(Y|X)\right]^2\right) \; . \tag{7}$$

Using the decomposition of variance into expected values (→ I/1.11.3), we finally have:

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)] \; . \tag{8}$$

■

Sources:
- Wikipedia (2021): "Law of total variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: https://en.wikipedia.org/wiki/Law_of_total_variance#Proof.

1.11.12 Precision

Definition: The precision of a random variable (→ I/1.2.2) X is defined as the inverse of the variance (→ I/1.11.1), i.e. one divided by the expected value (→ I/1.10.1) of the squared deviation from its expected value (→ I/1.10.1):

$$\mathrm{Prec}(X) = \mathrm{Var}(X)^{-1} = \frac{1}{\mathrm{E}\left[(X - \mathrm{E}(X))^2\right]} \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Precision (statistics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-04-21; URL: https://en.wikipedia.org/wiki/Precision_(statistics).

1.12 Skewness

1.12.1 Definition

Definition: Let X be a random variable (→ I/1.2.2) with expected value (→ I/1.10.1) μ and standard deviation (→ I/1.16.1) σ. Then, the skewness of X is defined as the third standardized moment (→ I/1.18.9) of X:

$$\mathrm{Skew}(X) = \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3} \; . \tag{1}$$

Sources:
- Wikipedia (2023): "Skewness"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-04-20; URL: https://en.wikipedia.org/wiki/Skewness.

1.12.2 Sample skewness

Definition: Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the sample skewness of x is given by

$$\hat{s} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right]^{3/2}} \; , \tag{1}$$

where $\bar{x}$ is the sample mean (→ I/1.10.2).

Sources:
- Joanes, D. N. and Gill, C. A. (1998): "Comparing measures of sample skewness and kurtosis"; in: *The Statistician*, vol. 47, part 1, pp. 183-189; URL: https://www.jstor.org/stable/2988433.

1.12.3 Partition into expected values

Theorem: Let X be a random variable (→ I/1.2.2) with expected value (→ I/1.10.1) μ and standard deviation (→ I/1.16.1) σ. Then, the skewness (→ I/1.12.1) of X can be computed as:

$$\mathrm{Skew}(X) = \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \; . \tag{1}$$

Proof: The skewness (→ I/1.12.1) of X is defined as

$$\mathrm{Skew}(X) = \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3} \; . \tag{2}$$

Because the expected value is a linear operator (→ I/1.10.5), we can rewrite (2) as

$$
\begin{aligned}
\mathrm{Skew}(X) &= \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3} \\
&= \frac{\mathrm{E}[X^3 - 3X^2\mu + 3X\mu^2 - \mu^3]}{\sigma^3} \\
&= \frac{\mathrm{E}(X^3) - 3\mathrm{E}(X^2)\mu + 3\mathrm{E}(X)\mu^2 - \mu^3}{\sigma^3} \\
&= \frac{\mathrm{E}(X^3) - 3\mu\left[\mathrm{E}(X^2) - \mathrm{E}(X)\mu\right] - \mu^3}{\sigma^3} \\
&= \frac{\mathrm{E}(X^3) - 3\mu\left[\mathrm{E}(X^2) - \mathrm{E}(X)^2\right] - \mu^3}{\sigma^3} \; .
\end{aligned} \tag{3}
$$

Because the variance can be written in terms of expected values (→ I/1.11.3) as

$$
\sigma^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; , \tag{4}
$$

we can rewrite (3) as

$$
\mathrm{Skew}(X) = \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \; . \tag{5}
$$

This finishes the proof of (1).

■

Sources:
- Wikipedia (2023): "Skewness"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-04-20; URL: https://en.wikipedia.org/wiki/Skewness.

1.13 Covariance

1.13.1 Definition

Definition: The covariance of two random variables (→ I/1.2.2) X and Y is defined as the expected value (→ I/1.10.1) of the product of their deviations from their individual expected values (→ I/1.10.1):

$$
\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \; . \tag{1}
$$

Sources:
- Wikipedia (2020): "Covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: https://en.wikipedia.org/wiki/Covariance#Definition.

1.13.2 Sample covariance

Definition: Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ be samples from random variables (→ I/1.2.2) X and Y. Then, the sample covariance of x and y is given by

$$
\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{1}
$$

and the unbiased sample covariance of x and y is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2).

Sources:
- Wikipedia (2021): "Covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-20; URL: https://en.wikipedia.org/wiki/Covariance#Calculating_the_sample_covariance.

1.13.3 Partition into expected values

Theorem: Let X and Y be random variables (→ I/1.2.2). Then, the covariance (→ I/1.13.1) of X and Y is equal to the mean (→ I/1.10.1) of the product of X and Y minus the product of the means (→ I/1.10.1) of X and Y:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \; . \tag{1}$$

Proof: The covariance (→ I/1.13.1) of X and Y is defined as

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \; . \tag{2}$$

which, due to the linearity of the expected value (→ I/1.10.5), can be rewritten as

$$\begin{split} \mathrm{Cov}(X,Y) &= \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \\ &= \mathrm{E}\left[XY - X\,\mathrm{E}(Y) - \mathrm{E}(X)\,Y + \mathrm{E}(X)\mathrm{E}(Y)\right] \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(X)\mathrm{E}(Y) + \mathrm{E}(X)\mathrm{E}(Y) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \; . \end{split} \tag{3}$$

■

Sources:
- Wikipedia (2020): "Covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-02; URL: https://en.wikipedia.org/wiki/Covariance#Definition.

1.13.4 Symmetry

Theorem: The covariance (→ I/1.13.1) of two random variables (→ I/1.2.2) is a symmetric function:

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) \; . \tag{1}$$

Proof: The covariance (→ I/1.13.1) of random variables (→ I/1.2.2) X and Y is defined as:

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \; . \tag{2}$$

Switching X and Y in (2), we can easily see:

$$
\begin{aligned}
\mathrm{Cov}(Y,X) &\overset{(2)}{=} \mathrm{E}\left[(Y-\mathrm{E}[Y])(X-\mathrm{E}[X])\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right] \\
&= \mathrm{Cov}(X,Y) \; .
\end{aligned} \tag{3}
$$

■

Sources:
- Wikipedia (2022): "Covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: https://en.wikipedia.org/wiki/Covariance#Covariance_of_linear_combinations.

1.13.5 Self-covariance

Theorem: The covariance (→ I/1.13.1) of a random variable (→ I/1.2.2) with itself is equal to the variance (→ I/1.11.1):

$$
\mathrm{Cov}(X,X) = \mathrm{Var}(X) \; . \tag{1}
$$

Proof: The covariance (→ I/1.13.1) of random variables (→ I/1.2.2) X and Y is defined as:

$$
\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right] \; . \tag{2}
$$

Inserting X for Y in (2), the result is the variance (→ I/1.11.1) of X:

$$
\begin{aligned}
\mathrm{Cov}(X,X) &\overset{(2)}{=} \mathrm{E}\left[(X-\mathrm{E}[X])(X-\mathrm{E}[X])\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}[X])^2\right] \\
&= \mathrm{Var}(X) \; .
\end{aligned} \tag{3}
$$

■

Sources:
- Wikipedia (2022): "Covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: https://en.wikipedia.org/wiki/Covariance#Covariance_with_itself.

1.13.6 Covariance under independence

Theorem: Let X and Y be independent (→ I/1.3.6) random variables (→ I/1.2.2). Then, the covariance (→ I/1.13.1) of X and Y is zero:

$$
X, Y \text{ independent} \quad \Rightarrow \quad \mathrm{Cov}(X,Y) = 0 \; . \tag{1}
$$

Proof: The covariance can be expressed in terms of expected values (→ I/1.13.3) as

$$
\mathrm{Cov}(X,Y) = \mathrm{E}(X\,Y) - \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{2}
$$

For independent random variables, the expected value of the product is equal to the product of the expected values (→ I/1.10.7):

$$\mathrm{E}(X\,Y) = \mathrm{E}(X)\,\mathrm{E}(Y) \; . \tag{3}$$

Taking (2) and (3) together, we have

$$\begin{split} \mathrm{Cov}(X,Y) &\overset{(2)}{=} \mathrm{E}(X\,Y) - \mathrm{E}(X)\,\mathrm{E}(Y) \\ &\overset{(3)}{=} \mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(X)\,\mathrm{E}(Y) \\ &= 0 \; . \end{split} \tag{4}$$

■

Sources:
- Wikipedia (2020): "Covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness_and_independence.

1.13.7 Relationship to correlation

Theorem: Let X and Y be random variables (→ I/1.2.2). Then, the covariance (→ I/1.13.1) of X and Y is equal to the product of their correlation (→ I/1.14.1) and the standard deviations (→ I/1.16.1) of X and Y:

$$\mathrm{Cov}(X,Y) = \sigma_X \, \mathrm{Corr}(X,Y) \, \sigma_Y \; . \tag{1}$$

Proof: The correlation (→ I/1.14.1) of X and Y is defined as

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \; . \tag{2}$$

which can be rearranged for the covariance (→ I/1.13.1) to give

$$\mathrm{Cov}(X,Y) = \sigma_X \, \mathrm{Corr}(X,Y) \, \sigma_Y \tag{3}$$

■

1.13.8 Law of total covariance

Theorem: (law of total covariance, also called "conditional covariance formula") Let X, Y and Z be random variables (→ I/1.2.2) defined on the same probability space (→ I/1.1.4) and assume that the covariance (→ I/1.13.1) of X and Y is finite. Then, the sum of the expectation (→ I/1.10.1) of the conditional covariance and the covariance (→ I/1.13.1) of the conditional expectations of X and Y given Z is equal to the covariance (→ I/1.13.1) of X and Y:

$$\mathrm{Cov}(X,Y) = \mathrm{E}[\mathrm{Cov}(X,Y|Z)] + \mathrm{Cov}[\mathrm{E}(X|Z), \mathrm{E}(Y|Z)] \; . \tag{1}$$

Proof: The covariance can be decomposed into expected values (→ I/1.13.3) as follows:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \; . \tag{2}$$

Then, conditioning on Z and applying the law of total expectation (→ I/1.10.11), we have:

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[\mathrm{E}(XY|Z)\right] - \mathrm{E}\left[\mathrm{E}(X|Z)\right]\mathrm{E}\left[\mathrm{E}(Y|Z)\right] \; . \tag{3}$$

Applying the decomposition of covariance into expected values (→ I/1.13.3) to the first term gives:

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[\mathrm{Cov}(X,Y|Z) + \mathrm{E}(X|Z)\mathrm{E}(Y|Z)\right] - \mathrm{E}\left[\mathrm{E}(X|Z)\right]\mathrm{E}\left[\mathrm{E}(Y|Z)\right] \; . \tag{4}$$

With the linearity of the expected value (→ I/1.10.5), the terms can be regrouped to give:

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[\mathrm{Cov}(X,Y|Z)\right] + \left(\mathrm{E}\left[\mathrm{E}(X|Z)\mathrm{E}(Y|Z)\right] - \mathrm{E}\left[\mathrm{E}(X|Z)\right]\mathrm{E}\left[\mathrm{E}(Y|Z)\right]\right) \; . \tag{5}$$

Once more using the decomposition of covariance into expected values (→ I/1.13.3), we finally have:

$$\mathrm{Cov}(X,Y) = \mathrm{E}[\mathrm{Cov}(X,Y|Z)] + \mathrm{Cov}[\mathrm{E}(X|Z),\mathrm{E}(Y|Z)] \; . \tag{6}$$

■

Sources:
- Wikipedia (2021): "Law of total covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-26; URL: https://en.wikipedia.org/wiki/Law_of_total_covariance#Proof.

1.13.9 Covariance matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ be a random vector (→ I/1.2.3). Then, the covariance matrix of X is defined as the $n \times n$ matrix in which the entry (i,j) is the covariance (→ I/1.13.1) of X_i and X_j:

$$\Sigma_{XX} = \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \ldots & \mathrm{Cov}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \ldots & \mathrm{Cov}(X_n,X_n) \end{bmatrix} = \begin{bmatrix} \mathrm{E}\,[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])] & \ldots & \mathrm{E}\,[(X_1-\mathrm{E}[X_1])(X_n-\mathrm{E} \\ \vdots & \ddots & \vdots \\ \mathrm{E}\,[(X_n-\mathrm{E}[X_n])(X_1-\mathrm{E}[X_1])] & \ldots & \mathrm{E}\,[(X_n-\mathrm{E}[X_n])(X_n-\mathrm{E} \end{bmatrix} \tag{1}$$

Sources:
- Wikipedia (2020): "Covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Definition.

1.13.10 Sample covariance matrix

Definition: Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random vector (→ I/1.2.3) $X \in \mathbb{R}^{p \times 1}$. Then, the sample covariance matrix of x is given by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\mathrm{T} \tag{1}$$

and the unbiased sample covariance matrix of x is given by

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\mathrm{T} \tag{2}$$

where $\bar{x}$ is the sample mean (→ I/1.10.2).

Sources:
- Wikipedia (2021): "Sample mean and covariance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-20; URL: https://en.wikipedia.org/wiki/Sample_mean_and_covariance#Definition_of_sample_covariance.

1.13.11 Covariance matrix and expected values

Theorem: Let X be a random vector (→ I/1.2.3). Then, the covariance matrix (→ I/1.13.9) of X is equal to the mean (→ I/1.10.1) of the outer product of X with itself minus the outer product of the mean (→ I/1.10.1) of X with itself:

$$\Sigma_{XX} = \mathrm{E}(XX^{\mathrm{T}}) - \mathrm{E}(X)\mathrm{E}(X)^{\mathrm{T}} \; . \tag{1}$$

Proof: The covariance matrix (→ I/1.13.9) of X is defined as

$$\Sigma_{XX} = \begin{bmatrix} \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])\right] & \ldots & \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])(X_n - \mathrm{E}[X_n])\right] \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left[(X_n - \mathrm{E}[X_n])(X_1 - \mathrm{E}[X_1])\right] & \ldots & \mathrm{E}\left[(X_n - \mathrm{E}[X_n])(X_n - \mathrm{E}[X_n])\right] \end{bmatrix} \tag{2}$$

which can also be expressed using matrix multiplication as

$$\Sigma_{XX} = \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}}\right] \tag{3}$$

Due to the linearity of the expected value (→ I/1.10.5), this can be rewritten as

$$\begin{split} \Sigma_{XX} &= \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}}\right] \\ &= \mathrm{E}\left[XX^{\mathrm{T}} - X\,\mathrm{E}(X)^{\mathrm{T}} - \mathrm{E}(X)\,X^{\mathrm{T}} + \mathrm{E}(X)\mathrm{E}(X)^{\mathrm{T}}\right] \\ &= \mathrm{E}(XX^{\mathrm{T}}) - \mathrm{E}(X)\mathrm{E}(X)^{\mathrm{T}} - \mathrm{E}(X)\mathrm{E}(X)^{\mathrm{T}} + \mathrm{E}(X)\mathrm{E}(X)^{\mathrm{T}} \\ &= \mathrm{E}(XX^{\mathrm{T}}) - \mathrm{E}(X)\mathrm{E}(X)^{\mathrm{T}} \; . \end{split} \tag{4}$$

■

Sources:
- Taboga, Marco (2010): "Covariance matrix"; in: *Lectures on probability and statistics*, retrieved on 2020-06-06; URL: https://www.statlect.com/fundamentals-of-probability/covariance-matrix.

1.13.12 Symmetry

Theorem: Each covariance matrix (→ I/1.13.9) is symmetric:

$$\Sigma_{XX}^{\mathrm{T}} = \Sigma_{XX} \; . \tag{1}$$

Proof: The covariance matrix (→ I/1.13.9) of a random vector (→ I/1.2.3) X is defined as

$$\Sigma_{XX} = \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \ldots & \mathrm{Cov}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \ldots & \mathrm{Cov}(X_n,X_n) \end{bmatrix} . \tag{2}$$

A symmetric matrix is a matrix whose transpose is equal to itself. The transpose of Σ_{XX} is

$$\Sigma_{XX}^{\mathrm{T}} = \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \ldots & \mathrm{Cov}(X_n,X_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1,X_n) & \ldots & \mathrm{Cov}(X_n,X_n) \end{bmatrix} . \tag{3}$$

Because the covariance is a symmetric function (→ I/1.13.4), i.e. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$, this matrix is equal to

$$\Sigma_{XX}^{\mathrm{T}} = \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \ldots & \mathrm{Cov}(X_1,X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \ldots & \mathrm{Cov}(X_n,X_n) \end{bmatrix} \tag{4}$$

which is equivalent to our original definition in (2).

■

Sources:

- Wikipedia (2022): "Covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties.

1.13.13 Positive semi-definiteness

Theorem: Each covariance matrix (→ I/1.13.9) is positive semi-definite:

$$a^{\mathrm{T}} \Sigma_{XX} a \geq 0 \quad \text{for all} \quad a \in \mathbb{R}^n . \tag{1}$$

Proof: The covariance matrix (→ I/1.13.9) of X can be expressed (→ I/1.13.11) in terms of expected values (→ I/1.10.1) as follows

$$\Sigma_{XX} = \Sigma(X) = \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}} \right] \tag{2}$$

A positive semi-definite matrix is a matrix whose eigenvalues are all non-negative or, equivalently,

$$M \text{ pos. semi-def.} \quad \Leftrightarrow \quad x^{\mathrm{T}} M x \geq 0 \quad \text{for all} \quad x \in \mathbb{R}^n . \tag{3}$$

Here, for an arbitrary real column vector $a \in \mathbb{R}^n$, we have:

$$a^{\mathrm{T}} \Sigma_{XX} a \overset{(2)}{=} a^{\mathrm{T}} \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}} \right] a . \tag{4}$$

Because the expected value is a linear operator (→ I/1.10.5), we can write:

$$a^{\mathrm{T}} \Sigma_{XX} a = \mathrm{E}\left[a^{\mathrm{T}} (X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}} a \right] . \tag{5}$$

Now define the scalar random variable (→ I/1.2.2)

$$Y = a^{\mathrm{T}}(X - \mu_X) \; . \tag{6}$$

where $\mu_X = \mathrm{E}[X]$ and note that

$$a^{\mathrm{T}}(X - \mu_X) = (X - \mu_X)^{\mathrm{T}} a \; . \tag{7}$$

Thus, combing (5) with (6), we have:

$$a^{\mathrm{T}} \Sigma_{XX} a = \mathrm{E}\left[Y^2\right] \; . \tag{8}$$

Because Y^2 is a random variable that cannot become negative and the expected value of a strictly non-negative random variable is also non-negative (→ I/1.10.4), we finally have

$$a^{\mathrm{T}} \Sigma_{XX} a \geq 0 \tag{9}$$

for any $a \in \mathbb{R}^n$.

■

Sources:
- hkBattousai (2013): "What is the proof that covariance matrices are always semi-definite?"; in: *StackExchange Mathematics*, retrieved on 2022-09-26; URL: https://math.stackexchange.com/a/327872.
- Wikipedia (2022): "Covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties.

1.13.14 Invariance under addition of vector

Theorem: The covariance matrix (→ I/1.13.9) Σ_{XX} of a random vector (→ I/1.2.3) X is invariant under addition of a constant vector (→ I/1.2.5) a:

$$\Sigma(X + a) = \Sigma(X) \; . \tag{1}$$

Proof: The covariance matrix (→ I/1.13.9) of X can be expressed (→ I/1.13.11) in terms of expected values (→ I/1.10.1) as follows:

$$\Sigma_{XX} = \Sigma(X) = \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}}\right] \; . \tag{2}$$

Using this and the linearity of the expected value (→ I/1.10.5), we can derive (1) as follows:

$$\begin{split} \Sigma(X + a) &\overset{(2)}{=} \mathrm{E}\left[([X + a] - \mathrm{E}[X + a])([X + a] - \mathrm{E}[X + a])^{\mathrm{T}}\right] \\ &= \mathrm{E}\left[(X + a - \mathrm{E}[X] - a)(X + a - \mathrm{E}[X] - a)^{\mathrm{T}}\right] \\ &= \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^{\mathrm{T}}\right] \\ &\overset{(2)}{=} \Sigma(X) \; . \end{split} \tag{3}$$

■

Sources:
- Wikipedia (2022): "Covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-22; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties.

1.13.15 Scaling upon multiplication with matrix

Theorem: The covariance matrix (→ I/1.13.9) Σ_{XX} of a random vector (→ I/1.2.3) X scales upon multiplication with a constant matrix (→ I/1.2.5) A:

$$\Sigma(AX) = A\,\Sigma(X)A^\mathrm{T} \; . \tag{1}$$

Proof: The covariance matrix (→ I/1.13.9) of X can be expressed (→ I/1.13.11) in terms of expected values (→ I/1.10.1) as follows:

$$\Sigma_{XX} = \Sigma(X) = \mathrm{E}\left[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^\mathrm{T} \right] \; . \tag{2}$$

Using this and the linearity of the expected value (→ I/1.10.5), we can derive (1) as follows:

$$\begin{split} \Sigma(AX) &\overset{(2)}{=} \mathrm{E}\left[([AX]-\mathrm{E}[AX])([AX]-\mathrm{E}[AX])^\mathrm{T} \right] \\ &= \mathrm{E}\left[(A[X-\mathrm{E}[X]])(A[X-\mathrm{E}[X]])^\mathrm{T} \right] \\ &= \mathrm{E}\left[A(X-\mathrm{E}[X])(X-\mathrm{E}[X])^\mathrm{T} A^\mathrm{T} \right] \\ &= A\,\mathrm{E}\left[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^\mathrm{T} \right] A^\mathrm{T} \\ &\overset{(2)}{=} A\,\Sigma(X)A^\mathrm{T} \; . \end{split} \tag{3}$$

■

Sources:
- Wikipedia (2022): "Covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-22; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties.

1.13.16 Cross-covariance matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ and $Y = [Y_1, \ldots, Y_m]^\mathrm{T}$ be two random vectors (→ I/1.2.3) that can or cannot be of equal size. Then, the cross-covariance matrix of X and Y is defined as the $n \times m$ matrix in which the entry (i,j) is the covariance (→ I/1.13.1) of X_i and Y_j:

$$\Sigma_{XY} = \begin{bmatrix} \mathrm{Cov}(X_1,Y_1) & \ldots & \mathrm{Cov}(X_1,Y_m) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,Y_1) & \ldots & \mathrm{Cov}(X_n,Y_m) \end{bmatrix} = \begin{bmatrix} \mathrm{E}\left[(X_1-\mathrm{E}[X_1])(Y_1-\mathrm{E}[Y_1])\right] & \ldots & \mathrm{E}\left[(X_1-\mathrm{E}[X_1])(Y_m-\mathrm{E}[Y \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left[(X_n-\mathrm{E}[X_n])(Y_1-\mathrm{E}[Y_1])\right] & \ldots & \mathrm{E}\left[(X_n-\mathrm{E}[X_n])(Y_m-\mathrm{E}[Y \end{bmatrix} \tag{1}$$

Sources:
- Wikipedia (2022): "Cross-covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: https://en.wikipedia.org/wiki/Cross-covariance_matrix#Definition.

1.13.17 Covariance matrix of a sum

Theorem: The covariance matrix (→ I/1.13.9) of the sum of two random vectors (→ I/1.2.3) of the same dimension equals the sum of the covariances of those random vectors, plus the sum of their cross-covariances (→ I/1.13.16):

$$\Sigma(X+Y) = \Sigma_{XX} + \Sigma_{YY} + \Sigma_{XY} + \Sigma_{YX} \; . \tag{1}$$

Proof: The covariance matrix (→ I/1.13.9) of X can be expressed (→ I/1.13.11) in terms of expected values (→ I/1.10.1) as follows

$$\Sigma_{XX} = \Sigma(X) = \mathrm{E}\left[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^\mathrm{T}\right] \tag{2}$$

and the cross-covariance matrix (→ I/1.13.16) of X and Y can similarly be written as

$$\Sigma_{XY} = \Sigma(X,Y) = \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])^\mathrm{T}\right] \tag{3}$$

Using this and the linearity of the expected value (→ I/1.10.5) as well as the definitions of covariance matrix (→ I/1.13.9) and cross-covariance matrix (→ I/1.13.16), we can derive (1) as follows:

$$\begin{split}
\Sigma(X+Y) &\overset{(2)}{=} \mathrm{E}\left[([X+Y]-\mathrm{E}[X+Y])([X+Y]-\mathrm{E}[X+Y])^\mathrm{T}\right] \\
&= \mathrm{E}\left[([X-\mathrm{E}(X)]+[Y-\mathrm{E}(Y)])([X-\mathrm{E}(X)]+[Y-\mathrm{E}(Y)])^\mathrm{T}\right] \\
&= \mathrm{E}\big[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^\mathrm{T} + (X-\mathrm{E}[X])(Y-\mathrm{E}[Y])^\mathrm{T} + (Y-\mathrm{E}[Y])(X-\mathrm{E}[X])^\mathrm{T} + (Y-\mathrm{E}[Y]) \\
&= \mathrm{E}\left[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^\mathrm{T}\right] + \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])^\mathrm{T}\right] + \mathrm{E}\left[(Y-\mathrm{E}[Y])(X-\mathrm{E}[X])^\mathrm{T}\right] + \mathrm{E} \\
&\overset{(2)}{=} \Sigma_{XX} + \Sigma_{YY} + \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])^\mathrm{T}\right] + \mathrm{E}\left[(Y-\mathrm{E}[Y])(X-\mathrm{E}[X])^\mathrm{T}\right] \\
&\overset{(3)}{=} \Sigma_{XX} + \Sigma_{YY} + \Sigma_{XY} + \Sigma_{YX} \; .
\end{split} \tag{4}$$

■

Sources:

- Wikipedia (2022): "Covariance matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-26; URL: https://en.wikipedia.org/wiki/Covariance_matrix#Basic_properties.

1.13.18 Covariance matrix and correlation matrix

Theorem: Let X be a random vector (→ I/1.2.3). Then, the covariance matrix (→ I/1.13.9) Σ_{XX} of X can be expressed in terms of its correlation matrix (→ I/1.14.5) C_{XX} as follows

$$\Sigma_{XX} = \mathrm{D}_X \cdot \mathrm{C}_{XX} \cdot \mathrm{D}_X \; , \tag{1}$$

where D_X is a diagonal matrix with the standard deviations (→ I/1.16.1) of $X_1, \ldots, X_n$ as entries on the diagonal:

$$\mathrm{D}_X = \mathrm{diag}(\sigma_{X_1}, \ldots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} \; . \tag{2}$$

Proof: Reiterating (1) and applying (2), we have:

$$\Sigma_{XX} = \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} \cdot \mathrm{C}_{XX} \cdot \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} . \tag{3}$$

Together with the definition of the correlation matrix (→ I/1.14.5), this gives

$$\begin{split} \Sigma_{XX} &= \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} \cdot \begin{bmatrix} \frac{\mathrm{E}[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])]}{\sigma_{X_1}\,\sigma_{X_1}} & \ldots & \frac{\mathrm{E}[(X_1-\mathrm{E}[X_1])(X_n-\mathrm{E}[X_n])]}{\sigma_{X_1}\,\sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{E}[(X_n-\mathrm{E}[X_n])(X_1-\mathrm{E}[X_1])]}{\sigma_{X_n}\,\sigma_{X_1}} & \ldots & \frac{\mathrm{E}[(X_n-\mathrm{E}[X_n])(X_n-\mathrm{E}[X_n])]}{\sigma_{X_n}\,\sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma_{X_1}\cdot\mathrm{E}[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])]}{\sigma_{X_1}\,\sigma_{X_1}} & \ldots & \frac{\sigma_{X_1}\cdot\mathrm{E}[(X_1-\mathrm{E}[X_1])(X_n-\mathrm{E}[X_n])]}{\sigma_{X_1}\,\sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n}\cdot\mathrm{E}[(X_n-\mathrm{E}[X_n])(X_1-\mathrm{E}[X_1])]}{\sigma_{X_n}\,\sigma_{X_1}} & \ldots & \frac{\sigma_{X_n}\cdot\mathrm{E}[(X_n-\mathrm{E}[X_n])(X_n-\mathrm{E}[X_n])]}{\sigma_{X_n}\,\sigma_{X_n}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma_{X_1}\cdot\mathrm{E}[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])]\cdot\sigma_{X_1}}{\sigma_{X_1}\,\sigma_{X_1}} & \ldots & \frac{\sigma_{X_1}\cdot\mathrm{E}[(X_1-\mathrm{E}[X_1])(X_n-\mathrm{E}[X_n])]\cdot\sigma_{X_n}}{\sigma_{X_1}\,\sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\sigma_{X_n}\cdot\mathrm{E}[(X_n-\mathrm{E}[X_n])(X_1-\mathrm{E}[X_1])]\cdot\sigma_{X_1}}{\sigma_{X_n}\,\sigma_{X_1}} & \ldots & \frac{\sigma_{X_n}\cdot\mathrm{E}[(X_n-\mathrm{E}[X_n])(X_n-\mathrm{E}[X_n])]\cdot\sigma_{X_n}}{\sigma_{X_n}\,\sigma_{X_n}} \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\left[(X_1-\mathrm{E}[X_1])(X_1-\mathrm{E}[X_1])\right] & \ldots & \mathrm{E}\left[(X_1-\mathrm{E}[X_1])(X_n-\mathrm{E}[X_n])\right] \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left[(X_n-\mathrm{E}[X_n])(X_1-\mathrm{E}[X_1])\right] & \ldots & \mathrm{E}\left[(X_n-\mathrm{E}[X_n])(X_n-\mathrm{E}[X_n])\right] \end{bmatrix} \end{split} \tag{4}$$

which is nothing else than the definition of the covariance matrix (→ I/1.13.9).

■

Sources:
- Penny, William (2006): "The correlation matrix"; in: *Mathematics for Brain Imaging*, ch. 1.4.5, p. 28, eq. 1.60; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.

1.13.19 Precision matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ be a random vector (→ I/1.2.3). Then, the precision matrix of X is defined as the inverse of the covariance matrix (→ I/1.13.9) of X:

$$\Lambda_{XX} = \Sigma_{XX}^{-1} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \ldots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}^{-1} . \tag{1}$$

Sources:

- Wikipedia (2020): "Precision (statistics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Precision_(statistics).

1.13.20 Precision matrix and correlation matrix

Theorem: Let X be a random vector (→ I/1.2.3). Then, the precision matrix (→ I/1.13.19) Λ_{XX} of X can be expressed in terms of its correlation matrix (→ I/1.14.5) C_{XX} as follows

$$\Lambda_{XX} = \mathrm{D}_X^{-1} \cdot \mathrm{C}_{XX}^{-1} \cdot \mathrm{D}_X^{-1} \; , \tag{1}$$

where D_X^{-1} is a diagonal matrix with the inverse standard deviations (→ I/1.16.1) of $X_1, \ldots, X_n$ as entries on the diagonal:

$$\mathrm{D}_X^{-1} = \mathrm{diag}(1/\sigma_{X_1}, \ldots, 1/\sigma_{X_n}) = \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \; . \tag{2}$$

Proof: The precision matrix (→ I/1.13.19) is defined as the inverse of the covariance matrix (→ I/1.13.9)

$$\Lambda_{XX} = \Sigma_{XX}^{-1} \tag{3}$$

and the relation between covariance matrix and correlation matrix (→ I/1.13.18) is given by

$$\Sigma_{XX} = \mathrm{D}_X \cdot \mathrm{C}_{XX} \cdot \mathrm{D}_X \tag{4}$$

where

$$\mathrm{D}_X = \mathrm{diag}(\sigma_{X_1}, \ldots, \sigma_{X_n}) = \begin{bmatrix} \sigma_{X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_{X_n} \end{bmatrix} \; . \tag{5}$$

Using the matrix product property

$$(A \cdot B \cdot C)^{-1} = C^{-1} \cdot B^{-1} \cdot A^{-1} \tag{6}$$

and the diagonal matrix property

$$\mathrm{diag}(a_1, \ldots, a_n)^{-1} = \begin{bmatrix} a_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & a_n \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{a_n} \end{bmatrix} = \mathrm{diag}(1/a_1, \ldots, 1/a_n) \; , \tag{7}$$

we obtain

$$
\begin{aligned}
\Lambda_{XX} &\overset{(3)}{=} \Sigma_{XX}^{-1} \\
&\overset{(4)}{=} (\mathrm{D}_X \cdot \mathrm{C}_{XX} \cdot \mathrm{D}_X)^{-1} \\
&\overset{(6)}{=} \mathrm{D}_X^{-1} \cdot \mathrm{C}_{XX}^{-1} \cdot \mathrm{D}_X^{-1} \\
&\overset{(7)}{=} \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \cdot \mathrm{C}_{XX}^{-1} \cdot \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sigma_{X_n}} \end{bmatrix}
\end{aligned} \tag{8}
$$

which conforms to equation (1).

■

1.14 Correlation

1.14.1 Definition

Definition: The correlation of two random variables (→ I/1.2.2) X and Y, also called Pearson product-moment correlation coefficient (PPMCC), is defined as the ratio of the covariance (→ I/1.13.1) of X and Y relative to the product of their standard deviations (→ I/1.16.1):

$$
\mathrm{Corr}(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right]}{\sqrt{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]}\sqrt{\mathrm{E}\left[(Y-\mathrm{E}[Y])^2\right]}} \; . \tag{1}
$$

Sources:

- Wikipedia (2020): "Correlation and dependence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-06; URL: https://en.wikipedia.org/wiki/Correlation_and_dependence#Pearson's_product-mom coefficient.

1.14.2 Range

Theorem: Let X and Y be two random variables (→ I/1.2.2). Then, the correlation of X and Y is between and including -1 and $+1$:

$$
-1 \leq \mathrm{Corr}(X,Y) \leq +1 \; . \tag{1}
$$

Proof: Consider the variance (→ I/1.11.1) of X plus or minus Y, each divided by their standard deviations (→ I/1.16.1):

$$
\mathrm{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) \; . \tag{2}
$$

Because the variance is non-negative (→ I/1.11.4), this term is larger than or equal to zero:

$$
0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y} \right) \; . \tag{3}
$$

Using the variance of a linear combination (→ I/1.11.9), it can also be written as:

$$
\begin{aligned}
\mathrm{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) &= \mathrm{Var}\left(\frac{X}{\sigma_X}\right) + \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) \pm 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\
&= \frac{1}{\sigma_X^2}\mathrm{Var}(X) + \frac{1}{\sigma_Y^2}\mathrm{Var}(Y) \pm 2\,\frac{1}{\sigma_X \sigma_Y}\,\mathrm{Cov}(X,Y) \\
&= \frac{1}{\sigma_X^2}\sigma_X^2 + \frac{1}{\sigma_Y^2}\sigma_Y^2 \pm 2\,\frac{1}{\sigma_X \sigma_Y}\,\sigma_{XY} \; .
\end{aligned} \tag{4}
$$

Using the relationship between covariance and correlation (→ I/1.13.7), we have:

$$
\mathrm{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) = 1 + 1 + \pm 2\,\mathrm{Corr}(X,Y) \; . \tag{5}
$$

Thus, the combination of (3) with (5) yields

$$
0 \leq 2 \pm 2\,\mathrm{Corr}(X,Y) \tag{6}
$$

which is equivalent to

$$
-1 \leq \mathrm{Corr}(X,Y) \leq +1 \; . \tag{7}
$$

■

Sources:

- Dor Leventer (2021): "How can I simply prove that the pearson correlation coefficient is between -1 and 1?"; in: *StackExchange Mathematics*, retrieved on 2021-12-14; URL: https://math.stackexchange.com/a/4260655/480910.

1.14.3 Sample correlation coefficient

Definition: Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ be samples from random variables (→ I/1.2.2) X and Y. Then, the sample correlation coefficient of x and y is given by

$$
r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{1}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2).

Sources:

- Wikipedia (2021): "Pearson correlation coefficient"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-14; URL: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#For_a_sample.

1.14.4 Relationship to standard scores

Theorem: Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ be samples from random variables (→ I/1.2.2) X and Y. Then, the sample correlation coefficient (→ I/1.14.3) r_{xy} can be expressed in terms of the standard scores of x and y:

$$
r_{xy} = \frac{1}{n-1}\sum_{i=1}^n z_i^{(x)} \cdot z_i^{(y)} = \frac{1}{n-1}\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \tag{1}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2) and s_x and s_y are the sample variances (→ I/1.11.2).

Proof: The sample correlation coefficient (→ I/1.14.3) is defined as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \; . \tag{2}$$

Using the sample variances (→ I/1.11.2) of x and y, we can write:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(n-1)s_x^2} \sqrt{(n-1)s_y^2}} \; . \tag{3}$$

Rearranging the terms, we arrive at:

$$r_{xy} = \frac{1}{(n-1)\, s_x\, s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \; . \tag{4}$$

Further simplifying, the result is:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \; . \tag{5}$$

■

Sources:
- Wikipedia (2021): "Peason correlation coefficient"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-14; URL: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#For_a_sample.

1.14.5 Correlation matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ be a random vector (→ I/1.2.3). Then, the correlation matrix of X is defined as the $n \times n$ matrix in which the entry (i,j) is the correlation (→ I/1.14.1) of X_i and X_j:

$$\mathrm{C}_{XX} = \begin{bmatrix} \mathrm{Corr}(X_1, X_1) & \ldots & \mathrm{Corr}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Corr}(X_n, X_1) & \ldots & \mathrm{Corr}(X_n, X_n) \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])]}{\sigma_{X_1}\, \sigma_{X_1}} & \ldots & \frac{\mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_n - \mathrm{E}[X_n])]}{\sigma_{X_1}\, \sigma_{X_n}} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{E}[(X_n - \mathrm{E}[X_n])(X_1 - \mathrm{E}[X_1])]}{\sigma_{X_n}\, \sigma_{X_1}} & \ldots & \frac{\mathrm{E}[(X_n - \mathrm{E}[X_n])(X_n - \mathrm{E}[X_n])]}{\sigma_{X_n}\, \sigma_{X_n}} \end{bmatrix} \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Correlation and dependence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-06; URL: https://en.wikipedia.org/wiki/Correlation_and_dependence#Correlation_matrices.

1.14.6 Sample correlation matrix

Definition: Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random vector (→ I/1.2.3) $X \in \mathbb{R}^{p \times 1}$. Then, the sample correlation matrix of x is the matrix whose entries are the sample correlation coefficients (→ I/1.14.3) between pairs of entries of $x_1, \ldots, x_n$:

$$\mathrm{R}_{xx} = \begin{bmatrix} r_{x^{(1)},x^{(1)}} & \ldots & r_{x^{(1)},x^{(n)}} \\ \vdots & \ddots & \vdots \\ r_{x^{(n)},x^{(1)}} & \ldots & r_{x^{(n)},x^{(n)}} \end{bmatrix} \tag{1}$$

where the $r_{x^{(j)},x^{(k)}}$ is the sample correlation (→ I/1.14.3) between the j-th and the k-th entry of X given by

$$r_{x^{(j)},x^{(k)}} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}^{(j)})(x_{ik} - \bar{x}^{(k)})}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}^{(j)})^2} \sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}^{(k)})^2}} \tag{2}$$

in which $\bar{x}^{(j)}$ and $\bar{x}^{(k)}$ are the sample means (→ I/1.10.2)

$$\begin{aligned} \bar{x}^{(j)} &= \frac{1}{n} \sum_{i=1}^n x_{ij} \\ \bar{x}^{(k)} &= \frac{1}{n} \sum_{i=1}^n x_{ik} \; . \end{aligned} \tag{3}$$

1.15 Measures of central tendency

1.15.1 Median

Definition: The median of a sample or random variable is the value separating the higher half from the lower half of its values.

1) Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the median of x is

$$\mathrm{median}(x) = \begin{cases} x_{(n+1)/2} \; , & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) \; , & \text{if } n \text{ is even} \; , \end{cases} \tag{1}$$

i.e. the median is the “middle” number when all numbers are sorted from smallest to largest.

2) Let X be a continuous random variable (→ I/1.2.2) with cumulative distribution function (→ I/1.8.1) $F_X(x)$. Then, the median of X is

$$\mathrm{median}(X) = x, \quad \text{s.t.} \quad F_X(x) = \frac{1}{2} \; , \tag{2}$$

i.e. the median is the value at which the CDF is 1/2.

Sources:

- Wikipedia (2020): “Median”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-15; URL: https://en.wikipedia.org/wiki/Median.

1.15.2 Mode

Definition: The mode of a sample or random variable is the value which occurs most often or with largest probability among all its values.

1) Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the mode of x is the value which occurs most often in the list $x_1, \ldots, x_n$.

2) Let X be a random variable (→ I/1.2.2) with probability mass function (→ I/1.6.1) or probability density function (→ I/1.7.1) $f_X(x)$. Then, the mode of X is the the value which maximizes the PMF or PDF:

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Mode (statistics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-15; URL: https://en.wikipedia.org/wiki/Mode_(statistics).

1.16 Measures of statistical dispersion

1.16.1 Standard deviation

Definition: The standard deviation σ of a random variable (→ I/1.2.2) X with expected value (→ I/1.10.1) μ is defined as the square root of the variance (→ I/1.11.1), i.e.

$$\sigma(X) = \sqrt{\mathrm{E}\left[(X-\mu)^2\right]} \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Standard deviation"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Standard_deviation#Definition_of_population_values.

1.16.2 Full width at half maximum

Definition: Let X be a continuous random variable (→ I/1.2.2) with a unimodal probability density function (→ I/1.7.1) $f_X(x)$ and mode (→ I/1.15.2) x_M. Then, the full width at half maximum of X is defined as

$$\mathrm{FHWM}(X) = \Delta x = x_2 - x_1 \tag{1}$$

where x_1 and x_2 are specified, such that

$$f_X(x_1) = f_X(x_2) = \frac{1}{2} f_X(x_M) \quad \text{and} \quad x_1 < x_M < x_2 \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Full width at half maximum"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: https://en.wikipedia.org/wiki/Full_width_at_half_maximum.

1.17 Further summary statistics

1.17.1 Minimum

Definition: The minimum of a sample or random variable is its lowest observed or possible value.

1) Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the minimum of x is

$$\min(x) = x_j, \quad \text{such that} \quad x_j \leq x_i \quad \text{for all} \quad i = 1, \ldots, n,\; i \neq j \;, \tag{1}$$

i.e. the minimum is the value which is smaller than or equal to all other observed values.

2) Let X be a random variable (→ I/1.2.2) with possible values $\mathcal{X}$. Then, the minimum of X is

$$\min(X) = \tilde{x}, \quad \text{such that} \quad \tilde{x} < x \quad \text{for all} \quad x \in \mathcal{X} \setminus \{\tilde{x}\} \;, \tag{2}$$

i.e. the minimum is the value which is smaller than all other possible values.

Sources:
- Wikipedia (2020): "Sample maximum and minimum"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: https://en.wikipedia.org/wiki/Sample_maximum_and_minimum.

1.17.2 Maximum

Definition: The maximum of a sample or random variable is its highest observed or possible value.

1) Let $x = \{x_1, \ldots, x_n\}$ be a sample from a random variable (→ I/1.2.2) X. Then, the maximum of x is

$$\max(x) = x_j, \quad \text{such that} \quad x_j \geq x_i \quad \text{for all} \quad i = 1, \ldots, n,\; i \neq j \;, \tag{1}$$

i.e. the maximum is the value which is larger than or equal to all other observed values.

2) Let X be a random variable (→ I/1.2.2) with possible values $\mathcal{X}$. Then, the maximum of X is

$$\max(X) = \tilde{x}, \quad \text{such that} \quad \tilde{x} > x \quad \text{for all} \quad x \in \mathcal{X} \setminus \{\tilde{x}\} \;, \tag{2}$$

i.e. the maximum is the value which is larger than all other possible values.

Sources:
- Wikipedia (2020): "Sample maximum and minimum"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-12; URL: https://en.wikipedia.org/wiki/Sample_maximum_and_minimum.

1.18 Further moments

1.18.1 Moment

Definition: Let X be a random variable (→ I/1.2.2), let c be a constant (→ I/1.2.5) and let n be a positive integer. Then, the n-th moment of X about c is defined as the expected value (→ I/1.10.1) of the n-th power of X minus c:

$$\mu_n(c) = \mathrm{E}[(X-c)^n] \; . \tag{1}$$

The "n-th moment of X" may also refer to:
- the n-th raw moment (→ I/1.18.3) $\mu_n' = \mu_n(0)$;
- the n-th central moment (→ I/1.18.6) $\mu_n = \mu_n(\mu)$;
- the n-th standardized moment (→ I/1.18.9) $\mu_n^* = \mu_n/\sigma^n$.

Sources:
- Wikipedia (2020): "Moment (mathematics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments.

1.18.2 Moment in terms of moment-generating function

Theorem: Let X be a scalar random variable (→ I/1.2.2) with the moment-generating function (→ I/1.9.5) $M_X(t)$. Then, the n-th raw moment (→ I/1.18.3) of X can be calculated from the moment-generating function via

$$\mathrm{E}(X^n) = M_X^{(n)}(0) \tag{1}$$

where n is a positive integer and $M_X^{(n)}(t)$ is the n-th derivative of $M_X(t)$.

Proof: Using the definition of the moment-generating function (→ I/1.9.5), we can write:

$$M_X^{(n)}(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathrm{E}(e^{tX}) \; . \tag{2}$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \; , \tag{3}$$

equation (2) becomes

$$M_X^{(n)}(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathrm{E}\left(\sum_{m=0}^{\infty} \frac{t^m X^m}{m!} \right) \; . \tag{4}$$

Because the expected value is a linear operator (→ I/1.10.5), we have:

$$\begin{split} M_X^{(n)}(t) &= \frac{\mathrm{d}^n}{\mathrm{d}t^n} \sum_{m=0}^{\infty} \mathrm{E}\left(\frac{t^m X^m}{m!} \right) \\ &= \sum_{m=0}^{\infty} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \frac{t^m}{m!} \mathrm{E}\left(X^m\right) \; . \end{split} \tag{5}$$

Using the n-th derivative of the m-th power

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n} x^m = \left\{ \begin{array}{rl} m^{\underline{n}} x^{m-n} \; , & \text{if} \; n \leq m \\ 0 \; , & \text{if} \; n > m \; . \end{array} \right. \tag{6}$$

with the falling factorial

$$m^{\underline{n}} = \prod_{i=0}^{n-1}(m-i) = \frac{m!}{(m-n)!} \; , \tag{7}$$

equation (5) becomes

$$\begin{aligned}
M_X^{(n)}(t) &= \sum_{m=n}^{\infty} \frac{m^{\underline{n}}\, t^{m-n}}{m!} \mathrm{E}\left(X^m\right) \\
&\overset{(7)}{=} \sum_{m=n}^{\infty} \frac{m!\, t^{m-n}}{(m-n)!\, m!} \mathrm{E}\left(X^m\right) \\
&= \sum_{m=n}^{\infty} \frac{t^{m-n}}{(m-n)!} \mathrm{E}\left(X^m\right) \\
&= \frac{t^{n-n}}{(n-n)!} \mathrm{E}\left(X^n\right) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m-n)!} \mathrm{E}\left(X^m\right) \\
&= \frac{t^0}{0!} \mathrm{E}\left(X^n\right) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m-n)!} \mathrm{E}\left(X^m\right) \\
&= \mathrm{E}\left(X^n\right) + \sum_{m=n+1}^{\infty} \frac{t^{m-n}}{(m-n)!} \mathrm{E}\left(X^m\right) \; .
\end{aligned} \tag{8}$$

Setting $t = 0$ in (8) yields

$$\begin{aligned}
M_X^{(n)}(0) &= \mathrm{E}\left(X^n\right) + \sum_{m=n+1}^{\infty} \frac{0^{m-n}}{(m-n)!} \mathrm{E}\left(X^m\right) \\
&= \mathrm{E}\left(X^n\right)
\end{aligned} \tag{9}$$

which conforms to equation (1).

■

Sources:

- ProofWiki (2020): "Moment in terms of Moment Generating Function"; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Moment_in_terms_of_Moment_Generating_Function.

1.18.3 Raw moment

Definition: Let X be a random variable (→ I/1.2.2) and let n be a positive integer. Then, the n-th raw moment of X, also called (n-th) "crude moment", is defined as the n-th moment (→ I/1.18.1) of X about the value 0:

$$\mu_n' = \mu_n(0) = \mathrm{E}[(X-0)^n] = \mathrm{E}[X^n] \; . \tag{1}$$

Sources:

- Wikipedia (2020): "Moment (mathematics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments.

1.18.4 First raw moment is mean

Theorem: The first raw moment (→ I/1.18.3) equals the mean (→ I/1.10.1), i.e.

$$\mu_1' = \mu \; . \tag{1}$$

Proof: The first raw moment (→ I/1.18.3) of a random variable (→ I/1.2.2) X is defined as

$$\mu_1' = \mathrm{E}\left[(X-0)^1\right] \tag{2}$$

which is equal to the expected value (→ I/1.10.1) of X:

$$\mu_1' = \mathrm{E}\left[X\right] = \mu \; . \tag{3}$$

■

1.18.5 Second raw moment and variance

Theorem: The second raw moment (→ I/1.18.3) can be expressed as

$$\mu_2' = \mathrm{Var}(X) + \mathrm{E}(X)^2 \tag{1}$$

where $\mathrm{Var}(X)$ is the variance (→ I/1.11.1) of X and $\mathrm{E}(X)$ is the expected value (→ I/1.10.1) of X.

Proof: The second raw moment (→ I/1.18.3) of a random variable (→ I/1.2.2) X is defined as

$$\mu_2' = \mathrm{E}\left[(X-0)^2\right] \; . \tag{2}$$

Using the partition of variance into expected values (→ I/1.11.3)

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; , \tag{3}$$

the second raw moment can be rearranged into:

$$\mu_2' \overset{(2)}{=} \mathrm{E}(X^2) \overset{(3)}{=} \mathrm{Var}(X) + \mathrm{E}(X)^2 \; . \tag{4}$$

■

1.18.6 Central moment

Definition: Let X be a random variable (→ I/1.2.2) with expected value (→ I/1.10.1) μ and let n be a positive integer. Then, the n-th central moment of X is defined as the n-th moment (→ I/1.18.1) of X about the value μ:

$$\mu_n = \mathrm{E}[(X-\mu)^n] \; . \tag{1}$$

Sources:

- Wikipedia (2020): "Moment (mathematics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments.

1.18.7 First central moment is zero

Theorem: The first central moment (→ I/1.18.6) is zero, i.e.

$$\mu_1 = 0 \; . \tag{1}$$

Proof: The first central moment (→ I/1.18.6) of a random variable (→ I/1.2.2) X with mean (→ I/1.10.1) μ is defined as

$$\mu_1 = \mathrm{E}\left[(X-\mu)^1\right] \; . \tag{2}$$

Due to the linearity of the expected value (→ I/1.10.5) and by plugging in $\mu = \mathrm{E}(X)$, we have

$$\begin{split} \mu_1 &= \mathrm{E}\left[X-\mu\right] \\ &= \mathrm{E}(X) - \mu \\ &= \mathrm{E}(X) - \mathrm{E}(X) \\ &= 0 \; . \end{split} \tag{3}$$

■

Sources:
- ProofWiki (2020): "First Central Moment is Zero"; in: *ProofWiki*, retrieved on 2020-09-09; URL: https://proofwiki.org/wiki/First_Central_Moment_is_Zero.

1.18.8 Second central moment is variance

Theorem: The second central moment (→ I/1.18.6) equals the variance (→ I/1.11.1), i.e.

$$\mu_2 = \mathrm{Var}(X) \; . \tag{1}$$

Proof: The second central moment (→ I/1.18.6) of a random variable (→ I/1.2.2) X with mean (→ I/1.10.1) μ is defined as

$$\mu_2 = \mathrm{E}\left[(X-\mu)^2\right] \tag{2}$$

which is equivalent to the definition of the variance (→ I/1.11.1):

$$\mu_2 = \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] = \mathrm{Var}(X) \; . \tag{3}$$

■

Sources:
- Wikipedia (2020): "Moment (mathematics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: https://en.wikipedia.org/wiki/Moment_(mathematics)#Significance_of_the_moments.

1.18.9 Standardized moment

Definition: Let X be a random variable (→ I/1.2.2) with expected value (→ I/1.10.1) μ and standard deviation (→ I/1.16.1) σ and let n be a positive integer. Then, the n-th standardized moment of X is defined as the n-th moment (→ I/1.18.1) of X about the value μ, divided by the n-th power of σ:

$$\mu_n^* = \frac{\mu_n}{\sigma^n} = \frac{\mathrm{E}[(X-\mu)^n]}{\sigma^n} \; . \tag{1}$$

Sources:

- Wikipedia (2020): "Moment (mathematics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-08; URL: https://en.wikipedia.org/wiki/Moment_(mathematics)#Standardized_moments.

2 Information theory

2.1 Shannon entropy

2.1.1 Definition

Definition: Let X be a discrete random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and the (observed or assumed) probability mass function (→ I/1.6.1) $p(x) = f_X(x)$. Then, the entropy (also referred to as "Shannon entropy") of X is defined as

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \tag{1}$$

where b is the base of the logarithm specifying in which unit the entropy is determined. By convention, $0 \cdot \log 0$ is taken to be zero when calculating the entropy of X.

Sources:
- Shannon CE (1948): "A Mathematical Theory of Communication"; in: *Bell System Technical Journal*, vol. 27, iss. 3, pp. 379-423; URL: https://ieeexplore.ieee.org/document/6773024; DOI: 10.1002/j.1538-7305.1948.tb01338.x.

2.1.2 Non-negativity

Theorem: The entropy of a discrete random variable (→ I/1.2.2) is a non-negative number:

$$\mathrm{H}(X) \geq 0 \; . \tag{1}$$

Proof: The entropy of a discrete random variable (→ I/2.1.1) is defined as

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \tag{2}$$

The minus sign can be moved into the sum:

$$\mathrm{H}(X) = \sum_{x \in \mathcal{X}} \left[p(x) \cdot (-\log_b p(x)) \right] \tag{3}$$

Because the co-domain of probability mass functions (→ I/1.6.1) is $[0, 1]$, we can deduce:

$$\begin{array}{rcccl} 0 & \leq & p(x) & \leq & 1 \\ -\infty & \leq & \log_b p(x) & \leq & 0 \\ 0 & \leq & -\log_b p(x) & \leq & +\infty \\ 0 & \leq & p(x) \cdot (-\log_b p(x)) & \leq & +\infty \; . \end{array} \tag{4}$$

By convention, $0 \cdot \log_b(0)$ is taken to be 0 when calculating entropy, consistent with

$$\lim_{p \to 0} \left[p \log_b(p) \right] = 0 \; . \tag{5}$$

Taking this together, each addend in (3) is positive or zero and thus, the entire sum must also be non-negative.

■

Sources:
- Cover TM, Thomas JA (1991): "Elements of Information Theory", p. 15; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

2.1.3 Concavity

Theorem: The entropy (→ I/2.1.1) is concave in the probability mass function (→ I/1.6.1) p, i.e.

$$\mathrm{H}[\lambda p_1 + (1-\lambda)p_2] \geq \lambda \mathrm{H}[p_1] + (1-\lambda)\mathrm{H}[p_2] \tag{1}$$

where p_1 and p_2 are probability mass functions and $0 \leq \lambda \leq 1$.

Proof: Let X be a discrete random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let $u(x)$ be the probability mass function (→ I/1.6.1) of a discrete uniform distribution (→ II/1.1.1) on $X \in \mathcal{X}$. Then, the entropy (→ I/2.1.1) of an arbitrary probability mass function (→ I/1.6.1) $p(x)$ can be rewritten as

$$\begin{aligned}
\mathrm{H}[p] &= -\sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} u(x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \cdot \log u(x) \\
&= -\mathrm{KL}[p||u] - \log \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(x) \\
&= \log |\mathcal{X}| - \mathrm{KL}[p||u] \\
\log |\mathcal{X}| - \mathrm{H}[p] &= \mathrm{KL}[p||u]
\end{aligned} \tag{2}$$

where we have applied the definition of the Kullback-Leibler divergence (→ I/2.5.1), the probability mass function of the discrete uniform distribution (→ II/1.1.2) and the total sum over the probability mass function (→ I/1.6.1).
Note that the KL divergence is convex (→ I/2.5.5) in the pair of probability distributions (→ I/1.5.1) (p,q):

$$\mathrm{KL}[\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2] \leq \lambda \mathrm{KL}[p_1||q_1] + (1-\lambda)\mathrm{KL}[p_2||q_2] \tag{3}$$

A special case of this is given by

$$\begin{aligned}
\mathrm{KL}[\lambda p_1 + (1-\lambda)p_2||\lambda u + (1-\lambda)u] &\leq \lambda \mathrm{KL}[p_1||u] + (1-\lambda)\mathrm{KL}[p_2||u] \\
\mathrm{KL}[\lambda p_1 + (1-\lambda)p_2||u] &\leq \lambda \mathrm{KL}[p_1||u] + (1-\lambda)\mathrm{KL}[p_2||u]
\end{aligned} \tag{4}$$

and applying equation (2), we have

$$\begin{aligned}
\log |\mathcal{X}| - \mathrm{H}[\lambda p_1 + (1-\lambda)p_2] &\leq \lambda \left(\log |\mathcal{X}| - \mathrm{H}[p_1]\right) + (1-\lambda)\left(\log |\mathcal{X}| - \mathrm{H}[p_2]\right) \\
\log |\mathcal{X}| - \mathrm{H}[\lambda p_1 + (1-\lambda)p_2] &\leq \log |\mathcal{X}| - \lambda \mathrm{H}[p_1] - (1-\lambda)\mathrm{H}[p_2] \\
-\mathrm{H}[\lambda p_1 + (1-\lambda)p_2] &\leq -\lambda \mathrm{H}[p_1] - (1-\lambda)\mathrm{H}[p_2] \\
\mathrm{H}[\lambda p_1 + (1-\lambda)p_2] &\geq \lambda \mathrm{H}[p_1] + (1-\lambda)\mathrm{H}[p_2]
\end{aligned} \tag{5}$$

which is equivalent to (1).

■

Sources:
- Wikipedia (2020): "Entropy (information theory)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Entropy_(information_theory)#Further_properties.
- Cover TM, Thomas JA (1991): "Elements of Information Theory", p. 30; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.
- Xie, Yao (2012): "Chain Rules and Inequalities"; in: *ECE587: Information Theory*, Lecture 3, Slide 25; URL: https://www2.isye.gatech.edu/~yxie77/ece587/Lecture3.pdf.
- Goh, Siong Thye (2016): "Understanding the proof of the concavity of entropy"; in: *StackExchange Mathematics*, retrieved on 2020-11-08; URL: https://math.stackexchange.com/questions/2000194/understanding-the-proof-of-the-concavity-of-entropy.

2.1.4 Conditional entropy

Definition: Let X and Y be discrete random variables (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and probability mass functions (→ I/1.6.1) $p(x)$ and $p(y)$. Then, the conditional entropy of Y given X or, entropy of Y conditioned on X, is defined as

$$\mathrm{H}(Y|X) = \sum_{x \in \mathcal{X}} p(x) \cdot \mathrm{H}(Y|X = x) \tag{1}$$

where $\mathrm{H}(Y|X = x)$ is the (marginal) entropy (→ I/2.1.1) of Y, evaluated at x.

Sources:
- Cover TM, Thomas JA (1991): "Joint Entropy and Conditional Entropy"; in: *Elements of Information Theory*, ch. 2.2, p. 15; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

2.1.5 Joint entropy

Definition: Let X and Y be discrete random variables (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and joint probability (→ I/1.3.2) mass function (→ I/1.6.1) $p(x, y)$. Then, the joint entropy of X and Y is defined as

$$\mathrm{H}(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) \tag{1}$$

where b is the base of the logarithm specifying in which unit the entropy is determined.

Sources:
- Cover TM, Thomas JA (1991): "Joint Entropy and Conditional Entropy"; in: *Elements of Information Theory*, ch. 2.2, p. 16; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

2.1.6 Cross-entropy

Definition: Let X be a discrete random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ I/1.5.1) on X with the probability mass functions (→ I/1.6.1) $p(x)$ and $q(x)$. Then, the cross-entropy of Q relative to P is defined as

$$\mathrm{H}(P,Q) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b q(x) \tag{1}$$

where b is the base of the logarithm specifying in which unit the cross-entropy is determined.

Sources:
- Wikipedia (2020): "Cross entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: https://en.wikipedia.org/wiki/Cross_entropy#Definition.

2.1.7 Convexity of cross-entropy

Theorem: The cross-entropy (→ I/2.1.6) is convex in the probability distribution (→ I/1.5.1) q, i.e.

$$\mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] \leq \lambda \mathrm{H}[p, q_1] + (1-\lambda) \mathrm{H}[p, q_2] \tag{1}$$

where p is a fixed and q_1 and q_2 are any two probability distributions and $0 \leq \lambda \leq 1$.

Proof: The relationship between Kullback-Leibler divergence, entropy and cross-entropy (→ I/2.5.8) is:

$$\mathrm{KL}[P||Q] = \mathrm{H}(P,Q) - \mathrm{H}(P) \; . \tag{2}$$

Note that the KL divergence is convex (→ I/2.5.5) in the pair of probability distributions (→ I/1.5.1) (p,q):

$$\mathrm{KL}[\lambda p_1 + (1-\lambda) p_2 || \lambda q_1 + (1-\lambda) q_2] \leq \lambda \mathrm{KL}[p_1||q_1] + (1-\lambda) \mathrm{KL}[p_2||q_2] \tag{3}$$

A special case of this is given by

$$\begin{split} \mathrm{KL}[\lambda p + (1-\lambda) p || \lambda q_1 + (1-\lambda) q_2] &\leq \lambda \mathrm{KL}[p||q_1] + (1-\lambda) \mathrm{KL}[p||q_2] \\ \mathrm{KL}[p || \lambda q_1 + (1-\lambda) q_2] &\leq \lambda \mathrm{KL}[p||q_1] + (1-\lambda) \mathrm{KL}[p||q_2] \end{split} \tag{4}$$

and applying equation (2), we have

$$\begin{split} \mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] - \mathrm{H}[p] &\leq \lambda \left(\mathrm{H}[p, q_1] - \mathrm{H}[p] \right) + (1-\lambda) \left(\mathrm{H}[p, q_2] - \mathrm{H}[p] \right) \\ \mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] - \mathrm{H}[p] &\leq \lambda \mathrm{H}[p, q_1] + (1-\lambda) \mathrm{H}[p, q_2] - \mathrm{H}[p] \\ \mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] &\leq \lambda \mathrm{H}[p, q_1] + (1-\lambda) \mathrm{H}[p, q_2] \end{split} \tag{5}$$

which is equivalent to (1).

■

Sources:
- Wikipedia (2020): "Cross entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Cross_entropy#Definition.
- gunes (2019): "Convexity of cross entropy"; in: *StackExchange CrossValidated*, retrieved on 2020-11-08; URL: https://stats.stackexchange.com/questions/394463/convexity-of-cross-entropy.

2.1.8 Gibbs' inequality

Theorem: Let X be a discrete random variable (→ I/1.2.2) and consider two probability distributions (→ I/1.5.1) with probability mass functions (→ I/1.6.1) $p(x)$ and $q(x)$. Then, Gibbs' inequality states that the entropy (→ I/2.1.1) of X according to P is smaller than or equal to the cross-entropy (→ I/2.1.6) of P and Q:

$$-\sum_{x \in \mathcal{X}} p(x) \, \log_b p(x) \leq -\sum_{x \in \mathcal{X}} p(x) \, \log_b q(x) \; . \tag{1}$$

Proof: Without loss of generality, we will use the natural logarithm, because a change in the base of the logarithm only implies multiplication by a constant:

$$\log_b a = \frac{\ln a}{\ln b} \; . \tag{2}$$

Let I be the set of all x for which $p(x)$ is non-zero. Then, proving (1) requires to show that

$$\sum_{x \in I} p(x) \, \ln \frac{p(x)}{q(x)} \geq 0 \; . \tag{3}$$

For all $x > 0$, it holds that $\ln x \leq x - 1$, with equality only if $x = 1$. Multiplying this with -1, we have $\ln \frac{1}{x} \geq 1 - x$. Applying this to (3), we can say about the left-hand side that

$$\begin{split} \sum_{x \in I} p(x) \, \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) \left(1 - \frac{q(x)}{p(x)} \right) \\ &= \sum_{x \in I} p(x) - \sum_{x \in I} q(x) \; . \end{split} \tag{4}$$

Finally, since $p(x)$ and $q(x)$ are probability mass functions (→ I/1.6.1), we have

$$\begin{split} 0 \leq p(x) \leq 1, \quad &\sum_{x \in I} p(x) = 1 \quad \text{and} \\ 0 \leq q(x) \leq 1, \quad &\sum_{x \in I} q(x) \leq 1 \; , \end{split} \tag{5}$$

such that it follows from (4) that

$$\begin{split} \sum_{x \in I} p(x) \, \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) - \sum_{x \in I} q(x) \\ &= 1 - \sum_{x \in I} q(x) \geq 0 \; . \end{split} \tag{6}$$

■

Sources:

• Wikipedia (2020): "Gibbs' inequality"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Gibbs%27_inequality#Proof.

2.1.9 Log sum inequality

Theorem: Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be non-negative real numbers and define $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$. Then, the log sum inequality states that

$$\sum_{i=1}^n a_i \, \log_c \frac{a_i}{b_i} \geq a \, \log_c \frac{a}{b} \; . \tag{1}$$

Proof: Without loss of generality, we will use the natural logarithm, because a change in the base of the logarithm only implies multiplication by a constant:

$$\log_c a = \frac{\ln a}{\ln c} \; . \tag{2}$$

Let $f(x) = x \ln x$. Then, the left-hand side of (1) can be rewritten as

$$\begin{aligned} \sum_{i=1}^n a_i \, \ln \frac{a_i}{b_i} &= \sum_{i=1}^n b_i \, f\left(\frac{a_i}{b_i}\right) \\ &= b \sum_{i=1}^n \frac{b_i}{b} \, f\left(\frac{a_i}{b_i}\right) \; . \end{aligned} \tag{3}$$

Because $f(x)$ is a convex function and

$$\begin{aligned} \frac{b_i}{b} &\geq 0 \\ \sum_{i=1}^n \frac{b_i}{b} &= 1 \; , \end{aligned} \tag{4}$$

applying Jensen's inequality yields

$$\begin{aligned} b \sum_{i=1}^n \frac{b_i}{b} \, f\left(\frac{a_i}{b_i}\right) &\geq b \, f\left(\sum_{i=1}^n \frac{b_i}{b} \frac{a_i}{b_i}\right) \\ &= b \, f\left(\frac{1}{b} \sum_{i=1}^n a_i\right) \\ &= b \, f\left(\frac{a}{b}\right) \\ &= a \, \ln \frac{a}{b} \; . \end{aligned} \tag{5}$$

Finally, combining (3) and (5), this demonstrates (1).

■

Sources:

- Wikipedia (2020): "Log sum inequality"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Log_sum_inequality#Proof.
- Wikipedia (2020): "Jensen's inequality"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Jensen%27s_inequality#Statements.

2.2 Differential entropy

2.2.1 Definition

Definition: Let X be a continuous random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and the (estimated or assumed) probability density function (→ I/1.7.1) $p(x) = f_X(x)$. Then, the differential entropy (also referred to as "continuous entropy") of X is defined as

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \log_b p(x) \, \mathrm{d}x \tag{1}$$

where b is the base of the logarithm specifying in which unit the entropy is determined.

Sources:

- Cover TM, Thomas JA (1991): "Differential Entropy"; in: *Elements of Information Theory*, ch. 8.1, p. 243; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

2.2.2 Negativity

Theorem: Unlike its discrete analogue (→ I/2.1.2), the differential entropy (→ I/2.2.1) can become negative.

Proof: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1) with minimum 0 and maximum 1/2:

$$X \sim \mathcal{U}(0, 1/2) \; . \tag{1}$$

Then, its probability density function (→ II/3.1.3) is:

$$f_X(x) = 2 \quad \text{for} \quad 0 \leq x \leq \frac{1}{2} \; . \tag{2}$$

Thus, the differential entropy (→ I/2.2.1) follows as

$$\begin{split} \mathrm{h}(X) &= -\int_{\mathcal{X}} f_X(x) \log_b f_X(x) \, \mathrm{d}x \\ &= -\int_0^{\frac{1}{2}} 2 \, \log_b(2) \, \mathrm{d}x \\ &= -\log_b(2) \int_0^{\frac{1}{2}} 2 \, \mathrm{d}x \\ &= -\log_b(2) \, [2x]_0^{\frac{1}{2}} \\ &= -\log_b(2) \end{split} \tag{3}$$

which is negative for any base $b > 1$.

■

Sources:

- Wikipedia (2020): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-02; URL: https://en.wikipedia.org/wiki/Differential_entropy#Definition.

2.2.3 Invariance under addition

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2). Then, the differential entropy (→ I/2.2.1) of X remains constant under addition of a constant:

$$\mathrm{h}(X + c) = \mathrm{h}(X) \; . \tag{1}$$

Proof: By definition, the differential entropy (→ I/2.2.1) of X is

$$\mathrm{h}(X) = - \int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \tag{2}$$

where $p(x) = f_X(x)$ is the probability density function (→ I/1.7.1) of X.
Define the mappings between X and $Y = X + c$ as

$$Y = g(X) = X + c \quad \Leftrightarrow \quad X = g^{-1}(Y) = Y - c \; . \tag{3}$$

Note that $g(X)$ is a strictly increasing function, such that the probability density function (→ I/1.7.3) of Y is

$$f_Y(y) = f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \overset{(3)}{=} f_X(y - c) \; . \tag{4}$$

Writing down the differential entropy for Y, we have:

$$\begin{split} \mathrm{h}(Y) &= - \int_{\mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y \\ &\overset{(4)}{=} - \int_{\mathcal{Y}} f_X(y - c) \log f_X(y - c) \, \mathrm{d}y \end{split} \tag{5}$$

Substituting $x = y - c$, such that $y = x + c$, this yields:

$$\begin{split} \mathrm{h}(Y) &= - \int_{\{y - c \,|\, y \in \mathcal{Y}\}} f_X(x + c - c) \log f_X(x + c - c) \, \mathrm{d}(x + c) \\ &= - \int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x \\ &\overset{(2)}{=} \mathrm{h}(X) \; . \end{split} \tag{6}$$

■

Sources:

- Wikipedia (2020): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-12; URL: https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy.

2.2.4 Addition upon multiplication

Theorem: Let X be a continuous (→ I/1.2.6) random variable (→ I/1.2.2). Then, the differential entropy (→ I/2.2.1) of X increases additively upon multiplication with a constant:

$$\mathrm{h}(aX) = \mathrm{h}(X) + \log |a| \; . \tag{1}$$

Proof: By definition, the differential entropy (→ I/2.2.1) of X is

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \tag{2}$$

where $p(x) = f_X(x)$ is the probability density function (→ I/1.7.1) of X.
Define the mappings between X and $Y = aX$ as

$$Y = g(X) = aX \quad \Leftrightarrow \quad X = g^{-1}(Y) = \frac{Y}{a} \; . \tag{3}$$

If $a > 0$, then $g(X)$ is a strictly increasing function, such that the probability density function (→ I/1.7.3) of Y is

$$f_Y(y) = f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \overset{(3)}{=} \frac{1}{a} \, f_X\left(\frac{y}{a}\right) \; ; \tag{4}$$

if $a < 0$, then $g(X)$ is a strictly decreasing function, such that the probability density function (→ I/1.7.4) of Y is

$$f_Y(y) = -f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \overset{(3)}{=} -\frac{1}{a} \, f_X\left(\frac{y}{a}\right) \; ; \tag{5}$$

thus, we can write

$$f_Y(y) = \frac{1}{|a|} \, f_X\left(\frac{y}{a}\right) \; . \tag{6}$$

Writing down the differential entropy for Y, we have:

$$\begin{split} \mathrm{h}(Y) &= -\int_{\mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y \\ &\overset{(6)}{=} -\int_{\mathcal{Y}} \frac{1}{|a|} \, f_X\left(\frac{y}{a}\right) \log\left[\frac{1}{|a|} \, f_X\left(\frac{y}{a}\right)\right] \, \mathrm{d}y \end{split} \tag{7}$$

Substituting $x = y/a$, such that $y = ax$, this yields:

$$\begin{split} \mathrm{h}(Y) &= -\int_{\{y/a \,|\, y \in \mathcal{Y}\}} \frac{1}{|a|} \, f_X\left(\frac{ax}{a}\right) \log\left[\frac{1}{|a|} \, f_X\left(\frac{ax}{a}\right)\right] \, \mathrm{d}(ax) \\ &= -\int_{\mathcal{X}} f_X(x) \log\left[\frac{1}{|a|} \, f_X(x)\right] \, \mathrm{d}x \\ &= -\int_{\mathcal{X}} f_X(x) \left[\log f_X(x) - \log|a|\right] \, \mathrm{d}x \\ &= -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x + \log|a| \int_{\mathcal{X}} f_X(x) \, \mathrm{d}x \\ &\overset{(2)}{=} \mathrm{h}(X) + \log|a| \; . \end{split} \tag{8}$$

■

Sources:

- Wikipedia (2020): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-12; URL: https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy.

2.2.5 Addition upon matrix multiplication

Theorem: Let X be a continuous (→ I/1.2.6) random vector (→ I/1.2.3). Then, the differential entropy (→ I/2.2.1) of X increases additively when multiplied with an invertible matrix A:

$$\mathrm{h}(AX) = \mathrm{h}(X) + \log |A| \; . \tag{1}$$

Proof: By definition, the differential entropy (→ I/2.2.1) of X is

$$\mathrm{h}(X) = -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x \tag{2}$$

where $f_X(x)$ is the probability density function (→ I/1.7.1) of X and $\mathcal{X}$ is the set of possible values of X.
The probability density function of a linear function of a continuous random vector (→ I/1.7.6) $Y = g(X) = \Sigma X + \mu$ is

$$f_Y(y) = \left\{ \begin{array}{rl} \frac{1}{|\Sigma|} f_X(\Sigma^{-1}(y-\mu)) \; , & \text{if} \; y \in \mathcal{Y} \\ 0 \; , & \text{if} \; y \notin \mathcal{Y} \end{array} \right. \tag{3}$$

where $\mathcal{Y} = \{y = \Sigma x + \mu : x \in \mathcal{X}\}$ is the set of possible outcomes of Y.
Therefore, with $Y = g(X) = AX$, i.e. $\Sigma = A$ and $\mu = 0_n$, the probability density function (→ I/1.7.1) of Y is given by

$$f_Y(y) = \left\{ \begin{array}{rl} \frac{1}{|A|} f_X(A^{-1}y) \; , & \text{if} \; y \in \mathcal{Y} \\ 0 \; , & \text{if} \; y \notin \mathcal{Y} \end{array} \right. \tag{4}$$

where $\mathcal{Y} = \{y = Ax : x \in \mathcal{X}\}$.
Thus, the differential entropy (→ I/2.2.1) of Y is

$$\begin{split} \mathrm{h}(Y) &\overset{(2)}{=} -\int_{\mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y \\ &\overset{(4)}{=} -\int_{\mathcal{Y}} \left[\frac{1}{|A|} f_X(A^{-1}y) \right] \log \left[\frac{1}{|A|} f_X(A^{-1}y) \right] \, \mathrm{d}y \; . \end{split} \tag{5}$$

Substituting $y = Ax$ into the integral, we obtain

$$\begin{split} \mathrm{h}(Y) &= -\int_{\mathcal{X}} \left[\frac{1}{|A|} f_X(A^{-1}Ax) \right] \log \left[\frac{1}{|A|} f_X(A^{-1}Ax) \right] \, \mathrm{d}(Ax) \\ &= -\frac{1}{|A|} \int_{\mathcal{X}} f_X(x) \log \left[\frac{1}{|A|} f_X(x) \right] \, \mathrm{d}(Ax) \; . \end{split} \tag{6}$$

Using the differential $\mathrm{d}(Ax) = |A|\mathrm{d}x$, this becomes

$$\begin{split} \mathrm{h}(Y) &= -\frac{|A|}{|A|} \int_{\mathcal{X}} f_X(x) \log \left[\frac{1}{|A|} f_X(x) \right] \, \mathrm{d}x \\ &= -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x - \int_{\mathcal{X}} f_X(x) \log \frac{1}{|A|} \, \mathrm{d}x \; . \end{split} \tag{7}$$

Finally, employing the fact (→ I/1.7.1) that $\int_{\mathcal{X}} f_X(x)\,\mathrm{d}x = 1$, we can derive the differential entropy (→ I/2.2.1) of Y as

$$
\begin{split}
\mathrm{h}(Y) &= -\int_{\mathcal{X}} f_X(x) \log f_X(x)\,\mathrm{d}x + \log|A| \int_{\mathcal{X}} f_X(x)\,\mathrm{d}x \\
&\overset{(2)}{=} \mathrm{h}(X) + \log|A| \; .
\end{split} \tag{8}
$$

■

Sources:
- Cover, Thomas M. & Thomas, Joy A. (1991): "Properties of Differential Entropy, Relative Entropy, and Mutual Information"; in: *Elements of Information Theory*, sect. 8.6, p. 253; URL: https://www.google.de/books/edition/Elements_of_Information_Theory/j0DBDwAAQBAJ.
- Wikipedia (2021): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy.

2.2.6 Non-invariance and transformation

Theorem: The differential entropy (→ I/2.2.1) is not invariant under change of variables, i.e. there exist random variables X and $Y = g(X)$, such that

$$
\mathrm{h}(Y) \neq \mathrm{h}(X) \; . \tag{1}
$$

In particular, for an invertible transformation $g : X \rightarrow Y$ from a random vector X to another random vector of the same dimension Y, it holds that

$$
\mathrm{h}(Y) = \mathrm{h}(X) + \int_{\mathcal{X}} f_X(x) \log |J_g(x)| \; \mathrm{d}x \; . \tag{2}
$$

where $J_g(x)$ is the Jacobian matrix of the vector-valued function g and $\mathcal{X}$ is the set of possible values of X.

Proof: By definition, the differential entropy (→ I/2.2.1) of X is

$$
\mathrm{h}(X) = -\int_{\mathcal{X}} f_X(x) \log f_X(x)\,\mathrm{d}x \tag{3}
$$

where $f_X(x)$ is the probability density function (→ I/1.7.1) of X.
The probability density function of an invertible function of a continuous random vector (→ I/1.7.5) $Y = g(X)$ is

$$
f_Y(y) = \left\{
\begin{array}{rl}
f_X(g^{-1}(y)) \; |J_{g^{-1}}(y)| \; , & \text{if} \; y \in \mathcal{Y} \\
0 \; , & \text{if} \; y \notin \mathcal{Y}
\end{array}
\right. \tag{4}
$$

where $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$ is the set of possible outcomes of Y and $J_{g^{-1}}(y)$ is the Jacobian matrix of $g^{-1}(y)$

$$
J_{g^{-1}}(y) = \left[
\begin{matrix}
\frac{\mathrm{d}x_1}{\mathrm{d}y_1} & \ldots & \frac{\mathrm{d}x_1}{\mathrm{d}y_n} \\
\vdots & \ddots & \vdots \\
\frac{\mathrm{d}x_n}{\mathrm{d}y_1} & \ldots & \frac{\mathrm{d}x_n}{\mathrm{d}y_n}
\end{matrix}
\right] \; . \tag{5}
$$

Thus, the differential entropy (→ I/2.2.1) of Y is

$$
\begin{aligned}
\mathrm{h}(Y) &\overset{(3)}{=} -\int_{\mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y \\
&\overset{(4)}{=} -\int_{\mathcal{Y}} \left[f_X(g^{-1}(y)) \, \left| J_{g^{-1}}(y) \right| \right] \log \left[f_X(g^{-1}(y)) \, \left| J_{g^{-1}}(y) \right| \right] \, \mathrm{d}y \; .
\end{aligned} \tag{6}
$$

Substituting $y = g(x)$ into the integral and applying $J_{f^{-1}}(y) = J_f^{-1}(x)$, we obtain

$$
\begin{aligned}
\mathrm{h}(Y) &= -\int_{\mathcal{X}} \left[f_X(g^{-1}(g(x))) \, \left| J_{g^{-1}}(y) \right| \right] \log \left[f_X(g^{-1}(g(x))) \, \left| J_{g^{-1}}(y) \right| \right] \, \mathrm{d}[g(x)] \\
&= -\int_{\mathcal{X}} \left[f_X(x) \, \left| J_g^{-1}(x) \right| \right] \log \left[f_X(x) \, \left| J_g^{-1}(x) \right| \right] \, \mathrm{d}[g(x)] \; .
\end{aligned} \tag{7}
$$

Using the relations $y = f(x) \Rightarrow \mathrm{d}y = |J_f(x)| \, \mathrm{d}x$ and $|A||B| = |AB|$, this becomes

$$
\begin{aligned}
\mathrm{h}(Y) &= -\int_{\mathcal{X}} \left[f_X(x) \, \left| J_g^{-1}(x) \right| \left| J_g(x) \right| \right] \log \left[f_X(x) \, \left| J_g^{-1}(x) \right| \right] \, \mathrm{d}x \\
&= -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x - \int_{\mathcal{X}} f_X(x) \log \left| J_g^{-1}(x) \right| \, \mathrm{d}x \; .
\end{aligned} \tag{8}
$$

Finally, employing the fact (→ I/1.7.1) that $\int_{\mathcal{X}} f_X(x) \, \mathrm{d}x = 1$ and the determinant property $|A^{-1}| = 1/|A|$, we can derive the differential entropy (→ I/2.2.1) of Y as

$$
\begin{aligned}
\mathrm{h}(Y) &= -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x - \int_{\mathcal{X}} f_X(x) \log \frac{1}{|J_g(x)|} \, \mathrm{d}x \\
&\overset{(3)}{=} \mathrm{h}(X) + \int_{\mathcal{X}} f_X(x) \log |J_g(x)| \, \mathrm{d}x \; .
\end{aligned} \tag{9}
$$

Because there exist X and Y, such that the integral term in (9) is non-zero, this also demonstrates that there exist X and Y, such that (1) is fulfilled.

■

Sources:

- Wikipedia (2021): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: https://en.wikipedia.org/wiki/Differential_entropy#Properties_of_differential_entropy.
- Bernhard (2016): "proof of upper bound on differential entropy of f(X)"; in: *StackExchange Mathematics*, retrieved on 2021-10-07; URL: https://math.stackexchange.com/a/1759531.
- peek-a-boo (2019): "How to come up with the Jacobian in the change of variables formula"; in: *StackExchange Mathematics*, retrieved on 2021-08-30; URL: https://math.stackexchange.com/a/3239222.
- Wikipedia (2021): "Jacobian matrix and determinant"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant#Inverse.
- Wikipedia (2021): "Inverse function theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: https://en.wikipedia.org/wiki/Inverse_function_theorem#Statement.
- Wikipedia (2021): "Determinant"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-07; URL: https://en.wikipedia.org/wiki/Determinant#Properties_of_the_determinant.

2.2.7 Conditional differential entropy

Definition: Let X and Y be continuous random variables (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and probability density functions (→ I/1.7.1) $p(x)$ and $p(y)$. Then, the conditional differential entropy of Y given X or, differential entropy of Y conditioned on X, is defined as

$$\mathrm{h}(Y|X) = \int_{x \in \mathcal{X}} p(x) \cdot \mathrm{h}(Y|X = x) \, \mathrm{d}x \tag{1}$$

where $\mathrm{h}(Y|X = x)$ is the (marginal) differential entropy (→ I/2.2.1) of Y, evaluated at x.

2.2.8 Joint differential entropy

Definition: Let X and Y be continuous random variables (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and joint probability (→ I/1.3.2) density function (→ I/1.7.1) $p(x, y)$. Then, the joint differential entropy of X and Y is defined as

$$\mathrm{h}(X, Y) = -\int_{x \in \mathcal{X}} \int_{y \in \mathcal{Y}} p(x, y) \cdot \log_b p(x, y) \, \mathrm{d}y \, \mathrm{d}x \tag{1}$$

where b is the base of the logarithm specifying in which unit the differential entropy is determined.

2.2.9 Differential cross-entropy

Definition: Let X be a continuous random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ I/1.5.1) on X with the probability density functions (→ I/1.7.1) $p(x)$ and $q(x)$. Then, the differential cross-entropy of Q relative to P is defined as

$$\mathrm{h}(P, Q) = -\int_{\mathcal{X}} p(x) \log_b q(x) \, \mathrm{d}x \tag{1}$$

where b is the base of the logarithm specifying in which unit the differential cross-entropy is determined.

Sources:

- Wikipedia (2020): "Cross entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: https://en.wikipedia.org/wiki/Cross_entropy#Definition.

2.3 Discrete mutual information

2.3.1 Definition

Definition:
1) The mutual information of two discrete random variables (→ I/1.2.2) X and Y is defined as

$$\mathrm{I}(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{x \in \mathcal{Y}} p(x, y) \cdot \log \frac{p(x, y)}{p(x) \cdot p(y)} \tag{1}$$

where $p(x)$ and $p(y)$ are the probability mass functions (→ I/1.6.1) of X and Y and $p(x, y)$ is the joint probability (→ I/1.3.2) mass function of X and Y.
2) The mutual information of two continuous random variables (→ I/1.2.2) X and Y is defined as

$$\mathrm{I}(X,Y) = -\int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \, \mathrm{d}y \, \mathrm{d}x \tag{2}$$

where $p(x)$ and $p(y)$ are the probability density functions (→ I/1.6.1) of X and Y and $p(x,y)$ is the joint probability (→ I/1.3.2) density function of X and Y.

Sources:
- Cover TM, Thomas JA (1991): "Relative Entropy and Mutual Information"; in: *Elements of Information Theory*, ch. 2.3/8.5, p. 20/251; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

2.3.2 Relation to marginal and conditional entropy

Theorem: Let X and Y be discrete random variables (→ I/1.2.2) with the joint probability (→ I/1.3.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ I/2.4.1) of X and Y can be expressed as

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &= \mathrm{H}(Y) - \mathrm{H}(Y|X) \end{split} \tag{1}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the marginal entropies (→ I/2.1.1) of X and Y and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies (→ I/2.1.4).

Proof: The mutual information (→ I/2.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)} - \sum_x \sum_y p(x,y) \log p(x) \; . \tag{3}$$

Applying the law of conditional probability (→ I/1.3.4), i.e. $p(x,y) = p(x|y)\,p(y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x|y)\,p(y) \log p(x|y) - \sum_x \sum_y p(x,y) \log p(x) \; . \tag{4}$$

Regrouping the variables, we have:

$$\mathrm{I}(X,Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) \; . \tag{5}$$

Applying the law of marginal probability (→ I/1.3.3), i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_y p(y) \sum_x p(x|y) \log p(x|y) - \sum_x p(x) \log p(x) \; . \tag{6}$$

Now considering the definitions of marginal (→ I/2.1.1) and conditional (→ I/2.1.4) entropy

$$\begin{aligned} \mathrm{H}(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X|Y) &= \sum_{y \in \mathcal{Y}} p(y) \, \mathrm{H}(X|Y=y) \; , \end{aligned} \tag{7}$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{H}(X|Y) + \mathrm{H}(X) \\ &= \mathrm{H}(X) - \mathrm{H}(X|Y) \; . \end{aligned} \tag{8}$$

The conditioning of X on Y in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional entropy of Y given X is obtained by simply switching x and y in the derivation.

■

Sources:
- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

2.3.3 Relation to marginal and joint entropy

Theorem: Let X and Y be discrete random variables (→ I/1.2.2) with the joint probability (→ I/1.3.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ I/2.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \tag{1}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are the marginal entropies (→ I/2.1.1) of X and Y and $\mathrm{H}(X,Y)$ is the joint entropy (→ I/2.1.5).

Proof: The mutual information (→ I/2.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x) \, p(y)} \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y) \; . \tag{3}$$

Regrouping the variables, this reads:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) - \sum_y \left(\sum_x p(x,y) \right) \log p(y) \; . \tag{4}$$

Applying the law of marginal probability (→ I/1.3.3), i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x p(x) \log p(x) - \sum_y p(y) \log p(y) \; . \tag{5}$$

Now considering the definitions of marginal (→ I/2.1.1) and joint (→ I/2.1.5) entropy

$$\begin{split} \mathrm{H}(X) &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X,Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \; , \end{split} \tag{6}$$

we can finally show:

$$\begin{split} \mathrm{I}(X,Y) &= -\mathrm{H}(X,Y) + \mathrm{H}(X) + \mathrm{H}(Y) \\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \end{split} \tag{7}$$

■

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

2.3.4 Relation to joint and conditional entropy

Theorem: Let X and Y be discrete random variables (→ I/1.2.2) with the joint probability (→ I/1.3.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ I/2.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \tag{1}$$

where $\mathrm{H}(X,Y)$ is the joint entropy (→ I/2.1.5) of X and Y and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies (→ I/2.1.4).

Proof: The existence of the joint probability mass function (→ I/1.6.1) ensures that the mutual information (→ I/2.4.1) is defined:

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \tag{2}$$

The relation of mutual information to conditional entropy (→ I/2.3.2) is:

$$\mathrm{I}(X,Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) \tag{3}$$

$$\mathrm{I}(X,Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) \tag{4}$$

The relation of mutual information to joint entropy (→ I/2.3.3) is:

$$\mathrm{I}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \tag{5}$$

It is true that

$$\mathrm{I}(X,Y) = \mathrm{I}(X,Y) + \mathrm{I}(X,Y) - \mathrm{I}(X,Y) \; . \tag{6}$$

Plugging in (3), (4) and (5) on the right-hand side, we have

$$\begin{aligned} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) + \mathrm{H}(Y) - \mathrm{H}(Y|X) - \mathrm{H}(X) - \mathrm{H}(Y) + \mathrm{H}(X,Y) \\ &= \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \end{aligned} \tag{7}$$

which proves the identity given above.

■

Sources:
- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

2.4 Continuous mutual information

2.4.1 Definition

Definition:
1) The mutual information of two discrete random variables (→ I/1.2.2) X and Y is defined as

$$\mathrm{I}(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{x\in\mathcal{Y}} p(x,y)\cdot\log\frac{p(x,y)}{p(x)\cdot p(y)} \tag{1}$$

where $p(x)$ and $p(y)$ are the probability mass functions (→ I/1.6.1) of X and Y and $p(x,y)$ is the joint probability (→ I/1.3.2) mass function of X and Y.
2) The mutual information of two continuous random variables (→ I/1.2.2) X and Y is defined as

$$\mathrm{I}(X,Y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\cdot\log\frac{p(x,y)}{p(x)\cdot p(y)}\,\mathrm{d}y\,\mathrm{d}x \tag{2}$$

where $p(x)$ and $p(y)$ are the probability density functions (→ I/1.6.1) of X and Y and $p(x,y)$ is the joint probability (→ I/1.3.2) density function of X and Y.

Sources:
- Cover TM, Thomas JA (1991): "Relative Entropy and Mutual Information"; in: *Elements of Information Theory*, ch. 2.3/8.5, p. 20/251; URL: https://www.wiley.com/en-us/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

2.4.2 Relation to marginal and conditional differential entropy

Theorem: Let X and Y be continuous random variables (→ I/1.2.2) with the joint probability (→ I/1.3.2) $p(x,y)$ for $x\in\mathcal{X}$ and $y\in\mathcal{Y}$. Then, the mutual information (→ I/2.4.1) of X and Y can be expressed as

$$\begin{aligned} \mathrm{I}(X,Y) &= \mathrm{h}(X) - \mathrm{h}(X|Y) \\ &= \mathrm{h}(Y) - \mathrm{h}(Y|X) \end{aligned} \tag{1}$$

where $\mathrm{h}(X)$ and $\mathrm{h}(Y)$ are the marginal differential entropies (→ I/2.2.1) of X and Y and $\mathrm{h}(X|Y)$ and $\mathrm{h}(Y|X)$ are the conditional differential entropies (→ I/2.2.7).

Proof: The mutual information (→ I/2.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \,\mathrm{d}y\,\mathrm{d}x \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(y)} \,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x) \,\mathrm{d}x\,\mathrm{d}y \; . \tag{3}$$

Applying the law of conditional probability (→ I/1.3.4), i.e. $p(x,y) = p(x|y)\,p(y)$, we get:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x|y)\,p(y) \log p(x|y) \,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x) \,\mathrm{d}y\,\mathrm{d}x \; . \tag{4}$$

Regrouping the variables, we have:

$$\mathrm{I}(X,Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} p(x,y) \,\mathrm{d}y \right) \log p(x) \,\mathrm{d}x \; . \tag{5}$$

Applying the law of marginal probability (→ I/1.3.3), i.e. $p(x) = \int_{\mathcal{Y}} p(x,y)\,\mathrm{d}y$, we get:

$$\mathrm{I}(X,Y) = \int_{\mathcal{Y}} p(y) \int_{\mathcal{X}} p(x|y) \log p(x|y) \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{X}} p(x) \log p(x) \,\mathrm{d}x \; . \tag{6}$$

Now considering the definitions of marginal (→ I/2.2.1) and conditional (→ I/2.2.7) differential entropy

$$\begin{split} \mathrm{h}(X) &= - \int_{\mathcal{X}} p(x) \log p(x) \,\mathrm{d}x \\ \mathrm{h}(X|Y) &= \int_{\mathcal{Y}} p(y)\, \mathrm{h}(X|Y=y) \,\mathrm{d}y \; , \end{split} \tag{7}$$

we can finally show:

$$\mathrm{I}(X,Y) = -\mathrm{h}(X|Y) + \mathrm{h}(X) = \mathrm{h}(X) - \mathrm{h}(X|Y) \; . \tag{8}$$

The conditioning of X on Y in this proof is without loss of generality. Thus, the proof for the expression using the reverse conditional differential entropy of Y given X is obtained by simply switching x and y in the derivation.

■

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

2.4.3 Relation to marginal and joint differential entropy

Theorem: Let X and Y be continuous random variables (→ I/1.2.2) with the joint probability (→ I/1.3.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ I/2.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \tag{1}$$

where $\mathrm{h}(X)$ and $\mathrm{h}(Y)$ are the marginal differential entropies (→ I/2.2.1) of X and Y and $\mathrm{h}(X,Y)$ is the joint differential entropy (→ I/2.2.8).

Proof: The mutual information (→ I/2.4.1) of X and Y is defined as

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \,\mathrm{d}y\,\mathrm{d}x \; . \tag{2}$$

Separating the logarithm, we have:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x) \,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(y) \,\mathrm{d}y\,\mathrm{d}x \; . \tag{3}$$

Regrouping the variables, this reads:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} p(x,y) \,\mathrm{d}y \right) \log p(x) \,\mathrm{d}x - \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} p(x,y) \,\mathrm{d}x \right) \log p(y) \,\mathrm{d}y \; . \tag{4}$$

Applying the law of marginal probability (→ I/1.3.3), i.e. $p(x) = \int_{\mathcal{Y}} p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \,\mathrm{d}y\,\mathrm{d}x - \int_{\mathcal{X}} p(x) \log p(x) \,\mathrm{d}x - \int_{\mathcal{Y}} p(y) \log p(y) \,\mathrm{d}y \; . \tag{5}$$

Now considering the definitions of marginal (→ I/2.2.1) and joint (→ I/2.2.8) differential entropy

$$\begin{aligned} \mathrm{h}(X) &= -\int_{\mathcal{X}} p(x) \log p(x) \,\mathrm{d}x \\ \mathrm{h}(X,Y) &= -\int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log p(x,y) \,\mathrm{d}y\,\mathrm{d}x \; , \end{aligned} \tag{6}$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{h}(X,Y) + \mathrm{h}(X) + \mathrm{h}(Y) \\ &= \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \; . \end{aligned} \tag{7}$$

■

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

2.4.4 Relation to joint and conditional differential entropy

Theorem: Let X and Y be continuous random variables (→ I/1.2.2) with the joint probability (→ I/1.3.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ I/2.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{h}(X,Y) - \mathrm{h}(X|Y) - \mathrm{h}(Y|X) \tag{1}$$

where $\mathrm{h}(X,Y)$ is the joint differential entropy (→ I/2.2.8) of X and Y and $\mathrm{h}(X|Y)$ and $\mathrm{h}(Y|X)$ are the conditional differential entropies (→ I/2.2.7).

Proof: The existence of the joint probability density function (→ I/1.7.1) ensures that the mutual information (→ I/2.4.1) is defined:

$$\mathrm{I}(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \,\mathrm{d}y\,\mathrm{d}x \; . \tag{2}$$

The relation of mutual information to conditional differential entropy (→ I/2.4.2) is:

$$\mathrm{I}(X,Y) = \mathrm{h}(X) - \mathrm{h}(X|Y) \tag{3}$$

$$\mathrm{I}(X,Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X) \tag{4}$$

The relation of mutual information to joint differential entropy (→ I/2.4.3) is:

$$\mathrm{I}(X,Y) = \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \; . \tag{5}$$

It is true that

$$\mathrm{I}(X,Y) = \mathrm{I}(X,Y) + \mathrm{I}(X,Y) - \mathrm{I}(X,Y) \; . \tag{6}$$

Plugging in (3), (4) and (5) on the right-hand side, we have

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{h}(X) - \mathrm{h}(X|Y) + \mathrm{h}(Y) - \mathrm{h}(Y|X) - \mathrm{h}(X) - \mathrm{h}(Y) + \mathrm{h}(X,Y) \\ &= \mathrm{h}(X,Y) - \mathrm{h}(X|Y) - \mathrm{h}(Y|X) \end{split} \tag{7}$$

which proves the identity given above.

■

Sources:

- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

2.5 Kullback-Leibler divergence

2.5.1 Definition

Definition: Let X be a random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ I/1.5.1) on X.

1) The Kullback-Leibler divergence of P from Q for a discrete random variable X is defined as

$$\mathrm{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{1}$$

where $p(x)$ and $q(x)$ are the probability mass functions (→ I/1.6.1) of P and Q.
2) The Kullback-Leibler divergence of P from Q for a continuous random variable X is defined as

$$\mathrm{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{2}$$

where $p(x)$ and $q(x)$ are the probability density functions (→ I/1.7.1) of P and Q.
By convention (→ I/2.1.1), $0 \cdot \log 0$ is taken to be zero when calculating the divergence between P and Q.

Sources:

- MacKay, David J.C. (2003): "Probability, Entropy, and Inference"; in: *Information Theory, Inference, and Learning Algorithms*, ch. 2.6, eq. 2.45, p. 34; URL: https://www.inference.org.uk/itprnn/book.pdf.

2.5.2 Non-negativity

Theorem: The Kullback-Leibler divergence (→ I/2.5.1) is always non-negative

$$\mathrm{KL}[P||Q] \geq 0 \tag{1}$$

with $\mathrm{KL}[P||Q] = 0$, if and only if $P = Q$.

Proof: The discrete Kullback-Leibler divergence (→ I/2.5.1) is defined as

$$\mathrm{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

which can be reformulated into

$$\mathrm{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log p(x) - \sum_{x \in \mathcal{X}} p(x) \cdot \log q(x) \; . \tag{3}$$

Gibbs' inequality (→ I/2.1.8) states that the entropy (→ I/2.1.1) of a probability distribution is always less than or equal to the cross-entropy (→ I/2.1.6) with another probability distribution – with equality only if the distributions are identical –,

$$-\sum_{i=1}^{n} p(x_i) \, \log p(x_i) \leq -\sum_{i=1}^{n} p(x_i) \, \log q(x_i) \tag{4}$$

which can be reformulated into

$$\sum_{i=1}^{n} p(x_i) \, \log p(x_i) - \sum_{i=1}^{n} p(x_i) \, \log q(x_i) \geq 0 \; . \tag{5}$$

Applying (5) to (3), this proves equation (1).

■

Sources:

- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.

2.5.3 Non-negativity

Theorem: The Kullback-Leibler divergence (→ I/2.5.1) is always non-negative

$$\mathrm{KL}[P||Q] \geq 0 \tag{1}$$

with $\mathrm{KL}[P||Q] = 0$, if and only if $P = Q$.

Proof: The discrete Kullback-Leibler divergence (→ I/2.5.1) is defined as

$$\mathrm{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \; . \tag{2}$$

The log sum inequality (→ I/2.1.9) states that

$$\sum_{i=1}^{n} a_i \, \log_c \frac{a_i}{b_i} \geq a \, \log_c \frac{a}{b} \; . \tag{3}$$

where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be non-negative real numbers and $a = \sum_{i=1}^{n} a_i$ and $b = \sum_{i=1}^{n} b_i$. Because $p(x)$ and $q(x)$ are probability mass functions (→ I/1.6.1), such that

$$\begin{split} p(x) \geq 0, \quad &\sum_{x \in \mathcal{X}} p(x) = 1 \quad \text{and} \\ q(x) \geq 0, \quad &\sum_{x \in \mathcal{X}} q(x) = 1 \; , \end{split} \tag{4}$$

theorem (1) is simply a special case of (3), i.e.

$$\mathrm{KL}[P||Q] \overset{(2)}{=} \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \overset{(3)}{\geq} 1 \, \log \frac{1}{1} = 0 \; . \tag{5}$$

■

Sources:
- Wikipedia (2020): "Log sum inequality"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-09; URL: https://en.wikipedia.org/wiki/Log_sum_inequality#Applications.

2.5.4 Non-symmetry

Theorem: The Kullback-Leibler divergence (→ I/2.5.1) is non-symmetric, i.e.

$$\mathrm{KL}[P||Q] \neq \mathrm{KL}[Q||P] \tag{1}$$

for some probability distributions (→ I/1.5.1) P and Q.

Proof: Let $X \in \mathcal{X} = \{0, 1, 2\}$ be a discrete random variable (→ I/1.2.2) and consider the two probability distributions (→ I/1.5.1)

$$\begin{aligned} P &: \; X \sim \mathrm{Bin}(2, 0.5) \\ Q &: \; X \sim \mathcal{U}(0, 2) \end{aligned} \tag{2}$$

where $\mathrm{Bin}(n, p)$ indicates a binomial distribution (→ II/1.3.1) and $\mathcal{U}(a, b)$ indicates a discrete uniform distribution (→ II/1.1.1).
Then, the probability mass function of the binomial distribution (→ II/1.3.2) entails that

$$p(x) = \begin{cases} 1/4 \; , & \text{if } x = 0 \\ 1/2 \; , & \text{if } x = 1 \\ 1/4 \; , & \text{if } x = 2 \end{cases} \tag{3}$$

and the probability mass function of the discrete uniform distribution (→ II/1.1.2) entails that

$$q(x) = \frac{1}{3} \; , \tag{4}$$

such that the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is

$$\begin{aligned} \mathrm{KL}[P||Q] &= \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \\ &= \frac{1}{4} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} + \frac{1}{4} \log \frac{3}{4} \\ &= \frac{1}{2} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} \\ &= \frac{1}{2} \left(\log \frac{3}{4} + \log \frac{3}{2} \right) \\ &= \frac{1}{2} \log \left(\frac{3}{4} \cdot \frac{3}{2} \right) \\ &= \frac{1}{2} \log \frac{9}{8} = 0.0589 \end{aligned} \tag{5}$$

and the Kullback-Leibler divergence (→ I/2.5.1) of Q from P is

$$\begin{aligned} \mathrm{KL}[Q||P] &= \sum_{x \in \mathcal{X}} q(x) \cdot \log \frac{q(x)}{p(x)} \\ &= \frac{1}{3} \log \frac{4}{3} + \frac{1}{3} \log \frac{2}{3} + \frac{1}{3} \log \frac{4}{3} \\ &= \frac{1}{3} \left(\log \frac{4}{3} + \log \frac{2}{3} + \log \frac{4}{3} \right) \\ &= \frac{1}{3} \log \left(\frac{4}{3} \cdot \frac{2}{3} \cdot \frac{4}{3} \right) \\ &= \frac{1}{3} \log \frac{32}{27} = 0.0566 \end{aligned} \tag{6}$$

which provides an example for

$$\mathrm{KL}[P||Q] \neq \mathrm{KL}[Q||P] \tag{7}$$

and thus proves the theorem.

■

Sources:

- Kullback, Solomon (1959): "Divergence"; in: *Information Theory and Statistics*, ch. 1.3, pp. 6ff.; URL: http://index-of.co.uk/Information-Theory/Information%20theory%20and%20statistics%20-%20Solomon%20Kullback.pdf.
- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Basic_example.

2.5.5 Convexity

Theorem: The Kullback-Leibler divergence (→ I/2.5.1) is convex in the pair of probability distributions (→ I/1.5.1) (p, q), i.e.

$$\mathrm{KL}[\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2] \leq \lambda \mathrm{KL}[p_1||q_1] + (1-\lambda)\mathrm{KL}[p_2||q_2] \tag{1}$$

where (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions and $0 \leq \lambda \leq 1$.

Proof: The Kullback-Leibler divergence (→ I/2.5.1) of P from Q is defined as

$$\mathrm{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

and the log sum inequality (→ I/2.1.9) states that

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \tag{3}$$

where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are non-negative real numbers.
Thus, we can rewrite the KL divergence of the mixture distribution as

$$\begin{aligned}
&\mathrm{KL}[\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2] \\
&\overset{(2)}{=} \sum_{x \in \mathcal{X}} \left[[\lambda p_1(x) + (1-\lambda)p_2(x)] \cdot \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \right] \\
&\overset{(3)}{\leq} \sum_{x \in \mathcal{X}} \left[\lambda p_1(x) \cdot \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \cdot \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \right] \\
&= \lambda \sum_{x \in \mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) \sum_{x \in \mathcal{X}} p_2(x) \cdot \log \frac{p_2(x)}{q_2(x)} \\
&\overset{(2)}{=} \lambda \, \mathrm{KL}[p_1||q_1] + (1-\lambda) \, \mathrm{KL}[p_2||q_2]
\end{aligned} \tag{4}$$

which is equivalent to (1).

■

Sources:

- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-11; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.
- Xie, Yao (2012): "Chain Rules and Inequalities"; in: *ECE587: Information Theory*, Lecture 3, Slides 22/24; URL: https://www2.isye.gatech.edu/~yxie77/ece587/Lecture3.pdf.

2.5.6 Additivity for independent distributions

Theorem: The Kullback-Leibler divergence (→ I/2.5.1) is additive for independent distributions, i.e.

$$\mathrm{KL}[P||Q] = \mathrm{KL}[P_1||Q_1] + \mathrm{KL}[P_2||Q_2] \tag{1}$$

where P_1 and P_2 are independent (→ I/1.3.6) distributions (→ I/1.5.1) with the joint distribution (→ I/1.5.2) P, such that $p(x,y) = p_1(x)\,p_2(y)$, and equivalently for Q_1, Q_2 and Q.

Proof: The continuous Kullback-Leibler divergence (→ I/2.5.1) is defined as

$$\mathrm{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{2}$$

which, applied to the joint distributions P and Q, yields

$$\mathrm{KL}[P||Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \cdot \log \frac{p(x,y)}{q(x,y)} \, \mathrm{d}y \, \mathrm{d}x \; . \tag{3}$$

Applying $p(x,y) = p_1(x)\,p_2(y)$ and $q(x,y) = q_1(x)\,q_2(y)$, we have

$$\mathrm{KL}[P||Q] = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x)\,p_2(y) \cdot \log \frac{p_1(x)\,p_2(y)}{q_1(x)\,q_2(y)} \, \mathrm{d}y \, \mathrm{d}x \; . \tag{4}$$

Now we can separate the logarithm and evaluate the integrals:

$$\begin{split}
\mathrm{KL}[P||Q] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x)\,p_2(y) \cdot \left(\log \frac{p_1(x)}{q_1(x)} + \log \frac{p_2(y)}{q_2(y)} \right) \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x)\,p_2(y) \cdot \log \frac{p_1(x)}{q_1(x)} \, \mathrm{d}y \, \mathrm{d}x + \int_{\mathcal{X}} \int_{\mathcal{Y}} p_1(x)\,p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} \int_{\mathcal{Y}} p_2(y) \, \mathrm{d}y \, \mathrm{d}x + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} \int_{\mathcal{X}} p_1(x) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_{\mathcal{X}} p_1(x) \cdot \log \frac{p_1(x)}{q_1(x)} \, \mathrm{d}x + \int_{\mathcal{Y}} p_2(y) \cdot \log \frac{p_2(y)}{q_2(y)} \, \mathrm{d}y \\
&\overset{(2)}{=} \mathrm{KL}[P_1||Q_1] + \mathrm{KL}[P_2||Q_2] \; .
\end{split} \tag{5}$$

■

Sources:

- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-31; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.

2.5.7 Invariance under parameter transformation

Theorem: The Kullback-Leibler divergence (→ I/2.5.1) is invariant under parameter transformation, i.e.

$$\mathrm{KL}[p(x)||q(x)] = \mathrm{KL}[p(y)||q(y)] \tag{1}$$

where $y(x) = mx + n$ is an affine transformation of x and $p(x)$ and $q(x)$ are the probability density functions (→ I/1.7.1) of the probability distributions (→ I/1.5.1) P and Q on the continuous random variable (→ I/1.2.2) X.

Proof: The continuous Kullback-Leibler divergence (→ I/2.5.1) (KL divergence) is defined as

$$\mathrm{KL}[p(x)||q(x)] = \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{2}$$

where $a = \min(\mathcal{X})$ and $b = \max(\mathcal{X})$ are the lower and upper bound of the possible outcomes $\mathcal{X}$ of X.
Due to the identity of the differentials

$$\begin{aligned} p(x) \, \mathrm{d}x &= p(y) \, \mathrm{d}y \\ q(x) \, \mathrm{d}x &= q(y) \, \mathrm{d}y \end{aligned} \tag{3}$$

which can be rearranged into

$$\begin{aligned} p(x) &= p(y) \, \frac{\mathrm{d}y}{\mathrm{d}x} \\ q(x) &= q(y) \, \frac{\mathrm{d}y}{\mathrm{d}x} \; , \end{aligned} \tag{4}$$

the KL divergence can be evaluated as follows:

$$\begin{aligned} \mathrm{KL}[p(x)||q(x)] &= \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)} \, \mathrm{d}x \\ &= \int_{y(a)}^{y(b)} p(y) \, \frac{\mathrm{d}y}{\mathrm{d}x} \cdot \log \left(\frac{p(y) \, \frac{\mathrm{d}y}{\mathrm{d}x}}{q(y) \, \frac{\mathrm{d}y}{\mathrm{d}x}} \right) \, \mathrm{d}x \\ &= \int_{y(a)}^{y(b)} p(y) \cdot \log \frac{p(y)}{q(y)} \, \mathrm{d}y \\ &= \mathrm{KL}[p(y)||q(y)] \; . \end{aligned} \tag{5}$$

■

Sources:

- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Properties.
- shimao (2018): "KL divergence invariant to affine transformation?"; in: *StackExchange CrossValidated*, retrieved on 2020-05-28; URL: https://stats.stackexchange.com/questions/341922/kl-divergence-inv

2.5.8 Relation to discrete entropy

Theorem: Let X be a discrete random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ I/1.5.1) on X. Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q can be expressed as

$$\mathrm{KL}[P||Q] = \mathrm{H}(P,Q) - \mathrm{H}(P) \tag{1}$$

where $\mathrm{H}(P,Q)$ is the cross-entropy (→ I/2.1.6) of P and Q and $\mathrm{H}(P)$ is the marginal entropy (→ I/2.1.1) of P.

Proof: The discrete Kullback-Leibler divergence (→ I/2.5.1) is defined as

$$\mathrm{KL}[P||Q] = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \tag{2}$$

where $p(x)$ and $q(x)$ are the probability mass functions (→ I/1.6.1) of P and Q.
Separating the logarithm, we have:

$$\mathrm{KL}[P||Q] = -\sum_{x \in \mathcal{X}} p(x) \, \log q(x) + \sum_{x \in \mathcal{X}} p(x) \, \log p(x) \; . \tag{3}$$

Now considering the definitions of marginal entropy (→ I/2.1.1) and cross-entropy (→ I/2.1.6)

$$\begin{aligned} \mathrm{H}(P) &= -\sum_{x \in \mathcal{X}} p(x) \, \log p(x) \\ \mathrm{H}(P,Q) &= -\sum_{x \in \mathcal{X}} p(x) \, \log q(x) \; , \end{aligned} \tag{4}$$

we can finally show:

$$\mathrm{KL}[P||Q] = \mathrm{H}(P,Q) - \mathrm{H}(P) \; . \tag{5}$$

■

Sources:
- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Motivation.

2.5.9 Relation to differential entropy

Theorem: Let X be a continuous random variable (→ I/1.2.2) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ I/1.5.1) on X. Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q can be expressed as

$$\mathrm{KL}[P||Q] = \mathrm{h}(P,Q) - \mathrm{h}(P) \tag{1}$$

where $\mathrm{h}(P,Q)$ is the differential cross-entropy (→ I/2.2.9) of P and Q and $\mathrm{h}(P)$ is the marginal differential entropy (→ I/2.2.1) of P.

Proof: The continuous Kullback-Leibler divergence (→ I/2.5.1) is defined as

$$\mathrm{KL}[P||Q] = \int_{\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{2}$$

where $p(x)$ and $q(x)$ are the probability density functions (→ I/1.7.1) of P and Q.
Separating the logarithm, we have:

$$\mathrm{KL}[P||Q] = -\int_{\mathcal{X}} p(x) \, \log q(x) \, \mathrm{d}x + \int_{\mathcal{X}} p(x) \, \log p(x) \, \mathrm{d}x \; . \tag{3}$$

Now considering the definitions of marginal differential entropy (→ I/2.2.1) and differential cross-entropy (→ I/2.2.9)

$$\begin{aligned} \mathrm{h}(P) &= -\int_{\mathcal{X}} p(x) \, \log p(x) \, \mathrm{d}x \\ \mathrm{h}(P,Q) &= -\int_{\mathcal{X}} p(x) \, \log q(x) \, \mathrm{d}x \; , \end{aligned} \tag{4}$$

we can finally show:

$$\mathrm{KL}[P||Q] = \mathrm{h}(P,Q) - \mathrm{h}(P) \; . \tag{5}$$

■

Sources:

- Wikipedia (2020): "Kullback-Leibler divergence"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-27; URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence#Motivation.

3 Estimation theory

3.1 Point estimates

3.1.1 Mean squared error

Definition: Let $\hat{\theta}$ be an estimator of an unknown parameter $\hat{\theta}$ based on measured data y. Then, the mean squared error is defined as the expected value (→ I/1.10.1) of the squared difference between the estimated value and the true value of the parameter:

$$\mathrm{MSE} = \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\theta\right)^2\right] \; . \tag{1}$$

where $\mathrm{E}_{\hat{\theta}}\left[\cdot\right]$ is expectation calculated over all possible samples y leading to values of $\hat{\theta}$.

Sources:
- Wikipedia (2022): "Estimator"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-27; URL: https://en.wikipedia.org/wiki/Estimator#Mean_squared_error.

3.1.2 Partition of the mean squared error into bias and variance

Theorem: The mean squared error (→ I/3.1.1) can be partitioned into variance and squared bias

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta},\theta)^2 \tag{1}$$

where the variance (→ I/1.11.1) is given by

$$\mathrm{Var}(\hat{\theta}) = \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] \tag{2}$$

and the bias is given by

$$\mathrm{Bias}(\hat{\theta},\theta) = \left(\mathrm{E}_{\hat{\theta}}(\hat{\theta})-\theta\right) \; . \tag{3}$$

Proof: The mean squared error (→ I/3.1.1) (MSE) is defined as the expected value (→ I/1.10.1) of the squared deviation of the estimated value $\hat{\theta}$ from the true value θ of a parameter, over all values $\hat{\theta}$:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\theta\right)^2\right] \; . \tag{4}$$

This formula can be evaluated in the following way:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}(\hat{\theta})+\mathrm{E}_{\hat{\theta}}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)^2 + 2\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathrm{E}_{\hat{\theta}}(\hat{\theta})-\theta\right) + \left(\mathrm{E}_{\hat{\theta}}(\hat{\theta})-\theta\right)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \mathrm{E}_{\hat{\theta}}\left[2\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathrm{E}_{\hat{\theta}}(\hat{\theta})-\theta\right)\right] + \mathrm{E}_{\hat{\theta}}\left[\left(\mathrm{E}_{\hat{\theta}}(\hat{\theta})-\theta\right)^2\right] \; .
\end{aligned} \tag{5}$$

Because $\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \theta$ is constant as a function of $\hat{\theta}$, we have:

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\mathrm{E}_{\hat{\theta}}\left[\hat{\theta} - \mathrm{E}_{\hat{\theta}}(\hat{\theta})\right] + \left(\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\left(\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \mathrm{E}_{\hat{\theta}}(\hat{\theta})\right) + \left(\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathrm{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \left(\mathrm{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \; .
\end{aligned} \tag{6}
$$

This proofs the partition given by (1).

■

Sources:
- Wikipedia (2019): "Mean squared error"; in: *Wikipedia, the free encyclopedia*, retrieved on 2019-11-27; URL: https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship.

3.2 Interval estimates

3.2.1 Confidence interval

Definition: Let y be a random sample from a probability distributions (→ I/1.5.1) governed by a parameter of interest θ and quantities not of interest φ. A confidence interval for θ is defined as an interval $[u(y), v(y)]$ determined by the random variables (→ I/1.2.2) $u(y)$ and $v(y)$ with the property

$$
\Pr(u(y) < \theta < v(y) \,|\, \theta, \varphi) = \gamma \quad \text{for all} \quad (\theta, \varphi) \; . \tag{1}
$$

where $\gamma = 1 - \alpha$ is called the confidence level.

Sources:
- Wikipedia (2022): "Confidence interval"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-03-27; URL: https://en.wikipedia.org/wiki/Confidence_interval#Definition.

3.2.2 Construction of confidence intervals using Wilks' theorem

Theorem: Let m be a generative model (→ I/5.1.1) for measured data y with model parameters $\theta \in \Theta$, consisting of a parameter of interest $\phi \in \Phi$ and nuisance parameters $\lambda \in \Lambda$:

$$
m : p(y|\theta) = \mathcal{D}(y;\theta), \quad \theta = \{\phi, \lambda\} \; . \tag{1}
$$

Further, let $\hat{\theta}$ be an estimate of θ, obtained using maximum-likelihood-estimation (→ I/4.1.3):

$$
\hat{\theta} = \operatorname*{arg\,max}_{\theta} \log p(y|\theta), \quad \hat{\theta} = \left\{\hat{\phi}, \hat{\lambda}\right\} \; . \tag{2}
$$

Then, an asymptotic confidence interval (→ I/3.2.1) for θ is given by

$$
\mathrm{CI}_{1-\alpha}(\hat{\phi}) = \left\{\phi \,|\, \log p(y|\phi, \hat{\lambda}) \geq \log p(y|\hat{\phi}, \hat{\lambda}) - \frac{1}{2}\chi^2_{1,1-\alpha}\right\} \tag{3}
$$

where $1-\alpha$ is the confidence level and $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-squared distribution (→ II/3.7.1) with 1 degree of freedom.

Proof: The confidence interval (→ I/3.2.1) is defined as the interval that, under infinitely repeated random experiments (→ I/1.1.1), contains the true parameter value with a certain probability.
Let us define the likelihood ratio (→ I/4.1.6)

$$\Lambda(\phi) = \frac{p(y|\phi,\hat{\lambda})}{p(y|\hat{\phi},\hat{\lambda})} \quad \text{for all} \quad \phi \in \Phi \tag{4}$$

and compute the log-likelihood ratio (→ I/4.1.7)

$$\log \Lambda(\phi) = \log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) \; . \tag{5}$$

Wilks' theorem states that, when comparing two statistical models with parameter spaces Θ_1 and $\Theta_0 \subset \Theta_1$, as the sample size approaches infinity, the quantity calculated as -2 times the log-ratio of maximum likelihoods follows a chi-squared distribution (→ II/3.7.1), if the null hypothesis is true:

$$H_0: \theta \in \Theta_0 \quad \Rightarrow \quad -2 \log \frac{\max_{\theta \in \Theta_0} p(y|\theta)}{\max_{\theta \in \Theta_1} p(y|\theta)} \sim \chi^2_{\Delta k} \quad \text{as} \quad n \rightarrow \infty \tag{6}$$

where Δk is the difference in dimensionality between Θ_0 and Θ_1. Applied to our example in (5), we note that $\Theta_1 = \left\{ \phi, \hat{\phi} \right\}$ and $\Theta_0 = \{\phi\}$, such that $\Delta k = 1$ and Wilks' theorem implies:

$$-2 \log \Lambda(\phi) \sim \chi^2_1 \; . \tag{7}$$

Using the quantile function (→ I/1.9.1) $\chi^2_{k,p}$ of the chi-squared distribution (→ II/3.7.1), an $(1-\alpha)$-confidence interval is therefore given by all values ϕ that satisfy

$$-2 \log \Lambda(\phi) \leq \chi^2_{1,1-\alpha} \; . \tag{8}$$

Applying (5) and rearranging, we can evaluate

$$\begin{split} -2 \left[\log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) \right] &\leq \chi^2_{1,1-\alpha} \\ \log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) &\geq -\frac{1}{2} \chi^2_{1,1-\alpha} \\ \log p(y|\phi,\hat{\lambda}) &\geq \log p(y|\hat{\phi},\hat{\lambda}) - \frac{1}{2} \chi^2_{1,1-\alpha} \end{split} \tag{9}$$

which is equivalent to the confidence interval given by (3).

■

Sources:

- Wikipedia (2020): "Confidence interval"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Confidence_interval#Methods_of_derivation.
- Wikipedia (2020): "Likelihood-ratio test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Likelihood-ratio_test#Definition.
- Wikipedia (2020): "Wilks' theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-19; URL: https://en.wikipedia.org/wiki/Wilks%27_theorem.

4 Frequentist statistics

4.1 Likelihood theory

4.1.1 Likelihood function

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ. Then, the probability density function (→ I/1.7.1) of the distribution of y given θ is called the likelihood function of m:

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) \; . \tag{1}$$

4.1.2 Log-likelihood function

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ. Then, the logarithm of the probability density function (→ I/1.7.1) of the distribution of y given θ is called the log-likelihood function (→ I/5.1.2) of m:

$$\mathrm{LL}_m(\theta) = \log p(y|\theta, m) = \log \mathcal{D}(y; \theta) \; . \tag{1}$$

4.1.3 Maximum likelihood estimation

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ. Then, the parameter values maximizing the likelihood function (→ I/5.1.2) or log-likelihood function (→ I/4.1.2) are called "maximum likelihood estimates" of θ:

$$\hat{\theta} = \operatorname*{arg\,max}_\theta \mathcal{L}_m(\theta) = \operatorname*{arg\,max}_\theta \mathrm{LL}_m(\theta) \; . \tag{1}$$

The process of calculating $\hat{\theta}$ is called "maximum likelihood estimation" and the functional form leading from y to $\hat{\theta}$ given m is called "maximum likelihood estimator". Maximum likelihood estimation, estimator and estimates may all be abbreviated as "MLE".

4.1.4 Maximum log-likelihood

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ. Then, the maximum log-likelihood (MLL) of m is the maximal value of the log-likelihood function (→ I/4.1.2) of this model:

$$\mathrm{MLL}(m) = \max_\theta \mathrm{LL}_m(\theta) \; . \tag{1}$$

The maximum log-likelihood can be obtained by plugging the maximum likelihood estimates (→ I/4.1.3) into the log-likelihood function (→ I/4.1.2).

4.1.5 MLE can be biased

Theorem: Maximum likelihood estimation (→ I/4.1.3) can result in biased estimates of model parameters, i.e. estimates whose long-term expected value is unequal to the quantities they estimate:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{MLE}}\right] = \mathrm{E}\left[\operatorname*{arg\,max}_\theta \mathrm{LL}_m(\theta)\right] \neq \theta \; . \tag{1}$$

Proof: Consider a set of independent and identical normally distributed (→ II/3.2.1) observations $x = \{x_1, \ldots, x_n\}$ with unknown mean (→ I/1.10.1) μ and variance (→ I/1.11.1) σ^2:

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{2}$$

Then, we know that the maximum likelihood estimator (→ I/4.1.3) for the variance (→ I/1.11.1) σ^2 is underestimating the true variance of the data distribution (→ IV/1.1.2):

$$\mathrm{E}\left[\hat{\sigma}^2_{\mathrm{MLE}}\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \; . \tag{3}$$

This proofs the existence of cases such as those stated by the theorem.

■

4.1.6 Likelihood ratio

Definition: Let m_0 and m_1 be two generative models (→ I/5.1.1) describing the same measured data y using different model parameters $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. Then, the quotient of the maximized likelihood functions (→ I/4.1.3) of these two models is denoted as Λ_{01} and is called the likelihood (→ I/5.1.2) ratio of m_0 relative to m_1:

$$\Lambda_{01} = \frac{\max_{\theta_0 \in \Theta_0} \mathcal{L}_{m_0}(\theta_0)}{\max_{\theta_1 \in \Theta_1} \mathcal{L}_{m_1}(\theta_1)} = \frac{p(y|\hat{\theta}_0, m_0)}{p(y|\hat{\theta}_1, m_1)} \; . \tag{1}$$

Sources:
- Wikipedia (2024): "Neyman-Pearson lemma"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-06-14; URL: https://en.wikipedia.org/wiki/Neyman%E2%80%93Pearson_lemma#Example.

4.1.7 Log-likelihood ratio

Definition: Let m_0 and m_1 be two generative models (→ I/5.1.1) describing the same measured data y using different model parameters $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. Then, the logarithmized quotient of the maximized likelihood functions (→ I/4.1.3) of these two models is denoted as $\log \Lambda_{01}$ and is called the log-likelihood (→ I/4.1.2) ratio of m_0 relative to m_1:

$$\log \Lambda_{01} = \log \frac{\max_{\theta_0 \in \Theta_0} \mathcal{L}_{m_0}(\theta_0)}{\max_{\theta_1 \in \Theta_1} \mathcal{L}_{m_1}(\theta_1)} = \log p(y|\hat{\theta}_0, m_0) - \log p(y|\hat{\theta}_1, m_1) \; . \tag{1}$$

Sources:
- Wikipedia (2024): "Likelihood-ratio test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-06-14; URL: https://en.wikipedia.org/wiki/Likelihood-ratio_test#Definition.

4.1.8 Method of moments

Definition: Let measured data y follow a probability distribution (→ I/1.5.1) with probability mass (→ I/1.6.1) or probability density (→ I/1.7.1) $p(y|\theta)$ governed by unknown parameters $\theta_1, \ldots, \theta_k$. Then, method-of-moments estimation, also referred to as "method of moments" or "matching the moments", consists in

1) expressing the first k moments (→ I/1.18.1) of y in terms of θ

$$
\begin{aligned}
\mu_1 &= f_1(\theta_1, \ldots, \theta_k) \\
&\vdots \\
\mu_k &= f_k(\theta_1, \ldots, \theta_k) \; ,
\end{aligned}
\tag{1}
$$

2) calculating the first k sample moments (→ I/1.18.1) from y

$$
\hat{\mu}_1(y), \ldots, \hat{\mu}_k(y) \tag{2}
$$

3) and solving the system of k equations

$$
\begin{aligned}
\hat{\mu}_1(y) &= f_1(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\
&\vdots \\
\hat{\mu}_k(y) &= f_k(\hat{\theta}_1, \ldots, \hat{\theta}_k)
\end{aligned}
\tag{3}
$$

for $\hat{\theta}_1, \ldots, \hat{\theta}_k$, which are subsequently refered to as "method-of-moments estimates".

Sources:
- Wikipedia (2021): "Method of moments (statistics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-29; URL: https://en.wikipedia.org/wiki/Method_of_moments_(statistics)#Method.

4.2 Statistical hypotheses

4.2.1 Statistical hypothesis

Definition: A statistical hypothesis is a statement about the parameters of a distribution describing a population from which observations can be sampled as measured data.
More precisely, let m be a generative model (→ I/5.1.1) describing measured data y in terms of a distribution $\mathcal{D}(\theta)$ with model parameters $\theta \in \Theta$. Then, a statistical hypothesis is formally specified as

$$
H: \; \theta \in \Theta^* \quad \text{where} \quad \Theta^* \subset \Theta \; . \tag{1}
$$

Sources:
- Wikipedia (2021): "Statistical hypothesis testing"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms.

4.2.2 Simple vs. composite

Definition: Let H be a statistical hypothesis (→ I/4.2.1). Then,

- H is called a simple hypothesis, if it completely specifies the population distribution; in this case, the sampling distribution (→ I/1.5.5) of the test statistic (→ I/4.3.5) is a function of sample size alone.
- H is called a composite hypothesis, if it does not completely specify the population distribution; for example, the hypothesis may only specify one parameter of the distribution and leave others unspecified.

Sources:
- Wikipedia (2021): "Exclusion of the null hypothesis"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Terminology.

4.2.3 Point/exact vs. set/inexact

Definition: Let H be a statistical hypothesis (→ I/4.2.1). Then,
- H is called a point hypothesis or exact hypothesis, if it specifies an exact parameter value:

$$H: \; \theta = \theta^* \; ; \tag{1}$$

- H is called a set hypothesis or inexact hypothesis, if it specifies a set of possible values with more than one element for the parameter value (e.g. a range or an interval):

$$H: \; \theta \in \Theta^* \; . \tag{2}$$

Sources:
- Wikipedia (2021): "Exclusion of the null hypothesis"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Terminology.

4.2.4 One-tailed vs. two-tailed

Definition: Let H_0 be a point (→ I/4.2.3) null hypothesis (→ I/4.3.2)

$$H_0: \; \theta = \theta_0 \tag{1}$$

and consider a set (→ I/4.2.3) alternative hypothesis (→ I/4.3.3) H_1. Then,
- H_1 is called a left-sided one-tailed hypothesis, if θ is assumed to be smaller than θ_0:

$$H_1: \; \theta < \theta_0 \; ; \tag{2}$$

- H_1 is called a right-sided one-tailed hypothesis, if θ is assumed to be larger than θ_0:

$$H_1: \; \theta > \theta_0 \; ; \tag{3}$$

- H_1 is called a two-tailed hypothesis, if θ is assumed to be unequal to θ_0:

$$H_1: \; \theta \neq \theta_0 \; . \tag{4}$$

Sources:

- Wikipedia (2021): "One- and two-tailed tests"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: https://en.wikipedia.org/wiki/One-_and_two-tailed_tests.

4.3 Hypothesis testing

4.3.1 Statistical test

Definition: Let y be a set of measured data. Then, a statistical hypothesis test consists of the following:
- an assumption about the distribution (→ I/1.5.1) of the data, often expressed in terms of a statistical model (→ I/5.1.1) m;
- a null hypothesis (→ I/4.3.2) H_0 and an alternative hypothesis (→ I/4.3.3) H_1 which make specific statements about the distribution of the data;
- a test statistic (→ I/4.3.5) $T(Y)$ which is a function of the data and whose distribution under the null hypothesis (→ I/4.3.2) is known;
- a significance level (→ I/4.3.8) α which imposes an upper bound on the probability (→ I/1.3.1) of rejecting H_0, given that H_0 is true.

Procedurally, the statistical hypothesis test works as follows:
- Given the null hypothesis H_0 and the significance level α, a critical value (→ I/4.3.9) t_{crit} is determined which partitions the set of possible values of $T(Y)$ into "acceptance region" and "rejection region".
- Then, the observed test statistic (→ I/4.3.5) $t_{\mathrm{obs}} = T(y)$ is calculated from the actually measured data y. If it is in the rejection region, H_0 is rejected in favor of H_1. Otherwise, the test fails to reject H_0.

Sources:
- Wikipedia (2021): "Statistical hypothesis testing"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#The_testing_process.

4.3.2 Null hypothesis

Definition: The statement which is tested in a statistical hypothesis test (→ I/4.3.1) is called the "null hypothesis", denoted as H_0. The test is designed to assess the strength of evidence against H_0 and possibly reject it. The opposite of H_0 is called the "alternative hypothesis (→ I/4.3.3)". Usually, H_0 is a statement that a particular parameter is zero, that there is no effect of a particular treatment or that there is no difference between particular conditions.

More precisely, let m be a generative model (→ I/5.1.1) describing measured data y using model parameters $\theta \in \Theta$. Then, a null hypothesis is formally specified as

$$H_0 : \; \theta \in \Theta_0 \quad \text{where} \quad \Theta_0 \subset \Theta \; . \tag{1}$$

Sources:
- Wikipedia (2021): "Exclusion of the null hypothesis"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Basic_definitions.

4.3.3 Alternative hypothesis

Definition: Let H_0 be a null hypothesis (→ I/4.3.2) of a statistical hypothesis test (→ I/4.3.1). Then, the corresponding alternative hypothesis, denoted as H_1, is either the negation of H_0 or an interesting sub-case in the negation of H_0, depending on context. The test is designed to assess the strength of evidence against H_0 and possibly reject it in favor of H_1. Usually, H_1 is a statement that a particular parameter is non-zero, that there is an effect of a particular treatment or that there is a difference between particular conditions.
More precisely, let m be a generative model (→ I/5.1.1) describing measured data y using model parameters $\theta \in \Theta$. Then, null and alternative hypothesis are formally specified as

$$\begin{aligned} H_0 &: \; \theta \in \Theta_0 \quad \text{where} \quad \Theta_0 \subset \Theta \\ H_1 &: \; \theta \in \Theta_1 \quad \text{where} \quad \Theta_1 = \Theta \setminus \Theta_0 \; . \end{aligned} \tag{1}$$

Sources:
- Wikipedia (2021): "Exclusion of the null hypothesis"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: https://en.wikipedia.org/wiki/Exclusion_of_the_null_hypothesis#Basic_definitions.

4.3.4 One-tailed vs. two-tailed

Definition: Let there be a statistical test (→ I/4.3.1) of an alternative hypothesis (→ I/4.3.3) H_1 against a null hypothesis (→ I/4.3.2) H_0. Then,
- the test is called a one-tailed test, if H_1 is a one-tailed hypothesis (→ I/4.2.4);
- the test is called a two-tailed test, if H_1 is a two-tailed hypothesis (→ I/4.2.4).

The fact whether a test (→ I/4.3.1) is one-tailed or two-tailed has consequences for the computation of critical value (→ I/4.3.9) and p-value (→ I/4.3.10).

Sources:
- Wikipedia (2021): "One- and two-tailed tests"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: https://en.wikipedia.org/wiki/One-_and_two-tailed_tests.

4.3.5 Test statistic

Definition: In a statistical hypothesis test (→ I/4.3.1), the test statistic $T(Y)$ is a scalar function of the measured data y whose distribution under the null hypothesis (→ I/4.3.2) H_0 can be established. Together with a significance level (→ I/4.3.8) α, this distribution implies a critical value (→ I/4.3.9) t_{crit} of the test statistic which determines whether the test rejects or fails to reject H_0.

Sources:
- Wikipedia (2021): "Statistical hypothesis testing"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#The_testing_process.

4.3.6 Size of a test

Definition: Let there be a statistical hypothesis test (→ I/4.3.1) with null hypothesis (→ I/4.3.2) H_0. Then, the size of the test is the probability of a false-positive result or making a type I error, i.e. the probability (→ I/1.3.1) of rejecting the null hypothesis (→ I/4.3.2) H_0, given that H_0 is actually true.
For a simple null hypothesis (→ I/4.2.2), the size is determined by the following conditional probability (→ I/1.3.4):

$$\mathrm{Pr}(\text{test rejects } H_0 | H_0) \; . \tag{1}$$

For a composite null hypothesis (→ I/4.2.2), the size is the supremum over all possible realizations of the null hypothesis (→ I/4.3.2):

$$\sup_{h \in H_0} \mathrm{Pr}(\text{test rejects } H_0 | h) \; . \tag{2}$$

Sources:
- Wikipedia (2021): "Size (statistics)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Size_(statistics).

4.3.7 Power of a test

Definition: Let there be a statistical hypothesis test (→ I/4.3.1) with null hypothesis (→ I/4.3.2) H_0 and alternative hypothesis (→ I/4.3.3) H_1. Then, the power of the test is the probability of a true-positive result or not making a type II error, i.e. the probability (→ I/1.3.1) of rejecting H_0, given that H_1 is actually true.
For given null (→ I/4.3.2) and alternative (→ I/4.3.3) hypothesis (→ I/4.2.1), the size is determined by the following conditional probability (→ I/1.3.4):

$$\mathrm{Pr}(\text{test rejects } H_0 | H_1) \; . \tag{1}$$

Sources:
- Wikipedia (2021): "Power of a test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-31; URL: https://en.wikipedia.org/wiki/Power_of_a_test#Description.

4.3.8 Significance level

Definition: Let the size (→ I/4.3.6) of a statistical hypothesis test (→ I/4.3.1) be the probability of a false-positive result or making a type I error, i.e. the probability (→ I/1.3.1) of rejecting the null hypothesis (→ I/4.3.2) H_0, given that H_0 is actually true:

$$\mathrm{Pr}(\text{test rejects } H_0 | H_0) \; . \tag{1}$$

Then, the test is said to have significance level α, if the size is less than or equal to α:

$$\mathrm{Pr}(\text{test rejects } H_0 | H_0) \leq \alpha \; . \tag{2}$$

Sources:

- Wikipedia (2021): "Statistical hypothesis testing"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms.

4.3.9 Critical value

Definition: In a statistical hypothesis test (→ I/4.3.1), the critical value (→ I/4.3.9) t_{crit} is that value of the test statistic (→ I/4.3.5) $T(Y)$ which partitions the set of possible test statistics into "acceptance region" and "rejection region" based on a significance level (→ I/4.3.8) α. Put differently, if the observed test statistic $t_{\mathrm{obs}} = T(y)$ computed from actually measured data y is as extreme or more extreme than the critical value, the test rejects the null hypothesis (→ I/4.3.2) H_0 in favor of the alternative hypothesis (→ I/4.3.3).

Sources:
- Wikipedia (2021): "Statistical hypothesis testing"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms.

4.3.10 p-value

Definition: Let there be a statistical test (→ I/4.3.1) of the null hypothesis (→ I/4.3.2) H_0 and the alternative hypothesis (→ I/4.3.3) H_1 using the test statistic (→ I/4.3.5) $T(Y)$. Let y be the measured data and let $t_{\mathrm{obs}} = T(y)$ be the observed test statistic computed from y. Moreover, assume that $F_T(t)$ is the cumulative distribution function (→ I/1.8.1) (CDF) of the distribution of $T(Y)$ under H_0.
Then, the p-value is the probability of obtaining a test statistic more extreme than or as extreme as t_{obs}, given that the null hypothesis H_0 is true:
- $p = F_T(t_{\mathrm{obs}})$, if H_1 is a left-sided one-tailed hypothesis (→ I/4.2.4);
- $p = 1 - F_T(t_{\mathrm{obs}})$, if H_1 is a right-sided one-tailed hypothesis (→ I/4.2.4);
- $p = 2 \cdot \min\left([F_T(t_{\mathrm{obs}}),\, 1 - F_T(t_{\mathrm{obs}})]\right)$, if H_1 is a two-tailed hypothesis (→ I/4.2.4).

Sources:
- Wikipedia (2021): "Statistical hypothesis testing"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-19; URL: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing#Definition_of_terms.

4.3.11 Distribution of p-value under null hypothesis

Theorem: Under the null hypothesis (→ I/4.3.2), the p-value (→ I/4.3.10) in a statistical test (→ I/4.3.1) follows a continuous uniform distribution (→ II/3.1.1):

$$p \sim \mathcal{U}(0,1) \; . \tag{1}$$

Proof: Without loss of generality, consider a left-sided one-tailed hypothesis test (→ I/4.2.4). Then, the p-value is a function of the test statistic (→ I/4.3.10)

$$
\begin{aligned}
P &= F_T(T) \\
p &= F_T(t_\mathrm{obs})
\end{aligned} \tag{2}
$$

where t_obs is the observed test statistic (→ I/4.3.5) and $F_T(t)$ is the cumulative distribution function (→ I/1.8.1) of the test statistic (→ I/4.3.5) under the null hypothesis (→ I/4.3.2).
Then, we can obtain the cumulative distribution function (→ I/1.8.1) of the p-value (→ I/4.3.10) as

$$
\begin{split}
F_P(p) &= \mathrm{Pr}(P < p) \\
&= \mathrm{Pr}(F_T(T) < p) \\
&= \mathrm{Pr}(T < F_T^{-1}(p)) \\
&= F_T(F_T^{-1}(p)) \\
&= p
\end{split} \tag{3}
$$

which is the cumulative distribution function of a continuous uniform distribution (→ II/3.1.4) over the interval $[0, 1]$:

$$
F_X(x) = \int_{-\infty}^{x} \mathcal{U}(z; 0, 1) \, \mathrm{d}z = x \quad \text{where} \quad 0 \leq x \leq 1 \; . \tag{4}
$$

■

Sources:

- jll (2018): "Why are p-values uniformly distributed under the null hypothesis?"; in: *StackExchange CrossValidated*, retrieved on 2022-03-18; URL: https://stats.stackexchange.com/a/345763/270304.

5 Bayesian statistics

5.1 Probabilistic modeling

5.1.1 Generative model

Definition: Consider measured data y and some unknown latent parameters θ. A statement about the distribution (→ I/1.5.1) of y given θ is called a generative model m

$$m : \, y \sim \mathcal{D}(\theta) \; , \tag{1}$$

where $\mathcal{D}$ denotes an arbitrary probability distribution (→ I/1.5.1) and θ are the parameters of this distribution.

Sources:

- Friston et al. (2008): "Bayesian decoding of brain images"; in: *NeuroImage*, vol. 39, pp. 181-205; URL: https://www.sciencedirect.com/science/article/abs/pii/S1053811907007203; DOI: 10.1016/j.neuroin

5.1.2 Likelihood function

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ. Then, the probability density function (→ I/1.7.1) of the distribution of y given θ is called the likelihood function of m:

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) \; . \tag{1}$$

5.1.3 Prior distribution

Definition: Consider measured data y and some unknown latent parameters θ. A distribution of θ unconditional on y is called a prior distribution:

$$\theta \sim \mathcal{D}(\lambda) \; . \tag{1}$$

The parameters λ of this distribution are called the prior hyperparameters and the probability density function (→ I/1.7.1) is called the prior density:

$$p(\theta|m) = \mathcal{D}(\theta; \lambda) \; . \tag{2}$$

5.1.4 Full probability model

Definition: Consider measured data y and some unknown latent parameters θ. The combination of a generative model (→ I/5.1.1) for y and a prior distribution (→ I/5.1.3) on θ is called a full probability model m:

$$m : \, y \sim \mathcal{D}(\theta), \, \theta \sim \mathcal{D}(\lambda) \; . \tag{1}$$

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Probability and inference"; in: *Bayesian Data Analysis*, ch. 1, p. 3; URL: http://www.stat.columbia.edu/~gelman/book/.

5.1.5 Joint likelihood

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (→ I/5.1.3) on θ. Then, the joint probability (→ I/1.3.2) distribution (→ I/1.5.1) of y and θ is called the joint likelihood:

$$p(y,\theta|m) = p(y|\theta,m)\, p(\theta|m) \; . \tag{1}$$

5.1.6 Joint likelihood is product of likelihood and prior

Theorem: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (→ I/5.1.3) on θ. Then, the joint likelihood (→ I/5.1.5) is equal to the product of likelihood function (→ I/5.1.2) and prior density (→ I/5.1.3):

$$p(y,\theta|m) = p(y|\theta,m)\, p(\theta|m) \; . \tag{1}$$

Proof: The joint likelihood (→ I/5.1.5) is defined as the joint probability (→ I/1.3.2) distribution (→ I/1.5.1) of data y and parameters θ:

$$p(y,\theta|m) \; . \tag{2}$$

Applying the law of conditional probability (→ I/1.3.4), we have:

$$\begin{gathered}
p(y|\theta,m) = \frac{p(y,\theta|m)}{p(\theta|m)} \\
\Leftrightarrow \\
p(y,\theta|m) = p(y|\theta,m)\, p(\theta|m) \; .
\end{gathered} \tag{3}$$

■

5.1.7 Posterior distribution

Definition: Consider measured data y and some unknown latent parameters θ. The distribution (→ I/1.5.1) of θ conditional (→ I/1.5.4) on y is called the posterior distribution:

$$\theta|y \sim \mathcal{D}(\phi) \; . \tag{1}$$

The parameters ϕ of this distribution are called the posterior hyperparameters and the probability density function (→ I/1.7.1) is called the posterior density:

$$p(\theta|y,m) = \mathcal{D}(\theta;\phi) \; . \tag{2}$$

5.1.8 Maximum-a-posteriori estimation

Definition: Consider a posterior distribution (→ I/5.1.7) of an unknown parameter θ, given measured data y, parametrized by posterior hyperparameters ϕ:

$$\theta|y \sim \mathcal{D}(\phi) \; . \tag{1}$$

Then, the value of θ at which the posterior density (→ I/5.1.7) attains its maximum is called the "maximum-a-posteriori estimate", "MAP estimate" or "posterior mode" of θ:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} \mathcal{D}(\theta;\phi) \; . \tag{2}$$

Sources:
- Wikipedia (2023): "Maximum a posteriori estimation"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-12-01; URL: https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation#Description.

5.1.9 Posterior density is proportional to joint likelihood

Theorem: In a full probability model (→ I/5.1.4) m describing measured data y using model parameters θ, the posterior density (→ I/5.1.7) over the model parameters is proportional to the joint likelihood (→ I/5.1.5):

$$p(\theta|y,m) \propto p(y,\theta|m) \; . \tag{1}$$

Proof: In a full probability model (→ I/5.1.4), the posterior distribution (→ I/5.1.7) can be expressed using Bayes' theorem (→ I/5.3.1):

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \; . \tag{2}$$

Applying the law of conditional probability (→ I/1.3.4) to the numerator, we have:

$$p(\theta|y,m) = \frac{p(y,\theta|m)}{p(y|m)} \; . \tag{3}$$

Because the denominator does not depend on θ, it is constant in θ and thus acts a proportionality factor between the posterior distribution and the joint likelihood:

$$p(\theta|y,m) \propto p(y,\theta|m) \; . \tag{4}$$

■

5.1.10 Combined posterior distribution from independent data

Theorem: Let $p(\theta|y_1)$ and $p(\theta|y_2)$ be posterior distributions (→ I/5.1.7), obtained using the same prior distribution (→ I/5.1.3) from conditionally independent (→ I/1.3.7) data sets y_1 and y_2:

$$p(y_1,y_2|\theta) = p(y_1|\theta) \cdot p(y_2|\theta) \; . \tag{1}$$

Then, the combined posterior distribution (→ I/1.5.1) is proportional to the product of the individual posterior densities (→ I/1.7.1), divided by the prior density:

$$p(\theta|y_1,y_2) \propto \frac{p(\theta|y_1) \cdot p(\theta|y_2)}{p(\theta)} \; . \tag{2}$$

Proof: Since $p(\theta|y_1)$ and $p(\theta|y_2)$ are posterior distributions (→ I/5.1.7), Bayes' theorem (→ I/5.3.1) holds for them:

$$
\begin{split}
p(\theta|y_1) &= \frac{p(y_1|\theta) \cdot p(\theta)}{p(y_1)} \\
p(\theta|y_2) &= \frac{p(y_2|\theta) \cdot p(\theta)}{p(y_2)} \; .
\end{split} \tag{3}
$$

Moreover, Bayes' theorem must also hold for the combined posterior distribution (→ I/5.1.9):

$$
p(\theta|y_1,y_2) = \frac{p(y_1,y_2|\theta) \cdot p(\theta)}{p(y_1,y_2)} \; . \tag{4}
$$

With that, we can express the combined posterior distribution as follows:

$$
\begin{split}
p(\theta|y_1,y_2) &\overset{(4)}{=} \frac{p(y_1,y_2|\theta) \cdot p(\theta)}{p(y_1,y_2)} \\
&\overset{(1)}{=} p(y_1|\theta) \cdot p(y_2|\theta) \cdot \frac{p(\theta)}{p(y_1,y_2)} \\
&\overset{(3)}{=} \frac{p(\theta|y_1) \cdot p(y_1)}{p(\theta)} \cdot \frac{p(\theta|y_2) \cdot p(y_2)}{p(\theta)} \cdot \frac{p(\theta)}{p(y_1,y_2)} \\
&= \frac{p(\theta|y_1) \cdot p(\theta|y_2)}{p(\theta)} \cdot \frac{p(y_1) \cdot p(y_2)}{p(y_1,y_2)} \; .
\end{split} \tag{5}
$$

Note that the second fraction does not depend on θ and thus, the posterior distribution over θ is proportional to the first fraction:

$$
p(\theta|y_1,y_2) \propto \frac{p(\theta|y_1) \cdot p(\theta|y_2)}{p(\theta)} \; . \tag{6}
$$

■

5.1.11 Marginal likelihood

Definition: Let there be a generative model (→ I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (→ I/5.1.3) on θ. Then, the marginal probability (→ I/1.3.3) distribution (→ I/1.5.1) of y across the parameter space Θ is called the marginal likelihood:

$$
p(y|m) = \int_{\Theta} p(y,\theta|m) \, \mathrm{d}\theta \; . \tag{1}
$$

5.1.12 Marginal likelihood is integral of joint likelihood

Theorem: In a full probability model (→ I/5.1.4) m describing measured data y using model parameters θ, the marginal likelihood (→ I/5.1.11) is the integral of the joint likelihood (→ I/5.1.5) across the parameter space Θ

$$
p(y|m) = \int_{\Theta} p(y,\theta|m) \, \mathrm{d}\theta \tag{1}
$$

and related to likelihood function (→ I/5.1.2) and prior distribution (→ I/5.1.3) as follows:

$$p(y|m) = \int_{\Theta} p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \; . \tag{2}$$

Proof: In a full probability model (→ I/5.1.4), the marginal likelihood (→ I/5.1.11) is defined as the marginal probability of the data y, given only the model m:

$$p(y|m) \; . \tag{3}$$

Using the law of marginal probabililty (→ I/1.3.3), this can be obtained by integrating the joint likelihood (→ I/5.1.5) function over the entire parameter space:

$$p(y|m) = \int_{\Theta} p(y, \theta|m) \, \mathrm{d}\theta \; . \tag{4}$$

Applying the law of conditional probability (→ I/1.3.4), the integrand can also be written as the product of likelihood function (→ I/5.1.2) and prior density (→ I/5.1.3):

$$p(y|m) = \int_{\Theta} p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \; . \tag{5}$$

■

5.2 Prior distributions

5.2.1 Flat vs. hard vs. soft

Definition: Let $p(\theta|m)$ be a prior distribution (→ I/5.1.3) for the parameter θ of a generative model (→ I/5.1.1) m. Then,

- the distribution is called a "flat prior", if its precision (→ I/1.11.12) is zero or variance (→ I/1.11.1) is infinite;
- the distribution is called a "hard prior", if its precision (→ I/1.11.12) is infinite or variance (→ I/1.11.1) is zero;
- the distribution is called a "soft prior", if its precision (→ I/1.11.12) and variance (→ I/1.11.1) are non-zero and finite.

Sources:

- Friston et al. (2002): "Classical and Bayesian Inference in Neuroimaging: Theory"; in: *NeuroImage*, vol. 16, iss. 2, pp. 465-483, fn. 1; URL: https://www.sciencedirect.com/science/article/pii/S1053811902910906; DOI: 10.1006/nimg.2002.1090.
- Friston et al. (2002): "Classical and Bayesian Inference in Neuroimaging: Applications"; in: *NeuroImage*, vol. 16, iss. 2, pp. 484-512, fn. 10; URL: https://www.sciencedirect.com/science/article/pii/S1053811902910918; DOI: 10.1006/nimg.2002.1091.

5.2.2 Uniform vs. non-uniform

Definition: Let $p(\theta|m)$ be a prior distribution (→ I/5.1.3) for the parameter θ of a generative model (→ I/5.1.1) m where θ belongs to the parameter space Θ. Then,

- the distribution is called a "uniform prior", if its density (→ I/1.7.1) or mass (→ I/1.6.1) is constant over Θ;

- the distribution is called a "non-uniform prior", if its density (→ I/1.7.1) or mass (→ I/1.6.1) is not constant over Θ.

Sources:
- Wikipedia (2020): "Lindley's paradox"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/Lindley%27s_paradox#Bayesian_approach.

5.2.3 Informative vs. non-informative

Definition: Let $p(\theta|m)$ be a prior distribution (→ I/5.1.3) for the parameter θ of a generative model (→ I/5.1.1) m. Then,
- the distribution is called an "informative prior", if it biases the parameter towards particular values;
- the distribution is called a "weakly informative prior", if it mildly influences the posterior distribution (→ I/5.1.9);
- the distribution is called a "non-informative prior", if it does not influence (→ I/5.1.9) the posterior hyperparameters (→ I/5.1.7).

Sources:
- Soch J, Allefeld C, Haynes JD (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection"; in: *NeuroImage*, vol. 141, pp. 469-489, eq. 15, p. 473; URL: https://www.sciencedirect.com/science/article/pii/S1053811916303615; DOI: 10.1016/j.neuroimage.2016.07.047.

5.2.4 Empirical vs. non-empirical

Definition: Let $p(\theta|m)$ be a prior distribution (→ I/5.1.3) for the parameter θ of a generative model (→ I/5.1.1) m. Then,
- the distribution is called an "empirical prior", if it has been derived from empirical data (→ I/5.1.9);
- the distribution is called a "theoretical prior", if it was specified without regard to empirical data.

Sources:
- Soch J, Allefeld C, Haynes JD (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection"; in: *NeuroImage*, vol. 141, pp. 469-489, eq. 13, p. 473; URL: https://www.sciencedirect.com/science/article/pii/S1053811916303615; DOI: 10.1016/j.neuroimage.2016.07.047.

5.2.5 Conjugate vs. non-conjugate

Definition: Let m be a generative model (→ I/5.1.1) with likelihood function (→ I/5.1.2) $p(y|\theta, m)$ and prior distribution (→ I/5.1.3) $p(\theta|m)$. Then,
- the prior distribution (→ I/5.1.3) is called "conjugate", if it, when combined with the likelihood function (→ I/5.1.2), leads to a posterior distribution (→ I/5.1.7) that belongs to the same family of probability distributions (→ I/1.5.1);
- the prior distribution is called "non-conjugate", if this is not the case.

Sources:
- Wikipedia (2020): "Conjugate prior"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: https://en.wikipedia.org/wiki/Conjugate_prior.

5.2.6 Maximum entropy priors

Definition: Let m be a generative model (→ I/5.1.1) with likelihood function (→ I/5.1.2) $p(y|\theta,m)$ and prior distribution (→ I/5.1.3) $p(\theta|\lambda,m)$ using prior hyperparameters (→ I/5.1.3) λ. Then, the prior distribution is called a "maximum entropy prior", if
1) when θ is a discrete random variable (→ I/1.2.6), it maximizes the entropy (→ I/2.1.1) of the prior probability mass function (→ I/1.6.1):

$$\lambda_{\mathrm{maxent}} = \operatorname*{arg\,max}_{\lambda} \mathrm{H}\left[p(\theta|\lambda,m)\right] \; ; \tag{1}$$

2) when θ is a continuous random variable (→ I/1.2.6), it maximizes the differential entropy (→ I/2.2.1) of the prior probability density function (→ I/1.7.1):

$$\lambda_{\mathrm{maxent}} = \operatorname*{arg\,max}_{\lambda} \mathrm{h}\left[p(\theta|\lambda,m)\right] \; . \tag{2}$$

Sources:
- Wikipedia (2020): "Prior probability"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors.

5.2.7 Empirical Bayes priors

Definition: Let m be a generative model (→ I/5.1.1) with likelihood function (→ I/5.1.2) $p(y|\theta,m)$ and prior distribution (→ I/5.1.3) $p(\theta|\lambda,m)$ using prior hyperparameters (→ I/5.1.3) λ. Let $p(y|\lambda,m)$ be the marginal likelihood (→ I/5.1.11) when integrating the parameters out of the joint likelihood (→ I/5.1.12). Then, the prior distribution is called an "Empirical Bayes (→ I/5.3.3) prior", if it maximizes the logarithmized marginal likelihood:

$$\lambda_{\mathrm{EB}} = \operatorname*{arg\,max}_{\lambda} \log p(y|\lambda,m) \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Empirical Bayes method"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: https://en.wikipedia.org/wiki/Empirical_Bayes_method#Introduction.

5.2.8 Reference priors

Definition: Let m be a generative model (→ I/5.1.1) with likelihood function (→ I/5.1.2) $p(y|\theta,m)$ and prior distribution (→ I/5.1.3) $p(\theta|\lambda,m)$ using prior hyperparameters (→ I/5.1.3) λ. Let $p(\theta|y,\lambda,m)$ be the posterior distribution (→ I/5.1.7) that is proportional to the the joint likelihood (→ I/5.1.9). Then, the prior distribution is called a "reference prior", if it maximizes the expected (→ I/1.10.1) Kullback-Leibler divergence (→ I/2.5.1) of the posterior distribution relative to the prior distribution:

$$\lambda_{\mathrm{ref}} = \arg\max_{\lambda} \left\langle \mathrm{KL}\left[p(\theta|y,\lambda,m) \,||\, p(\theta|\lambda,m)\right] \right\rangle \ . \tag{1}$$

Sources:
- Wikipedia (2020): "Prior probability"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-12-02; URL: https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors.

5.3 Bayesian inference

5.3.1 Bayes' theorem

Theorem: Let A and B be two arbitrary statements about random variables (→ I/1.2.2), such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then, the conditional probability that A is true, given that B is true, is equal to

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)} \ . \tag{1}$$

Proof: The conditional probability (→ I/1.3.4) is defined as the ratio of joint probability (→ I/1.3.2), i.e. the probability of both statements being true, and marginal probability (→ I/1.3.3), i.e. the probability of only the second one being true:

$$p(A|B) = \frac{p(A,B)}{p(B)} \ . \tag{2}$$

It can also be written down for the reverse situation, i.e. to calculate the probability that B is true, given that A is true:

$$p(B|A) = \frac{p(A,B)}{p(A)} \ . \tag{3}$$

Both equations can be rearranged for the joint probability

$$p(A|B)\,p(B) \overset{(2)}{=} p(A,B) \overset{(3)}{=} p(B|A)\,p(A) \tag{4}$$

from which Bayes' theorem can be directly derived:

$$p(A|B) \overset{(4)}{=} \frac{p(B|A)\,p(A)}{p(B)} \ . \tag{5}$$

■

Sources:
- Koch, Karl-Rudolf (2007): "Rules of Probability"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, pp. 6/13, eqs. 2.12/2.38; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

5.3.2 Bayes' rule

Theorem: Let A_1, A_2 and B be arbitrary statements about random variables (→ I/1.2.2) where A_1 and A_2 are mutually exclusive. Then, Bayes' rule states that the posterior odds are equal to the Bayes factor (→ IV/**??**) times the prior odds, i.e.

$$\frac{p(A_1|B)}{p(A_2|B)} = \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} \; . \tag{1}$$

Proof: Using Bayes' theorem (→ I/5.3.1), the conditional probabilities (→ I/1.3.4) on the left are given by

$$p(A_1|B) = \frac{p(B|A_1) \cdot p(A_1)}{p(B)} \tag{2}$$

$$p(A_2|B) = \frac{p(B|A_2) \cdot p(A_2)}{p(B)} \; . \tag{3}$$

Dividing the two conditional probabilities by each other

$$\begin{split} \frac{p(A_1|B)}{p(A_2|B)} &= \frac{p(B|A_1) \cdot p(A_1)/p(B)}{p(B|A_2) \cdot p(A_2)/p(B)} \\ &= \frac{p(B|A_1)}{p(B|A_2)} \cdot \frac{p(A_1)}{p(A_2)} \; , \end{split} \tag{4}$$

one obtains the posterior odds ratio as given by the theorem.

■

Sources:
- Wikipedia (2019): "Bayes' theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-06; URL: https://en.wikipedia.org/wiki/Bayes%27_theorem#Bayes%E2%80%99_rule.

5.3.3 Empirical Bayes

Definition: Let m be a generative model (→ I/5.1.1) with model parameters θ and hyper-parameters λ implying the likelihood function (→ I/5.1.2) $p(y|\theta,\lambda,m)$ and prior distribution (→ I/5.1.3) $p(\theta|\lambda,m)$. Then, an Empirical Bayes treatment of m, also referred to as "type II maximum likelihood (→ I/4.1.3)" or "evidence (→ IV/**??**) approximation", consists in

1) evaluating the marginal likelihood (→ I/5.1.11) of the model m

$$p(y|\lambda,m) = \int p(y|\theta,\lambda,m) \, (\theta|\lambda,m) \, \mathrm{d}\theta \; , \tag{1}$$

2) maximizing the log model evidence (→ IV/**??**) with respect to λ

$$\hat{\lambda} = \operatorname*{arg\,max}_{\lambda} \log p(y|\lambda,m) \tag{2}$$

3) and using the prior distribution (→ I/5.1.3) at this maximum

$$p(\theta|m) = p(\theta|\hat{\lambda}, m) \tag{3}$$

for Bayesian inference (→ I/5.3.1), i.e. obtaining the posterior distribution (→ I/5.1.9) and computing the marginal likelihood (→ I/5.1.12).

Sources:
- Wikipedia (2021): "Empirical Bayes method"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-29; URL: https://en.wikipedia.org/wiki/Empirical_Bayes_method#Introduction.
- Bishop CM (2006): "The Evidence Approximation"; in: *Pattern Recognition for Machine Learning*, ch. 3.5, pp. 165-172; URL: https://www.springer.com/gp/book/9780387310732.

5.3.4 Variational Bayes

Definition: Let m be a generative model (→ I/5.1.1) with model parameters θ implying the likelihood function (→ I/5.1.2) $p(y|\theta, m)$ and prior distribution (→ I/5.1.3) $p(\theta|m)$. Then, a Variational Bayes treatment of m, also referred to as "approximate inference" or "variational inference", consists in

1) constructing an approximate posterior distribution (→ I/5.1.7)

$$q(\theta) \approx p(\theta|y, m) \; , \tag{1}$$

2) evaluating the variational free energy (→ IV/**??**)

$$F_q(m) = \int q(\theta) \log p(y|\theta, m) \, \mathrm{d}\theta - \int q(\theta) \frac{q(\theta)}{p(\theta|m)} \, \mathrm{d}\theta \tag{2}$$

3) and maximizing this function with respect to $q(\theta)$

$$\hat{q}(\theta) = \operatorname*{arg\,max}_q F_q(m) \; . \tag{3}$$

for Bayesian inference (→ I/5.3.1), i.e. obtaining the posterior distribution (→ I/5.1.7) (from eq. (3)) and approximating the marginal likelihood (→ I/5.1.11) (by plugging eq. (3) into eq. (2)).

Sources:
- Wikipedia (2021): "Variational Bayesian methods"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-29; URL: https://en.wikipedia.org/wiki/Variational_Bayesian_methods#Evidence_lower_bound.
- Penny W, Flandin G, Trujillo-Barreto N (2007): "Bayesian Comparison of Spatially Regularised General Linear Models"; in: *Human Brain Mapping*, vol. 28, pp. 275–293, eqs. 2-9; URL: https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327; DOI: 10.1002/hbm.20327.

6 Machine learning

6.1 Scoring rules

6.1.1 Scoring rule

Definition: A scoring rule is any extended real-valued function $\mathbf{S} : \mathcal{Q} \times \Omega \rightarrow \mathbb{R}$ where $\mathcal{Q}$ is a family of probability distributions over the space Ω, such that $\mathbf{S}(Q, \cdot)$ is $\mathcal{Q}$-quasi-integrable for all $Q \in \mathcal{Q}$. Output of the function $\mathbf{S}(Q, y)$ represents the loss or penalty when the forecast $Q \in \mathcal{Q}$ is issued and the observation $y \in \Omega$ is realized.

Sources:
- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.
- Wikipedia (2024): “Scoring rule”; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-02-28; URL: https://en.wikipedia.org/wiki/Scoring_rule.

6.1.2 Proper scoring rule

Definition: A scoring rule (→ I/6.1.1) $\mathbf{S}$ is called a proper scoring rule, if and only if

$$\max_{Q \in \mathcal{Q}} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = \mathbb{E}_{Y \sim P}[\mathbf{S}(P, Y)] \; . \tag{1}$$

In other words, score function $\mathbf{S}$ is a proper scoring rule, if it is maximized when the forecaster gives exactly the ground truth distribution $P(Y)$ as its probabilistic forecast $Q \in \mathcal{Q}$.

Sources:
- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.

6.1.3 Strictly proper scoring rule

Definition: A scoring rule (→ I/6.1.1) $\mathbf{S}$ is called a strictly proper scoring rule, if and only if
- $\mathbf{S}$ is a proper scoring rule (→ I/6.1.2), and
- $\arg\max \_Q \in \mathcal{Q} \mathbb{E}\_Y \sim P[\mathbf{S}(Q, Y)] = P$ is the unique maximizer of $\mathbf{S}$ in Q.

In other words, a strictly proper scoring rule is maximized only when the the forecaster gives exactly the ground truth distribution $P(Y)$ as its probabilistic forecast $Q \in \mathcal{Q}$.

Sources:
- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): “Proper/Strictly Proper Scoring Rule”; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.

6.1.4 Log probability scoring rule

Definition: A log (probability) scoring rule (→ I/6.1.1) $S(q, y)$ is as a scoring rule that measures the quality of a probabilistic forecast in decision theory. Formally, it can be defined in discrete or continuous form as follows:
1) Log scoring rule for binary classification:

$$S(q, y) = \begin{cases} \log q \;, & \text{if } y = 1 \\ \log(1-q) \;, & \text{if } y = 0 \end{cases} \tag{1}$$

which can be expressed as

$$S(q, y) = y \log q + (1-y) \log(1-q) \tag{2}$$

Note that the expressions given above have slightly different domains. For the first equation, the domain is $D_1 = ([0,1) \times \{0\}) \cup ((0,1] \times \{1\})$, while for the second equation, the domain is $D_2 = (0,1) \times \{0,1\}$.
2) Log scoring rule for multiclass classification:

$$S(q, y) = \sum_k y_k \log q_k(x) = \log q_{y^*}(x) \tag{3}$$

where y^* is the true class and q is the predicted probability distribution over the classes. We have $y_k = 1$, if the true class is k and $y_k = 0$ otherwise.
3) Log scoring rule for regression (continuous case):

$$S(q, y) = \log q(y) \tag{4}$$

where q is the predicted probability distribution over the continuous space and y is the true value.

Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): "Proper/Strictly Proper Scoring Rule"; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.

6.1.5 Log probability is strictly proper scoring rule

Theorem: The log (probability) scoring rule (→ I/6.1.4) is a strictly proper scoring rule (→ I/6.1.3).

Proof: We will show that all versions of the log probability scoring rule (binary/multiclass/regression) are strictly proper scoring rules.
1) Binary log probability scoring rule:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = P(Y=1) \log q + P(Y=0) \log(1-q) \tag{1}$$

Let p be the true probability of the event $Y = 1$. Then, the expected score is:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = p \log q + (1-p) \log(1-q) \tag{2}$$

To find the maxima, take the derivative with respect to q and set it to zero:

$$
\begin{aligned}
\frac{\partial}{\partial q}\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= \frac{p}{q} - \frac{1-p}{1-q} \\
0 &= \frac{p-pq-q+pq}{q(1-q)} \\
0 &= \frac{p-q}{q(1-q)} \\
\Rightarrow p-q &= 0 \\
\Rightarrow p &= q
\end{aligned} \tag{3}
$$

Now, we need to check the second derivative to see, if it is a maximum for the properness condition and if it is the only maximizer for the strictness condition:

$$
\begin{aligned}
\frac{\partial^2}{\partial q^2}\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= -\frac{p}{q^2} - \frac{1-p}{(1-q)^2} \\
&= -\left(\underbrace{\frac{p}{q^2}}_{>0} + \underbrace{\frac{1-p}{(1-q)^2}}_{>0}\right) < 0
\end{aligned} \tag{4}
$$

Except for the cases $q = 0$ and $q = 1$, the second derivative is always negative, which means that the function is concave and the maximum is unique. For $q = 1$, maximum is achieved only if $p = 1$, and similarly for $q = 0$, maximum is achieved only if $p = 0$. Therefore, $p = q$ is the only maximizer and the log probability scoring rule for binary classification is strictly proper.
2a) Multiclass log probability scoring rule (Proof 1):

$$
S(q,y) = \sum_{k}^{K} y_k \log q_k(x) \tag{5}
$$

Let p_k be the true probability of the event $Y = k$. Since q_k is the predicted probability for class k, we know that $\sum_i q_i = 1$. Then, the expected score is:

$$
\begin{aligned}
\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= \sum_{k} P(Y=k|x)\log(q_k(x)) \\
&= p_1\log(q_1(x)) + p_2\log(q_2(x)) + ... + p_K\log(q_K(x)) \\
&= p_1\log(q_1(x)) + p_2\log(q_2(x)) + ... + p_K\log(1-\sum_{i\neq K} q_i(x))
\end{aligned} \tag{6}
$$

Taking the derivative with respect to q_j and setting it to zero:

$$
\begin{aligned}
\frac{\partial}{\partial q_j}\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= \frac{p_j}{q_j} - \frac{p_K}{1-\sum_{i\neq K} q_i(x)} \\
0 &= \frac{p_j}{q_j} - \frac{p_K}{q_K} \\
\Rightarrow \frac{p_j}{q_j} &= \frac{p_K}{q_K}
\end{aligned} \tag{7}
$$

This equality holds for any j:

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} = ... = \frac{p_K}{q_K} = \lambda \tag{8}$$

Each q_i can be represented as a constant multiple of p_i as follows: $q_i = \lambda\ p_i$

$$\begin{aligned} \sum_i q_i &= 1 \\ \sum_i \lambda\ p_i &= 1 \\ \lambda \sum_i p_i &= 1 \\ \lambda &= 1 \end{aligned} \tag{9}$$

Since $\lambda = 1$, we have $p_i = q_i$ for all i. Now, we need to check the second derivative to see, if it is a maximum for the properness condition and if it is the only maximizer for the strictness condition:

$$\begin{aligned} \frac{\partial^2}{\partial q_j^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= -\frac{p_j}{q_j^2} - \frac{p_K}{(1 - \sum_{i \neq K} q_i(x))^2} \\ &= -\left(\underbrace{\frac{p_j}{q_j^2}}_{>0} + \underbrace{\frac{p_K}{(1 - \sum_{i \neq K} q_i(x))^2}}_{>0} \right) < 0 \end{aligned} \tag{10}$$

Except for the cases $q_j = 0$ and $q_j = 1$, the second derivative is always negative, which means that the function is concave and the maximum is unique. For $q_j = 1$, maximum is achieved only if $p_j = 1$, and similarly for $q_j = 0$ maximum is achieved only if $p_j = 0$. Therefore, $p_j = q_j$ is the only maximizer and the log probability scoring rule for multiclass classification is strictly proper.
2b) Multiclass log probability scoring rule (Proof 2):
Alternatively, we can solve the optimization problem with Lagrange multipliers. The Lagrangian is:

$$\mathcal{L}(q, \lambda) = \sum_k P(Y = k|x) \log(q_k(x)) + \lambda \left(1 - \sum_k q_k(x) \right) \tag{11}$$

Taking the derivative with respect to q_j and setting it to zero:

$$\begin{aligned} \frac{\partial}{\partial q_j} \mathcal{L}(q, \lambda) &= \frac{p_j}{q_j} - \lambda \\ 0 &= \frac{p_j}{q_j} - \lambda \\ \Rightarrow \frac{p_j}{q_j} &= \lambda \end{aligned} \tag{12}$$

The rest of the proof follows as in the first proof.
3) Continuous log probability scoring rule:

$$S(q, y) = \log q(y) \tag{13}$$

Let $p(y)$ be the true probability density function (→ I/1.7.1) of the event $Y = y$. Then, the expected (→ I/1.10.1) score is:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = \int p(y) \log q(y) dy \tag{14}$$

Let $X = \frac{q(y)}{p(y)}$ and $\phi = \log(\cdot)$ (a concave function). By Jensen's inequality, we know that $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$, if f is concave. Therefore, we have:

$$\begin{aligned}
\int p(y) \log \frac{q(y)}{p(y)} dy &\leq \log \int p(y) \frac{q(y)}{p(y)} dy \\
\int p(y) \log \frac{q(y)}{p(y)} dy &\leq \log \int q(y) dy \\
\int p(y) \log \frac{q(y)}{p(y)} dy &\leq \log(1) \\
\int p(y) \log \frac{q(y)}{p(y)} dy &\leq 0
\end{aligned} \tag{15}$$

The same result can be obtained by using the Kullback-Leibler divergence (→ I/2.5.1). The Kullback-Leibler divergence is always non-negative (→ I/2.5.3), therefore $\mathrm{E} - \mathrm{CE} = \mathrm{KL} \geq 0$. The resulting expression is $-\mathrm{KL}$, which is always non-positive. It is maximized only when $q(y) = p(y)$ which means that the log probability scoring rule for continuous classification is strictly proper.
An alternative argument for uniqueness of the maximum point can be proposed as follows: $\int p(y) \log \frac{q(y)}{p(y)} dy$ can be equal to 0 in two cases: Either $\frac{q(y)}{p(y)}$ is equal to 1 for each value or the expression $\log(\frac{q(y)}{p(y)})$ takes positive and negative values summing up to 0 at the end. The second case cannot occur, because it means that there exists a y_0 such that $q(y_0) > p(y_0)$, implying that Jensen's inequality is violated. Therefore, the maximum is achieved, if and only if $q = p$.

■

Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): "Proper/Strictly Proper Scoring Rule"; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.

6.1.6 Brier scoring rule

Definition: A Brier scoring rule (→ I/6.1.1) $S(q, y)$ is as a scoring rule that measures the quality of a probabilistic forecast in decision theory. Formally, it can be defined for binary or multiclass classification as follows:
1) Brier scoring rule for binary classification:

$$S(q, y) = -(q - y)^2 \tag{1}$$

q represents the predicted probability of the positive class ($Y = 1$) and y is the true class label. Since we want the output of the scoring rule to be maximized when the predicted probability is close to the true class label, we use the negative of the squared difference between the predicted probability and the true class label.
2) Brier scoring rule for multiclass classification:

$$S(q, y) = -\sum_k (q_k - y_k)^2 = -(q_{y^*} - 1)^2 - \sum_{k \neq y^*} q_k^2 \tag{2}$$

where q_k is the predicted probability of class k and y^* is the true class label. Similar to the log probability scoring rule, we have $y_k = 1$, if the true class is k and $y_k = 0$ otherwise.
3) Regression (continuous case):
Although there is no direct version of Brier score for regression, we can use the squared error loss as a scoring rule for regression problems as well.

Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): "Proper/Strictly Proper Scoring Rule"; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.

6.1.7 Brier scoring rule is strictly proper scoring rule

Theorem: The brier scoring rule (→ I/6.1.6) is a strictly proper scoring rule (→ I/6.1.3).

Proof: We will show that both versions of the brier scoring rule (binary/multiclass) are strictly proper scoring rules.
1) Brier scoring rule for binary classification:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -P(Y = 1)(q - 1)^2 + P(Y = 0) - q^2 \tag{1}$$

Let p be the true probability of the event $Y = 1$. Then, the expected score is:

$$\mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -p(q - 1)^2 - (1 - p)q^2 \tag{2}$$

To find the maxima, take the derivative with respect to q and set it to zero:

$$\begin{aligned} \frac{\partial}{\partial q} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] &= -2p(q - 1) - 2(1 - p)q \\ &= -2pq + 2p - 2q + 2pq \\ &= 2p - 2q \\ 0 &= 2p - 2q \\ \Rightarrow p &= q \end{aligned} \tag{3}$$

We need to check the second derivative to see if it is a maximum (for the properness condition) and if it is the only maximizer (for the strictness condition):

$$\frac{\partial^2}{\partial q^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -2 < 0 \tag{4}$$

The second derivative is always negative which means that the function is concave and the maximum is unique. Therefore, $p = q$ is the only maximizer and the Brier scoring rule for binary classification is strictly proper.
2) Brier scoring rule for multiclass classification:

$$
\begin{aligned}
\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= \sum_k P(Y=k)\Big[-\sum_i (q_i-y_i)^2\Big] \\
&= \sum_k P(Y=k)\Big[-(q_k-1)^2-\sum_{i\neq k} q_i^2\Big] \\
&= \sum_k P(Y=k)\Big[-(q_k-1)^2+q_k^2-\sum_i q_i^2\Big] \\
&= \sum_k P(Y=k)\Big[-q_k^2-1+2q_k+q_k^2-\sum_i q_i^2\Big] \\
&= \sum_k P(Y=k)\Big[2q_k-1-\sum_i q_i^2\Big] \\
&= \sum_k P(Y=k)(2q_k-1)-\sum_k P(Y=k)\Big(\sum_i q_i^2\Big) \\
&= \sum_k P(Y=k)(2q_k-1)-\sum_i q_i^2\Big(\underbrace{\sum_k P(Y=k)}_{1}\Big) \\
&= \sum_k P(Y=k)(2q_k-1)-\sum_i q_i^2 \\
&= \sum_k P(Y=k)(2q_k-1)-q_k^2
\end{aligned} \tag{5}
$$

Similar to what we did for log probability (→ I/6.1.5), this expression can be expressed as follows (replacing q_K with $1-\sum_{i\neq K} q_i$):

$$
\begin{aligned}
\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= p_1(2q_1-1)-q_1^2+p_2(2q_2-1)-q_2^2+\ldots+p_K(2q_K-1)-q_K^2 \\
&= p_1(2q_1-1)-q_1^2+p_2(2q_2-1)-q_2^2+\ldots+p_K\left(1-2\sum_{i\neq K} q_i\right)-\left(1-\sum_{i\neq K} q_i\right)^2
\end{aligned} \tag{6}
$$

Taking the derivative with respect to q_j and setting it to zero, we obtain:

$$
\begin{aligned}
\frac{\partial}{\partial q_j}\mathbb{E}_{Y\sim P}[\mathbf{S}(Q,Y)] &= 2p_j-2q_j-2p_K+2\left(1-\sum_{i\neq K} q_i\right) \\
&= 2p_j-2q_j-2p_K+2q_K \\
&= (p_j-q_j)+(q_K-p_K) \\
(p_j-q_j) &= (p_K-q_K)
\end{aligned} \tag{7}
$$

We know that $\sum_i p_i = 1$ and $\sum_i q_i = 1$, therefore:

$$
\begin{aligned}
p_1 - q_1 = p_2 - q_2 = \ldots = p_K - q_K = \lambda \\
\sum_i p_i - q_i = K \cdot \lambda = 0 \\
\Rightarrow \lambda = 0 \quad \text{since } K \neq 0 \\
\Rightarrow p_i = q_i \quad \text{for all } i = 1, \ldots, K
\end{aligned} \tag{8}
$$

Now, we need to check the second derivative to see, if it is a maximum for the properness condition and if it is the only maximizer for the strictness condition:

$$
\frac{\partial^2}{\partial q_j^2} \mathbb{E}_{Y \sim P}[\mathbf{S}(Q, Y)] = -2 - 2 = -4 < 0 \tag{9}
$$

The second derivative is always negative which means that the function is concave and the maximum is unique. Therefore, $p = q$ is the only maximizer and the Brier scoring rule for multiclass classification is strictly proper.

■

Sources:

- Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, Seong Joon Oh (2023): "Proper/Strictly Proper Scoring Rule"; in: *Trustworthy Machine Learning*; URL: https://trustworthyml.io/; DOI: 10.48550/arXiv.2310.08215.

Chapter II

Probability Distributions

1 Univariate discrete distributions

1.1 Discrete uniform distribution

1.1.1 Definition

Definition: Let X be a discrete random variable (→ I/1.2.2). Then, X is said to be uniformly distributed with minimum a and maximum b

$$X \sim \mathcal{U}(a, b) \; , \tag{1}$$

if and only if each integer between and including a and b occurs with the same probability.

Sources:
- Wikipedia (2020): "Discrete uniform distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-28; URL: https://en.wikipedia.org/wiki/Discrete_uniform_distribution.

1.1.2 Probability mass function

Theorem: Let X be a random variable (→ I/1.2.2) following a discrete uniform distribution (→ II/1.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \frac{1}{b-a+1} \quad \text{where} \quad x \in \{a, a+1, \ldots, b-1, b\} \; . \tag{2}$$

Proof: A discrete uniform variable is defined as (→ II/1.1.1) having the same probability for each integer between and including a and b. The number of integers between and including a and b is

$$n = b - a + 1 \tag{3}$$

and because the sum across all probabilities (→ I/1.6.1) is

$$\sum_{x=a}^{b} f_X(x) = 1 \; , \tag{4}$$

we have

$$f_X(x) = \frac{1}{n} = \frac{1}{b-a+1} \; . \tag{5}$$

■

1.1.3 Cumulative distribution function

Theorem: Let X be a random variable (→ I/1.2.2) following a discrete uniform distribution (→ II/1.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \begin{cases} 0\;, & \text{if } x < a \\ \frac{\lfloor x \rfloor - a + 1}{b-a+1}\;, & \text{if } a \le x \le b \\ 1\;, & \text{if } x > b\;. \end{cases} \tag{2}$$

Proof: The probability mass function of the discrete uniform distribution (→ II/1.1.2) is

$$\mathcal{U}(x; a, b) = \frac{1}{b-a+1} \quad \text{where} \quad x \in \{a, a+1, \ldots, b-1, b\}\;. \tag{3}$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$F_X(x) = \int_{-\infty}^{x} \mathcal{U}(z; a, b)\,\mathrm{d}z \tag{4}$$

From (3), it follows that the cumulative probability increases step-wise by $1/n$ at each integer between and including a and b where

$$n = b - a + 1 \tag{5}$$

is the number of integers between and including a and b. This can be expressed by noting that

$$F_X(x) \overset{(3)}{=} \frac{\lfloor x \rfloor - a + 1}{n},\; \text{if } a \le x \le b\;. \tag{6}$$

Also, because $\mathrm{Pr}(X < a) = 0$, we have

$$F_X(x) \overset{(4)}{=} \int_{-\infty}^{x} 0\,\mathrm{d}z = 0,\; \text{if } x < a \tag{7}$$

and because $\mathrm{Pr}(X > b) = 0$, we have

$$\begin{split} F_X(x) &\overset{(4)}{=} \int_{-\infty}^{x} \mathcal{U}(z; a, b)\,\mathrm{d}z \\ &= \int_{-\infty}^{b} \mathcal{U}(z; a, b)\,\mathrm{d}z + \int_{b}^{x} \mathcal{U}(z; a, b)\,\mathrm{d}z \\ &= F_X(b) + \int_{b}^{x} 0\,\mathrm{d}z \overset{(6)}{=} 1 + 0 \\ &= 1,\; \text{if } x > b\;. \end{split} \tag{8}$$

This completes the proof.

■

1.1.4 Quantile function

Theorem: Let X be a random variable (→ I/1.2.2) following a discrete uniform distribution (→ II/1.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the quantile function (→ I/1.9.1) of X is

$$Q_X(p) = \left\{ \begin{array}{rl} -\infty \; , & \text{if} \; p = 0 \\ a(1-p) + (b+1)p - 1 \; , & \text{when} \; p \in \left\{ \frac{1}{n}, \frac{2}{n}, \ldots, \frac{b-a}{n}, 1 \right\} \; . \end{array} \right. \tag{2}$$

with $n = b-a+1$.

Proof: The cumulative distribution function of the discrete uniform distribution (→ II/1.1.3) is:

$$F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < a \\ \frac{\lfloor x \rfloor - a + 1}{b-a+1} \; , & \text{if} \; a \leq x \leq b \\ 1 \; , & \text{if} \; x > b \; . \end{array} \right. \tag{3}$$

The quantile function (→ I/1.9.1) $Q_X(p)$ is defined as the smallest x, such that $F_X(x) = p$:

$$Q_X(p) = \min \left\{ x \in \mathbb{R} \, | \, F_X(x) = p \right\} \; . \tag{4}$$

Because the CDF only returns (→ II/1.1.3) multiples of $1/n$ with $n = b-a+1$, the quantile function (→ I/1.9.1) is only defined for such values. First, we have $Q_X(p) = -\infty$, if $p = 0$. Second, since the cumulative probability increases step-wise (→ II/1.1.3) by $1/n$ at each integer between and including a and b, the minimum x at which

$$F_X(x) = \frac{c}{n} \quad \text{where} \quad c \in \{1, \ldots, n\} \tag{5}$$

is given by

$$Q_X\left(\frac{c}{n}\right) = a + \frac{c}{n} \cdot n - 1 \; . \tag{6}$$

Substituting $p = c/n$ and $n = b-a+1$, we can finally show:

$$\begin{split} Q_X(p) &= a + p \cdot (b-a+1) - 1 \\ &= a + pb - pa + p - 1 \\ &= a(1-p) + (b+1)p - 1 \; . \end{split} \tag{7}$$

■

1.1.5 Shannon entropy

Theorem: Let X be a random variable (→ I/1.2.2) following a discrete uniform distribution (→ II/1.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the (Shannon) entropy (→ I/2.1.1) of X in nats is

$$\mathrm{H}(X) = \ln(b-a+1) \; . \tag{2}$$

Proof: The entropy (→ I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \; . \tag{3}$$

Entropy is measured in nats by setting $b = e$. Then, with the probability mass function of the discrete uniform distribution (→ II/1.1.2), we have:

$$\begin{split} \mathrm{H}(X) &= -\sum_{x \in \mathcal{X}} p(x) \cdot \log_e p(x) \\ &= -\sum_{x=a}^{b} p(x) \cdot \ln p(x) \\ &= -\sum_{x=a}^{b} \frac{1}{b-a+1} \cdot \ln \frac{1}{b-a+1} \\ &= -(b-a+1) \cdot \frac{1}{b-a+1} \cdot \ln \frac{1}{b-a+1} \\ &= -\ln \frac{1}{b-a+1} \\ &= \ln(b-a+1) \; . \end{split} \tag{4}$$

■

1.1.6 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ I/1.2.2). Assume two discrete uniform distributions (→ II/**??**) P and Q specifying the probability distribution of X as

$$\begin{split} P: \; X &\sim \mathcal{U}(a_1, b_1) \\ Q: \; X &\sim \mathcal{U}(a_2, b_2) \; . \end{split} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P \,||\, Q] = \ln \frac{b_2 - a_2 + 1}{b_1 - a_1 + 1} \; . \tag{2}$$

Proof: The KL divergence for a discrete random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P \,||\, Q] = \sum_{x \in \mathcal{X}} p(x) \, \ln \frac{p(x)}{q(x)} \; . \tag{3}$$

This means that the KL divergence of P from Q is only defined, if for all $x \in \mathcal{X}$, $q(x) = 0$ implies $p(x) = 0$. Thus, $\mathrm{KL}[P \,||\, Q]$ only exists, if $a_2 \leq a_1$ and $b_1 \leq b_2$, i.e. if P only places non-zero probability

where Q also places non-zero probability, such that $q(x)$ is not zero for any $x \in \mathcal{X}$ where $p(x)$ is positive.
If this requirement is fulfilled, we can write

$$\mathrm{KL}[P\,||\,Q] = \sum_{x=-\infty}^{a_1} p(x)\,\ln \frac{p(x)}{q(x)} + \sum_{x=a_1}^{b_1} p(x)\,\ln \frac{p(x)}{q(x)} + \sum_{x=b_1}^{+\infty} p(x)\,\ln \frac{p(x)}{q(x)} \tag{4}$$

and because $p(x) = 0$ for any $x < a_1$ and any $x > b_1$, we have

$$\mathrm{KL}[P\,||\,Q] = \sum_{x=-\infty}^{a_1} 0 \cdot \ln \frac{0}{q(x)} + \sum_{x=a_1}^{b_1} p(x)\,\ln \frac{p(x)}{q(x)} + \sum_{x=b_1}^{+\infty} 0 \cdot \ln \frac{0}{q(x)} \; . \tag{5}$$

Now, $(0 \cdot \ln 0)$ is taken to be 0 by convention (→ I/2.1.1), such that

$$\mathrm{KL}[P\,||\,Q] = \sum_{x=a_1}^{b_1} p(x)\,\ln \frac{p(x)}{q(x)} \tag{6}$$

and we can use the probability mass function of the discrete uniform distribution (→ II/1.1.2) to evaluate:

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \sum_{x=a_1}^{b_1} \frac{1}{b_1-a_1+1} \cdot \ln \frac{\frac{1}{b_1-a_1+1}}{\frac{1}{b_2-a_2+1}} \\
&= \frac{1}{b_1-a_1+1} \cdot \ln \frac{b_2-a_2+1}{b_1-a_1+1} \sum_{x=a_1}^{b_1} 1 \\
&= \frac{1}{b_1-a_1+1} \cdot \ln \frac{b_2-a_2+1}{b_1-a_1+1} \cdot (b_1-a_1+1) \\
&= \ln \frac{b_2-a_2+1}{b_1-a_1+1} \; .
\end{split} \tag{7}$$

■

1.1.7 Maximum entropy distribution

Theorem: The discrete uniform distribution (→ II/1.1.1) maximizes (Shannon) entropy (→ I/2.1.1) for a random variable (→ I/1.2.2) with finite support.

Proof: A random variable with finite support is a discrete random variable (→ I/1.2.6). Let X be such a random variable. Without loss of generality, we can assume that the possible values of the X can be enumerated from 1 to n.
Let $g(x)$ be the discrete uniform distribution with minimum $a = 1$ and maximum $b = n$ which assigns to equal probability to all n possible values and let $f(x)$ be an arbitrary discrete (→ I/1.2.6) probability distribution (→ I/1.5.1) on the set $\{1, 2, \ldots, n-1, n\}$.
For a discrete random variable (→ I/1.2.6) X with set of possible values $\mathcal{X}$ and probability mass function (→ I/1.6.1) $p(x)$, the Shannon entropy (→ I/2.1.1) is defined as:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{1}$$

Consider the Kullback-Leibler divergence (→ I/2.5.1) of distribution $f(x)$ from distribution $g(x)$ which is non-negative (→ I/2.5.2):

$$
\begin{split}
0 \leq \mathrm{KL}[f||g] &= \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \\
&= \sum_{x \in \mathcal{X}} f(x) \log f(x) - \sum_{x \in \mathcal{X}} f(x) \log g(x) \\
&\overset{(1)}{=} -\mathrm{H}[f(x)] - \sum_{x \in \mathcal{X}} f(x) \log g(x) \; .
\end{split} \tag{2}
$$

By plugging the probability mass function of the discrte uniform distribution (→ II/1.1.2) into the second term, we obtain:

$$
\begin{split}
\sum_{x \in \mathcal{X}} f(x) \log g(x) &= \sum_{x=1}^{n} f(x) \log \frac{1}{n-1+1} \\
&= \log \frac{1}{n} \sum_{x=1}^{n} f(x) \\
&= -\log(n) \; .
\end{split} \tag{3}
$$

This is actually the negative of the entropy of the discrete uniform distribution (→ II/1.1.5), such that:

$$
\sum_{x \in \mathcal{X}} f(x) \log g(x) = -\mathrm{H}[\mathcal{U}(1,n)] = -\mathrm{H}[g(x)] \; . \tag{4}
$$

Combining (2) with (4), we can show that

$$
\begin{split}
0 &\leq \mathrm{KL}[f||g] \\
0 &\leq -\mathrm{H}[f(x)] - (-\mathrm{H}[g(x)]) \\
\mathrm{H}[g(x)] &\geq \mathrm{H}[f(x)]
\end{split} \tag{5}
$$

which means that the entropy (→ I/2.1.1) of the discrete uniform distribution (→ II/1.1.1) $\mathcal{U}(a,b)$ will be larger than or equal to any other distribution (→ I/1.5.1) defined on the same set of values $\{a, \ldots, b\}$.

■

Sources:

- Probability Fact (2023): "The entropy of a distribution with finite domain"; in: *Twitter*, retrieved on 2023-08-18; URL: https://twitter.com/ProbFact/status/1673787091610750980.

1.2 Bernoulli distribution

1.2.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to follow a Bernoulli distribution with success probability p

$$X \sim \mathrm{Bern}(p) \; , \tag{1}$$

if $X = 1$ with probability (→ I/1.3.1) p and $X = 0$ with probability (→ I/1.3.1) $q = 1 - p$.

Sources:

- Wikipedia (2020): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution.

1.2.2 Probability mass function

Theorem: Let X be a random variable (→ I/1.2.2) following a Bernoulli distribution (→ II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \left\{ \begin{array}{rl} p \; , & \text{if} \; x = 1 \\ 1 - p \; , & \text{if} \; x = 0 \; . \end{array} \right. \tag{2}$$

Proof: This follows directly from the definition of the Bernoulli distribution (→ II/1.2.1).

■

1.2.3 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a Bernoulli distribution (→ II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = p \; . \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average of all possible values:

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x) \; . \tag{3}$$

Since there are only two possible outcomes for a Bernoulli random variable (→ II/1.2.2), we have:

$$\begin{split} \mathrm{E}(X) &= 0 \cdot \mathrm{Pr}(X = 0) + 1 \cdot \mathrm{Pr}(X = 1) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p \; . \end{split} \tag{4}$$

■

Sources:

- Wikipedia (2020): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-16; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution#Mean.

1.2.4 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a Bernoulli distribution (→ II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the variance (→ I/1.10.1) of X is

$$\mathrm{Var}(X) = p \, (1-p) \; . \tag{2}$$

Proof: The variance (→ I/1.11.1) is the probability-weighted average of the squared deviation from the expected value (→ I/1.10.1) across all possible values

$$\mathrm{Var}(X) = \sum_{x \in \mathcal{X}} (x - \mathrm{E}(X))^2 \cdot \mathrm{Pr}(X = x) \tag{3}$$

and can also be written in terms of the expected values (→ I/1.11.3):

$$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 \; . \tag{4}$$

The mean of a Bernoulli random variable (→ II/1.2.3) is

$$X \sim \mathrm{Bern}(p) \quad \Rightarrow \quad \mathrm{E}(X) = p \tag{5}$$

and the mean of a squared Bernoulli random variable is

$$\mathrm{E}\left(X^2\right) = 0^2 \cdot \mathrm{Pr}(X = 0) + 1^2 \cdot \mathrm{Pr}(X = 1) = 0 \cdot (1-p) + 1 \cdot p = p \; . \tag{6}$$

Combining (4), (5) and (6), we have:

$$\mathrm{Var}(X) = p - p^2 = p \, (1-p) \; . \tag{7}$$

■

Sources:
- Wikipedia (2022): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-01-20; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution#Variance.

1.2.5 Range of variance

Theorem: Let X be a random variable (→ I/1.2.2) following a Bernoulli distribution (→ II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is necessarily between 0 and 1/4:

$$0 \leq \mathrm{Var}(X) \leq \frac{1}{4} \; . \tag{2}$$

Proof: The variance of a Bernoulli random variable (→ II/1.2.4) is

$$X \sim \mathrm{Bern}(p) \quad \Rightarrow \quad \mathrm{Var}(X) = p \, (1-p) \tag{3}$$

which can also be understood as a function of the success probability (→ II/1.2.1) p:

$$\mathrm{Var}(X) = \mathrm{Var}(p) = -p^2 + p \; . \tag{4}$$

The first derivative of this function is

$$\frac{\mathrm{dVar}(p)}{\mathrm{d}p} = -2\,p + 1 \tag{5}$$

and setting this deriative to zero

$$\begin{aligned} \frac{\mathrm{dVar}(p_M)}{\mathrm{d}p} &= 0 \\ 0 &= -2\,p_M + 1 \\ p_M &= \frac{1}{2} \; , \end{aligned} \tag{6}$$

we obtain the maximum possible variance

$$\max\left[\mathrm{Var}(X)\right] = \mathrm{Var}(p_M) = -\left(\frac{1}{2}\right)^2 + \frac{1}{2} = \frac{1}{4} \; . \tag{7}$$

The function $\mathrm{Var}(p)$ is monotonically increasing for $0 < p < p_M$ as $\mathrm{dVar}(p)/\mathrm{d}p > 0$ in this interval and it is monotonically decreasing for $p_M < p < 1$ as $\mathrm{dVar}(p)/\mathrm{d}p < 0$ in this interval. Moreover, as variance is always non-negative (→ I/1.11.4), the minimum variance is

$$\min\left[\mathrm{Var}(X)\right] = \mathrm{Var}(0) = \mathrm{Var}(1) = 0 \; . \tag{8}$$

Thus, we have:

$$\mathrm{Var}(p) \in \left[0, \; \frac{1}{4}\right] \; . \tag{9}$$

■

Sources:
- Wikipedia (2022): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-01-27; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution#Variance.

1.2.6 Shannon entropy

Theorem: Let X be a random variable (→ I/1.2.2) following a Bernoulli distribution (→ II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the (Shannon) entropy (→ I/2.1.1) of X in bits is

$$\mathrm{H}(X) = -p \log_2 p - (1-p) \log_2 (1-p) \; . \tag{2}$$

Proof: The entropy (→ I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \; . \tag{3}$$

Entropy is measured in bits by setting $b = 2$. Since there are only two possible outcomes for a Bernoulli random variable (→ II/1.2.2), we have:

$$
\begin{split}
\mathrm{H}(X) &= -\mathrm{Pr}(X=0) \cdot \log_2 \mathrm{Pr}(X=0) - \mathrm{Pr}(X=1) \cdot \log_2 \mathrm{Pr}(X=1) \\
&= -p \log_2 p - (1-p) \log_2 (1-p) \; .
\end{split} \tag{4}
$$

■

Sources:
- Wikipedia (2022): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-02; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution.
- Wikipedia (2022): "Binary entropy function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-02; URL: https://en.wikipedia.org/wiki/Binary_entropy_function.

1.2.7 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ I/1.2.2). Assume two Bernoulli distributions (→ II/1.2.1) P and Q specifying the probability distribution of X as

$$
\begin{split}
P: \; X &\sim \mathrm{Bern}(p_1) \\
Q: \; X &\sim \mathrm{Bern}(p_2) \; .
\end{split} \tag{1}
$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$
\mathrm{KL}[P \,||\, Q] = \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1 \, (1-p_2)}{p_2 \, (1-p_1)} \; . \tag{2}
$$

Proof: The KL divergence for a discrete random variable (→ I/2.5.1) is given by

$$
\mathrm{KL}[P \,||\, Q] = \sum_{x \in \mathcal{X}} p(x) \, \ln \frac{p(x)}{q(x)} \tag{3}
$$

which, applied to the Bernoulli distributions (→ II/1.2.1) in (1), yields

$$
\begin{split}
\mathrm{KL}[P \,||\, Q] &= \sum_{x \in \{0,1\}} p(x) \, \ln \frac{p(x)}{q(x)} \\
&= p(X=0) \cdot \ln \frac{p(X=0)}{q(X=0)} + p(X=1) \cdot \ln \frac{p(X=1)}{q(X=1)} \; .
\end{split} \tag{4}
$$

Using the probability mass function of the Bernoulli distribution (→ II/1.2.2), this becomes:

$$
\begin{split}
\mathrm{KL}[P \,||\, Q] &= (1-p_1) \cdot \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1}{p_2} \\
&= \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1}{p_2} - p_1 \cdot \ln \frac{1-p_1}{1-p_2} \\
&= \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \left(\ln \frac{p_1}{p_2} + \ln \frac{1-p_2}{1-p_1} \right) \\
&= \ln \frac{1-p_1}{1-p_2} + p_1 \cdot \ln \frac{p_1 \, (1-p_2)}{p_2 \, (1-p_1)}
\end{split} \tag{5}
$$

■

1.3 Binomial distribution

1.3.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to follow a binomial distribution with number of trials n and success probability p

$$X \sim \mathrm{Bin}(n, p) \; , \tag{1}$$

if X is the number of successes observed in n independent (→ I/1.3.6) trials, where each trial has two possible outcomes (→ II/1.2.1) (success/failure) and the probability of success and failure are identical across trials ($p/q = 1 - p$).

Sources:

- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Binomial_distribution.

1.3.2 Probability mass function

Theorem: Let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1):

$$X \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \binom{n}{x} \, p^x \, (1-p)^{n-x} \; . \tag{2}$$

Proof: A binomial variable (→ II/1.3.1) is defined as the number of successes observed in n independent (→ I/1.3.6) trials, where each trial has two possible outcomes (→ II/1.2.1) (success/failure) and the probability (→ I/1.3.1) of success and failure are identical across trials (p, $q = 1 - p$).
If one has obtained x successes in n trials, one has also obtained $(n - x)$ failures. The probability of a particular series of x successes and $(n - x)$ failures, when order does matter, is

$$p^x \, (1-p)^{n-x} \; . \tag{3}$$

When order does not matter, there is a number of series consisting of x successes and $(n-x)$ failures. This number is equal to the number of possibilities in which x objects can be choosen from n objects which is given by the binomial coefficient:

$$\binom{n}{x} \; . \tag{4}$$

In order to obtain the probability of x successes and $(n - x)$ failures, when order does not matter, the probability in (3) has to be multiplied with the number of possibilities in (4) which gives

$$p(X = x|n, p) = \binom{n}{x} \, p^x \, (1-p)^{n-x} \tag{5}$$

which is equivalent to the expression above.

■

1.3.3 Probability-generating function

Theorem: Let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the probability-generating function (→ I/1.9.9) of X is

$$G_X(z) = (q+pz)^n \tag{2}$$

where $q = 1-p$.

Proof: The probability-generating function (→ I/1.9.9) of X is defined as

$$G_X(z) = \sum_{x=0}^{\infty} f_X(x) \, z^x \tag{3}$$

With the probability mass function of the binomial distribution (→ II/1.3.2)

$$f_X(x) = \binom{n}{x} \, p^x \, (1-p)^{n-x} \; , \tag{4}$$

we obtain:

$$\begin{split} G_X(z) &= \sum_{x=0}^{n} \binom{n}{x} \, p^x \, (1-p)^{n-x} \, z^x \\ &= \sum_{x=0}^{n} \binom{n}{x} \, (pz)^x \, (1-p)^{n-x} \; . \end{split} \tag{5}$$

According to the binomial theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} \, x^{n-k} \, y^k \; , \tag{6}$$

the sum in equation (5) equals

$$G_X(z) = ((1-p)+(pz))^n \tag{7}$$

which is equivalent to the result in (2).

■

Sources:

- ProofWiki (2022): "Probability Generating Function of Binomial Distribution"; in: *ProofWiki*, retrieved on 2022-10-11; URL: https://proofwiki.org/wiki/Probability_Generating_Function_of_Binomial_Distribution.

1.3.4 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = np \; . \tag{2}$$

Proof: By definition, a binomial random variable (→ II/1.3.1) is the sum of n independent and identical Bernoulli trials (→ II/1.2.1) with success probability p. Therefore, the expected value is

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \ldots + X_n) \tag{3}$$

and because the expected value is a linear operator (→ I/1.10.5), this is equal to

$$\mathrm{E}(X) = \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) = \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \tag{4}$$

With the expected value of the Bernoulli distribution (→ II/1.2.3), we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} p = np \; . \tag{5}$$

■

Sources:
- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-16; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Expected_value_and_variance.

1.3.5 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = np\,(1-p) \; . \tag{2}$$

Proof: By definition, a binomial random variable (→ II/1.3.1) is the sum of n independent and identical Bernoulli trials (→ II/1.2.1) with success probability p. Therefore, the variance is

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + \ldots + X_n) \tag{3}$$

and because variances add up under independence (→ I/1.11.10), this is equal to

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \; . \tag{4}$$

With the variance of the Bernoulli distribution (→ II/1.2.4), we have:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p\,(1-p) = np\,(1-p) \; . \tag{5}$$

■

Sources:
- Wikipedia (2022): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-01-20; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Expected_value_and_variance.

1.3.6 Range of variance

Theorem: Let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is necessarily between 0 and $n/4$:

$$0 \leq \mathrm{Var}(X) \leq \frac{n}{4} \; . \tag{2}$$

Proof: By definition, a binomial random variable (→ II/1.3.1) is the sum of n independent and identical Bernoulli trials (→ II/1.2.1) with success probability p. Therefore, the variance is

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + \ldots + X_n) \tag{3}$$

and because variances add up under independence (→ I/1.11.10), this is equal to

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \; . \tag{4}$$

As the variance of a Bernoulli random variable is always between 0 and 1/4 (→ II/1.2.5)

$$0 \leq \mathrm{Var}(X_i) \leq \frac{1}{4} \quad \text{for all} \quad i = 1, \ldots, n \; , \tag{5}$$

the minimum variance of X is

$$\min\left[\mathrm{Var}(X)\right] = n \cdot 0 = 0 \tag{6}$$

and the maximum variance of X is

$$\max\left[\mathrm{Var}(X)\right] = n \cdot \frac{1}{4} = \frac{n}{4} \; . \tag{7}$$

Thus, we have:

$$\mathrm{Var}(X) \in \left[0, \; \frac{n}{4}\right] \; . \tag{8}$$

■

1.3.7 Shannon entropy

Theorem: Let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the (Shannon) entropy (→ I/2.1.1) of X in bits is

$$\mathrm{H}(X) = n \cdot \mathrm{H}_\mathrm{bern}(p) - \mathrm{E}_\mathrm{lbc}(n,p) \tag{2}$$

where $\mathrm{H}_\mathrm{bern}(p)$ is the binary entropy function, i.e. the (Shannon) entropy of the Bernoulli distribution (→ II/1.2.6) with success probability p

$$\mathrm{H}_\mathrm{bern}(p) = -p \cdot \log_2 p - (1-p) \log_2 (1-p) \tag{3}$$

and $\mathrm{E}_\mathrm{lbc}(n,p)$ is the expected value (→ I/1.10.1) of the logarithmized binomial coefficient (→ II/1.3.2) with superset size n

$$\mathrm{E}_\mathrm{lbc}(n,p) = \mathrm{E}\left[\log_2 \binom{n}{X} \right] \quad \text{where} \quad X \sim \mathrm{Bin}(n,p) \; . \tag{4}$$

Proof: The entropy (→ I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$\mathrm{H}(X) = - \sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \; . \tag{5}$$

Entropy is measured in bits by setting $b=2$. Then, with the probability mass function of the binomial distribution (→ II/1.3.2), we have:

$$\begin{split}
\mathrm{H}(X) &= - \sum_{x \in \mathcal{X}} f_X(x) \cdot \log_2 f_X(x) \\
&= - \sum_{x=0}^{n} \binom{n}{x} \, p^x \, (1-p)^{n-x} \cdot \log_2 \left[\binom{n}{x} \, p^x \, (1-p)^{n-x} \right] \\
&= - \sum_{x=0}^{n} \binom{n}{x} \, p^x \, (1-p)^{n-x} \cdot \left[\log_2 \binom{n}{x} + x \cdot \log_2 p + (n-x) \cdot \log_2 (1-p) \right] \\
&= - \sum_{x=0}^{n} \binom{n}{x} \, p^x \, (1-p)^{n-x} \cdot \left[\log_2 \binom{n}{x} + x \cdot \log_2 p + n \cdot \log_2 (1-p) - x \cdot \log_2 (1-p) \right] \; .
\end{split} \tag{6}$$

Since the first factor in the sum corresponds to the probability mass (→ I/1.6.1) of $X = x$, we can rewrite this as the sum of the expected values (→ I/1.10.1) of the functions (→ I/1.10.12) of the discrete random variable (→ I/1.2.6) x in the square bracket:

$$\begin{split}
\mathrm{H}(X) &= - \left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} - \left\langle x \cdot \log_2 p \right\rangle_{p(x)} - \left\langle n \cdot \log_2 (1-p) \right\rangle_{p(x)} + \left\langle x \cdot \log_2 (1-p) \right\rangle_{p(x)} \\
&= - \left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} - \log_2 p \cdot \left\langle x \right\rangle_{p(x)} - n \cdot \log_2 (1-p) + \log_2 (1-p) \cdot \left\langle x \right\rangle_{p(x)} \; .
\end{split} \tag{7}$$

Using the expected value of the binomial distribution (→ II/1.3.4), i.e. $X \sim \mathrm{Bin}(n,p) \Rightarrow \langle x \rangle = np$, this gives:

$$\begin{aligned} \mathrm{H}(X) &= -\left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} - np \cdot \log_2 p - n \cdot \log_2(1-p) + np \cdot \log_2(1-p) \\ &= -\left\langle \log_2 \binom{n}{x} \right\rangle_{p(x)} + n\left[-p \cdot \log_2 p - (1-p)\log_2(1-p)\right] \; . \end{aligned} \tag{8}$$

Finally, we note that the first term is the negative expected value (→ I/1.10.1) of the logarithm of a binomial coefficient (→ II/1.3.2) and that the term in square brackets is the entropy of the Bernoulli distribution (→ II/1.3.7), such that we finally get:

$$\mathrm{H}(X) = n \cdot \mathrm{H}_{\mathrm{bern}}(p) - \mathrm{E}_{\mathrm{lbc}}(n,p) \; . \tag{9}$$

Note that, because $0 \leq$ H_bern$(p) \leq 1$, we have $0 \leq n \cdot$ H_bern$(p) \leq n$, and because the entropy is non-negative (→ I/2.1.2), it must hold that $n \geq$ E_lbc$(n,p) \geq 0$.

■

1.3.8 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ I/1.2.2). Assume two binomial distributions (→ II/1.3.1) P and Q specifying the probability distribution of X as

$$\begin{aligned} P &: \; X \sim \mathrm{Bin}(n, p_1) \\ Q &: \; X \sim \mathrm{Bin}(n, p_2) \; . \end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P \,||\, Q] = np_1 \cdot \ln \frac{p_1}{p_2} + n(1-p_1) \cdot \ln \frac{1-p_1}{1-p_2} \; . \tag{2}$$

Proof: The KL divergence for a discrete random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P \,||\, Q] = \sum_{x \in \mathcal{X}} p(x) \, \ln \frac{p(x)}{q(x)} \tag{3}$$

which, applied to the binomial distributions (→ II/1.3.1) in (1), yields

$$\begin{aligned} \mathrm{KL}[P \,||\, Q] &= \sum_{x=0}^{n} p(x) \, \ln \frac{p(x)}{q(x)} \\ &= p(X=0) \cdot \ln \frac{p(X=0)}{q(X=0)} + \ldots + p(X=n) \cdot \ln \frac{p(X=n)}{q(X=n)} \; . \end{aligned} \tag{4}$$

Using the probability mass function of the binomial distribution (→ II/1.3.2), this becomes:

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \sum_{x=0}^{n} \binom{n}{x} p_1^x \, (1-p_1)^{n-x} \cdot \ln \frac{\binom{n}{x} p_1^x \, (1-p_1)^{n-x}}{\binom{n}{x} p_2^x \, (1-p_2)^{n-x}} \\
&= \sum_{x=0}^{n} \binom{n}{x} p_1^x \, (1-p_1)^{n-x} \cdot \left[x \cdot \ln \frac{p_1}{p_2} + (n-x) \cdot \ln \frac{1-p_1}{1-p_2} \right] \\
&= \ln \frac{p_1}{p_2} \cdot \sum_{x=0}^{n} \binom{n}{x} p_1^x \, (1-p_1)^{n-x} x + \ln \frac{1-p_1}{1-p_2} \cdot \sum_{x=0}^{n} \binom{n}{x} p_1^x \, (1-p_1)^{n-x} (n-x) \; .
\end{aligned} \tag{5}
$$

We can now see that some terms in this sum are expected values (→ I/1.10.1) with respect to binomial distributions (→ II/1.3.1):

$$
\begin{aligned}
\sum_{x=0}^{n} \binom{n}{x} p_1^x \, (1-p_1)^{n-x} x &= \mathrm{E}\left[x\right]_{\mathrm{Bin}(n,p_1)} \\
\sum_{x=0}^{n} \binom{n}{x} p_1^x \, (1-p_1)^{n-x} (n-x) &= \mathrm{E}\left[n-x\right]_{\mathrm{Bin}(n,p_1)} \; .
\end{aligned} \tag{6}
$$

Using the expected value of the binomial distribution (→ II/1.3.4), these can be simplified to

$$
\begin{aligned}
\mathrm{E}\left[x\right]_{\mathrm{Bin}(n,p_1)} &= np_1 \\
\mathrm{E}\left[n-x\right]_{\mathrm{Bin}(n,p_1)} &= n - np_1 \; ,
\end{aligned} \tag{7}
$$

such that the Kullback-Leibler divergence finally becomes:

$$
\mathrm{KL}[P\,||\,Q] = np_1 \cdot \ln \frac{p_1}{p_2} + n(1-p_1) \cdot \ln \frac{1-p_1}{1-p_2} \; . \tag{8}
$$

■

Sources:

- PSPACEhard (2017): "Kullback-Leibler divergence for binomial distributions P and Q"; in: *Stack-Exchange Mathematics*, retrieved on 2023-10-20; URL: https://math.stackexchange.com/a/2215384/480910.

1.3.9 Conditional binomial

Theorem: Let X and Y be two random variables (→ I/1.2.2) where Y is binomially distributed (→ II/1.3.1) conditional on (→ I/1.5.4) X

$$
Y|X \sim \mathrm{Bin}(X,q) \tag{1}
$$

and X also follows a binomial distribution (→ II/1.3.1), but with different success frequency (→ II/1.3.1):

$$
X \sim \mathrm{Bin}(n,p) \; . \tag{2}
$$

Then, the maginal distribution (→ I/1.5.3) of Y unconditional on X is again a binomial distribution (→ II/1.3.1):

$$Y \sim \mathrm{Bin}(n, pq) \; . \tag{3}$$

Proof: We are interested in the probability that Y equals a number m. According to the law of marginal probability (→ I/1.3.3) or the law of total probability (→ I/1.4.7), this probability can be expressed as:

$$\mathrm{Pr}(Y = m) = \sum_{k=0}^{\infty} \mathrm{Pr}(Y = m|X = k) \cdot \mathrm{Pr}(X = k) \; . \tag{4}$$

Since, by definitions (2) and (1), $\mathrm{Pr}(X = k) = 0$ when $k > n$ and $\mathrm{Pr}(Y = m|X = k) = 0$ when $k < m$, we have:

$$\mathrm{Pr}(Y = m) = \sum_{k=m}^{n} \mathrm{Pr}(Y = m|X = k) \cdot \mathrm{Pr}(X = k) \; . \tag{5}$$

Now we can take the probability mass function of the binomial distribution (→ II/1.3.2) and plug it in for the terms in the sum of (5) to get:

$$\mathrm{Pr}(Y = m) = \sum_{k=m}^{n} \binom{k}{m} q^m \, (1-q)^{k-m} \cdot \binom{n}{k} p^k \, (1-p)^{n-k} \; . \tag{6}$$

Applying the binomial coefficient identity $\binom{n}{k}\binom{k}{m} = \binom{n}{m}\binom{n-m}{k-m}$ and rearranging the terms, we have:

$$\mathrm{Pr}(Y = m) = \sum_{k=m}^{n} \binom{n}{m} \binom{n-m}{k-m} p^k \, q^m \, (1-p)^{n-k} \, (1-q)^{k-m} \; . \tag{7}$$

Now we partition $p^k = p^m \cdot p^{k-m}$ and pull all terms dependent on k out of the sum:

$$\begin{split} \mathrm{Pr}(Y = m) &= \binom{n}{m} p^m \, q^m \sum_{k=m}^{n} \binom{n-m}{k-m} p^{k-m} \, (1-p)^{n-k} \, (1-q)^{k-m} \\ &= \binom{n}{m} (pq)^m \sum_{k=m}^{n} \binom{n-m}{k-m} \, (p(1-q))^{k-m} \, (1-p)^{n-k} \; . \end{split} \tag{8}$$

Then we subsititute $i = k - m$, such that $k = i + m$:

$$\mathrm{Pr}(Y = m) = \binom{n}{m} (pq)^m \sum_{i=0}^{n-m} \binom{n-m}{i} \, (p - pq)^i \, (1-p)^{n-m-i} \; . \tag{9}$$

According to the binomial theorem

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} \, y^k \; , \tag{10}$$

the sum in equation (9) is equal to

$$\sum_{i=0}^{n-m} \binom{n-m}{i} \, (p - pq)^i \, (1-p)^{n-m-i} = ((p - pq) + (1-p))^{n-m} \; . \tag{11}$$

Thus, (9) develops into

$$\begin{aligned}\Pr(Y=m) &= \binom{n}{m}(pq)^m(p-pq+1-p)^{n-m} \\ &= \binom{n}{m}(pq)^m(1-pq)^{n-m}\end{aligned} \tag{12}$$

which is the probability mass function of the binomial distribution (→ II/1.3.2) with parameters n and pq, such that

$$Y \sim \mathrm{Bin}(n, pq) \; . \tag{13}$$

■

Sources:
- Wikipedia (2022): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-07; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Conditional_binomials.

1.4 Beta-binomial distribution

1.4.1 Definition

Definition: Let p be a random variable (→ I/1.2.2) following a beta distribution (→ II/3.9.1)

$$p \sim \mathrm{Bet}(\alpha, \beta) \tag{1}$$

and let X be a random variable (→ I/1.2.2) following a binomial distribution (→ II/1.3.1) conditional on p

$$X \mid p \sim \mathrm{Bin}(n, p) \; . \tag{2}$$

Then, the marginal distribution (→ I/1.5.3) of X is called a beta-binomial distribution

$$X \sim \mathrm{BetBin}(n, \alpha, \beta) \tag{3}$$

with number of trials (→ II/1.3.1) n and shape parameters (→ II/3.9.1) α and β.

Sources:
- Wikipedia (2022): "Beta-binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-20; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#Motivation_and_derivation.

1.4.2 Probability mass function

Theorem: Let X be a random variable (→ I/1.2.2) following a beta-binomial distribution (→ II/1.4.1):

$$X \sim \mathrm{BetBin}(n, \alpha, \beta) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \binom{n}{x} \cdot \frac{\mathrm{B}(\alpha+x, \beta+n-x)}{\mathrm{B}(\alpha, \beta)} \tag{2}$$

where $\mathrm{B}(x, y)$ is the beta function.

Proof: A beta-binomial random variable (→ II/1.4.1) is defined as a binomial variate (→ II/1.3.1) for which the success probability is following a beta distribution (→ II/3.9.1):

$$\begin{aligned} X \mid p &\sim \mathrm{Bin}(n, p) \\ p &\sim \mathrm{Bet}(\alpha, \beta) \; . \end{aligned} \tag{3}$$

Thus, we can combine the law of marginal probability (→ I/1.3.3) and the law of conditional probability (→ I/1.3.4) to derive the probability (→ I/1.3.1) of X as

$$\begin{aligned} p(x) &= \int_{\mathcal{P}} \mathrm{p}(x, p) \, \mathrm{d}p \\ &= \int_{\mathcal{P}} \mathrm{p}(x|p) \, \mathrm{p}(p) \, \mathrm{d}p \; . \end{aligned} \tag{4}$$

Now, we can plug in the probability mass function of the binomial distribution (→ II/1.3.2) and the probability density function of the beta distribution (→ II/3.9.3) to get

$$\begin{aligned} p(x) &= \int_0^1 \binom{n}{x} p^x \, (1-p)^{n-x} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \, p^{\alpha-1} \, (1-p)^{\beta-1} \, \mathrm{d}p \\ &= \binom{n}{x} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 p^{\alpha+x-1} \, (1-p)^{\beta+n-x-1} \, \mathrm{d}p \\ &= \binom{n}{x} \cdot \frac{\mathrm{B}(\alpha+x, \beta+n-x)}{\mathrm{B}(\alpha, \beta)} \int_0^1 \frac{1}{\mathrm{B}(\alpha+x, \beta+n-x)} \, p^{\alpha+x-1} \, (1-p)^{\beta+n-x-1} \, \mathrm{d}p \; . \end{aligned} \tag{5}$$

Finally, we recognize that the integrand is equal to the probability density function of a beta distribution (→ II/3.9.3) and because probability density integrates to one (→ I/1.7.1), we have

$$p(x) = \binom{n}{x} \cdot \frac{\mathrm{B}(\alpha+x, \beta+n-x)}{\mathrm{B}(\alpha, \beta)} = f_X(x) \; . \tag{6}$$

This completes the proof.

■

Sources:

- Wikipedia (2022): "Beta-binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-20; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#As_a_compound_distribution.

1.4.3 Probability mass function in terms of gamma function

Theorem: Let X be a random variable (→ I/1.2.2) following a beta-binomial distribution (→ II/1.4.1):

$$X \sim \mathrm{BetBin}(n, \alpha, \beta) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X can be expressed as

$$f_X(x) = \frac{\Gamma(n+1)}{\Gamma(x+1)\,\Gamma(n-x+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\,\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)} \tag{2}$$

where $\Gamma(x)$ is the gamma function.

Proof: The probability mass function of the beta-binomial distribution (→ II/1.4.2) is given by

$$f_X(x) = \binom{n}{x} \cdot \frac{\mathrm{B}(\alpha+x, \beta+n-x)}{\mathrm{B}(\alpha, \beta)} \; . \tag{3}$$

Note that the binomial coefficient can be expressed in terms of factorials

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \; , \tag{4}$$

that factorials are related to the gamma function via $n! = \Gamma(n+1)$

$$\frac{n!}{x!\,(n-x)!} = \frac{\Gamma(n+1)}{\Gamma(x+1)\,\Gamma(n-x+1)} \tag{5}$$

and that the beta function is related to the gamma function via

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} \; . \tag{6}$$

Applying (4), (5) and (6) to (3), we get

$$f_X(x) = \frac{\Gamma(n+1)}{\Gamma(x+1)\,\Gamma(n-x+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+x)\,\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)} \; . \tag{7}$$

This completes the proof.

■

Sources:
- Wikipedia (2022): "Beta-binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-20; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#As_a_compound_distribution.

1.4.4 Cumulative distribution function

Theorem: Let X be a random variable (→ I/1.2.2) following a beta-binomial distribution (→ II/1.4.1):

$$X \sim \mathrm{BetBin}(n, \alpha, \beta) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \cdot \frac{\Gamma(n+1)}{\Gamma(\alpha+\beta+n)} \cdot \sum_{i=0}^{x} \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{\Gamma(i+1) \cdot \Gamma(n-i+1)} \tag{2}$$

where $\mathrm{B}(x, y)$ is the beta function and $\Gamma(x)$ is the gamma function.

Proof: The cumulative distribution function (→ I/1.8.1) is defined as

$$F_X(x) = \Pr(X \leq x) \tag{3}$$

which, for a discrete random variable (→ I/1.2.6), evaluates to

$$F_X(x) = \sum_{i=-\infty}^{x} f_X(i) \; . \tag{4}$$

With the probability mass function of the beta-binomial distribution (→ II/1.4.2), this becomes

$$F_X(x) = \sum_{i=0}^{x} \binom{n}{i} \cdot \frac{\mathrm{B}(\alpha+i, \beta+n-i)}{\mathrm{B}(\alpha, \beta)} \; . \tag{5}$$

Using the expression of binomial coefficients in terms of factorials

$$\binom{n}{k} = \frac{n!}{k! \, (n-k)!} \; , \tag{6}$$

the relationship between factorials and the gamma function

$$n! = \Gamma(n+1) \tag{7}$$

and the link between gamma function and beta function

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha) \, \Gamma(\beta)}{\Gamma(\alpha+\beta)} \; , \tag{8}$$

equation (5) can be further developped as follows:

$$\begin{split}
F_X(x) &\overset{(6)}{=} \frac{1}{\mathrm{B}(\alpha, \beta)} \cdot \sum_{i=0}^{x} \frac{n!}{i! \, (n-i)!} \cdot \mathrm{B}(\alpha+i, \beta+n-i) \\
&\overset{(8)}{=} \frac{1}{\mathrm{B}(\alpha, \beta)} \cdot \sum_{i=0}^{x} \frac{n!}{i! \, (n-i)!} \cdot \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{\Gamma(\alpha+\beta+n)} \\
&= \frac{1}{\mathrm{B}(\alpha, \beta)} \cdot \frac{n!}{\Gamma(\alpha+\beta+n)} \cdot \sum_{i=0}^{x} \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{i! \, (n-i)!} \\
&\overset{(7)}{=} \frac{1}{\mathrm{B}(\alpha, \beta)} \cdot \frac{\Gamma(n+1)}{\Gamma(\alpha+\beta+n)} \cdot \sum_{i=0}^{x} \frac{\Gamma(\alpha+i) \cdot \Gamma(\beta+n-i)}{\Gamma(i+1) \cdot \Gamma(n-i+1)} \; .
\end{split} \tag{9}$$

This completes the proof.

■

1.5 Poisson distribution

1.5.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to follow a Poisson distribution with rate λ

$$X \sim \mathrm{Poiss}(\lambda) \; , \tag{1}$$

if and only if its probability mass function (→ I/1.6.1) is given by

$$\mathrm{Poiss}(x; \lambda) = \frac{\lambda^x \, e^{-\lambda}}{x!} \tag{2}$$

where $x \in \mathbb{N}_0$ and $\lambda > 0$.

Sources:

- Wikipedia (2020): "Poisson distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-25; URL: https://en.wikipedia.org/wiki/Poisson_distribution#Definitions.

1.5.2 Probability mass function

Theorem: Let X be a random variable (→ I/1.2.2) following a Poisson distribution (→ II/1.5.1):

$$X \sim \mathrm{Poiss}(\lambda) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \frac{\lambda^x \, e^{-\lambda}}{x!}, \; x \in \mathbb{N}_0 \; . \tag{2}$$

Proof: This follows directly from the definition of the Poisson distribution (→ II/1.5.1).

■

1.5.3 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a Poisson distribution (→ II/1.5.1):

$$X \sim \mathrm{Poiss}(\lambda) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \lambda \; . \tag{2}$$

Proof: The expected value of a discrete random variable (→ I/1.10.1) is defined as

$$\mathrm{E}(X) = \sum_{x \in \mathcal{X}} x \cdot f_X(x) \; , \tag{3}$$

such that, with the probability mass function of the Poisson distribution (→ II/1.5.2), we have:

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x \, e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\lambda^x \, e^{-\lambda}}{x!} \\
&= e^{-\lambda} \cdot \sum_{x=1}^{\infty} \frac{x}{x!} \lambda^x \\
&= \lambda e^{-\lambda} \cdot \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \; .
\end{aligned} \tag{4}
$$

Substituting $z = x - 1$, such that $x = z + 1$, we get:

$$
\mathrm{E}(X) = \lambda e^{-\lambda} \cdot \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \; . \tag{5}
$$

Using the power series expansion of the exponential function

$$
e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \; , \tag{6}
$$

the expected value of X finally becomes

$$
\begin{aligned}
\mathrm{E}(X) &= \lambda e^{-\lambda} \cdot e^{\lambda} \\
&= \lambda \; .
\end{aligned} \tag{7}
$$

■

Sources:

- ProofWiki (2020): "Expectation of Poisson Distribution"; in: *ProofWiki*, retrieved on 2020-08-19; URL: https://proofwiki.org/wiki/Expectation_of_Poisson_Distribution.

1.5.4 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a Poisson distribution (→ II/1.5.1):

$$
X \sim \mathrm{Poiss}(\lambda) \; . \tag{1}
$$

Then, the variance (→ I/1.11.1) of X is

$$
\mathrm{Var}(X) = \lambda \; . \tag{2}
$$

Proof: The variance (→ I/1.11.1) can be expressed in terms of expected values (→ I/1.11.3) as

$$
\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; . \tag{3}
$$

The expected value of a Poisson random variable (→ II/1.5.3) is

$$
\mathrm{E}(X) = \lambda \; . \tag{4}
$$

Let us now consider the expectation (→ I/1.10.1) of $X\,(X-1)$ which is defined as

$$\mathrm{E}[X\,(X-1)] = \sum_{x \in \mathcal{X}} x\,(x-1) \cdot f_X(x) \; , \tag{5}$$

such that, with the probability mass function of the Poisson distribution (→ II/1.5.2), we have:

$$\begin{split}
\mathrm{E}[X\,(X-1)] &= \sum_{x=0}^{\infty} x\,(x-1) \cdot \frac{\lambda^x \, e^{-\lambda}}{x!} \\
&= \sum_{x=2}^{\infty} x\,(x-1) \cdot \frac{\lambda^x \, e^{-\lambda}}{x!} \\
&= e^{-\lambda} \cdot \sum_{x=2}^{\infty} x\,(x-1) \cdot \frac{\lambda^x}{x \cdot (x-1) \cdot (x-2)!} \\
&= \lambda^2 \cdot e^{-\lambda} \cdot \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \; .
\end{split} \tag{6}$$

Substituting $z = x-2$, such that $x = z+2$, we get:

$$\mathrm{E}[X\,(X-1)] = \lambda^2 \cdot e^{-\lambda} \cdot \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \; . \tag{7}$$

Using the power series expansion of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \; , \tag{8}$$

the expected value of $X\,(X-1)$ finally becomes

$$\mathrm{E}[X\,(X-1)] = \lambda^2 \cdot e^{-\lambda} \cdot e^{\lambda} = \lambda^2 \; . \tag{9}$$

Note that this expectation can be written as

$$\mathrm{E}[X\,(X-1)] = \mathrm{E}(X^2 - X) = \mathrm{E}(X^2) - \mathrm{E}(X) \; , \tag{10}$$

such that, with (9) and (4), we have:

$$\mathrm{E}(X^2) - \mathrm{E}(X) = \lambda^2 \quad \Rightarrow \quad \mathrm{E}(X^2) = \lambda^2 + \lambda \; . \tag{11}$$

Plugging (11) and (4) into (3), the variance of a Poisson random variable finally becomes

$$\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda \; . \tag{12}$$

■

Sources:

- jbstatistics (2013): "The Poisson Distribution: Mathematically Deriving the Mean and Variance"; in: *YouTube*, retrieved on 2021-04-29; URL: https://www.youtube.com/watch?v=65n_v92JZeE.

2 Multivariate discrete distributions

2.1 Categorical distribution

2.1.1 Definition

Definition: Let X be a random vector (→ I/1.2.3). Then, X is said to follow a categorical distribution with success probability $p_1, \ldots, p_k$

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; , \tag{1}$$

if $X = e_i$ with probability (→ I/1.3.1) p_i for all $i = 1, \ldots, k$, where e_i is the i-th elementary row vector, i.e. a $1 \times k$ vector of zeros with a one in i-th position.

Sources:
- Wikipedia (2020): "Categorical distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Categorical_distribution.

2.1.2 Probability mass function

Theorem: Let X be a random vector (→ I/1.2.3) following a categorical distribution (→ II/2.1.1):

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \left\{ \begin{array}{rl} p_1 \; , & \text{if} \; x = e_1 \\ \vdots & \quad \vdots \\ p_k \; , & \text{if} \; x = e_k \; . \end{array} \right. \tag{2}$$

where $e_1, \ldots, e_k$ are the $1 \times k$ elementary row vectors.

Proof: This follows directly from the definition of the categorical distribution (→ II/2.1.1).

■

2.1.3 Mean

Theorem: Let X be a random vector (→ I/1.2.3) following a categorical distribution (→ II/2.1.1):

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = [p_1, \ldots, p_k] \; . \tag{2}$$

Proof: If we conceive the outcome of a categorical distribution (→ II/2.1.1) to be a $1 \times k$ vector, then the elementary row vectors $e_1 = [1, 0, \ldots, 0]$, ..., $e_k = [0, \ldots, 0, 1]$ are all the possible outcomes and they occur with probabilities $\Pr(X = e_1) = p_1$, ..., $\Pr(X = e_k) = p_k$. Consequently, the expected value (→ I/1.10.1) is

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{x \in \mathcal{X}} x \cdot \Pr(X = x) \\
&= \sum_{i=1}^{k} e_i \cdot \Pr(X = e_i) \\
&= \sum_{i=1}^{k} e_i \cdot p_i \\
&= [p_1, \ldots, p_k] \; .
\end{aligned} \tag{3}
$$

■

2.1.4 Covariance

Theorem: Let X be a random vector (→ I/1.2.3) following a categorical distribution (→ II/2.1.1):

$$X \sim \mathrm{Cat}(n, p) \; . \tag{1}$$

Then, the covariance matrix (→ I/1.13.9) of X is

$$\mathrm{Cov}(X) = \mathrm{diag}(p) - pp^{\mathrm{T}} \; . \tag{2}$$

Proof: The categorical distribution (→ II/2.1.1) is a special case of the multinomial distribution (→ II/2.2.1) in which $n = 1$:

$$X \sim \mathrm{Mult}(n, p) \quad \text{and} \quad n = 1 \quad \Rightarrow \quad X \sim \mathrm{Cat}(p) \; . \tag{3}$$

The covariance matrix of the multinomial distribution (→ II/2.2.4) is

$$\mathrm{Cov}(X) = n \left(\mathrm{diag}(p) - pp^{\mathrm{T}} \right) \; , \tag{4}$$

thus the covariance matrix of the categorical distribution is

$$\mathrm{Cov}(X) = \mathrm{diag}(p) - pp^{\mathrm{T}} \; . \tag{5}$$

■

2.1.5 Shannon entropy

Theorem: Let X be a random vector (→ I/1.2.3) following a categorical distribution (→ II/2.1.1):

$$X \sim \mathrm{Cat}(p) \; . \tag{1}$$

Then, the (Shannon) entropy (→ I/2.1.1) of X is

$$\mathrm{H}(X) = -\sum_{i=1}^{k} p_i \cdot \log p_i \; . \tag{2}$$

Proof: The entropy (→ I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \; . \tag{3}$$

Since there are k possible values for a categorical random vector (→ II/2.1.1) with probabilities given by the entries (→ II/2.1.2) of the $1 \times k$ vector p, we have:

$$\begin{aligned}
\mathrm{H}(X) &= -\mathrm{Pr}(X = e_1) \cdot \log \mathrm{Pr}(X = e_1) - \ldots - \mathrm{Pr}(X = e_k) \cdot \log \mathrm{Pr}(X = e_k) \\
\mathrm{H}(X) &= -\sum_{i=1}^{k} \mathrm{Pr}(X = e_i) \cdot \log \mathrm{Pr}(X = e_i) \\
\mathrm{H}(X) &= -\sum_{i=1}^{k} p_i \cdot \log p_i \; .
\end{aligned} \tag{4}$$

■

2.2 Multinomial distribution

2.2.1 Definition

Definition: Let X be a random vector (→ I/1.2.3). Then, X is said to follow a multinomial distribution with number of trials n and category probabilities $p_1, \ldots, p_k$

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; , \tag{1}$$

if X are the numbers of observations belonging to k distinct categories in n independent (→ I/1.3.6) trials, where each trial has k possible outcomes (→ II/2.1.1) and the category probabilities are identical across trials.

Sources:

- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Multinomial_distribution.

2.2.2 Probability mass function

Theorem: Let X be a random vector (→ I/1.2.3) following a multinomial distribution (→ II/2.2.1):

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the probability mass function (→ I/1.6.1) of X is

$$f_X(x) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} {p_i}^{x_i} \; . \tag{2}$$

Proof: A multinomial variable (→ II/2.2.1) is defined as a vector of the numbers of observations belonging to k distinct categories in n independent (→ I/1.3.6) trials, where each trial has k possible outcomes (→ II/2.1.1) and the category probabilities (→ I/1.3.1) are identical across trials.

Since the individual trials are independent (→ II/2.2.1) and joint probability factorizes under independence (→ I/1.3.8), the probability of a particular series of x_1 observations for category 1, x_2 observations for category 2, ... etc., when order does matter, is

$$\prod_{i=1}^{k} {p_i}^{x_i} \; . \tag{3}$$

When order does not matter, there is a number of series consisting of x_1 observations for category 1, x_2 observations for category 2, ... etc. This number is equal to the number of possibilities in which x_1 category 1 objects, x_2 category 2 objects, ... etc. can be distributed in a sequence of n objects which is given by the multinomial coefficient that can be expressed in terms of factorials:

$$\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1! \cdot \ldots \cdot x_k!} \; . \tag{4}$$

In order to obtain the probability of x_1 observations for category 1, x_2 observations for category 2, ... etc., when order does not matter, the probability in (3) has to be multiplied with the number of possibilities in (4) which gives

$$p(X = x|n, [p_1, \ldots, p_k]) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} {p_i}^{x_i} \tag{5}$$

which is equivalent to the expression above.

■

2.2.3 Mean

Theorem: Let X be a random vector (→ I/1.2.3) following a multinomial distribution (→ II/2.2.1):

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = [np_1, \ldots, np_k] \; . \tag{2}$$

Proof: By definition, a multinomial random variable (→ II/2.2.1) is the sum of n independent and identical categorical trials (→ II/2.1.1) with category probabilities $p_1, \ldots, p_k$. Therefore, the expected value is

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \ldots + X_n) \tag{3}$$

and because the expected value is a linear operator (→ I/1.10.5), this is equal to

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) \\ &= \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \end{aligned} \tag{4}$$

With the expected value of the categorical distribution (→ II/2.1.3), we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} [p_1, \ldots, p_k] = n \cdot [p_1, \ldots, p_k] = [np_1, \ldots, np_k] \; . \tag{5}$$

■

2.2.4 Covariance

Theorem: Let X be a random vector (→ I/1.2.3) following a multinomial distribution (→ II/2.2.1):

$$[X_1, \ldots, X_k] = X \sim \mathrm{Mult}(n, p), \; n \in \mathbb{N}, \; p = [p_1, \ldots, p_k]^\mathrm{T} \; . \tag{1}$$

Then, the covariance matrix (→ I/1.13.9) of X is

$$\mathrm{Cov}(X) = n \left(\mathrm{diag}(p) - pp^\mathrm{T} \right) \; . \tag{2}$$

Proof: We first observe that the sample space (→ I/1.1.2) of each coordinate X_i is $\{0, 1, \ldots, n\}$ and X_i is the sum of independent draws of category i, which is drawn with probability p_i. Thus each coordinate follows a binomial distribution (→ II/1.3.1):

$$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(n, p_i), \; i = 1, \ldots, k \; , \tag{3}$$

which has the variance (→ II/1.3.5) $\mathrm{Var}(X_i) = np_i(1-p_i) = n(p_i - p_i^2)$, constituting the elements of the main diagonal in $\mathrm{Cov}(X)$ in (2). To prove $\mathrm{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$ (which constitutes the off-diagonal elements of the covariance matrix), we first recognize that

$$X_i = \sum_{k=1}^{n} \mathbb{I}_i(k), \quad \text{with} \quad \mathbb{I}_i(k) = \begin{cases} 1 & \text{if } k\text{-th draw was of category } i, \\ 0 & \text{otherwise} \; , \end{cases} \tag{4}$$

where the indicator function $\mathbb{I}_i$ is a Bernoulli-distributed (→ II/1.2.1) random variable with the expected value (→ II/1.2.3) p_i. Then, we have

$$\begin{split} \mathrm{Cov}(X_i, X_j) &= \mathrm{Cov}\left(\sum_{k=1}^{n} \mathbb{I}_i(k), \sum_{l=1}^{n} \mathbb{I}_j(l) \right) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} \mathrm{Cov}\left(\mathbb{I}_i(k), \mathbb{I}_j(l) \right) \\ &= \sum_{k=1}^{n} \left[\mathrm{Cov}\left(\mathbb{I}_i(k), \mathbb{I}_j(k) \right) + \sum_{\substack{l=1 \\ l \neq k}}^{n} \underbrace{\mathrm{Cov}\left(\mathbb{I}_i(k), \mathbb{I}_j(l) \right)}_{=0} \right] \\ &\overset{i \neq j}{=} \sum_{k=1}^{n} \left(\mathrm{E}\big(\underbrace{\mathbb{I}_i(k) \, \mathbb{I}_j(k)}_{=0} \big) - \mathrm{E}\big(\mathbb{I}_i(k)\big) \mathrm{E}\big(\mathbb{I}_j(k)\big) \right) \\ &= -\sum_{k=1}^{n} \mathrm{E}\big(\mathbb{I}_i(k)\big) \mathrm{E}\big(\mathbb{I}_j(k)\big) \\ &= -np_ip_j \; , \end{split} \tag{5}$$

as desired.

■

Sources:

- Tutz G (2012): "Multinomial Response Models"; in: *Regression for Categorical Data*, pp. 209ff.; URL: https://www.cambridge.org/core/books/regression-for-categorical-data/B71F71F2A484E2DF88256 DOI: 10.1017/CBO9780511842061.

2.2.5 Shannon entropy

Theorem: Let X be a random vector (→ I/1.2.3) following a multinomial distribution (→ II/2.2.1):

$$X \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Then, the (Shannon) entropy (→ I/2.1.1) of X is

$$\mathrm{H}(X) = n \cdot \mathrm{H}_{\mathrm{cat}}(p) - \mathrm{E}_{\mathrm{lmc}}(n, p) \tag{2}$$

where $\mathrm{H}_{\mathrm{cat}}(p)$ is the categorical entropy function, i.e. the (Shannon) entropy of the categorical distribution (→ II/2.1.5) with category probabilities p

$$\mathrm{H}_{\mathrm{cat}}(p) = -\sum_{i=1}^{k} p_i \cdot \log p_i \tag{3}$$

and $\mathrm{E}_{\mathrm{lmc}}(n, p)$ is the expected value (→ I/1.10.1) of the logarithmized multinomial coefficient (→ II/2.2.2) with superset size n

$$\mathrm{E}_{\mathrm{lmf}}(n, p) = \mathrm{E}\left[\log \binom{n}{X_1, \ldots, X_k}\right] \quad \text{where} \quad X \sim \mathrm{Mult}(n, p) \; . \tag{4}$$

Proof: The entropy (→ I/2.1.1) is defined as the probability-weighted average of the logarithmized probabilities for all possible values:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log_b p(x) \; . \tag{5}$$

The probability mass function of the multinomial distribution (→ II/2.2.2) is

$$f_X(x) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} {p_i}^{x_i} \tag{6}$$

Let $\mathcal{X}_{n,k}$ be the set of all vectors $x \in \mathbb{N}^{1 \times k}$ satisfying $\sum_{i=1}^{k} x_i = n$. Then, we have:

$$\begin{split} \mathrm{H}(X) &= -\sum_{x \in \mathcal{X}_{n,k}} f_X(x) \cdot \log f_X(x) \\ &= -\sum_{x \in \mathcal{X}_{n,k}} f_X(x) \cdot \log \left[\binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} {p_i}^{x_i} \right] \\ &= -\sum_{x \in \mathcal{X}_{n,k}} f_X(x) \cdot \left[\log \binom{n}{x_1, \ldots, x_k} + \sum_{i=1}^{k} x_i \cdot \log p_i \right] \; . \end{split} \tag{7}$$

Since the first factor in the sum corresponds to the probability mass (→ I/1.6.1) of $X = x$, we can rewrite this as the sum of the expected values (→ I/1.10.1) of the functions (→ I/1.10.12) of the discrete random variable (→ I/1.2.6) x in the square bracket:

$$
\begin{split}
\mathrm{H}(X) &= -\left\langle \log \binom{n}{x_1, \ldots, x_k} \right\rangle_{p(x)} - \left\langle \sum_{i=1}^{k} x_i \cdot \log p_i \right\rangle_{p(x)} \\
&= -\left\langle \log \binom{n}{x_1, \ldots, x_k} \right\rangle_{p(x)} - \sum_{i=1}^{k} \left\langle x_i \cdot \log p_i \right\rangle_{p(x)} \; .
\end{split} \tag{8}
$$

Using the expected value of the multinomial distribution (→ II/2.2.3), i.e. $X \sim \mathrm{Mult}(n, p) \Rightarrow \langle x_i \rangle = np_i$, this gives:

$$
\begin{split}
\mathrm{H}(X) &= -\left\langle \log \binom{n}{x_1, \ldots, x_k} \right\rangle_{p(x)} - \sum_{i=1}^{k} np_i \cdot \log p_i \\
&= -\left\langle \log \binom{n}{x_1, \ldots, x_k} \right\rangle_{p(x)} - n \sum_{i=1}^{k} p_i \cdot \log p_i \; .
\end{split} \tag{9}
$$

Finally, we note that the first term is the negative expected value (→ I/1.10.1) of the logarithm of a multinomial coefficient (→ II/2.2.2) and that the second term is the entropy of the categorical distribution (→ II/2.1.5), such that we finally get:

$$
\mathrm{H}(X) = n \cdot \mathrm{H}_{\mathrm{cat}}(p) - \mathrm{E}_{\mathrm{lmc}}(n, p) \; . \tag{10}
$$

■

3 Univariate continuous distributions

3.1 Continuous uniform distribution

3.1.1 Definition

Definition: Let X be a continuous random variable (→ I/1.2.2). Then, X is said to be uniformly distributed with minimum a and maximum b

$$X \sim \mathcal{U}(a,b) \; , \tag{1}$$

if and only if each value between and including a and b occurs with the same probability.

Sources:

- Wikipedia (2020): "Uniform distribution (continuous)"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: https://en.wikipedia.org/wiki/Uniform_distribution_(continuous).

3.1.2 Standard uniform distribution

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to be standard uniformly distributed, if X follows a continuous uniform distribution (→ II/3.1.1) with minimum $a = 0$ and maximum $b = 1$:

$$X \sim \mathcal{U}(0,1) \; . \tag{1}$$

Sources:

- Wikipedia (2021): "Continuous uniform distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-23; URL: https://en.wikipedia.org/wiki/Continuous_uniform_distribution#Standard_uniform.

3.1.3 Probability density function

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \begin{cases} \frac{1}{b-a} \; , & \text{if } a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{cases} \tag{2}$$

Proof: A continuous uniform variable is defined as (→ II/3.1.1) having a constant probability density between minimum a and maximum b. Therefore,

$$\begin{aligned} &f_X(x) \propto 1 \quad \text{for all} \quad x \in [a,b] \quad \text{and} \\ &f_X(x) = 0, \quad \text{if} \quad x < a \quad \text{or} \quad x > b \; . \end{aligned} \tag{3}$$

To ensure that $f_X(x)$ is a proper probability density function (→ I/1.7.1), the integral over all non-zero probabilities has to sum to 1. Therefore,

$$f_X(x) = \frac{1}{c(a,b)} \quad \text{for all} \quad x \in [a,b] \tag{4}$$

where the normalization factor $c(a,b)$ is specified, such that

$$\frac{1}{c(a,b)} \int_a^b 1 \, \mathrm{d}x = 1 \; . \tag{5}$$

Solving this for $c(a,b)$, we obtain:

$$\begin{aligned} \int_a^b 1 \, \mathrm{d}x &= c(a,b) \\ [x]_a^b &= c(a,b) \\ c(a,b) &= b - a \; . \end{aligned} \tag{6}$$

■

3.1.4 Cumulative distribution function

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \begin{cases} 0 \; , & \text{if } x < a \\ \frac{x-a}{b-a} \; , & \text{if } a \leq x \leq b \\ 1 \; , & \text{if } x > b \; . \end{cases} \tag{2}$$

Proof: The probability density function of the continuous uniform distribution (→ II/3.1.3) is:

$$\mathcal{U}(x;a,b) = \begin{cases} \frac{1}{b-a} \; , & \text{if } a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{cases} \tag{3}$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$F_X(x) = \int_{-\infty}^{x} \mathcal{U}(z;a,b) \, \mathrm{d}z \tag{4}$$

First of all, if $x < a$, we have

$$F_X(x) = \int_{-\infty}^{x} 0 \, \mathrm{d}z = 0 \; . \tag{5}$$

Moreover, if $a \leq x \leq b$, we have using (3)

$$
\begin{split}
F_X(x) &= \int_{-\infty}^{a} \mathcal{U}(z; a, b) \, \mathrm{d}z + \int_{a}^{x} \mathcal{U}(z; a, b) \, \mathrm{d}z \\
&= \int_{-\infty}^{a} 0 \, \mathrm{d}z + \int_{a}^{x} \frac{1}{b-a} \, \mathrm{d}z \\
&= 0 + \frac{1}{b-a} [z]_a^x \\
&= \frac{x-a}{b-a} \; .
\end{split} \tag{6}
$$

Finally, if $x > b$, we have

$$
\begin{split}
F_X(x) &= \int_{-\infty}^{b} \mathcal{U}(z; a, b) \, \mathrm{d}z + \int_{b}^{x} \mathcal{U}(z; a, b) \, \mathrm{d}z \\
&= F_X(b) + \int_{b}^{x} 0 \, \mathrm{d}z \\
&= \frac{b-a}{b-a} + 0 \\
&= 1 \; .
\end{split} \tag{7}
$$

This completes the proof.

■

3.1.5 Quantile function

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$
X \sim \mathcal{U}(a, b) \; . \tag{1}
$$

Then, the quantile function (→ I/1.9.1) of X is

$$
Q_X(p) = \left\{
\begin{array}{rl}
-\infty \; , & \text{if} \; p = 0 \\
bp + a(1-p) \; , & \text{if} \; p > 0 \; .
\end{array}
\right. \tag{2}
$$

Proof: The cumulative distribution function of the continuous uniform distribution (→ II/3.1.4) is:

$$
F_X(x) = \left\{
\begin{array}{rl}
0 \; , & \text{if} \; x < a \\
\frac{x-a}{b-a} \; , & \text{if} \; a \leq x \leq b \\
1 \; , & \text{if} \; x > b \; .
\end{array}
\right. \tag{3}
$$

The quantile function (→ I/1.9.1) $Q_X(p)$ is defined as the smallest x, such that $F_X(x) = p$:

$$
Q_X(p) = \min \left\{ x \in \mathbb{R} \, | \, F_X(x) = p \right\} \; . \tag{4}
$$

Thus, we have $Q_X(p) = -\infty$, if $p = 0$. When $p > 0$, it holds that (→ I/1.9.2)

$$Q_X(p) = F_X^{-1}(x) \; . \tag{5}$$

This can be derived by rearranging equation (3):

$$\begin{aligned}
p &= \frac{x-a}{b-a} \\
x &= p(b-a) + a \\
x &= bp + a(1-p) \; .
\end{aligned} \tag{6}$$

■

3.1.6 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \frac{1}{2}(a+b) \; . \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \; . \tag{3}$$

With the probability density function of the continuous uniform distribution (→ II/3.1.3), this becomes:

$$\begin{aligned}
\mathrm{E}(X) &= \int_a^b x \cdot \frac{1}{b-a} \, \mathrm{d}x \\
&= \left[\frac{1}{2} \frac{x^2}{b-a} \right]_a^b \\
&= \frac{1}{2} \frac{b^2-a^2}{b-a} \\
&= \frac{1}{2} \frac{(b+a)(b-a)}{b-a} \\
&= \frac{1}{2}(a+b) \; .
\end{aligned} \tag{4}$$

■

3.1.7 Median

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the median (→ I/1.15.1) of X is

$$\mathrm{median}(X) = \frac{1}{2}(a+b) \; . \tag{2}$$

Proof: The median (→ I/1.15.1) is the value at which the cumulative distribution function (→ I/1.8.1) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the continuous uniform distribution (→ II/3.1.4) is

$$F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < a \\ \frac{x-a}{b-a} \; , & \text{if} \; a \leq x \leq b \\ 1 \; , & \text{if} \; x > b \; . \end{array} \right. \tag{4}$$

Thus, the inverse CDF (→ II/3.1.5) is

$$x = bp + a(1-p) \; . \tag{5}$$

Setting $p = 1/2$, we obtain:

$$\mathrm{median}(X) = b \cdot \frac{1}{2} + a \cdot \left(1 - \frac{1}{2} \right) = \frac{1}{2}(a+b) \; . \tag{6}$$

■

3.1.8 Mode

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the mode (→ I/1.15.2) of X is

$$\mathrm{mode}(X) \in [a, b] \; . \tag{2}$$

Proof: The mode (→ I/1.15.2) is the value which maximizes the probability density function (→ I/1.7.1):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the continuous uniform distribution (→ II/3.1.3) is:

$$f_X(x) = \begin{cases} \frac{1}{b-a} \; , & \text{if } a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{cases} \tag{4}$$

Since the PDF attains its only non-zero value whenever $a \leq x \leq b$,

$$\max_x f_X(x) = \frac{1}{b-a} \; , \tag{5}$$

any value in the interval $[a, b]$ may be considered the mode of X.

■

3.1.9 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = \frac{1}{12}(b-a)^2 \; . \tag{2}$$

Proof: The variance (→ I/1.11.1) is the probability-weighted average of the squared deviation from the mean (→ I/1.10.1):

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mathrm{E}(X))^2 \cdot f_{\mathrm{X}}(x) \, \mathrm{d}x \; . \tag{3}$$

With the expected value (→ II/3.1.6) and probability density function (→ II/3.1.3) of the continuous uniform distribution, this reads:

$$
\begin{aligned}
\mathrm{Var}(X) &= \int_a^b \left(x - \frac{1}{2}(a+b) \right)^2 \cdot \frac{1}{b-a} \, \mathrm{d}x \\
&= \frac{1}{b-a} \cdot \int_a^b \left(x - \frac{a+b}{2} \right)^2 \mathrm{d}x \\
&= \frac{1}{b-a} \cdot \left[\frac{1}{3} \left(x - \frac{a+b}{2} \right)^3 \right]_a^b \\
&= \frac{1}{3(b-a)} \cdot \left[\left(\frac{2x-(a+b)}{2} \right)^3 \right]_a^b \\
&= \frac{1}{3(b-a)} \cdot \left[\frac{1}{8} (2x-a-b)^3 \right]_a^b \\
&= \frac{1}{24(b-a)} \cdot \left[(2x-a-b)^3 \right]_a^b \\
&= \frac{1}{24(b-a)} \cdot \left[(2b-a-b)^3 - (2a-a-b)^3 \right] \\
&= \frac{1}{24(b-a)} \cdot \left[(b-a)^3 - (a-b)^3 \right] \\
&= \frac{1}{24(b-a)} \cdot \left[(b-a)^3 + (-1)^3 (a-b)^3 \right] \\
&= \frac{1}{24(b-a)} \cdot \left[(b-a)^3 + (b-a)^3 \right] \\
&= \frac{2}{24(b-a)} (b-a)^3 \\
&= \frac{1}{12} (b-a)^2 \; .
\end{aligned} \tag{4}
$$

■

3.1.10 Differential entropy

Theorem: Let X be a random variable (→ I/1.2.2) following a continuous uniform distribution (→ II/3.1.1):

$$
X \sim \mathcal{U}(a,b) \; . \tag{1}
$$

Then, the differential entropy (→ I/2.2.1) of X is

$$
\mathrm{h}(X) = \ln(b-a) \; . \tag{2}
$$

Proof: The differential entropy (→ I/2.2.1) of a random variable is defined as

$$
\mathrm{h}(X) = - \int_{\mathcal{X}} p(x) \, \log_b p(x) \, \mathrm{d}x \; . \tag{3}
$$

To measure $h(X)$ in nats, we set $b = e$, such that

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \, \ln p(x) \, \mathrm{d}x \; . \tag{4}$$

With the probability density function of the continuous uniform distribution (→ II/3.1.3), the differential entropy of X is:

$$\begin{split}
\mathrm{h}(X) &= -\int_a^b \frac{1}{b-a} \, \ln\left(\frac{1}{b-a}\right) \, \mathrm{d}x \\
&= \frac{1}{b-a} \cdot \int_a^b \ln(b-a) \, \mathrm{d}x \\
&= \frac{1}{b-a} \cdot \left[x \cdot \ln(b-a)\right]_a^b \\
&= \frac{1}{b-a} \cdot \left[b \cdot \ln(b-a) - a \cdot \ln(b-a)\right] \\
&= \frac{1}{b-a} (b-a) \ln(b-a) \\
&= \ln(b-a) \; .
\end{split} \tag{5}$$

■

3.1.11 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ I/1.2.2). Assume two continuous uniform distributions (→ II/3.1.1) P and Q specifying the probability distribution of X as

$$\begin{split}
P &: \; X \sim \mathcal{U}(a_1, b_1) \\
Q &: \; X \sim \mathcal{U}(a_2, b_2) \; .
\end{split} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P \,||\, Q] = \ln \frac{b_2 - a_2}{b_1 - a_1} \; . \tag{2}$$

Proof: The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P \,||\, Q] = \int_{\mathcal{X}} p(x) \, \ln \frac{p(x)}{q(x)} \, \mathrm{d}x \; . \tag{3}$$

This means that the KL divergence of P from Q is only defined, if for all $x \in \mathcal{X}$, $q(x) = 0$ implies $p(x) = 0$. Thus, $\mathrm{KL}[P \,||\, Q]$ only exists, if $a_2 \leq a_1$ and $b_1 \leq b_2$, i.e. if P only places non-zero probability where Q also places non-zero probability, such that $q(x)$ is not zero for any $x \in \mathcal{X}$ where $p(x)$ is positive.
If this requirement is fulfilled, we can write

$$\mathrm{KL}[P \,||\, Q] = \int_{-\infty}^{a_1} p(x) \, \ln \frac{p(x)}{q(x)} \, \mathrm{d}x + \int_{a_1}^{b_1} p(x) \, \ln \frac{p(x)}{q(x)} \, \mathrm{d}x + \int_{b_1}^{+\infty} p(x) \, \ln \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{4}$$

and because $p(x) = 0$ for any $x < a_1$ and any $x > b_1$, we have

$$\mathrm{KL}[P\,||\,Q] = \int_{-\infty}^{a_1} 0 \cdot \ln \frac{0}{q(x)}\,\mathrm{d}x + \int_{a_1}^{b_1} p(x)\,\ln \frac{p(x)}{q(x)}\,\mathrm{d}x + \int_{b_1}^{+\infty} 0 \cdot \ln \frac{0}{q(x)}\,\mathrm{d}x\ . \tag{5}$$

Now, $(0 \cdot \ln 0)$ is taken to be zero by convention (→ I/2.1.1), such that

$$\mathrm{KL}[P\,||\,Q] = \int_{a_1}^{b_1} p(x)\,\ln \frac{p(x)}{q(x)}\,\mathrm{d}x \tag{6}$$

and we can use the probability density function of the continuous uniform distribution (→ II/3.1.3) to evaluate:

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \int_{a_1}^{b_1} \frac{1}{b_1-a_1}\,\ln \frac{\frac{1}{b_1-a_1}}{\frac{1}{b_2-a_2}}\,\mathrm{d}x \\
&= \frac{1}{b_1-a_1}\,\ln \frac{b_2-a_2}{b_1-a_1} \int_{a_1}^{b_1} \mathrm{d}x \\
&= \frac{1}{b_1-a_1}\,\ln \frac{b_2-a_2}{b_1-a_1}\,[x]_{a_1}^{b_1} \\
&= \frac{1}{b_1-a_1}\,\ln \frac{b_2-a_2}{b_1-a_1}(b_1-a_1) \\
&= \ln \frac{b_2-a_2}{b_1-a_1}\ .
\end{split} \tag{7}$$

■

3.1.12 Maximum entropy distribution

Theorem: The continuous uniform distribution (→ II/3.1.1) maximizes differential entropy (→ I/2.2.1) for a random variable (→ I/1.2.2) with a fixed range.

Proof: Without loss of generality, let us assume that the random variable X is in the following range: $a \leq X \leq b$.
Let $g(x)$ be the probability density function (→ I/1.7.1) of a continuous uniform distribution (→ II/3.1.1) with minimum a and maximum b and let $f(x)$ be an arbitrary probability density function (→ I/1.7.1) defined on the same support $\mathcal{X} = [a, b]$.
For a random variable (→ I/1.2.2) X with set of possible values $\mathcal{X}$ and probability density function (→ I/1.7.1) $p(x)$, the differential entropy (→ I/2.2.1) is defined as:

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \log p(x)\,\mathrm{d}x \tag{1}$$

Consider the Kullback-Leibler divergence (→ I/2.5.1) of distribution $f(x)$ from distribution $g(x)$ which is non-negative (→ I/2.5.2):

$$\begin{split}
0 \leq \mathrm{KL}[f||g] &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)}\,\mathrm{d}x \\
&= \int_{\mathcal{X}} f(x) \log f(x)\,\mathrm{d}x - \int_{\mathcal{X}} f(x) \log g(x)\,\mathrm{d}x \\
&\overset{(1)}{=} -\mathrm{h}[f(x)] - \int_{\mathcal{X}} f(x) \log g(x)\,\mathrm{d}x\ .
\end{split} \tag{2}$$

By plugging the probability density function of the continuous uniform distribution (→ II/3.1.3) into the second term, we obtain:

$$\begin{aligned} \int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x &= \int_{\mathcal{X}} f(x) \log \frac{1}{b-a} \, \mathrm{d}x \\ &= \log \frac{1}{b-a} \int_{\mathcal{X}} f(x) \, \mathrm{d}x \\ &= -\log(b-a) \; . \end{aligned} \tag{3}$$

This is actually the negative of the differential entropy of the continuous uniform distribution (→ II/3.1.10), such that:

$$\int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x = -\mathrm{h}[\mathcal{U}(a,b)] = -\mathrm{h}[g(x)] \; . \tag{4}$$

Combining (2) with (4), we can show that

$$\begin{aligned} 0 &\leq \mathrm{KL}[f||g] \\ 0 &\leq -\mathrm{h}[f(x)] - (-\mathrm{h}[g(x)]) \\ \mathrm{h}[g(x)] &\geq \mathrm{h}[f(x)] \end{aligned} \tag{5}$$

which means that the differential entropy (→ I/2.2.1) of the continuous uniform distribution (→ II/3.1.1) $\mathcal{U}(a,b)$ will be larger than or equal to any other distribution (→ I/1.5.1) defined in the same range.

■

Sources:

- Wikipedia (2023): "Maximum entropy probability distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-08-25; URL: https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution#Uniform_and_piecewise_uniform_distributions.

3.2 Normal distribution

3.2.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to be normally distributed with mean μ and variance σ^2 (or, standard deviation σ)

$$X \sim \mathcal{N}(\mu, \sigma^2) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{2}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Sources:

- Wikipedia (2020): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: https://en.wikipedia.org/wiki/Normal_distribution.

3.2.2 Special case of multivariate normal distribution

Theorem: The normal distribution (→ II/3.2.1) is a special case of the multivariate normal distribution (→ II/4.1.1) with number of variables $n = 1$, i.e. random vector (→ I/1.2.3) $x \in \mathbb{R}$, mean $\mu \in \mathbb{R}$ and covariance matrix $\Sigma = \sigma^2$.

Proof: The probability density function of the multivariate normal distribution (→ II/4.1.7) is

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \; . \tag{1}$$

Setting $n = 1$, such that $x, \mu \in \mathbb{R}$, and $\Sigma = \sigma^2$, we obtain

$$\begin{split} \mathcal{N}(x; \mu, \sigma^2) &= \frac{1}{\sqrt{(2\pi)^1 |\sigma^2|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} (\sigma^2)^{-1} (x-\mu) \right] \\ &= \frac{1}{\sqrt{(2\pi) \sigma^2}} \cdot \exp\left[-\frac{1}{2\sigma^2} (x-\mu)^2 \right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \end{split} \tag{2}$$

which is equivalent to the probability density function of the normal distribution (→ II/3.2.10).

■

Sources:
- Wikipedia (2022): "Multivariate normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-08-19; URL: https://en.wikipedia.org/wiki/Multivariate_normal_distribution.

3.2.3 Standard normal distribution

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to be standard normally distributed, if X follows a normal distribution (→ II/3.2.1) with mean $\mu = 0$ and variance $\sigma^2 = 1$:

$$X \sim \mathcal{N}(0, 1) \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-26; URL: https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution.

3.2.4 Relationship to standard normal distribution

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1) with mean μ and variance σ^2:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the quantity $Z = (X - \mu)/\sigma$ will have a standard normal distribution (→ II/3.2.3) with mean 0 and variance 1:

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \; . \tag{2}$$

Proof: Note that Z is a function of X

$$Z = g(X) = \frac{X-\mu}{\sigma} \tag{3}$$

with the inverse function

$$X = g^{-1}(Z) = \sigma Z + \mu \; . \tag{4}$$

Because σ is positive, $g(X)$ is strictly increasing and we can calculate the cumulative distribution function of a strictly increasing function (→ I/1.8.3) as

$$F_Y(y) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; y < \mathrm{min}(\mathcal{Y}) \\ F_X(g^{-1}(y)) \; , & \text{if} \; y \in \mathcal{Y} \\ 1 \; , & \text{if} \; y > \mathrm{max}(\mathcal{Y}) \; . \end{array} \right. \tag{5}$$

The cumulative distribution function of the normally distributed (→ II/3.2.12) X is

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2 \pi} \sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{t-\mu}{\sigma} \right)^2 \right] \, \mathrm{d}t \; . \tag{6}$$

Applying (5) to (6), we have:

$$\begin{split} F_Z(z) &\overset{(5)}{=} F_X(g^{-1}(z)) \\ &\overset{(6)}{=} \int_{-\infty}^{\sigma z + \mu} \frac{1}{\sqrt{2 \pi} \sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{t-\mu}{\sigma} \right)^2 \right] \, \mathrm{d}t \; . \end{split} \tag{7}$$

Substituting $s = (t-\mu)/\sigma$, such that $t = \sigma s + \mu$, we obtain

$$\begin{split} F_Z(z) &= \int_{(-\infty-\mu)/\sigma}^{([\sigma z + \mu]-\mu)/\sigma} \frac{1}{\sqrt{2 \pi} \sigma} \cdot \exp \left[-\frac{1}{2} \left(\frac{(\sigma s + \mu)-\mu}{\sigma} \right)^2 \right] \, \mathrm{d}(\sigma s + \mu) \\ &= \int_{-\infty}^{z} \frac{\sigma}{\sqrt{2 \pi} \sigma} \cdot \exp \left[-\frac{1}{2} s^2 \right] \, \mathrm{d}s \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2 \pi}} \cdot \exp \left[-\frac{1}{2} s^2 \right] \, \mathrm{d}s \end{split} \tag{8}$$

which is the cumulative distribution function (→ I/1.8.1) of the standard normal distribution (→ II/3.2.3).

■

3.2.5 Relationship to standard normal distribution

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1) with mean μ and variance σ^2:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the quantity $Z = (X-\mu)/\sigma$ will have a standard normal distribution (→ II/3.2.3) with mean 0 and variance 1:

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \; . \tag{2}$$

Proof: Note that Z is a function of X

$$Z = g(X) = \frac{X-\mu}{\sigma} \tag{3}$$

with the inverse function

$$X = g^{-1}(Z) = \sigma Z + \mu \; . \tag{4}$$

Because σ is positive, $g(X)$ is strictly increasing and we can calculate the probability density function of a strictly increasing function (→ I/1.7.3) as

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{5}$$

where $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$. With the probability density function of the normal distribution (→ II/3.2.10), we have

$$\begin{split} f_Z(z) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{g^{-1}(z)-\mu}{\sigma}\right)^2\right] \cdot \frac{\mathrm{d}g^{-1}(z)}{\mathrm{d}z} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)^2\right] \cdot \frac{\mathrm{d}(\sigma z+\mu)}{\mathrm{d}z} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}z^2\right] \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}z^2\right] \end{split} \tag{6}$$

which is the probability density function (→ I/1.7.1) of the standard normal distribution (→ II/3.2.3).

■

3.2.6 Relationship to standard normal distribution

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1) with mean μ and variance σ^2:

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the quantity $Z = (X - \mu)/\sigma$ will have a standard normal distribution (→ II/3.2.3) with mean 0 and variance 1:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \; . \tag{2}$$

Proof: The linear transformation theorem for multivariate normal distribution (→ II/4.1.13) states

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathrm{T}) \tag{3}$$

where x is an $n \times 1$ random vector (→ I/1.2.3) following a multivariate normal distribution (→ II/4.1.1) with mean μ and covariance Σ, A is an $m \times n$ matrix and b is an $m \times 1$ vector. Note that

$$Z = \frac{X - \mu}{\sigma} = \frac{X}{\sigma} - \frac{\mu}{\sigma} \tag{4}$$

is a special case of (3) with $x = X$, $\mu = \mu$, $\Sigma = \sigma^2$, $A = 1/\sigma$ and $b = \mu/\sigma$. Applying theorem (3) to Z as a function of X, we have

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X}{\sigma} - \frac{\mu}{\sigma} \sim \mathcal{N}\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{1}{\sigma} \cdot \sigma^2 \cdot \frac{1}{\sigma} \right) \tag{5}$$

which results in the distribution:

$$Z \sim \mathcal{N}(0, 1) \; . \tag{6}$$

■

3.2.7 Relationship to chi-squared distribution

Theorem: Let $X_1, \ldots, X_n$ be independent (→ I/1.3.6) random variables (→ I/1.2.2) where each of them is following a normal distribution (→ II/3.2.1) with mean μ and variance σ^2:

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for} \quad i = 1, \ldots, n \; . \tag{1}$$

Define the sample mean (→ I/1.10.2)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2}$$

and the unbiased sample variance (→ I/1.11.2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \; . \tag{3}$$

Then, the sampling distribution (→ I/1.5.5) of the sample variance is given by a chi-squared distribution (→ II/3.7.1) with $n - 1$ degrees of freedom:

$$V = (n-1) \, \frac{s^2}{\sigma^2} \sim \chi^2(n-1) \; . \tag{4}$$

Proof: Consider the random variable (→ I/1.2.2) U_i defined as

$$U_i = \frac{X_i - \mu}{\sigma} \tag{5}$$

which follows a standard normal distribution (→ II/3.2.4)

$$U_i \sim \mathcal{N}(0,1) \; . \tag{6}$$

Then, the sum of squared random variables U_i can be rewritten as

$$\begin{split}
\sum_{i=1}^n U_i^2 &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \\
&= \sum_{i=1}^n \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \\
&= \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \sum_{i=1}^n \frac{(\bar{X} - \mu)^2}{\sigma^2} + 2 \sum_{i=1}^n \frac{(X_i - \bar{X})(\bar{X} - \mu)}{\sigma^2} \\
&= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 + 2 \frac{(\bar{X} - \mu)}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}) \; .
\end{split} \tag{7}$$

Because the following sum is zero

$$\begin{split}
\sum_{i=1}^n (X_i - \bar{X}) &= \sum_{i=1}^n X_i - n\bar{X} \\
&= \sum_{i=1}^n X_i - n \cdot \frac{1}{n} \sum_{i=1}^n X_i \\
&= \sum_{i=1}^n X_i - \sum_{i=1}^n X_i \\
&= 0 \; ,
\end{split} \tag{8}$$

the third term disappears, i.e.

$$\sum_{i=1}^n U_i^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \; . \tag{9}$$

Cochran's theorem states that, if a sum of squared standard normal (→ II/3.2.3) random variables (→ I/1.2.2) can be written as a sum of squared forms

$$\begin{split}
\sum_{i=1}^n U_i^2 = \sum_{j=1}^m Q_j \quad &\text{where} \quad Q_j = \sum_{k=1}^n \sum_{l=1}^n U_k B_{kl}^{(j)} U_l \\
&\text{with} \quad \sum_{j=1}^m B^{(j)} = I_n \\
&\text{and} \quad r_j = \mathrm{rank}(B^{(j)}) \; ,
\end{split} \tag{10}$$

then the terms Q_j are independent (→ I/1.3.6) and each term Q_j follows a chi-squared distribution (→ II/3.7.1) with r_j degrees of freedom:

$$Q_j \sim \chi^2(r_j) \; . \tag{11}$$

We observe that (9) can be represented as

$$\begin{aligned} \sum_{i=1}^n U_i^2 &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \sum_{i=1}^n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\ = Q_1 + Q_2 &= \sum_{i=1}^n \left(U_i - \frac{1}{n} \sum_{j=1}^n U_j \right)^2 + \frac{1}{n} \left(\sum_{i=1}^n U_i \right)^2 \end{aligned} \tag{12}$$

where, with the $n \times n$ matrix of ones J_n, the matrices $B^{(j)}$ are

$$B^{(1)} = I_n - \frac{J_n}{n} \quad \text{and} \quad B^{(2)} = \frac{J_n}{n} \; . \tag{13}$$

Because all columns of $B^{(2)}$ are identical, it has rank $r_2 = 1$. Because the n columns of $B^{(1)}$ add up to zero, it has rank $r_1 = n - 1$. Thus, the conditions of Cochran's theorem are met and the squared form

$$Q_1 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = (n-1) \, \frac{1}{\sigma^2} \, \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \bar{X} \right)^2 = (n-1) \, \frac{s^2}{\sigma^2} \tag{14}$$

follows a chi-squared distribution (→ II/3.7.1) with $n-1$ degrees of freedom:

$$(n-1) \, \frac{s^2}{\sigma^2} \sim \chi^2(n-1) \; . \tag{15}$$

■

Sources:

- Glen-b (2014): "Why is the sampling distribution of variance a chi-squared distribution?"; in: *StackExchange CrossValidated*, retrieved on 2021-05-20; URL: https://stats.stackexchange.com/questions/121662/why-is-the-sampling-distribution-of-variance-a-chi-squared-distribution.
- Wikipedia (2021): "Cochran's theorem"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-20; URL: https://en.wikipedia.org/wiki/Cochran%27s_theorem#Sample_mean_and_sample_variance.

3.2.8 Relationship to t-distribution

Theorem: Let $X_1, \ldots, X_n$ be independent (→ I/1.3.6) random variables (→ I/1.2.2) where each of them is following a normal distribution (→ II/3.2.1) with mean μ and variance σ^2:

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for} \quad i = 1, \ldots, n \; . \tag{1}$$

Define the sample mean (→ I/1.10.2)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \tag{2}$$

and the unbiased sample variance (→ I/1.11.2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \; . \tag{3}$$

Then, subtracting μ from the sample mean (→ I/1.10.1), dividing by the sample standard deviation (→ I/1.16.1) and multiplying with $\sqrt{n}$ results in a qunatity that follows a t-distribution (→ II/3.3.1) with $n-1$ degrees of freedom:

$$t = \sqrt{n}\, \frac{\bar{X}-\mu}{s} \sim \mathrm{t}(n-1) \; . \tag{4}$$

Proof: Note that $\bar{X}$ is a linear combination of $X_1, \ldots, X_n$:

$$\bar{X} = \frac{1}{n} X_1, + \ldots + \frac{1}{n} X_n \; . \tag{5}$$

Because the linear combination of independent normal random variables is also normally distributed (→ II/3.2.26), we have:

$$\bar{X} \sim \mathcal{N}\left(\frac{1}{n} n\mu, \left(\frac{1}{n}\right)^2 n\sigma^2 \right) = \mathcal{N}\left(\mu, \sigma^2/n\right) \; . \tag{6}$$

Let $Z = \sqrt{n}\,(\bar{X}-\mu)/\sigma$. Because Z is a linear transformation (→ II/4.1.13) of $\bar{X}$, it also follows a normal distribution:

$$Z = \sqrt{n} \frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}\left(\frac{\sqrt{n}}{\sigma}(\mu-\mu), \left(\frac{\sqrt{n}}{\sigma}\right)^2 \sigma^2/n \right) = \mathcal{N}(0,1) \; . \tag{7}$$

Let $V = (n-1)\, s^2/\sigma^2$. We know that this function of the sample variance follows a chi-squared distribution (→ II/3.2.7) with $n-1$ degrees of freedom:

$$V = (n-1) \frac{s^2}{\sigma^2} \sim \chi^2(n-1) \; . \tag{8}$$

Observe that t is the ratio of a standard normal random variable (→ II/3.2.3) and the square root of a chi-squared random variable (→ II/3.7.1), divided by its degrees of freedom:

$$t = \sqrt{n}\, \frac{\bar{X}-\mu}{s} = \frac{\sqrt{n} \frac{\bar{X}-\mu}{\sigma}}{\sqrt{(n-1)\frac{s^2}{\sigma^2}/(n-1)}} = \frac{Z}{\sqrt{V/(n-1)}} \; . \tag{9}$$

Thus, by definition of the t-distribution (→ II/3.3.1), this ratio follows a t-distribution with $n-1$ degrees of freedom:

$$t \sim \mathrm{t}(n-1) \; . \tag{10}$$

■

Sources:
- Wikipedia (2021): "Student's t-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-05-27; URL: https://en.wikipedia.org/wiki/Student%27s_t-distribution#Characterization.
- Wikipedia (2021): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-05-27; URL: https://en.wikipedia.org/wiki/Normal_distribution#Operations_on_multiple_independent_normal_variables.

3.2.9 Gaussian integral

Theorem: The definite integral of $\exp\left[-x^2\right]$ from $-\infty$ to $+\infty$ is equal to the square root of π:

$$\int_{-\infty}^{+\infty} \exp\left[-x^2\right] \, \mathrm{d}x = \sqrt{\pi} \; . \tag{1}$$

Proof: Let

$$I = \int_0^{\infty} \exp\left[-x^2\right] \, \mathrm{d}x \tag{2}$$

and

$$I_P = \int_0^{P} \exp\left[-x^2\right] \, \mathrm{d}x = \int_0^{P} \exp\left[-y^2\right] \, \mathrm{d}y \; . \tag{3}$$

Then, we have

$$\lim_{P \to \infty} I_P = I \tag{4}$$

and

$$\lim_{P \to \infty} I_P^2 = I^2 \; . \tag{5}$$

Moreover, we can write

$$\begin{split} I_P^2 &\overset{(3)}{=} \left(\int_0^{P} \exp\left[-x^2\right] \, \mathrm{d}x \right) \left(\int_0^{P} \exp\left[-y^2\right] \, \mathrm{d}y \right) \\ &= \int_0^{P} \int_0^{P} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \\ &= \iint_{S_P} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \end{split} \tag{6}$$

where S_P is the square with corners $(0,0)$, $(0,P)$, (P,P) and $(P,0)$. For this integral, we can write down the following inequality

$$\iint_{C_1} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \leq I_P^2 \leq \iint_{C_2} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \tag{7}$$

where C_1 and C_2 are the regions in the first quadrant bounded by circles with center at $(0,0)$ and going through the points $(0,P)$ and (P,P), respectively. The radii of these two circles are $r_1 = \sqrt{P^2} = P$ and $r_2 = \sqrt{2P^2} = P\sqrt{2}$, such that we can rewrite equation (7) using polar coordinates as

$$\int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp\left[-r^2\right] \, r \, \mathrm{d}r \, \mathrm{d}\theta \leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp\left[-r^2\right] \, r \, \mathrm{d}r \, \mathrm{d}\theta \; . \tag{8}$$

Solving the definite integrals yields:

$$
\begin{aligned}
\int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp\left[-r^2\right] r \, \mathrm{d}r \, \mathrm{d}\theta \leq I_P^2 &\leq \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp\left[-r^2\right] r \, \mathrm{d}r \, \mathrm{d}\theta \\
\int_0^{\frac{\pi}{2}} \left[-\frac{1}{2}\exp\left[-r^2\right]\right]_0^{r_1} \mathrm{d}\theta \leq I_P^2 &\leq \int_0^{\frac{\pi}{2}} \left[-\frac{1}{2}\exp\left[-r^2\right]\right]_0^{r_2} \mathrm{d}\theta \\
-\frac{1}{2}\int_0^{\frac{\pi}{2}} \left(\exp\left[-r_1^2\right]-1\right) \mathrm{d}\theta \leq I_P^2 &\leq -\frac{1}{2}\int_0^{\frac{\pi}{2}} \left(\exp\left[-r_2^2\right]-1\right) \mathrm{d}\theta \\
-\frac{1}{2}\left[\left(\exp\left[-r_1^2\right]-1\right)\theta\right]_0^{\frac{\pi}{2}} \leq I_P^2 &\leq -\frac{1}{2}\left[\left(\exp\left[-r_2^2\right]-1\right)\theta\right]_0^{\frac{\pi}{2}} \\
\frac{1}{2}\left(1-\exp\left[-r_1^2\right]\right)\frac{\pi}{2} \leq I_P^2 &\leq \frac{1}{2}\left(1-\exp\left[-r_2^2\right]\right)\frac{\pi}{2} \\
\frac{\pi}{4}\left(1-\exp\left[-P^2\right]\right) \leq I_P^2 &\leq \frac{\pi}{4}\left(1-\exp\left[-2P^2\right]\right)
\end{aligned} \tag{9}
$$

Calculating the limit for $P \to \infty$, we obtain

$$
\begin{aligned}
\lim_{P\to\infty} \frac{\pi}{4}\left(1-\exp\left[-P^2\right]\right) \leq \lim_{P\to\infty} I_P^2 &\leq \lim_{P\to\infty} \frac{\pi}{4}\left(1-\exp\left[-2P^2\right]\right) \\
\frac{\pi}{4} \leq I^2 &\leq \frac{\pi}{4} \; ,
\end{aligned} \tag{10}
$$

such that we have a preliminary result for I:

$$
I^2 = \frac{\pi}{4} \quad \Rightarrow \quad I = \frac{\sqrt{\pi}}{2} \; . \tag{11}
$$

Because the integrand in (1) is an even function, we can calculate the final result as follows:

$$
\begin{aligned}
\int_{-\infty}^{+\infty} \exp\left[-x^2\right] \mathrm{d}x &= 2 \int_0^{\infty} \exp\left[-x^2\right] \mathrm{d}x \\
&\overset{(11)}{=} 2\,\frac{\sqrt{\pi}}{2} \\
&= \sqrt{\pi} \; .
\end{aligned} \tag{12}
$$

■

Sources:

- ProofWiki (2020): "Gaussian Integral"; in: *ProofWiki*, retrieved on 2020-11-25; URL: https://proofwiki.org/wiki/Gaussian_Integral.
- ProofWiki (2020): "Integral to Infinity of Exponential of minus t squared"; in: *ProofWiki*, retrieved on 2020-11-25; URL: https://proofwiki.org/wiki/Integral_to_Infinity_of_Exponential_of_-t%5E2.

3.2.10 Probability density function

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \; . \tag{2}$$

Proof: This follows directly from the definition of the normal distribution (→ II/3.2.1).

■

3.2.11 Moment-generating function

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the moment-generating function (→ I/1.9.5) of X is

$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \; . \tag{2}$$

Proof: The probability density function of the normal distribution (→ II/3.2.10) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3}$$

and the moment-generating function (→ I/1.9.5) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] \; . \tag{4}$$

Using the expected value for continuous random variables (→ I/1.10.1), the moment-generating function of X therefore is

$$\begin{split} M_X(t) &= \int_{-\infty}^{+\infty} \exp[tx] \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left[tx - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \; . \end{split} \tag{5}$$

Substituting $u = (x-\mu)/(\sqrt{2}\sigma)$, i.e. $x = \sqrt{2}\sigma u + \mu$, we have

$$
\begin{aligned}
M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(+\infty-\mu)/(\sqrt{2}\sigma)} \exp\left[t\left(\sqrt{2}\sigma u + \mu\right) - \frac{1}{2}\left(\frac{\sqrt{2}\sigma u + \mu - \mu}{\sigma}\right)^2 \right] \mathrm{d}\left(\sqrt{2}\sigma u + \mu\right) \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left[\left(\sqrt{2}\sigma u + \mu\right) t - u^2 \right] \mathrm{d}u \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[\sqrt{2}\sigma u t - u^2 \right] \mathrm{d}u \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u^2 - \sqrt{2}\sigma u t\right) \right] \mathrm{d}u \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u - \frac{\sqrt{2}}{2}\sigma t\right)^2 + \frac{1}{2}\sigma^2 t^2 \right] \mathrm{d}u \\
&= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u - \frac{\sqrt{2}}{2}\sigma t\right)^2 \right] \mathrm{d}u
\end{aligned} \tag{6}
$$

Now substituting $v = u - \sqrt{2}/2\,\sigma t$, i.e. $u = v + \sqrt{2}/2\,\sigma t$, we have

$$
\begin{aligned}
M_X(t) &= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty-\sqrt{2}/2\,\sigma t}^{+\infty-\sqrt{2}/2\,\sigma t} \exp\left[-v^2\right] \mathrm{d}\left(v + \sqrt{2}/2\,\sigma t\right) \\
&= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-v^2\right] \mathrm{d}v \; .
\end{aligned} \tag{7}
$$

With the Gaussian integral (→ II/3.2.9)

$$
\int_{-\infty}^{+\infty} \exp\left[-x^2\right] \mathrm{d}x = \sqrt{\pi} \; , \tag{8}
$$

this finally becomes

$$
M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \; . \tag{9}
$$

■

Sources:

- ProofWiki (2020): "Moment Generating Function of Gaussian Distribution"; in: *ProofWiki*, retrieved on 2020-03-03; URL: https://proofwiki.org/wiki/Moment_Generating_Function_of_Gaussian_Distribution.

3.2.12 Cumulative distribution function

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{2}$$

where $\mathrm{erf}(x)$ is the error function defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,\mathrm{d}t \; . \tag{3}$$

Proof: The probability density function of the normal distribution (→ II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \; . \tag{4}$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x \mathcal{N}(z;\mu,\sigma^2)\,\mathrm{d}z \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right]\,\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x \exp\left[-\left(\frac{z-\mu}{\sqrt{2}\sigma}\right)^2\right]\,\mathrm{d}z \; . \end{aligned} \tag{5}$$

Substituting $t = (z-\mu)/(\sqrt{2}\sigma)$, i.e. $z = \sqrt{2}\sigma t + \mu$, this becomes:

$$\begin{aligned} F_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(x-\mu)/(\sqrt{2}\sigma)} \exp(-t^2)\,\mathrm{d}\left(\sqrt{2}\sigma t + \mu\right) \\ &= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}} \exp(-t^2)\,\mathrm{d}t \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}} \exp(-t^2)\,\mathrm{d}t \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} \exp(-t^2)\,\mathrm{d}t + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x-\mu}{\sqrt{2}\sigma}} \exp(-t^2)\,\mathrm{d}t \\ &= \frac{1}{\sqrt{\pi}} \int_0^{\infty} \exp(-t^2)\,\mathrm{d}t + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x-\mu}{\sqrt{2}\sigma}} \exp(-t^2)\,\mathrm{d}t \; . \end{aligned} \tag{6}$$

Applying (3) to (6), we have:

$$\begin{aligned} F_X(x) &= \frac{1}{2}\lim_{x\to\infty} \mathrm{erf}(x) + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \\ &= \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \\ &= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \; . \end{aligned} \tag{7}$$

■

Sources:

- Wikipedia (2020): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Normal_distribution#Cumulative_distribution_function.
- Wikipedia (2020): "Error function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Error_function.

3.2.13 Cumulative distribution function without error function

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X can be expressed as

$$F_X(x) = \Phi_{\mu,\sigma}(x) = \varphi\left(\frac{x-\mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x-\mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2} \tag{2}$$

where $\varphi(x)$ is the probability density function (→ I/1.7.1) of the standard normal distribution (→ II/3.2.3) and $n!!$ is a double factorial.

Proof:

1) First, consider the standard normal distribution (→ II/3.2.3) $\mathcal{N}(0,1)$ which has the probability density function (→ II/3.2.10)

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \; . \tag{3}$$

Let $T(x)$ be the indefinite integral of this function. It can be obtained using infinitely repeated integration by parts as follows:

$$\begin{aligned}
T(x) &= \int \varphi(x) \, \mathrm{d}x \\
&= \int \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int 1 \cdot e^{-\frac{1}{2}x^2} \, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \int x^2 \cdot e^{-\frac{1}{2}x^2} \, \mathrm{d}x \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{3}x^4 \cdot e^{-\frac{1}{2}x^2} \, \mathrm{d}x \right]\right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[x \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{3}x^3 \cdot e^{-\frac{1}{2}x^2} + \left[\frac{1}{15}x^5 \cdot e^{-\frac{1}{2}x^2} + \int \frac{1}{15}x^6 \cdot e^{-\frac{1}{2}x^2} \, \mathrm{d}x \right]\right]\right] \\
&= \ldots \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[\sum_{i=1}^{n} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \mathrm{d}x \right] \\
&= \frac{1}{\sqrt{2\pi}} \cdot \left[\sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + \lim_{n \to \infty} \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \mathrm{d}x \right] \; .
\end{aligned} \tag{4}$$

Since $(2n-1)!!$ grows faster than x^{2n}, it holds that

$$\frac{1}{\sqrt{2\pi}} \cdot \lim_{n \to \infty} \int \left(\frac{x^{2n}}{(2n-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) \mathrm{d}x = \int 0 \, \mathrm{d}x = c \tag{5}$$

for constant c, such that the indefinite integral becomes

$$\begin{split} T(x) &= \frac{1}{\sqrt{2\pi}} \cdot \sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + c \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c \\ &\overset{(3)}{=} \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c \; . \end{split} \tag{6}$$

2) Next, let $\Phi(x)$ be the cumulative distribution function (→ I/1.8.1) of the standard normal distribution (→ II/3.2.3):

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x) \, \mathrm{d}x \; . \tag{7}$$

It can be obtained by matching $T(0)$ to $\Phi(0)$ which is 1/2, because the standard normal distribution is symmetric around zero:

$$\begin{split} T(0) = \varphi(0) \cdot \sum_{i=1}^{\infty} \frac{0^{2i-1}}{(2i-1)!!} + c &= \frac{1}{2} = \Phi(0) \\ \Leftrightarrow c &= \frac{1}{2} \\ \Rightarrow \Phi(x) = \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} &+ \frac{1}{2} \; . \end{split} \tag{8}$$

3) Finally, the cumulative distribution functions (→ I/1.8.1) of the standard normal distribution (→ II/3.2.3) and the general normal distribution (→ II/3.2.1) are related to each other (→ II/3.2.4) as

$$\Phi_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \; . \tag{9}$$

Combining (9) with (8), we have:

$$\Phi_{\mu,\sigma}(x) = \varphi\left(\frac{x-\mu}{\sigma}\right) \cdot \sum_{i=1}^{\infty} \frac{\left(\frac{x-\mu}{\sigma}\right)^{2i-1}}{(2i-1)!!} + \frac{1}{2} \; . \tag{10}$$

■

Sources:

- Soch J (2015): "Solution for the Indefinite Integral of the Standard Normal Probability Density Function"; in: *arXiv stat.OT*, 1512.04858; URL: https://arxiv.org/abs/1512.04858.
- Wikipedia (2020): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Normal_distribution#Cumulative_distribution_function.

3.2.14 Probability of being within standard deviations from mean

Theorem: (also called "68-95-99.7 rule") Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1) with mean (→ I/1.10.1) μ and variance (→ I/1.11.1) σ^2. Then, about 68%, 95% and 99.7% of the values of X will fall within 1, 2 and 3 standard deviations (→ I/1.16.1) from the mean (→ I/1.10.1), respectively:

$$
\begin{aligned}
\mathrm{Pr}(\mu - 1\sigma \leq X \leq \mu + 1\sigma) &\approx 68\% \\
\mathrm{Pr}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 95\% \\
\mathrm{Pr}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 99.7\% \; .
\end{aligned}
\tag{1}
$$

Proof: The cumulative distribution function of a normally distributed (→ II/3.2.12) random variable X is

$$
F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{2}
$$

where $\mathrm{erf}(x)$ is the error function defined as

$$
\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, \mathrm{d}t \tag{3}
$$

which exhibits a point-symmetry property:

$$
\mathrm{erf}(-x) = -\mathrm{erf}(x) \; . \tag{4}
$$

Thus, the probability that X falls between $\mu - a \cdot \sigma$ and $\mu + a \cdot \sigma$ is equal to:

$$
\begin{aligned}
p(a) &= \mathrm{Pr}(\mu - a\sigma \leq X \leq \mu + a\sigma) \\
&= F_X(\mu + a\sigma) - F_X(\mu - a\sigma) \\
&\overset{(2)}{=} \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu + a\sigma - \mu}{\sqrt{2}\sigma}\right)\right] - \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu - a\sigma - \mu}{\sqrt{2}\sigma}\right)\right] \\
&= \frac{1}{2}\left[\mathrm{erf}\left(\frac{\mu + a\sigma - \mu}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{\mu - a\sigma - \mu}{\sqrt{2}\sigma}\right)\right] \\
&= \frac{1}{2}\left[\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right) - \mathrm{erf}\left(-\frac{a}{\sqrt{2}}\right)\right] \\
&\overset{(4)}{=} \frac{1}{2}\left[\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right) + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right] \\
&= \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)
\end{aligned}
\tag{5}
$$

With that, we can use numerical implementations of the error function to calculate:

$$
\begin{aligned}
\mathrm{Pr}(\mu - 1\sigma \leq X \leq \mu + 1\sigma) &= p(1) = 68.27\% \\
\mathrm{Pr}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= p(2) = 95.45\% \\
\mathrm{Pr}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &= p(3) = 99.73\% \; .
\end{aligned}
\tag{6}
$$

■

Sources:
- Wikipedia (2022): "68-95-99.7 rule"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-05.08; URL: https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule.

3.2.15 Quantile function

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distributions (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the quantile function (→ I/1.9.1) of X is

$$Q_X(p) = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p-1) + \mu \tag{2}$$

where $\mathrm{erf}^{-1}(x)$ is the inverse error function.

Proof: The cumulative distribution function of the normal distribution (→ II/3.2.12) is:

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] . \tag{3}$$

Because the cumulative distribution function (CDF) is strictly monotonically increasing, the quantile function is equal to the inverse of the CDF (→ I/1.9.2):

$$Q_X(p) = F_X^{-1}(x) . \tag{4}$$

This can be derived by rearranging equation (3):

$$\begin{aligned}
p &= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \\
2p - 1 &= \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \\
\mathrm{erf}^{-1}(2p-1) &= \frac{x-\mu}{\sqrt{2}\sigma} \\
x &= \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p-1) + \mu .
\end{aligned} \tag{5}$$

■

Sources:
- Wikipedia (2020): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Normal_distribution#Quantile_function.

3.2.16 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \mu \; . \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \; . \tag{3}$$

With the probability density function of the normal distribution (→ II/3.2.10), this reads:

$$\begin{split} \mathrm{E}(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \mathrm{d}x \; . \end{split} \tag{4}$$

Substituting $z = x - \mu$, we have:

$$\begin{split} \mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} (z+\mu) \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}(z+\mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (z+\mu) \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}z \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \mathrm{d}z \right) \; . \end{split} \tag{5}$$

The general antiderivatives are

$$\begin{split} \int x \cdot \exp\left[-ax^2 \right] \mathrm{d}x &= -\frac{1}{2a} \cdot \exp\left[-ax^2 \right] \\ \int \exp\left[-ax^2 \right] \mathrm{d}x &= \frac{1}{2} \sqrt{\frac{\pi}{a}} \cdot \mathrm{erf}\left[\sqrt{a}x \right] \end{split} \tag{6}$$

where $\mathrm{erf}(x)$ is the error function. Using this, the integrals can be calculated as:

$$\begin{split} \mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right]_{-\infty}^{+\infty} + \mu \left[\sqrt{\frac{\pi}{2}}\sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z \right] \right]_{-\infty}^{+\infty} \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[\lim_{z \to \infty} \left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right) - \lim_{z \to -\infty} \left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2 \right] \right) \right] \right. \\ &\qquad \left. + \mu \left[\lim_{z \to \infty} \left(\sqrt{\frac{\pi}{2}}\sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z \right] \right) - \lim_{z \to -\infty} \left(\sqrt{\frac{\pi}{2}}\sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z \right] \right) \right] \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \left([0 - 0] + \mu \left[\sqrt{\frac{\pi}{2}}\sigma - \left(-\sqrt{\frac{\pi}{2}}\sigma \right) \right] \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \mu \cdot 2\sqrt{\frac{\pi}{2}}\sigma \\ &= \mu \; . \end{split} \tag{7}$$

■

Sources:

- Papadopoulos, Alecos (2013): "How to derive the mean and variance of Gaussian random variable?"; in: *StackExchange Mathematics*, retrieved on 2020-01-09; URL: https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-variable.

3.2.17 Median

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the median (→ I/1.15.1) of X is

$$\mathrm{median}(X) = \mu \; . \tag{2}$$

Proof: The median (→ I/1.15.1) is the value at which the cumulative distribution function (→ I/1.8.1) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the normal distribution (→ II/3.2.12) is

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right] \tag{4}$$

where $\mathrm{erf}(x)$ is the error function. Thus, the inverse CDF is

$$x = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p-1) + \mu \tag{5}$$

where $\mathrm{erf}^{-1}(x)$ is the inverse error function. Setting $p = 1/2$, we obtain:

$$\mathrm{median}(X) = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(0) + \mu = \mu \; . \tag{6}$$

■

3.2.18 Mode

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the mode (→ I/1.15.2) of X is

$$\mathrm{mode}(X) = \mu \; . \tag{2}$$

Proof: The mode (→ I/1.15.2) is the value which maximizes the probability density function (→ I/1.7.1):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the normal distribution (→ II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \; . \tag{4}$$

The first two deriatives of this function are:

$$f'_X(x) = \frac{\mathrm{d}f_X(x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{5}$$

$$f''_X(x) = \frac{\mathrm{d}^2 f_X(x)}{\mathrm{d}x^2} = -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (-x+\mu)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \; . \tag{6}$$

We now calculate the root of the first derivative (5):

$$\begin{aligned} f'_X(x) = 0 &= \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ 0 &= -x+\mu \\ x &= \mu \; . \end{aligned} \tag{7}$$

By plugging this value into the second deriative (6),

$$\begin{aligned} f''_X(\mu) &= -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp(0) + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (0)^2 \cdot \exp(0) \\ &= -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 \; , \end{aligned} \tag{8}$$

we confirm that it is in fact a maximum which shows that

$$\mathrm{mode}(X) = \mu \; . \tag{9}$$

■

3.2.19 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = \sigma^2 \; . \tag{2}$$

Proof: The variance (→ I/1.11.1) is the probability-weighted average of the squared deviation from the mean (→ I/1.10.1):

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mathrm{E}(X))^2 \cdot f_X(x) \, \mathrm{d}x \; . \tag{3}$$

With the expected value (→ II/3.2.16) and probability density function (→ II/3.2.10) of the normal distribution, this reads:

$$
\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \; .
\end{aligned} \tag{4}
$$

Substituting $z = x - \mu$, we have:

$$
\begin{aligned}
\mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} z^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}(z+\mu) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} z^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \mathrm{d}z \; .
\end{aligned} \tag{5}
$$

Now substituting $z = \sqrt{2}\sigma x$, we have:

$$
\begin{aligned}
\mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma x)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}\sigma x}{\sigma}\right)^2\right] \mathrm{d}(\sqrt{2}\sigma x) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot 2\sigma^2 \cdot \sqrt{2}\sigma \int_{-\infty}^{+\infty} x^2 \cdot \exp\left[-x^2\right] \mathrm{d}x \\
&= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 \cdot e^{-x^2} \, \mathrm{d}x \; .
\end{aligned} \tag{6}
$$

Since the integrand is symmetric with respect to $x = 0$, we can write:

$$
\mathrm{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} x^2 \cdot e^{-x^2} \, \mathrm{d}x \; . \tag{7}
$$

If we define $z = x^2$, then $x = \sqrt{z}$ and $\mathrm{d}x = 1/2\, z^{-1/2}\, \mathrm{d}z$. Substituting this into the integral

$$
\mathrm{Var}(X) = \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z \cdot e^{-z} \cdot \frac{1}{2} z^{-\frac{1}{2}} \, \mathrm{d}z = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z^{\frac{3}{2}-1} \cdot e^{-z} \, \mathrm{d}z \tag{8}
$$

and using the definition of the gamma function

$$
\Gamma(x) = \int_0^{\infty} z^{x-1} \cdot e^{-z} \, \mathrm{d}z \; , \tag{9}
$$

we can finally show that

$$
\mathrm{Var}(X) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \Gamma\left(\frac{3}{2}\right) = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} = \sigma^2 \; . \tag{10}
$$

■

Sources:

- Papadopoulos, Alecos (2013): "How to derive the mean and variance of Gaussian random variable?"; in: *StackExchange Mathematics*, retrieved on 2020-01-09; URL: https://math.stackexchange.com/questions/518281/how-to-derive-the-mean-and-variance-of-a-gaussian-random-variable.

3.2.20 Full width at half maximum

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the full width at half maximum (→ I/1.16.2) (FWHM) of X is

$$\mathrm{FWHM}(X) = 2\sqrt{2\ln 2}\sigma \; . \tag{2}$$

Proof: The probability density function of the normal distribution (→ II/3.2.10) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3}$$

and the mode of the normal distribution (→ II/3.2.18) is

$$\mathrm{mode}(X) = \mu \; , \tag{4}$$

such that

$$f_{\mathrm{max}} = f_X(\mathrm{mode}(X)) \overset{(4)}{=} f_X(\mu) \overset{(3)}{=} \frac{1}{\sqrt{2\pi}\sigma} \; . \tag{5}$$

The FWHM bounds satisfy the equation (→ I/1.16.2)

$$f_X(x_{\mathrm{FWHM}}) = \frac{1}{2} f_{\mathrm{max}} \overset{(5)}{=} \frac{1}{2\sqrt{2\pi}\sigma} \; . \tag{6}$$

Using (3), we can develop this equation as follows:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x_{\mathrm{FWHM}}-\mu}{\sigma}\right)^2\right] &= \frac{1}{2\sqrt{2\pi}\sigma} \\
\exp\left[-\frac{1}{2}\left(\frac{x_{\mathrm{FWHM}}-\mu}{\sigma}\right)^2\right] &= \frac{1}{2} \\
-\frac{1}{2}\left(\frac{x_{\mathrm{FWHM}}-\mu}{\sigma}\right)^2 &= \ln\frac{1}{2} \\
\left(\frac{x_{\mathrm{FWHM}}-\mu}{\sigma}\right)^2 &= -2\ln\frac{1}{2} \\
\frac{x_{\mathrm{FWHM}}-\mu}{\sigma} &= \pm\sqrt{2\ln 2} \\
x_{\mathrm{FWHM}}-\mu &= \pm\sqrt{2\ln 2}\sigma \\
x_{\mathrm{FWHM}} &= \pm\sqrt{2\ln 2}\sigma + \mu \; .
\end{aligned} \tag{7}$$

This implies the following two solutions for x_{FWHM}

$$\begin{aligned}
x_1 &= \mu - \sqrt{2\ln 2}\sigma \\
x_2 &= \mu + \sqrt{2\ln 2}\sigma \; ,
\end{aligned} \tag{8}$$

such that the full width at half maximum (→ I/1.16.2) of X is

$$
\begin{aligned}
\mathrm{FWHM}(X) &= \Delta x = x_2 - x_1 \\
&\overset{(8)}{=} \left(\mu + \sqrt{2\ln 2}\sigma\right) - \left(\mu - \sqrt{2\ln 2}\sigma\right) \\
&= 2\sqrt{2\ln 2}\sigma \; .
\end{aligned} \tag{9}
$$

■

Sources:
- Wikipedia (2020): "Full width at half maximum"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-08-19; URL: https://en.wikipedia.org/wiki/Full_width_at_half_maximum.

3.2.21 Extreme points

Theorem: The probability density function (→ I/1.7.1) of the normal distribution (→ II/3.2.1) with mean μ and variance σ^2 has a maximum at $x = \mu$ and no other extrema. Consequently, the normal distribution (→ II/3.2.1) is a unimodal probability distribution.

Proof: The probability density function of the normal distribution (→ II/3.2.10) is:

$$
f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \; . \tag{1}
$$

The first two deriatives of this function (→ II/3.2.18) are:

$$
f_X'(x) = \frac{\mathrm{d}f_X(x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{2}
$$

$$
f_X''(x) = \frac{\mathrm{d}^2 f_X(x)}{\mathrm{d}x^2} = -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (-x+\mu)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \; . \tag{3}
$$

The first derivative is zero, if and only if

$$
-x + \mu = 0 \quad \Leftrightarrow \quad x = \mu \; . \tag{4}
$$

Since the second derivative is negative at this value

$$
f_X''(\mu) = -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 \; , \tag{5}
$$

there is a maximum at $x = \mu$. From (2), it can be seen that $f_X'(x)$ is positive for $x < \mu$ and negative for $x > \mu$. Thus, there are no further extrema and $\mathcal{N}(\mu, \sigma^2)$ is unimodal (→ II/3.2.18).

■

Sources:
- Wikipedia (2021): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-25; URL: https://en.wikipedia.org/wiki/Normal_distribution#Symmetries_and_derivatives.

3.2.22 Inflection points

Theorem: The probability density function (→ I/1.7.1) of the normal distribution (→ II/3.2.1) with mean μ and variance σ^2 has two inflection points at $x = \mu - \sigma$ and $x = \mu + \sigma$, i.e. exactly one standard deviation (→ I/1.16.1) away from the expected value (→ I/1.10.1).

Proof: The probability density function of the normal distribution (→ II/3.2.10) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] . \tag{1}$$

The first three deriatives of this function are:

$$f_X'(x) = \frac{\mathrm{d}f_X(x)}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}\sigma} \cdot \left(-\frac{x-\mu}{\sigma^2}\right) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{2}$$

$$\begin{aligned} f_X''(x) = \frac{\mathrm{d}^2 f_X(x)}{\mathrm{d}x^2} &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left(-\frac{1}{\sigma^2}\right) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] + \frac{1}{\sqrt{2\pi}\sigma} \cdot \left(\frac{x-\mu}{\sigma^2}\right)^2 \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right] \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \end{aligned} \tag{3}$$

$$\begin{aligned} f_X'''(x) = \frac{\mathrm{d}^3 f_X(x)}{\mathrm{d}x^3} &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[\frac{2}{\sigma^2}\left(\frac{x-\mu}{\sigma^2}\right)\right] \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] - \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right] \cdot \left(\frac{x-\mu}{\sigma^2}\right) \cdot \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[-\left(\frac{x-\mu}{\sigma^2}\right)^3 + 3\left(\frac{x-\mu}{\sigma^4}\right)\right] \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] . \end{aligned} \tag{4}$$

The second derivative is zero, if and only if

$$\begin{aligned} 0 &= \left[\left(\frac{x-\mu}{\sigma^2}\right)^2 - \frac{1}{\sigma^2}\right] \\ 0 &= \frac{x^2}{\sigma^4} - \frac{2\mu x}{\sigma^4} + \frac{\mu^2}{\sigma^4} - \frac{1}{\sigma^2} \\ 0 &= x^2 - 2\mu x + (\mu^2 - \sigma^2) \\ x_{1/2} &= -\frac{-2\mu}{2} \pm \sqrt{\left(\frac{-2\mu}{2}\right)^2 - (\mu^2 - \sigma^2)} \\ x_{1/2} &= \mu \pm \sqrt{\mu^2 - \mu^2 + \sigma^2} \\ x_{1/2} &= \mu \pm \sigma . \end{aligned} \tag{5}$$

Since the third derivative is non-zero at this value

$$
\begin{aligned}
f_X'''(\mu \pm \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left[-\left(\frac{\pm\sigma}{\sigma^2} \right)^3 + 3 \left(\frac{\pm\sigma}{\sigma^4} \right) \right] \cdot \exp\left[-\frac{1}{2} \left(\frac{\pm\sigma}{\sigma} \right)^2 \right] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \left(\pm\frac{2}{\sigma^3} \right) \cdot \exp\left(-\frac{1}{2} \right) \neq 0 \; ,
\end{aligned} \tag{6}
$$

there are inflection points at $x_{1/2} = \mu \pm \sigma$. Because μ is the mean and σ^2 is the variance of a normal distribution (→ II/3.2.1), these points are exactly one standard deviation (→ I/1.16.1) away from the mean.

■

Sources:

- Wikipedia (2021): "Normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-25; URL: https://en.wikipedia.org/wiki/Normal_distribution#Symmetries_and_derivatives.

3.2.23 Differential entropy

Theorem: Let X be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the differential entropy (→ I/2.2.1) of X is

$$
\mathrm{h}(X) = \frac{1}{2} \ln\left(2\pi\sigma^2 e \right) \; . \tag{2}
$$

Proof: The differential entropy (→ I/2.2.1) of a random variable is defined as

$$
\mathrm{h}(X) = - \int_{\mathcal{X}} p(x) \, \log_b p(x) \, \mathrm{d}x \; . \tag{3}
$$

To measure $h(X)$ in nats, we set $b = e$, such that (→ I/1.10.1)

$$
\mathrm{h}(X) = -\mathrm{E}\left[\ln p(x) \right] \; . \tag{4}
$$

With the probability density function of the normal distribution (→ II/3.2.10), the differential entropy of X is:

$$
\begin{aligned}
\mathrm{h}(X) &= -\mathrm{E}\left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \right) \right] \\
&= -\mathrm{E}\left[-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \\
&= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \mathrm{E}\left[\left(\frac{x-\mu}{\sigma} \right)^2 \right] \\
&= \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \mathrm{E}\left[(x-\mu)^2 \right] \; .
\end{aligned} \tag{5}
$$

Note that $\mathrm{E}\left[(x-\mu)^2\right]$ corresponds to the variance (→ I/1.11.1) of X and the variance of the normal distribution (→ II/3.2.19) is σ^2. Thus, we can proceed:

$$\begin{aligned}
\mathrm{h}(X) &= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\cdot\frac{1}{\sigma^2}\cdot\sigma^2 \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}\ln(2\pi\sigma^2 e) \; .
\end{aligned} \tag{6}$$

■

Sources:

- Wang, Peng-Hua (2012): "Differential Entropy"; in: *National Taipei University*; URL: https://web.ntpu.edu.tw/~phwang/teaching/2012s/IT/slides/chap08.pdf.

3.2.24 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ I/1.2.2). Assume two normal distributions (→ II/3.2.1) P and Q specifying the probability distribution of X as

$$\begin{aligned}
P &: \; X \sim \mathcal{N}(\mu_1, \sigma_1^2) \\
Q &: \; X \sim \mathcal{N}(\mu_2, \sigma_2^2) \; .
\end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[\frac{(\mu_2-\mu_1)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2} - 1\right] \; . \tag{2}$$

Proof: The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P\,||\,Q] = \int_{\mathcal{X}} p(x)\,\ln\frac{p(x)}{q(x)}\,\mathrm{d}x \tag{3}$$

which, applied to the normal distributions (→ II/3.2.1) in (1), yields

$$\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \int_{-\infty}^{+\infty} \mathcal{N}(x;\mu_1,\sigma_1^2)\,\ln\frac{\mathcal{N}(x;\mu_1,\sigma_1^2)}{\mathcal{N}(x;\mu_2,\sigma_2^2)}\,\mathrm{d}x \\
&= \left\langle \ln\frac{\mathcal{N}(x;\mu_1,\sigma_1^2)}{\mathcal{N}(x;\mu_2,\sigma_2^2)} \right\rangle_{p(x)} \; .
\end{aligned} \tag{4}$$

Using the probability density function of the normal distribution (→ II/3.2.10), this becomes:

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{2\pi}\sigma_1} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]}{\frac{1}{\sqrt{2\pi}\sigma_2} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]} \right\rangle_{p(x)} \\
&= \left\langle \ln \left(\sqrt{\frac{\sigma_2^2}{\sigma_1^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 \right] \right) \right\rangle_{p(x)} \\
&= \left\langle \frac{1}{2} \ln \frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 \right\rangle_{p(x)} \\
&= \frac{1}{2} \left\langle -\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \ln \frac{\sigma_1^2}{\sigma_2^2} \right\rangle_{p(x)} \\
&= \frac{1}{2} \left\langle -\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{x^2 - 2\mu_2 x + \mu_2^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right\rangle_{p(x)} \; .
\end{aligned} \tag{5}
$$

Because the expected value (→ I/1.10.1) is a linear operator (→ I/1.10.5), the expectation can be moved into the sum:

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \frac{1}{2} \left[-\frac{\langle (x-\mu_1)^2 \rangle}{\sigma_1^2} + \frac{\langle x^2 - 2\mu_2 x + \mu_2^2 \rangle}{\sigma_2^2} - \left\langle \ln \frac{\sigma_1^2}{\sigma_2^2} \right\rangle \right] \\
&= \frac{1}{2} \left[-\frac{\langle (x-\mu_1)^2 \rangle}{\sigma_1^2} + \frac{\langle x^2 \rangle - \langle 2\mu_2 x \rangle + \langle \mu_2^2 \rangle}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right] \; .
\end{aligned} \tag{6}
$$

The first expectation corresponds to the variance (→ I/1.11.1)

$$
\left\langle (X-\mu)^2 \right\rangle = \mathrm{E}[(X-\mathrm{E}(X))^2] = \mathrm{Var}(X) \tag{7}
$$

and the variance of a normally distributed random variable (→ II/3.2.19) is

$$
X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \mathrm{Var}(X) = \sigma^2 \; . \tag{8}
$$

Additionally applying the raw moments of the normal distribution (→ II/3.2.11)

$$
X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \langle x \rangle = \mu \quad \text{and} \quad \left\langle x^2 \right\rangle = \mu^2 + \sigma^2 \; , \tag{9}
$$

the Kullback-Leibler divergence in (6) becomes

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \frac{1}{2} \left[-\frac{\sigma_1^2}{\sigma_1^2} + \frac{\mu_1^2 + \sigma_1^2 - 2\mu_2\mu_1 + \mu_2^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right] \\
&= \frac{1}{2} \left[\frac{\mu_1^2 - 2\mu_1\mu_2 + \mu_2^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right] \\
&= \frac{1}{2} \left[\frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right]
\end{aligned} \tag{10}
$$

which is equivalent to (2).

■

3.2.25 Maximum entropy distribution

Theorem: The normal distribution (→ II/3.2.1) maximizes differential entropy (→ I/2.2.1) for a random variable (→ I/1.2.2) with fixed variance (→ I/1.11.1).

Proof: For a random variable (→ I/1.2.2) X with set of possible values $\mathcal{X}$ and probability density function (→ I/1.7.1) $p(x)$, the differential entropy (→ I/2.2.1) is defined as:

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \tag{1}$$

Let $g(x)$ be the probability density function (→ I/1.7.1) of a normal distribution (→ II/3.2.1) with mean (→ I/1.10.1) μ and variance (→ I/1.11.1) σ^2 and let $f(x)$ be an arbitrary probability density function (→ I/1.7.1) with the same variance (→ I/1.11.1). Since differential entropy (→ I/2.2.1) is translation-invariant (→ I/2.2.3), we can assume that $f(x)$ has the same mean as $g(x)$.
Consider the Kullback-Leibler divergence (→ I/2.5.1) of distribution $f(x)$ from distribution $g(x)$ which is non-negative (→ I/2.5.2):

$$\begin{split}
0 \leq \mathrm{KL}[f||g] &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \, \mathrm{d}x \\
&= \int_{\mathcal{X}} f(x) \log f(x) \, \mathrm{d}x - \int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x \\
&\overset{(1)}{=} -\mathrm{h}[f(x)] - \int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x \; .
\end{split} \tag{2}$$

By plugging the probability density function of the normal distribution (→ II/3.2.10) into the second term, we obtain:

$$\begin{split}
\int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x &= \int_{\mathcal{X}} f(x) \log \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \right) \mathrm{d}x \\
&= \int_{\mathcal{X}} f(x) \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \mathrm{d}x + \int_{\mathcal{X}} f(x) \log \left(\exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \right) \mathrm{d}x \\
&= -\frac{1}{2} \log\left(2\pi\sigma^2 \right) \int_{\mathcal{X}} f(x) \, \mathrm{d}x - \frac{\log(e)}{2\sigma^2} \int_{\mathcal{X}} f(x) (x-\mu)^2 \, \mathrm{d}x \; .
\end{split} \tag{3}$$

Because the entire integral over a probability density function is one (→ I/1.7.1) and the second central moment is equal to the variance (→ I/1.18.8), we have:

$$\begin{split}
\int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x &= -\frac{1}{2} \log\left(2\pi\sigma^2 \right) - \frac{\log(e)\sigma^2}{2\sigma^2} \\
&= -\frac{1}{2} \left[\log\left(2\pi\sigma^2 \right) + \log(e) \right] \\
&= -\frac{1}{2} \log\left(2\pi\sigma^2 e \right) \; .
\end{split} \tag{4}$$

This is actually the negative of the differential entropy of the normal distribution (→ II/3.2.23), such that:

$$\int_{\mathcal{X}} f(x) \log g(x) \, \mathrm{d}x = -\mathrm{h}[\mathcal{N}(\mu, \sigma^2)] = -\mathrm{h}[g(x)] \; . \tag{5}$$

Combining (2) with (5), we can show that

$$\begin{split} 0 &\leq \mathrm{KL}[f||g] \\ 0 &\leq -\mathrm{h}[f(x)] - (-\mathrm{h}[g(x)]) \\ \mathrm{h}[g(x)] &\geq \mathrm{h}[f(x)] \end{split} \tag{6}$$

which means that the differential entropy (→ I/2.2.1) of the normal distribution (→ II/3.2.1) $\mathcal{N}(\mu, \sigma^2)$ will be larger than or equal to any other distribution (→ I/1.5.1) with the same variance (→ I/1.11.1) σ^2.

■

Sources:
- Wikipedia (2021): "Differential entropy"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-08-25; URL: https://en.wikipedia.org/wiki/Differential_entropy#Maximization_in_the_normal_distribution.

3.2.26 Linear combination

Theorem: Let $X_1, \ldots, X_n$ be independent (→ I/1.3.6) normally distributed (→ II/3.2.1) random variables (→ I/1.2.2) with means (→ I/1.10.1) $\mu_1, \ldots, \mu_n$ and variances (→ I/1.11.1) $\sigma_1^2, \ldots, \sigma_n^2$:

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \quad \text{for} \quad i = 1, \ldots, n \; . \tag{1}$$

Then, any linear combination of those random variables

$$Y = \sum_{i=1}^{n} a_i X_i \quad \text{where} \quad a_1, \ldots, a_n \in \mathbb{R} \tag{2}$$

also follows a normal distribution

$$Y \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \; \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right) \tag{3}$$

with mean and variance which are functions of the individual means and variances.

Proof: A set of n independent normal random variables $X_1, \ldots, X_n$ is equivalent (→ II/4.1.16) to an $n \times 1$ random vector (→ I/1.2.3) x following a multivariate normal distribution (→ II/4.1.1) with a diagonal covariance matrix (→ I/1.13.9). Therefore, we can write

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \; i = 1, \ldots, n \quad \Rightarrow \quad x = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) \tag{4}$$

with mean vector and covariance matrix

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} = \mathrm{diag}\left(\left[\sigma_1^2, \ldots, \sigma_n^2\right]\right) \; . \tag{5}$$

Thus, we can apply the linear transformation theorem for the multivariate normal distribution (→ II/4.1.13)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathrm{T}) \tag{6}$$

with the constant matrix and vector

$$A = [a_1, \ldots, a_n] \quad \text{and} \quad b = 0 \; . \tag{7}$$

This implies the following distribution the linear combination given by equation (2):

$$Y = Ax + b \sim \mathcal{N}(A\mu, A\Sigma A^\mathrm{T}) \; . \tag{8}$$

Finally, we note that

$$\begin{split} A\mu &= [a_1, \ldots, a_n] \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \sum_{i=1}^n a_i \mu_i \quad \text{and} \\ A\Sigma A^\mathrm{T} &= [a_1, \ldots, a_n] \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \sum_{i=1}^n a_i^2 \sigma_i^2 \; . \end{split} \tag{9}$$

■

3.3 t-distribution

3.3.1 Definition

Definition: Let Z and V be independent (→ I/1.3.6) random variables (→ I/1.2.2) following a standard normal distribution (→ II/3.2.3) and a chi-squared distribution (→ II/3.7.1) with ν degrees of freedom, respectively:

$$\begin{split} Z &\sim \mathcal{N}(0,1) \\ V &\sim \chi^2(\nu) \; . \end{split} \tag{1}$$

Then, the ratio of Z to the square root of V, divided by the respective degrees of freedom, is said to be t-distributed with degrees of freedom ν:

$$Y = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu) \; . \tag{2}$$

The t-distribution is also called "Student's t-distribution", after William S. Gosset a.k.a. "Student".

Sources:

- Wikipedia (2021): "Student's t-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-21; URL: https://en.wikipedia.org/wiki/Student%27s_t-distribution#Characterization.

3.3.2 Special case of multivariate t-distribution

Theorem: The t-distribution (→ II/3.3.1) is a special case of the multivariate t-distribution (→ II/4.2.1) with number of variables $n = 1$, i.e. random vector (→ I/1.2.3) $x \in \mathbb{R}$, mean $\mu = 0$ and covariance matrix $\Sigma = 1$.

Proof: The probability density function of the multivariate t-distribution (→ II/4.2.2) is

$$t(x; \mu, \Sigma, \nu) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \frac{\Gamma([\nu+n]/2)}{\Gamma(\nu/2)} \left[1 + \frac{1}{\nu}(x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu)\right]^{-(\nu+n)/2} . \tag{1}$$

Setting $n = 1$, such that $x \in \mathbb{R}$, as well as $\mu = 0$ and $\Sigma = 1$, we obtain

$$\begin{aligned} t(x; 0, 1, \nu) &= \sqrt{\frac{1}{(\nu\pi)^1 |1|}} \frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)} \left[1 + \frac{1}{\nu}(x-0)^\mathrm{T} 1^{-1} (x-0)\right]^{-(\nu+1)/2} \\ &= \sqrt{\frac{1}{\nu\pi}} \frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)} \left[1 + \frac{x^2}{\nu}\right]^{-(\nu+1)/2} \\ &= \frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \cdot \left[1 + \frac{x^2}{\nu}\right]^{-\frac{\nu+1}{2}} . \end{aligned} \tag{2}$$

which is equivalent to the probability density function of the t-distribution (→ II/3.3.5).

■

Sources:

- Wikipedia (2022): "Multivariate t-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-08-25; URL: https://en.wikipedia.org/wiki/Multivariate_t-distribution#Derivation.

3.3.3 Non-standardized t-distribution

Definition: Let X be a random variable (→ I/1.2.2) following a Student's t-distribution (→ II/3.3.1) with ν degrees of freedom. Then, the random variable (→ I/1.2.2)

$$Y = \sigma X + \mu \tag{1}$$

is said to follow a non-standardized t-distribution with non-centrality μ, scale σ^2 and degrees of freedom ν:

$$Y \sim \mathrm{nst}(\mu, \sigma^2, \nu) . \tag{2}$$

Sources:

- Wikipedia (2021): "Student's t-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-05-20; URL: https://en.wikipedia.org/wiki/Student%27s_t-distribution#Generalized_Student's_t-distribution.

3.3.4 Relationship to non-standardized t-distribution

Theorem: Let X be a random variable (→ I/1.2.2) following a non-standardized t-distribution (→ II/3.3.3) with mean μ, scale σ^2 and degrees of freedom ν:

$$X \sim \mathrm{nst}(\mu, \sigma^2, \nu) \; . \tag{1}$$

Then, subtracting the mean and dividing by the square root of the scale results in a random variable (→ I/1.2.2) following a t-distribution (→ II/3.3.1) with degrees of freedom ν:

$$Y = \frac{X-\mu}{\sigma} \sim t(\nu) \; . \tag{2}$$

Proof: The non-standardized t-distribution is a special case of the multivariate t-distribution (→ II/4.2.1) in which the mean vector and scale matrix are scalars:

$$X \sim \mathrm{nst}(\mu, \sigma^2, \nu) \quad \Rightarrow \quad X \sim t(\mu, \sigma^2, \nu) \; . \tag{3}$$

Therefore, we can apply the linear transformation theorem for the multivariate t-distribution for an $n \times 1$ random vector x:

$$x \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad y = Ax + b \sim t(A\mu + b, A \Sigma A^\mathrm{T}, \nu) \; . \tag{4}$$

Comparing with equation (2), we have $A = 1/\sigma$, $b = -\mu/\sigma$ and the variable Y is distributed as:

$$\begin{split} Y = \frac{X-\mu}{\sigma} &= \frac{X}{\sigma} - \frac{\mu}{\sigma} \\ &\sim t\left(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \left(\frac{1}{\sigma} \right)^2 \sigma^2, \nu \right) \\ &= t\left(0, 1, \nu \right) \; . \end{split} \tag{5}$$

Plugging $\mu = 0$, $\Sigma = 1$ and $n = 1$ into the probability density function of the multivariate t-distribution (→ II/4.2.2),

$$p(x) = \sqrt{\frac{1}{(\nu \pi)^n |\Sigma|}} \, \frac{\Gamma([\nu+n]/2)}{\Gamma(\nu/2)} \, \left[1 + \frac{1}{\nu} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \; , \tag{6}$$

we get

$$p(x) = \sqrt{\frac{1}{\nu \pi}} \, \frac{\Gamma([\nu+1]/2)}{\Gamma(\nu/2)} \, \left[1 + \frac{x^2}{\nu} \right] \tag{7}$$

which is the probability density function of Student's t-distribution (→ II/3.3.5) with ν degrees of freedom.

■

3.3.5 Probability density function

Theorem: Let T be a random variable (→ I/1.2.2) following a t-distribution (→ II/3.3.1):

$$T \sim t(\nu) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of T is

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \cdot \sqrt{\nu\pi}} \cdot \left(\frac{t^2}{\nu}+1\right)^{-\frac{\nu+1}{2}} \; . \tag{2}$$

Proof: A t-distributed random variable (→ II/3.3.1) is defined as the ratio of a standard normal random variable (→ II/3.2.3) and the square root of a chi-squared random variable (→ II/3.7.1), divided by its degrees of freedom

$$X \sim \mathcal{N}(0,1), \; Y \sim \chi^2(\nu) \quad \Rightarrow \quad T = \frac{X}{\sqrt{Y/\nu}} \sim t(\nu) \tag{3}$$

where X and Y are independent of each other (→ I/1.3.6).
The probability density function (→ II/3.2.10) of the standard normal distribution (→ II/3.2.3) is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \tag{4}$$

and the probability density function of the chi-squared distribution (→ II/3.7.3) is

$$f_Y(y) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot y^{\frac{\nu}{2}-1} \cdot e^{-\frac{y}{2}} \; . \tag{5}$$

Define the random variables T and W as functions of X and Y

$$\begin{aligned} T &= X \cdot \sqrt{\frac{\nu}{Y}} \\ W &= Y \; , \end{aligned} \tag{6}$$

such that the inverse functions X and Y in terms of T and W are

$$\begin{aligned} X &= T \cdot \sqrt{\frac{W}{\nu}} \\ Y &= W \; . \end{aligned} \tag{7}$$

This implies the following Jacobian matrix and determinant:

$$\begin{aligned} J &= \begin{bmatrix} \frac{\mathrm{d}X}{\mathrm{d}T} & \frac{\mathrm{d}X}{\mathrm{d}W} \\ \frac{\mathrm{d}Y}{\mathrm{d}T} & \frac{\mathrm{d}Y}{\mathrm{d}W} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{W}{\nu}} & \frac{T}{2\nu\sqrt{W/\nu}} \\ 0 & 1 \end{bmatrix} \\ |J| &= \sqrt{\frac{W}{\nu}} \; . \end{aligned} \tag{8}$$

Because X and Y are independent (→ I/1.3.6), the joint density (→ I/1.5.2) of X and Y is equal to the product (→ I/1.3.8) of the marginal densities (→ I/1.5.3):

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \; . \tag{9}$$

With the probability density function of an invertible function (→ I/1.7.5), the joint density (→ I/1.5.2) of T and W can be derived as:

$$f_{T,W}(t,w) = f_{X,Y}(x,y) \cdot |J| \; . \tag{10}$$

Substituting (7) into (4) and (5), and then with (8) into (10), we get:

$$\begin{split} f_{T,W}(t,w) &= f_X\left(t \cdot \sqrt{\frac{w}{\nu}}\right) \cdot f_Y(w) \cdot |J| \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{\left(t \cdot \sqrt{\frac{w}{\nu}}\right)^2}{2}} \cdot \frac{1}{\Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot w^{\frac{\nu}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \sqrt{\frac{w}{\nu}} \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot w^{\frac{\nu+1}{2}-1} \cdot e^{-\frac{w}{2}\left(\frac{t^2}{\nu}+1\right)} \; . \end{split} \tag{11}$$

The marginal density (→ I/1.5.3) of T can now be obtained by integrating out (→ I/1.3.3) W:

$$\begin{split} f_T(t) &= \int_0^\infty f_{T,W}(t,w) \, \mathrm{d}w \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \int_0^\infty w^{\frac{\nu+1}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{t^2}{\nu}+1\right) w\right] \mathrm{d}w \\ &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left[\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)\right]^{(\nu+1)/2}} \cdot \int_0^\infty \frac{\left[\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)\right]^{(\nu+1)/2}}{\Gamma\left(\frac{\nu+1}{2}\right)} \cdot w^{\frac{\nu+1}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{t^2}{\nu}+1\right) w\right] \mathrm{d}w \end{split} \tag{12}$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution (→ II/3.4.7) with

$$a = \frac{\nu+1}{2} \quad \text{and} \quad b = \frac{1}{2}\left(\frac{t^2}{\nu}+1\right) \; , \tag{13}$$

and because a probability density function integrates to one (→ I/1.7.1), we finally have:

$$\begin{split} f_T(t) &= \frac{1}{\sqrt{2\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right) \cdot 2^{\nu/2}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\left[\frac{1}{2}\left(\frac{t^2}{\nu}+1\right)\right]^{(\nu+1)/2}} \\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \cdot \sqrt{\nu\pi}} \cdot \left(\frac{t^2}{\nu}+1\right)^{-\frac{\nu+1}{2}} \; . \end{split} \tag{14}$$

■

Sources:

- Computation Empire (2021): "Student's t Distribution: Derivation of PDF"; in: *YouTube*, retrieved on 2021-10-11; URL: https://www.youtube.com/watch?v=6BraaGEVRY8.

3.4 Gamma distribution

3.4.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to follow a gamma distribution with shape a and rate b

$$X \sim \mathrm{Gam}(a, b) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathrm{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx], \quad x > 0 \tag{2}$$

where $a > 0$ and $b > 0$, and the density is zero, if $x \leq 0$.

Sources:

- Koch, Karl-Rudolf (2007): "Gamma Distribution"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, p. 47, eq. 2.172; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

3.4.2 Special case of Wishart distribution

Theorem: The gamma distribution (→ II/3.4.1) is a special case of the Wishart distribution (→ II/5.2.1) where the number of columns of the random matrix (→ I/1.2.4) is $p = 1$.

Proof: Let X be a $p \times p$ positive-definite symmetric matrix, such that X follows a Wishart distribution (→ II/5.2.1):

$$Y \sim \mathcal{W}(V, n) \; . \tag{1}$$

Then, Y is described by the probability density function

$$p(Y) = \frac{1}{\Gamma_p\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2^{np}|V|^n}} \cdot |X|^{(n-p-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1}X\right)\right] \tag{2}$$

where $|A|$ is a matrix determinant, A^{-1} is a matrix inverse and $\Gamma_p(x)$ is the multivariate gamma function of order p. If $p = 1$, then $\Gamma_p(x) = \Gamma(x)$ is the ordinary gamma function, $x = X$ and $v = V$ are real numbers. Thus, the probability density function (→ I/1.7.1) of x can be developed as

$$\begin{aligned} p(x) &= \frac{1}{\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2^n \, v^n}} \cdot x^{(n-2)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(v^{-1}x\right)\right] \\ &= \frac{(2v)^{-n/2}}{\Gamma\left(\frac{n}{2}\right)} \cdot x^{n/2-1} \cdot \exp\left[-\frac{1}{2v}x\right] \end{aligned} \tag{3}$$

Finally, substituting $a = \frac{n}{2}$ and $b = \frac{1}{2v}$, we get

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \tag{4}$$

which is the probability density function of the gamma distribution (→ II/3.4.7).

■

3.4.3 Standard gamma distribution

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to have a standard gamma distribution, if X follows a gamma distribution (→ II/3.4.1) with shape $a > 0$ and rate $b = 1$:

$$X \sim \mathrm{Gam}(a, 1) \; . \tag{1}$$

Sources:
- JoramSoch (2017): "Gamma-distributed random numbers"; in: *MACS – a new SPM toolbox for model assessment, comparison and selection*, retrieved on 2020-05-26; URL: https://github.com/JoramSoch/MACS/blob/master/MD_gamrnd.m; DOI: 10.5281/zenodo.845404.
- NIST/SEMATECH (2012): "Gamma distribution"; in: *e-Handbook of Statistical Methods*, ch. 1.3.6.6.11; URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366b.htm; DOI: 10.18434/M

3.4.4 Relationship to standard gamma distribution

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1) with shape a and rate b:

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the quantity $Y = bX$ will have a standard gamma distribution (→ II/3.4.3) with shape a and rate 1:

$$Y = bX \sim \mathrm{Gam}(a, 1) \; . \tag{2}$$

Proof: Note that Y is a function of X

$$Y = g(X) = bX \tag{3}$$

with the inverse function

$$X = g^{-1}(Y) = \frac{1}{b} Y \; . \tag{4}$$

Because b is positive, $g(X)$ is strictly increasing and we can calculate the cumulative distribution function of a strictly increasing function (→ I/1.8.3) as

$$F_Y(y) = \left\{ \begin{array}{rl} 0 \; , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 1 \; , & \text{if } y > \max(\mathcal{Y}) \; . \end{array} \right. \tag{5}$$

The cumulative distribution function of the gamma-distributed (→ II/3.4.9) X is

$$F_X(x) = \int_{-\infty}^{x} \frac{b^a}{\Gamma(a)} t^{a-1} \exp[-bt] \, \mathrm{d}t \; . \tag{6}$$

Applying (5) to (6), we have:

$$
\begin{aligned}
F_Y(y) &\overset{(5)}{=} F_X(g^{-1}(y)) \\
&\overset{(6)}{=} \int_{-\infty}^{y/b} \frac{b^a}{\Gamma(a)} t^{a-1} \exp[-bt] \, \mathrm{d}t \; .
\end{aligned} \tag{7}
$$

Substituting $s = bt$, such that $t = s/b$, we obtain

$$
\begin{aligned}
F_Y(y) &= \int_{-b\infty}^{b(y/b)} \frac{b^a}{\Gamma(a)} \left(\frac{s}{b}\right)^{a-1} \exp\left[-b\left(\frac{s}{b}\right)\right] \, \mathrm{d}\left(\frac{s}{b}\right) \\
&= \int_{-\infty}^{y} \frac{b^a}{\Gamma(a)} \frac{1}{b^{a-1}\, b} \, s^{a-1} \exp[-s] \, \mathrm{d}s \\
&= \int_{-\infty}^{y} \frac{1}{\Gamma(a)} s^{a-1} \exp[-s] \, \mathrm{d}s
\end{aligned} \tag{8}
$$

which is the cumulative distribution function (→ I/1.8.1) of the standard gamma distribution (→ II/3.4.3).

■

3.4.5 Relationship to standard gamma distribution

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1) with shape a and rate b:

$$
X \sim \mathrm{Gam}(a, b) \; . \tag{1}
$$

Then, the quantity $Y = bX$ will have a standard gamma distribution (→ II/3.4.3) with shape a and rate 1:

$$
Y = bX \sim \mathrm{Gam}(a, 1) \; . \tag{2}
$$

Proof: Note that Y is a function of X

$$
Y = g(X) = bX \tag{3}
$$

with the inverse function

$$
X = g^{-1}(Y) = \frac{1}{b} Y \; . \tag{4}
$$

Because b is positive, $g(X)$ is strictly increasing and we can calculate the probability density function of a strictly increasing function (→ I/1.7.3) as

$$
f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{cases} \tag{5}
$$

where $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$. With the probability density function of the gamma distribution (→ II/3.4.7), we have

$$
\begin{aligned}
f_Y(y) &= \frac{b^a}{\Gamma(a)} [g^{-1}(y)]^{a-1} \exp[-b\, g^{-1}(y)] \cdot \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{b}y\right)^{a-1} \exp\left[-b\left(\frac{1}{b}y\right)\right] \cdot \frac{\mathrm{d}\left(\frac{1}{b}y\right)}{\mathrm{d}y} \\
&= \frac{b^a}{\Gamma(a)} \frac{1}{b^{a-1}}\, y^{a-1} \exp[-y] \cdot \frac{1}{b} \\
&= \frac{1}{\Gamma(a)}\, y^{a-1} \exp[-y]
\end{aligned} \tag{6}
$$

which is the probability density function (→ I/1.7.1) of the standard gamma distribution (→ II/3.4.3).

■

3.4.6 Scaling of a gamma random variable

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1) with shape a and rate b:

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the quantity $Y = cX$ will also be gamma-distributed with shape a and rate b/c:

$$Y = cX \sim \mathrm{Gam}\left(a, \frac{b}{c}\right) \; . \tag{2}$$

Proof: Note that Y is a function of X

$$Y = g(X) = cX \tag{3}$$

with the inverse function

$$X = g^{-1}(Y) = \frac{1}{c}Y \; . \tag{4}$$

Because the parameters of a gamma distribution are positive (→ II/3.4.1), c must also be positive. Thus, $g(X)$ is strictly increasing and we can calculate the probability density function of a strictly increasing function (→ I/1.7.3) as

$$f_Y(y) = \left\{ \begin{array}{rl} f_X(g^{-1}(y))\, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{array} \right. \tag{5}$$

The probability density function of the gamma-distributed (→ II/3.4.7) X is

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{6}$$

Applying (5) to (6), we have:

$$
\begin{aligned}
f_Y(y) &= \frac{b^a}{\Gamma(a)} [g^{-1}(y)]^{a-1} \exp[-b g^{-1}(y)] \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{c} y \right)^{a-1} \exp\left[-b \left(\frac{1}{c} y \right) \right] \, \frac{\mathrm{d}\left(\frac{1}{c}y\right)}{\mathrm{d}y} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{c} \right)^{a} \left(\frac{1}{c} \right)^{-1} y^{a-1} \exp\left[-\frac{b}{c} y \right] \frac{1}{c} \\
&= \frac{(b/c)^a}{\Gamma(a)} y^{a-1} \exp\left[-\frac{b}{c} y \right]
\end{aligned} \tag{7}
$$

which is the probability density function (→ I/1.7.1) of a gamma distribution (→ II/3.4.1) with shape a and rate b/c.

■

3.4.7 Probability density function

Theorem: Let X be a positive random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$
X \sim \mathrm{Gam}(a, b) \; . \tag{1}
$$

Then, the probability density function (→ I/1.7.1) of X is

$$
f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{2}
$$

Proof: This follows directly from the definition of the gamma distribution (→ II/3.4.1).

■

3.4.8 Moment-generating function

Theorem: Let X follow a gamma distribution (→ II/3.4.1):

$$
X \sim \mathrm{Gam}(a, b) \; . \tag{1}
$$

Then, the moment-generating function (→ I/1.9.5) of X is

$$
M_X(t) = \left(1 - \frac{t}{b} \right)^{-a} \; . \tag{2}
$$

Proof: The moment-generating function of a random variable (→ I/1.9.5) X is defined as:

$$
M_X(t) = \mathrm{E}\left[e^{tX} \right] , \quad t \in \mathbb{R} \; . \tag{3}
$$

Applying the law of the unconscious statistician (→ I/1.10.12), we have:

$$
M_X(t) = \int_{\mathcal{X}} e^{tx} \cdot f_X(x) \, \mathrm{d}x \; . \tag{4}
$$

With the probability density function of the gamma distribution (→ II/3.4.7), we have:

$$M_X(t) = \int_{\mathbb{R}} \exp[tx] \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \, \mathrm{d}x \; . \tag{5}$$

Now we summarize the two exponential functions inside the integral:

$$\begin{aligned}
M_X(t) &= \int_{\mathbb{R}} \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] \, \mathrm{d}x \\
&= \int_{\mathbb{R}} \frac{(b-t)^a}{(b-t)^a} \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] \, \mathrm{d}x \\
&= \int_{\mathbb{R}} \frac{b^a}{(b-t)^a} \cdot \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] \, \mathrm{d}x \\
&= \left(\frac{b}{b-t} \right)^a \int_{\mathbb{R}} \frac{(b-t)^a}{\Gamma(a)} x^{a-1} \exp[-(b-t)x] \, \mathrm{d}x \; .
\end{aligned} \tag{6}$$

The integrand is equal to the probability density function of a gamma distribution (→ II/3.4.7):

$$M_X(t) = \left(\frac{b}{b-t} \right)^a \int_{\mathbb{R}} \mathrm{Gam}(x; a, b-t) \, \mathrm{d}x \; . \tag{7}$$

Because the entire probability density integrates to one (→ I/1.7.1), we finally have:

$$M_X(t) = \left(\frac{b}{b-t} \right)^a = \left(\frac{b-t}{b} \right)^{-a} = \left(\frac{b}{b} - \frac{t}{b} \right)^{-a} = \left(1 - \frac{t}{b} \right)^{-a} \; . \tag{8}$$

■

3.4.9 Cumulative distribution function

Theorem: Let X be a positive random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \frac{\gamma(a, bx)}{\Gamma(a)} \tag{2}$$

where $\Gamma(x)$ is the gamma function and $\gamma(s, x)$ is the lower incomplete gamma function.

Proof: The probability density function of the gamma distribution (→ II/3.4.7) is:

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{3}$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$\begin{aligned} F_X(x) &= \int_0^x \mathrm{Gam}(z; a, b) \, \mathrm{d}z \\ &= \int_0^x \frac{b^a}{\Gamma(a)} z^{a-1} \exp[-bz] \, \mathrm{d}z \\ &= \frac{b^a}{\Gamma(a)} \int_0^x z^{a-1} \exp[-bz] \, \mathrm{d}z \; . \end{aligned} \tag{4}$$

Substituting $t = bz$, i.e. $z = t/b$, this becomes:

$$\begin{aligned} F_X(x) &= \frac{b^a}{\Gamma(a)} \int_{b \cdot 0}^{bx} \left(\frac{t}{b}\right)^{a-1} \exp\left[-b\left(\frac{t}{b}\right)\right] \mathrm{d}\left(\frac{t}{b}\right) \\ &= \frac{b^a}{\Gamma(a)} \cdot \frac{1}{b^{a-1}} \cdot \frac{1}{b} \int_0^{bx} t^{a-1} \exp[-t] \, \mathrm{d}t \\ &= \frac{1}{\Gamma(a)} \int_0^{bx} t^{a-1} \exp[-t] \, \mathrm{d}t \; . \end{aligned} \tag{5}$$

With the definition of the lower incomplete gamma function

$$\gamma(s, x) = \int_0^x t^{s-1} \exp[-t] \, \mathrm{d}t \; , \tag{6}$$

we arrive at the final result given by equation (2):

$$F_X(x) = \frac{\gamma(a, bx)}{\Gamma(a)} \; . \tag{7}$$

■

Sources:

- Wikipedia (2020): "Incomplete gamma function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-29; URL: https://en.wikipedia.org/wiki/Incomplete_gamma_function#Definition.

3.4.10 Quantile function

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the quantile function (→ I/1.9.1) of X is

$$Q_X(p) = \begin{cases} -\infty \; , & \text{if } p = 0 \\ \gamma^{-1}(a, \Gamma(a) \cdot p)/b \; , & \text{if } p > 0 \end{cases} \tag{2}$$

where $\gamma^{-1}(s, y)$ is the inverse of the lower incomplete gamma function $\gamma(s, x)$

Proof: The cumulative distribution function of the gamma distribution (→ II/3.4.9) is:

$$F_X(x) = \begin{cases} 0 \; , & \text{if} \; x < 0 \\ \frac{\gamma(a,bx)}{\Gamma(a)} \; , & \text{if} \; x \geq 0 \; . \end{cases} \tag{3}$$

The quantile function (→ I/1.9.1) $Q_X(p)$ is defined as the smallest x, such that $F_X(x) = p$:

$$Q_X(p) = \min \left\{ x \in \mathbb{R} \, | \, F_X(x) = p \right\} \; . \tag{4}$$

Thus, we have $Q_X(p) = -\infty$, if $p = 0$. When $p > 0$, it holds that (→ I/1.9.2)

$$Q_X(p) = F_X^{-1}(x) \; . \tag{5}$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{\gamma(a,bx)}{\Gamma(a)} \\ \Gamma(a) \cdot p &= \gamma(a,bx) \\ \gamma^{-1}(a, \Gamma(a) \cdot p) &= bx \\ x &= \frac{\gamma^{-1}(a, \Gamma(a) \cdot p)}{b} \; . \end{aligned} \tag{6}$$

■

Sources:

- Wikipedia (2020): "Incomplete gamma function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: https://en.wikipedia.org/wiki/Incomplete_gamma_function#Definition.

3.4.11 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$X \sim \mathrm{Gam}(a,b) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \frac{a}{b} \; . \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \; . \tag{3}$$

With the probability density function of the gamma distribution (→ II/3.4.7), this reads:

$$\begin{aligned} \mathrm{E}(X) &= \int_0^\infty x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \, \mathrm{d}x \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] \, \mathrm{d}x \\ &= \int_0^\infty \frac{1}{b} \cdot \frac{b^{a+1}}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] \, \mathrm{d}x \; . \end{aligned} \tag{4}$$

Employing the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$\mathrm{E}(X) = \int_0^\infty \frac{a}{b} \cdot \frac{b^{a+1}}{\Gamma(a+1)} x^{(a+1)-1} \exp[-bx] \, \mathrm{d}x \tag{5}$$

and again using the density of the gamma distribution (→ II/3.4.7), we get

$$\begin{split} \mathrm{E}(X) &= \frac{a}{b} \int_0^\infty \mathrm{Gam}(x; a+1, b) \, \mathrm{d}x \\ &= \frac{a}{b} \; . \end{split} \tag{6}$$

■

Sources:
- Turlapaty, Anish (2013): "Gamma random variable: mean & variance"; in: *YouTube*, retrieved on 2020-05-19; URL: https://www.youtube.com/watch?v=Sy4wP-Y2dmA.

3.4.12 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = \frac{a}{b^2} \; . \tag{2}$$

Proof: The variance (→ I/1.11.1) can be expressed in terms of expected values (→ I/1.11.3) as

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; . \tag{3}$$

The expected value of a gamma random variable (→ II/3.4.11) is

$$\mathrm{E}(X) = \frac{a}{b} \; . \tag{4}$$

With the probability density function of the gamma distribution (→ II/3.4.7), the expected value of a squared gamma random variable is

$$\begin{split} \mathrm{E}(X^2) &= \int_0^\infty x^2 \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \, \mathrm{d}x \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+2)-1} \exp[-bx] \, \mathrm{d}x \\ &= \int_0^\infty \frac{1}{b^2} \cdot \frac{b^{a+2}}{\Gamma(a)} x^{(a+2)-1} \exp[-bx] \, \mathrm{d}x \; . \end{split} \tag{5}$$

Twice-applying the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$\mathrm{E}(X^2) = \int_0^\infty \frac{a\,(a+1)}{b^2} \cdot \frac{b^{a+2}}{\Gamma(a+2)} x^{(a+2)-1} \exp[-bx] \, \mathrm{d}x \tag{6}$$

and again using the density of the gamma distribution (→ II/3.4.7), we get

$$
\begin{aligned}
\mathrm{E}(X^2) &= \frac{a\,(a+1)}{b^2} \int_0^\infty \mathrm{Gam}(x; a+2, b)\,\mathrm{d}x \\
&= \frac{a^2+a}{b^2} \; .
\end{aligned} \tag{7}
$$

Plugging (7) and (4) into (3), the variance of a gamma random variable finally becomes

$$
\begin{aligned}
\mathrm{Var}(X) &= \frac{a^2+a}{b^2} - \left(\frac{a}{b}\right)^2 \\
&= \frac{a}{b^2} \; .
\end{aligned} \tag{8}
$$

■

Sources:

- Turlapaty, Anish (2013): "Gamma random variable: mean & variance"; in: *YouTube*, retrieved on 2020-05-19; URL: https://www.youtube.com/watch?v=Sy4wP-Y2dmA.

3.4.13 Logarithmic expectation

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$
X \sim \mathrm{Gam}(a, b) \; . \tag{1}
$$

Then, the expectation (→ I/1.10.1) of the natural logarithm of X is

$$
\mathrm{E}(\ln X) = \psi(a) - \ln(b) \tag{2}
$$

where $\psi(x)$ is the digamma function.

Proof: Let $Y = \ln(X)$, such that $\mathrm{E}(Y) = \mathrm{E}(\ln X)$ and consider the special case that $b = 1$. In this case, the probability density function of the gamma distribution (→ II/3.4.7) is

$$
f_X(x) = \frac{1}{\Gamma(a)}\, x^{a-1} \exp[-x] \; . \tag{3}
$$

Multiplying this function with $\mathrm{d}x$, we obtain

$$
f_X(x)\,\mathrm{d}x = \frac{1}{\Gamma(a)}\, x^{a} \exp[-x]\, \frac{\mathrm{d}x}{x} \; . \tag{4}
$$

Substituting $y = \ln x$, i.e. $x = e^y$, such that $\mathrm{d}x/\mathrm{d}y = x$, i.e. $\mathrm{d}x/x = \mathrm{d}y$, we get

$$
\begin{aligned}
f_Y(y)\,\mathrm{d}y &= \frac{1}{\Gamma(a)}\, (e^y)^a \, \exp[-e^y]\,\mathrm{d}y \\
&= \frac{1}{\Gamma(a)} \exp\left[ay - e^y\right]\, \mathrm{d}y \; .
\end{aligned} \tag{5}
$$

Because $f_Y(y)$ integrates to one, we have

$$
\begin{aligned}
1 &= \int_{\mathbb{R}} f_Y(y) \, \mathrm{d}y \\
1 &= \int_{\mathbb{R}} \frac{1}{\Gamma(a)} \exp\left[ay - e^y\right] \, \mathrm{d}y \\
\Gamma(a) &= \int_{\mathbb{R}} \exp\left[ay - e^y\right] \, \mathrm{d}y \; .
\end{aligned} \tag{6}
$$

Note that the integrand in (6) is differentiable with respect to a:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}a} \exp\left[ay - e^y\right] \, \mathrm{d}y &= y \exp\left[ay - e^y\right] \, \mathrm{d}y \\
&\overset{(5)}{=} \Gamma(a) \, y \, f_Y(y) \, \mathrm{d}y \; .
\end{aligned} \tag{7}
$$

Now we can calculate the expected value of $Y = \ln(X)$:

$$
\begin{aligned}
\mathrm{E}(Y) &= \int_{\mathbb{R}} y \, f_Y(y) \, \mathrm{d}y \\
&\overset{(7)}{=} \frac{1}{\Gamma(a)} \int_{\mathbb{R}} \frac{\mathrm{d}}{\mathrm{d}a} \exp\left[ay - e^y\right] \, \mathrm{d}y \\
&= \frac{1}{\Gamma(a)} \frac{\mathrm{d}}{\mathrm{d}a} \int_{\mathbb{R}} \exp\left[ay - e^y\right] \, \mathrm{d}y \\
&\overset{(6)}{=} \frac{1}{\Gamma(a)} \frac{\mathrm{d}}{\mathrm{d}a} \Gamma(a) \\
&= \frac{\Gamma'(a)}{\Gamma(a)} \; .
\end{aligned} \tag{8}
$$

Using the derivative of a logarithmized function

$$
\frac{\mathrm{d}}{\mathrm{d}x} \ln f(x) = \frac{f'(x)}{f(x)} \tag{9}
$$

and the definition of the digamma function

$$
\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x) \; , \tag{10}
$$

we have

$$
\mathrm{E}(Y) = \psi(a) \; . \tag{11}
$$

Finally, noting that $1/b$ acts as a scaling parameter (→ II/3.4.4) on a gamma-distributed (→ II/3.4.1) random variable (→ I/1.2.2),

$$
X \sim \mathrm{Gam}(a, 1) \quad \Rightarrow \quad \frac{1}{b} X \sim \mathrm{Gam}(a, b) \; , \tag{12}
$$

and that a scaling parameter acts additively on the logarithmic expectation of a random variable,

$$
\mathrm{E}\left[\ln(cX)\right] = \mathrm{E}\left[\ln(X) + \ln(c)\right] = \mathrm{E}\left[\ln(X)\right] + \ln(c) \; , \tag{13}
$$

it follows that

$$X \sim \mathrm{Gam}(a, b) \quad \Rightarrow \quad \mathrm{E}(\ln X) = \psi(a) - \ln(b) \; . \tag{14}$$

■

Sources:

- whuber (2018): "What is the expected value of the logarithm of Gamma distribution?"; in: *StackExchange CrossValidated*, retrieved on 2020-05-25; URL: https://stats.stackexchange.com/questions/370880/what-is-the-expected-value-of-the-logarithm-of-gamma-distribution.

3.4.14 Expectation of x ln x

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of $(X \cdot \ln X)$ is

$$\mathrm{E}(X \ln X) = \frac{a}{b} \left[\psi(a) - \ln(b) \right] \; . \tag{2}$$

Proof: With the definition of the expected value (→ I/1.10.1), the law of the unconscious statistician (→ I/1.10.12) and the probability density function of the gamma distribution (→ II/3.4.7), we have:

$$\begin{aligned} \mathrm{E}(X \ln X) &= \int_0^\infty x \ln x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \, \mathrm{d}x \\ &= \frac{1}{\Gamma(a)} \int_0^\infty \ln x \cdot \frac{b^{a+1}}{b} x^a \exp[-bx] \, \mathrm{d}x \\ &= \frac{\Gamma(a+1)}{\Gamma(a)\, b} \int_0^\infty \ln x \cdot \frac{b^{a+1}}{\Gamma(a+1)} x^{(a+1)-1} \exp[-bx] \, \mathrm{d}x \end{aligned} \tag{3}$$

The integral now corresponds to the logarithmic expectation of a gamma distribution (→ II/3.4.13) with shape $a+1$ and rate b

$$\mathrm{E}(\ln Y) \quad \text{where} \quad Y \sim \mathrm{Gam}(a+1, b) \tag{4}$$

which is given by (→ II/3.4.13)

$$\mathrm{E}(\ln Y) = \psi(a+1) - \ln(b) \tag{5}$$

where $\psi(x)$ is the digamma function. Additionally employing the relation

$$\Gamma(x+1) = \Gamma(x) \cdot x \quad \Leftrightarrow \quad \frac{\Gamma(x+1)}{\Gamma(x)} = x \; , \tag{6}$$

the expression in equation (3) develops into:

$$\mathrm{E}(X \ln X) = \frac{a}{b} \left[\psi(a) - \ln(b) \right] \; . \tag{7}$$

■

Sources:

- gunes (2020): "What is the expected value of x log(x) of the gamma distribution?"; in: *StackExchange Cross Validated*, retrieved on 2020-10-15; URL: https://stats.stackexchange.com/questions/457357/what-is-the-expected-value-of-x-logx-of-the-gamma-distribution.

3.4.15 Differential entropy

Theorem: Let X be a random variable (→ I/1.2.2) following a gamma distribution (→ II/3.4.1):

$$X \sim \mathrm{Gam}(a, b) \tag{1}$$

Then, the differential entropy (→ I/2.2.1) of X in nats is

$$\mathrm{h}(X) = a + \ln \Gamma(a) + (1-a) \cdot \psi(a) + \ln b \; . \tag{2}$$

Proof: The differential entropy (→ I/2.2.1) of a random variable is defined as

$$\mathrm{h}(X) = - \int_{\mathcal{X}} p(x) \, \log_b p(x) \, \mathrm{d}x \; . \tag{3}$$

To measure $h(X)$ in nats, we set $b = e$, such that (→ I/1.10.1)

$$\mathrm{h}(X) = -\mathrm{E}\left[\ln p(x) \right] \; . \tag{4}$$

With the probability density function of the gamma distribution (→ II/3.4.7), the differential entropy of X is:

$$\begin{split}
\mathrm{h}(X) &= -\mathrm{E}\left[\ln \left(\frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \right) \right] \\
&= -\mathrm{E}\left[a \cdot \ln b - \ln \Gamma(a) + (a-1) \ln x - bx \right] \\
&= -a \cdot \ln b + \ln \Gamma(a) - (a-1) \cdot \mathrm{E}(\ln x) + b \cdot \mathrm{E}(x) \; .
\end{split} \tag{5}$$

Using the mean (→ II/3.4.11) and logarithmic expectation (→ II/3.4.13) of the gamma distribution (→ II/3.4.1)

$$X \sim \mathrm{Gam}(a, b) \quad \Rightarrow \quad \mathrm{E}(X) = \frac{a}{b} \quad \text{and} \quad \mathrm{E}(\ln X) = \psi(a) - \ln(b) \; , \tag{6}$$

the differential entropy (→ I/2.2.1) of X becomes:

$$\begin{split}
\mathrm{h}(X) &= -a \cdot \ln b + \ln \Gamma(a) - (a-1) \cdot (\psi(a) - \ln b) + b \cdot \frac{a}{b} \\
&= -a \cdot \ln b + \ln \Gamma(a) + (1-a) \cdot \psi(a) + a \cdot \ln b - \ln b + a \\
&= a + \ln \Gamma(a) + (1-a) \cdot \psi(a) - \ln b \; .
\end{split} \tag{7}$$

■

Sources:

- Wikipedia (2021): "Gamma distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-07-14; URL: https://en.wikipedia.org/wiki/Gamma_distribution#Information_entropy.

3.4.16 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ I/1.2.2). Assume two gamma distributions (→ II/3.4.1) P and Q specifying the probability distribution of X as

$$\begin{aligned} P &: \; X \sim \mathrm{Gam}(a_1, b_1) \\ Q &: \; X \sim \mathrm{Gam}(a_2, b_2) \; . \end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P \,||\, Q] = a_2 \, \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2) \, \psi(a_1) - (b_1 - b_2) \, \frac{a_1}{b_1} \; . \tag{2}$$

Proof: The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P \,||\, Q] = \int_{\mathcal{X}} p(x) \, \ln \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{3}$$

which, applied to the gamma distributions (→ II/3.4.1) in (1), yields

$$\begin{aligned} \mathrm{KL}[P \,||\, Q] &= \int_{-\infty}^{+\infty} \mathrm{Gam}(x; a_1, b_1) \, \ln \frac{\mathrm{Gam}(x; a_1, b_1)}{\mathrm{Gam}(x; a_2, b_2)} \, \mathrm{d}x \\ &= \left\langle \ln \frac{\mathrm{Gam}(x; a_1, b_1)}{\mathrm{Gam}(x; a_2, b_2)} \right\rangle_{p(x)} \; . \end{aligned} \tag{4}$$

Using the probability density function of the gamma distribution (→ II/3.4.7), this becomes:

$$\begin{aligned} \mathrm{KL}[P \,||\, Q] &= \left\langle \ln \frac{\frac{{b_1}^{a_1}}{\Gamma(a_1)} x^{a_1-1} \exp[-b_1 x]}{\frac{{b_2}^{a_2}}{\Gamma(a_2)} x^{a_2-1} \exp[-b_2 x]} \right\rangle_{p(x)} \\ &= \left\langle \ln \left(\frac{{b_1}^{a_1}}{{b_2}^{a_2}} \cdot \frac{\Gamma(a_2)}{\Gamma(a_1)} \cdot x^{a_1-a_2} \cdot \exp[-(b_1-b_2)x] \right) \right\rangle_{p(x)} \\ &= \left\langle a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \ln x - (b_1 - b_2) \cdot x \right\rangle_{p(x)} \; . \end{aligned} \tag{5}$$

Using the mean of the gamma distribution (→ II/3.4.11) and the expected value of a logarithmized gamma variate (→ II/3.4.13)

$$\begin{aligned} x \sim \mathrm{Gam}(a,b) \quad \Rightarrow \quad & \langle x \rangle = \frac{a}{b} \quad \text{and} \\ & \langle \ln x \rangle = \psi(a) - \ln(b) \; , \end{aligned} \tag{6}$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{aligned} \mathrm{KL}[P \,||\, Q] &= a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot (\psi(a_1) - \ln(b_1)) - (b_1 - b_2) \cdot \frac{a_1}{b_1} \\ &= a_2 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \psi(a_1) - (b_1 - b_2) \cdot \frac{a_1}{b_1} \; . \end{aligned} \tag{7}$$

Finally, combining the logarithms, we get:

$$\mathrm{KL}[P\,||\,Q] = a_2\, \ln \frac{b_1}{b_2} - \ln \frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1 - a_2)\, \psi(a_1) - (b_1 - b_2)\, \frac{a_1}{b_1} \; . \tag{8}$$

■

Sources:
- Penny, William D. (2001): "KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities"; in: *University College, London*; URL: https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps.

3.5 Exponential distribution

3.5.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to be exponentially distributed with rate (or, inverse scale) λ

$$X \sim \mathrm{Exp}(\lambda) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathrm{Exp}(x; \lambda) = \lambda \exp[-\lambda x], \quad x \geq 0 \tag{2}$$

where $\lambda > 0$, and the density is zero, if $x < 0$.

Sources:
- Wikipedia (2020): "Exponential distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-08; URL: https://en.wikipedia.org/wiki/Exponential_distribution#Definitions.

3.5.2 Special case of gamma distribution

Theorem: The exponential distribution (→ II/3.5.1) is a special case of the gamma distribution (→ II/3.4.1) with shape $a = 1$ and rate $b = \lambda$.

Proof: The probability density function of the gamma distribution (→ II/3.4.7) is

$$\mathrm{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{1}$$

Setting $a = 1$ and $b = \lambda$, we obtain

$$\begin{split} \mathrm{Gam}(x; 1, \lambda) &= \frac{\lambda^1}{\Gamma(1)} x^{1-1} \exp[-\lambda x] \\ &= \frac{x^0}{\Gamma(1)} \lambda \exp[-\lambda x] \\ &= \lambda \exp[-\lambda x] \end{split} \tag{2}$$

which is equivalent to the probability density function of the exponential distribution (→ II/3.5.3).

■

3.5.3 Probability density function

Theorem: Let X be a non-negative random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \lambda \exp[-\lambda x] \; . \tag{2}$$

Proof: This follows directly from the definition of the exponential distribution (→ II/3.5.1).

■

3.5.4 Moment-generating function

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the moment generating function (→ I/1.9.5) of X is

$$M_X(t) = \frac{\lambda}{\lambda - t} \tag{2}$$

which is well-defined for $t < \lambda$.

Proof: Suppose X follows an exponential distribution (→ II/3.5.1) with rate λ; that is, $X \sim \mathrm{Exp}(\lambda)$. Then, the probability density function (→ II/3.5.3) is given by

$$f_X(x) = \lambda e^{-\lambda x} \tag{3}$$

and the moment-generating function (→ I/1.9.5) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] \; . \tag{4}$$

Using the definition of expected value for continuous random variables (→ I/1.10.1), the moment-generating function of X is thus:

$$\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx} \cdot f_X(x) dx \\
&= \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x} dx \\
&= \int_0^\infty \lambda e^{x(t-\lambda)} dx \\
&= \frac{\lambda}{t-\lambda} e^{x(t-\lambda)} \Big|_{x=0}^{x=\infty} \\
&= \lim_{x\to\infty} \left[\frac{\lambda}{t-\lambda} e^{x(t-\lambda)} - \frac{\lambda}{t-\lambda} \right] \\
&= \frac{\lambda}{t-\lambda} \left[\lim_{x\to\infty} e^{x(t-\lambda)} - 1 \right] \; .
\end{aligned} \tag{5}$$

Note that t cannot be equal to λ, else $M_X(t)$ is undefined. Further, if $t > \lambda$, then $\lim_{x\to\infty} e^{x(t-\lambda)} = \infty$, which implies that $M_X(t)$ diverges for $t \geq \lambda$. So, we must restrict the domain of $M_X(t)$ to $t < \lambda$. Assuming this, we can further simplify (5):

$$\begin{aligned} M_X(t) &= \frac{\lambda}{t-\lambda} \left[\lim_{x\to\infty} e^{x(t-\lambda)} - 1 \right] \\ &= \frac{\lambda}{t-\lambda} [0-1] \\ &= \frac{\lambda}{\lambda - t} \; . \end{aligned} \tag{6}$$

This completes the proof of (2).

■

3.5.5 Cumulative distribution function

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if } x < 0 \\ 1 - \exp[-\lambda x] \; , & \text{if } x \geq 0 \; . \end{array} \right. \tag{2}$$

Proof: The probability density function of the exponential distribution (→ II/3.5.3) is:

$$\mathrm{Exp}(x; \lambda) = \left\{ \begin{array}{rl} 0 \; , & \text{if } x < 0 \\ \lambda \exp[-\lambda x] \; , & \text{if } x \geq 0 \; . \end{array} \right. \tag{3}$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$F_X(x) = \int_{-\infty}^{x} \mathrm{Exp}(z; \lambda) \, \mathrm{d}z \; . \tag{4}$$

If $x < 0$, we have:

$$F_X(x) = \int_{-\infty}^{x} 0 \, \mathrm{d}z = 0 \; . \tag{5}$$

If $x \geq 0$, we have using (3):

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{0} \mathrm{Exp}(z;\lambda)\,\mathrm{d}z + \int_{0}^{x} \mathrm{Exp}(z;\lambda)\,\mathrm{d}z \\
&= \int_{-\infty}^{0} 0\,\mathrm{d}z + \int_{0}^{x} \lambda \exp[-\lambda z]\,\mathrm{d}z \\
&= 0 + \lambda \left[-\frac{1}{\lambda} \exp[-\lambda z] \right]_0^x \\
&= \lambda \left[\left(-\frac{1}{\lambda} \exp[-\lambda x] \right) - \left(-\frac{1}{\lambda} \exp[-\lambda \cdot 0] \right) \right] \\
&= 1 - \exp[-\lambda x] \; .
\end{aligned} \tag{6}
$$

■

3.5.6 Quantile function

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$
X \sim \mathrm{Exp}(\lambda) \; . \tag{1}
$$

Then, the quantile function (→ I/1.9.1) of X is

$$
Q_X(p) = \left\{ \begin{array}{rl} -\infty \; , & \text{if} \; p = 0 \\ -\frac{\ln(1-p)}{\lambda} \; , & \text{if} \; p > 0 \; . \end{array} \right. \tag{2}
$$

Proof: The cumulative distribution function of the exponential distribution (→ II/3.5.5) is:

$$
F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < 0 \\ 1 - \exp[-\lambda x] \; , & \text{if} \; x \geq 0 \; . \end{array} \right. \tag{3}
$$

The quantile function (→ I/1.9.1) $Q_X(p)$ is defined as the smallest x, such that $F_X(x) = p$:

$$
Q_X(p) = \min \left\{ x \in \mathbb{R} \,|\, F_X(x) = p \right\} \; . \tag{4}
$$

Thus, we have $Q_X(p) = -\infty$, if $p = 0$. When $p > 0$, it holds that (→ I/1.9.2)

$$
Q_X(p) = F_X^{-1}(x) \; . \tag{5}
$$

This can be derived by rearranging equation (3):

$$
\begin{aligned}
p &= 1 - \exp[-\lambda x] \\
\exp[-\lambda x] &= 1 - p \\
-\lambda x &= \ln(1-p) \\
x &= -\frac{\ln(1-p)}{\lambda} \; .
\end{aligned} \tag{6}
$$

■

3.5.7 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \frac{1}{\lambda} \; . \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_{\mathrm{X}}(x) \, \mathrm{d}x \; . \tag{3}$$

With the probability density function of the exponential distribution (→ II/3.5.3), this reads:

$$\begin{split} \mathrm{E}(X) &= \int_0^{+\infty} x \cdot \lambda \exp(-\lambda x) \, \mathrm{d}x \\ &= \lambda \int_0^{+\infty} x \cdot \exp(-\lambda x) \, \mathrm{d}x \; . \end{split} \tag{4}$$

Using the following anti-derivative

$$\int x \cdot \exp(-\lambda x) \, \mathrm{d}x = \left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \; , \tag{5}$$

the expected value becomes

$$\begin{split} \mathrm{E}(X) &= \lambda \left[\left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \right]_0^{+\infty} \\ &= \lambda \left[\lim_{x \to \infty} \left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) - \left(-\frac{1}{\lambda} \cdot 0 - \frac{1}{\lambda^2} \right) \exp(-\lambda \cdot 0) \right] \\ &= \lambda \left[0 + \frac{1}{\lambda^2} \right] \\ &= \frac{1}{\lambda} \; . \end{split} \tag{6}$$

■

Sources:

- Koch, Karl-Rudolf (2007): "Expected Value"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, p. 39, eq. 2.142a; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

3.5.8 Median

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the median (→ I/1.15.1) of X is

$$\mathrm{median}(X) = \frac{\ln 2}{\lambda} \; . \tag{2}$$

Proof: The median (→ I/1.15.1) is the value at which the cumulative distribution function (→ I/1.8.1) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the exponential distribution (→ II/3.5.5) is

$$F_X(x) = 1 - \exp[-\lambda x], \quad x \geq 0 \; . \tag{4}$$

Thus, the inverse CDF is

$$x = -\frac{\ln(1-p)}{\lambda} \tag{5}$$

and setting $p = 1/2$, we obtain:

$$\mathrm{median}(X) = -\frac{\ln(1-\frac{1}{2})}{\lambda} = \frac{\ln 2}{\lambda} \; . \tag{6}$$

■

3.5.9 Mode

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the mode (→ I/1.15.2) of X is

$$\mathrm{mode}(X) = 0 \; . \tag{2}$$

Proof: The mode (→ I/1.15.2) is the value which maximizes the probability density function (→ I/1.7.1):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the exponential distribution (→ II/3.5.3) is:

$$f_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < 0 \\ \lambda e^{-\lambda x} \; , & \text{if} \; x \geq 0 \; . \end{array} \right. \tag{4}$$

Since

$$f_X(0) = \lambda \tag{5}$$

and

$$0 < e^{-\lambda x} < 1 \quad \text{for any} \quad x > 0 \; , \tag{6}$$

it follows that

$$\text{mode}(X) = 0 \; . \tag{7}$$

■

3.5.10 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \text{Exp}(\lambda). \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\text{Var}(X) = \frac{1}{\lambda^2} \; . \tag{2}$$

Proof: The variance (→ I/1.11.1) of a random variable is defined as

$$\text{Var}(X) = \text{E}\left[(X - \text{E}(X))^2\right] \tag{3}$$

which, partitioned into expected values (→ I/1.11.3), reads:

$$\text{Var}(X) = \text{E}\left[X^2\right] - \text{E}\left[X\right]^2 \; . \tag{4}$$

The expected value of the exponential distribution (→ II/3.5.7) is:

$$\text{E}[X] = \frac{1}{\lambda} \tag{5}$$

The second moment $\text{E}[X^2]$ can be derived as follows:

$$\begin{aligned}
\text{E}[X^2] &= \int_{-\infty}^{+\infty} x^2 \cdot f_\text{X}(x) \, \mathrm{d}x \\
&= \int_{0}^{+\infty} x^2 \cdot \lambda \exp(-\lambda x) \, \mathrm{d}x \\
&= \lambda \int_{0}^{+\infty} x^2 \cdot \exp(-\lambda x) \, \mathrm{d}x
\end{aligned} \tag{6}$$

Using the following anti-derivative

$$
\begin{split}
\int x^2 \cdot \exp(-\lambda x) \, \mathrm{d}x &= \left[-\frac{1}{\lambda} x^2 \cdot \exp(-\lambda x) \right]_0^{+\infty} - \int 2x \left(-\frac{1}{\lambda} x \cdot \exp(-\lambda x) \right) \mathrm{d}x \\
&= \left[-\frac{1}{\lambda} x^2 \cdot \exp(-\lambda x) \right]_0^{+\infty} - \left(\left[\frac{1}{\lambda^2} 2x \cdot \exp(-\lambda x) \right]_0^{+\infty} - \int 2 \left(\frac{1}{\lambda^2} \cdot \exp(-\lambda x) \right) \mathrm{d}x \right) \\
&= \left[-\frac{x^2}{\lambda} \cdot \exp(-\lambda x) \right]_0^{+\infty} - \left(\left[\frac{2x}{\lambda^2} \cdot \exp(-\lambda x) \right]_0^{+\infty} - \left[-\frac{2}{\lambda^3} \cdot \exp(-\lambda x) \right]_0^{+\infty} \right) \\
&= \left[\left(-\frac{x^2}{\lambda} - \frac{2x}{\lambda^2} - \frac{2}{\lambda^3} \right) \exp(-\lambda x) \right]_0^{+\infty} \; ,
\end{split} \tag{7}
$$

the second moment becomes

$$
\begin{split}
\mathrm{E}[X^2] &= \lambda \left[\left(-\frac{x^2}{\lambda} - \frac{2x}{\lambda^2} - \frac{2}{\lambda^3} \right) \exp(-\lambda x) \right]_0^{+\infty} \\
&= \lambda \left[\lim_{x \to \infty} \left(-\frac{x^2}{\lambda} - \frac{2x}{\lambda^2} - \frac{2}{\lambda^3} \right) \exp(-\lambda x) - \left(0 - 0 - \frac{2}{\lambda^3} \right) \exp(-\lambda \cdot 0) \right] \\
&= \lambda \left[0 + \frac{2}{\lambda^3} \right] \\
&= \frac{2}{\lambda^2} \; .
\end{split} \tag{8}
$$

Plugging (8) and (5) into (4), we have:

$$
\begin{split}
\mathrm{Var}(X) &= \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 \\
&= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 \\
&= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
&= \frac{1}{\lambda^2} \; .
\end{split} \tag{9}
$$

■

Sources:

- Taboga, Marco (2023): "Exponential distribution"; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2023-01-23; URL: https://www.statlect.com/probability-distributions/exponential-distribution.
- Wikipedia (2023): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-01-23; URL: https://en.wikipedia.org/wiki/Variance#Definition.

3.5.11 Skewness

Theorem: Let X be a random variable (→ I/1.2.2) following an exponential distribution (→ II/3.5.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then the skewness (→ I/1.12.1) of X is

$$\mathrm{Skew}(X) = 2 \; . \tag{2}$$

Proof:
To compute the skewness of X, we partition the skewness into expected values (→ I/1.12.3):

$$\mathrm{Skew}(X) = \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \; , \tag{3}$$

where μ and σ are the mean and standard deviation of X, respectively. Since X follows an exponential distribution (→ II/3.5.1), the mean (→ II/3.5.7) of X is given by

$$\mu = \mathrm{E}(X) = \frac{1}{\lambda} \tag{4}$$

and the standard deviation (→ II/3.5.10) of X is given by

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda} \; . \tag{5}$$

Substituting (4) and (5) into (3) gives:

$$\begin{split}
\mathrm{Skew}(X) &= \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \\
&= \frac{\mathrm{E}(X^3)}{\sigma^3} - \frac{3\mu\sigma^2 + \mu^3}{\sigma^3} \\
&= \frac{\mathrm{E}(X^3)}{\left(\frac{1}{\lambda}\right)^3} - \frac{3\left(\frac{1}{\lambda}\right)\left(\frac{1}{\lambda}\right)^2 + \left(\frac{1}{\lambda}\right)^3}{\left(\frac{1}{\lambda}\right)^3} \\
&= \lambda^3 \cdot \mathrm{E}(X^3) - \frac{\frac{3}{\lambda^3} + \frac{1}{\lambda^3}}{\frac{1}{\lambda^3}} \\
&= \lambda^3 \cdot \mathrm{E}(X^3) - 4 \; .
\end{split} \tag{6}$$

Thus, the remaining work is to compute $\mathrm{E}(X^3)$. To do this, we use the moment-generating function of the exponential distribution (→ II/3.5.4) to calculate

$$\mathrm{E}(X^3) = M_X'''(0) \tag{7}$$

based on the relationship between raw moment and moment-generating function (→ I/1.18.2).
First, we differentiate the moment-generating function of the exponential distribution (→ II/3.5.4)

$$M_X(t) = \frac{\lambda}{\lambda - t} = \lambda(\lambda - t)^{-1} \tag{8}$$

with respect to t. Using the chain rule gives:

$$\begin{split}
M_X'(t) &= -1 \cdot \lambda(\lambda - t)^{-2} \cdot (-1) \\
&= \lambda(\lambda - t)^{-2} \; .
\end{split} \tag{9}$$

We continue using the chain rule to obtain the second derivative:

$$
\begin{aligned}
M_X''(t) &= -2 \cdot \lambda(\lambda - t)^{-3} \cdot (-1) \\
&= 2\lambda(\lambda - t)^{-3} \; .
\end{aligned} \tag{10}
$$

Finally, one more application of the chain rule gives us the third derivative:

$$
\begin{aligned}
M_X'''(t) &= -3 \cdot 2\lambda(\lambda - t)^{-4} \cdot (-1) \\
&= 6\lambda(\lambda - t)^{-4} \\
&= \frac{6\lambda}{(\lambda - t)^4} \; .
\end{aligned} \tag{11}
$$

Applying (7), together with (11), yields

$$
\begin{aligned}
\mathrm{E}(X^3) &= M_X'''(0) \\
&= \frac{6\lambda}{(\lambda - 0)^4} \\
&= \frac{6\lambda}{\lambda^4} \\
&= \frac{6}{\lambda^3} \; .
\end{aligned} \tag{12}
$$

We now substitute (12) into (6), giving

$$
\begin{aligned}
\mathrm{Skew}(X) &= \lambda^3 \cdot \mathrm{E}(X^3) - 4 \\
&= \lambda^3 \cdot \left(\frac{6}{\lambda^3} \right) - 4 \\
&= 6 - 4 \\
&= 2 \; .
\end{aligned} \tag{13}
$$

This completes the proof of (2).

■

3.6 Log-normal distribution

3.6.1 Definition

Definition: Let $\ln X$ be a random variable (→ I/1.2.2) following a normal distribution (→ II/3.2.1) with mean μ and variance σ^2 (or, standard deviation σ):

$$
Y = \ln(X) \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the exponential function of Y is said to have a log-normal distribution with location parameter μ and scale parameter σ

$$
X = \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2) \tag{2}
$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Sources:

- Wikipedia (2022): "Log-normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-02-07; URL: https://en.wikipedia.org/wiki/Log-normal_distribution.

3.6.2 Probability density function

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is given by:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \; . \tag{2}$$

Proof: A log-normally distributed random variable (→ II/3.6.1) is defined as the exponential function of a normal random variable (→ II/3.2.1):

$$Y \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad X = \exp(Y) \sim \ln \mathcal{N}(\mu, \sigma^2) \; . \tag{3}$$

The probability density function of the normal distribution (→ II/3.2.10) is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{\left(x - \mu\right)^2}{2\sigma^2}\right] \; . \tag{4}$$

Writing X as a function of Y, we have

$$X = g(Y) = \exp(Y) \tag{5}$$

with the inverse function

$$Y = g^{-1}(X) = \ln(X) \; . \tag{6}$$

Because the derivative of $\exp(Y)$ is always positive, $g(Y)$ is strictly increasing and we can calculate the probability density function of a strictly increasing function (→ I/1.7.3) as

$$f_X(x) = \left\{ \begin{array}{rl} f_Y(g^{-1}(x)) \, \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x} \; , & \text{if} \; x \in \mathcal{X} \\ 0 \; , & \text{if} \; x \notin \mathcal{X} \end{array} \right. \tag{7}$$

where $\mathcal{X} = \{x = g(y) : y \in \mathcal{Y}\}$. With the probability density function of the normal distribution (→ II/3.2.10), we have

$$
\begin{aligned}
f_X(x) &= f_Y(g^{-1}(x)) \cdot \frac{dg^{-1}(x)}{dx} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{g^{-1}(x)-\mu}{\sigma}\right)^2\right] \cdot \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{(\ln x)-\mu}{\sigma}\right)^2\right] \cdot \frac{\mathrm{d}(\ln x)}{\mathrm{d}x} \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right] \cdot \frac{1}{x} \\
&= \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]
\end{aligned} \tag{8}
$$

which is the probability density function (→ I/1.7.1) of the log-normal distribution (→ II/3.6.1).

■

Sources:

- Taboga, Marco (2021): "Log-normal distribution"; in: *Lectures on probability and statistics*, retrieved on 2022-02-13; URL: https://www.statlect.com/probability-distributions/log-normal-distribution.

3.6.3 Cumulative distribution function

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$
X \sim \ln\mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$
F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma}\right)\right] \tag{2}
$$

where $\mathrm{erf}(x)$ is the error function defined as

$$
\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, \mathrm{d}t \; . \tag{3}
$$

Proof: The probability density function of the log-normal distribution (→ II/3.6.2) is:

$$
f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\left[-\left(\frac{\ln x - \mu}{\sqrt{2}\sigma}\right)^2\right] \; . \tag{4}
$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$
\begin{split}
F_X(x) &= \int_{-\infty}^{x} \ln\mathcal{N}(z; \mu, \sigma^2) \, \mathrm{d}z \\
&= \int_{-\infty}^{x} \frac{1}{z\sigma\sqrt{2\pi}} \cdot \exp\left[-\left(\frac{\ln z - \mu}{\sqrt{2}\sigma} \right)^2 \right] \, \mathrm{d}z \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \frac{1}{z} \cdot \exp\left[-\left(\frac{\ln z - \mu}{\sqrt{2}\sigma} \right)^2 \right] \, \mathrm{d}z \; .
\end{split} \tag{5}
$$

From this point forward, the proof is similar to the derivation of the cumulative distribution function for the normal distribution (→ II/3.2.12). Substituting $t = (\ln z - \mu)/(\sqrt{2}\sigma)$, i.e. $\ln z = \sqrt{2}\sigma t + \mu$, $z = \exp(\sqrt{2}\sigma t + \mu)$ this becomes:

$$
\begin{split}
F_X(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(\ln x-\mu)/(\sqrt{2}\sigma)} \frac{1}{\exp(\sqrt{2}\sigma t + \mu)} \cdot \exp\left(-t^2\right) \, \mathrm{d}\left[\exp\left(\sqrt{2}\sigma t + \mu \right) \right] \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{\ln x-\mu}{\sqrt{2}\sigma}} \frac{1}{\exp(\sqrt{2}\sigma t + \mu)} \cdot \exp(-t^2) \cdot \exp\left(\sqrt{2}\sigma t + \mu \right) \, \mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{\ln x-\mu}{\sqrt{2}\sigma}} \exp(-t^2) \, \mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} \exp(-t^2) \, \mathrm{d}t + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{\ln x-\mu}{\sqrt{2}\sigma}} \exp(-t^2) \, \mathrm{d}t \\
&= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \exp(-t^2) \, \mathrm{d}t + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{\ln x-\mu}{\sqrt{2}\sigma}} \exp(-t^2) \, \mathrm{d}t \; .
\end{split} \tag{6}
$$

Applying (3) to (6), we have:

$$
\begin{split}
F_X(x) &= \frac{1}{2} \lim_{x \to \infty} \mathrm{erf}(x) + \frac{1}{2} \, \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \\
&= \frac{1}{2} + \frac{1}{2} \, \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \\
&= \frac{1}{2} \left[1 + \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma} \right) \right] \; .
\end{split} \tag{7}
$$

■

Sources:

- skdhfgeq2134 (2015): "How to derive the cdf of a lognormal distribution from its pdf"; in: *StackExchange*, retrieved on 2022-06-29; URL: https://stats.stackexchange.com/questions/151398/how-to-derive-t 151404#151404.

3.6.4 Quantile function

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$
X \sim \ln\mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the quantile function (→ I/1.9.1) of X is

$$Q_X(p) = \exp(\mu + \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p-1)) \tag{2}$$

where $\mathrm{erf}^{-1}(x)$ is the inverse error function.

Proof: The cumulative distribution function of the log-normal distribution (→ II/3.6.3) is:

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma}\right)\right] \; . \tag{3}$$

From this point forward, the proof is similar to the derivation of the quantile function for the normal distribution (→ II/3.2.15). Because the cumulative distribution function (CDF) is strictly monotonically increasing, the quantile function is equal to the inverse of the CDF (→ I/1.9.2):

$$Q_X(p) = F_X^{-1}(x) \; . \tag{4}$$

This can be derived by rearranging equation (3):

$$\begin{aligned}
p &= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma}\right)\right] \\
2p - 1 &= \mathrm{erf}\left(\frac{\ln x - \mu}{\sqrt{2}\sigma}\right) \\
\mathrm{erf}^{-1}(2p-1) &= \frac{\ln x - \mu}{\sqrt{2}\sigma} \\
x &= \exp(\mu + \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(2p-1)) \; .
\end{aligned} \tag{5}$$

■

Sources:
- Wikipedia (2022): "Log-normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-07-08; URL: https://en.wikipedia.org/wiki/Log-normal_distribution#Mode,_median,_quantiles.

3.6.5 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$X \sim \ln\mathcal{N}(\mu, \sigma^2) \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \tag{3}$$

With the probability density function of the log-normal distribution (→ II/3.6.2), this is:

$$
\begin{aligned}
\mathrm{E}(X) &= \int_0^{+\infty} x \cdot \frac{1}{x\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\frac{(\ln x-\mu)^2}{\sigma^2}\right] \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{+\infty} \exp\left[-\frac{1}{2}\frac{(\ln x-\mu)^2}{\sigma^2}\right] \mathrm{d}x
\end{aligned} \tag{4}
$$

Substituting $z = \frac{\ln x - \mu}{\sigma}$, i.e. $x = \exp(\mu + \sigma z)$, we have:

$$
\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{(-\infty-\mu)/(\sigma)}^{(\ln x-\mu)/(\sigma)} \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}\left[\exp(\mu+\sigma z)\right] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}z^2\right) \sigma \exp(\mu+\sigma z) \,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}z^2 + \sigma z + \mu\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(z^2 - 2\sigma z - 2\mu\right)\right] \mathrm{d}z
\end{aligned} \tag{5}
$$

Now multiplying $\exp\left(\frac{1}{2}\sigma^2\right)$ and $\exp\left(-\frac{1}{2}\sigma^2\right)$, we have:

$$
\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(z^2 - 2\sigma z + \sigma^2 - 2\mu - \sigma^2\right)\right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(z^2 - 2\sigma z + \sigma^2\right)\right] \exp\left(\mu + \frac{1}{2}\sigma^2\right) \mathrm{d}z \\
&= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z-\sigma)^2\right] \mathrm{d}z
\end{aligned} \tag{6}
$$

The probability density function of a normal distribution (→ II/3.2.10) is given by

$$
f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{7}
$$

and, with unit variance $\sigma^2 = 1$, this reads:

$$
f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^2\right] \tag{8}
$$

Using the definition of the probability density function (→ I/1.7.1), we get

$$
\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^2\right] \mathrm{d}x = 1 \tag{9}
$$

and applying (9) to (6), we have:

$$
\mathrm{E}(X) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) . \tag{10}
$$

■

Sources:

- Taboga, Marco (2022): "Log-normal distribution"; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2022-10-01; URL: https://www.statlect.com/probability-distributions/log-normal-distribution.

3.6.6 Median

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the median (→ I/1.15.1) of X is

$$\mathrm{median}(X) = e^{\mu} \; . \tag{2}$$

Proof: The median (→ I/1.15.1) is the value at which the cumulative distribution function is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the lognormal distribution (→ II/3.6.3) is

$$F_X(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right)\right] \tag{4}$$

where $\mathrm{erf}(x)$ is the error function defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, \mathrm{d}t \; . \tag{5}$$

Thus, the inverse CDF is

$$\begin{aligned} \ln(x) &= \sigma\sqrt{2} \cdot \mathrm{erf}^{-1}(2p-1) + \mu \\ x &= \exp\left[\sigma\sqrt{2} \cdot \mathrm{erf}^{-1}(2p-1) + \mu\right] \end{aligned} \tag{6}$$

where $\mathrm{erf}^{-1}(x)$ is the inverse error function. Setting $p = 1/2$, we obtain:

$$\begin{aligned} \ln\left[\mathrm{median}(X)\right] &= \sigma\sqrt{2} \cdot \mathrm{erf}^{-1}(0) + \mu \\ \mathrm{median}(X) &= e^{\mu} \; . \end{aligned} \tag{7}$$

■

3.6.7 Mode

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$X \sim \ln \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the mode (→ I/1.15.2) of X is

$$\mathrm{mode}(X) = e^{\left(\mu-\sigma^2\right)} \; . \tag{2}$$

Proof: The mode (→ I/1.15.2) is the value which maximizes the probability density function (→ I/1.7.1):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the log-normal distribution (→ II/3.6.2) is:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \; . \tag{4}$$

The first two derivatives of this function are:

$$f'_X(x) = -\frac{1}{x^2\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \tag{5}$$

$$\begin{split} f''_X(x) = &\frac{1}{\sqrt{2\pi}\sigma^2 x^3}\exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \cdot (\ln x - \mu) \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \\ &+ \frac{\sqrt{2}}{\sqrt{\pi}x^3}\exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \\ &- \frac{1}{\sqrt{2\pi}\sigma^2 x^3}\exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \; . \end{split} \tag{6}$$

We now calculate the root of the first derivative (5):

$$\begin{split} f'_X(x) = 0 &= -\frac{1}{x^2\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{\left(\ln x - \mu\right)^2}{2\sigma^2}\right] \cdot \left(1 + \frac{\ln x - \mu}{\sigma^2}\right) \\ -1 &= \frac{\ln x - \mu}{\sigma^2} \\ x &= e^{\left(\mu-\sigma^2\right)} \; . \end{split} \tag{7}$$

By plugging this value into the second derivative (6),

$$\begin{split} f''_X(e^{(\mu-\sigma^2)}) = &\frac{1}{\sqrt{2\pi}\sigma^2 (e^{(\mu-\sigma^2)})^3}\exp\left[-\frac{\sigma^2}{2}\right] \cdot \left(\sigma^2\right) \cdot (0) \\ &+ \frac{\sqrt{2}}{\sqrt{\pi}(e^{(\mu-\sigma^2)})^3}\exp\left[-\frac{\sigma^2}{2}\right] \cdot (0) \\ &- \frac{1}{\sqrt{2\pi}\sigma^2 (e^{(\mu-\sigma^2)})^3}\exp\left[-\frac{\sigma^2}{2}\right] \\ = &-\frac{1}{\sqrt{2\pi}\sigma^2 (e^{(\mu-\sigma^2)})^3}\exp\left[-\frac{\sigma^2}{2}\right] < 0 \; , \end{split} \tag{8}$$

we confirm that it is a maximum, showing that

$$\mathrm{mode}(X) = e^{\left(\mu-\sigma^2\right)} \; . \tag{9}$$

■

Sources:

- Wikipedia (2022): "Log-normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-02-12; URL: https://en.wikipedia.org/wiki/Log-normal_distribution#Mode.
- Mdoc (2015): "Mode of lognormal distribution"; in: *Mathematics Stack Exchange*, retrieved on 2022-02-12; URL: https://math.stackexchange.com/questions/1321221/mode-of-lognormal-distribution/1321626.

3.6.8 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a log-normal distribution (→ II/3.6.1):

$$X \sim \ln\mathcal{N}(\mu, \sigma^2). \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = \exp\left(2\mu + 2\sigma^2\right) - \exp\left(2\mu + \sigma^2\right) \; . \tag{2}$$

Proof: The variance (→ I/1.11.1) of a random variable is defined as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \tag{3}$$

which, partitioned into expected values (→ I/1.11.3), reads:

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 \; . \tag{4}$$

The expected value of the log-normal distribution (→ II/3.6.5) is:

$$\mathrm{E}[X] = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \tag{5}$$

The second moment $\mathrm{E}[X^2]$ can be derived as follows:

$$\begin{split} \mathrm{E}[X^2] &= \int_{-\infty}^{+\infty} x^2 \cdot f_\mathrm{X}(x) \, \mathrm{d}x \\ &= \int_0^{+\infty} x^2 \cdot \frac{1}{x\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\frac{\left(\ln x - \mu\right)^2}{\sigma^2}\right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{+\infty} x \cdot \exp\left[-\frac{1}{2}\frac{\left(\ln x - \mu\right)^2}{\sigma^2}\right] \mathrm{d}x \end{split} \tag{6}$$

Substituting $z = \frac{\ln x - \mu}{\sigma}$, i.e. $x = \exp\left(\mu + \sigma z\right)$, we have:

$$
\begin{aligned}
\mathrm{E}[X^2] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{(-\infty-\mu)/(\sigma)}^{(\ln x-\mu)/(\sigma)} \exp\left(\mu+\sigma z\right) \exp\left(-\frac{1}{2}z^2\right) \mathrm{d}\left[\exp\left(\mu+\sigma z\right)\right] \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}z^2\right) \sigma \exp\left(2\mu+2\sigma z\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(z^2-4\sigma z-4\mu\right)\right] \mathrm{d}z
\end{aligned} \tag{7}
$$

Now multiplying by $\exp\left(2\sigma^2\right)$ and $\exp\left(-2\sigma^2\right)$, this becomes:

$$
\begin{aligned}
\mathrm{E}[X^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(z^2-4\sigma z+4\sigma^2-4\sigma^2-4\mu\right)\right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(z^2-4\sigma z+4\sigma^2\right)\right] \exp\left(2\sigma^2+2\mu\right) \mathrm{d}z \\
&= \exp\left(2\sigma^2+2\mu\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(z-2\sigma\right)^2\right] \mathrm{d}z
\end{aligned} \tag{8}
$$

The probability density function of a normal distribution (→ II/3.2.10) is given by

$$
f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{9}
$$

and, with $\mu = 2\sigma$ and unit variance, this reads:

$$
f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(x-2\sigma\right)^2\right] \; . \tag{10}
$$

Using the definition of the probability density function (→ I/1.7.1), we get

$$
\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}\left(x-2\sigma\right)^2\right] \mathrm{d}x = 1 \tag{11}
$$

and applying (11) to (8), we have:

$$
\mathrm{E}[X]^2 = \exp\left(2\sigma^2+2\mu\right) \; . \tag{12}
$$

Finally, plugging (12) and (5) into (4), we have:

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 \\
&= \exp\left(2\sigma^2+2\mu\right) - \left[\exp\left(\mu+\frac{1}{2}\sigma^2\right)\right]^2 \\
&= \exp\left(2\sigma^2+2\mu\right) - \exp\left(2\mu+\sigma^2\right) \; .
\end{aligned} \tag{13}
$$

■

Sources:

- Taboga, Marco (2022): "Log-normal distribution"; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2022-10-01; URL: https://www.statlect.com/probability-distributions/log-normal-distribution.
- Wikipedia (2022): "Variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-01; URL: https://en.wikipedia.org/wiki/Variance#Definition.

3.7 Chi-squared distribution

3.7.1 Definition

Definition: Let $X_1, ..., X_k$ be independent (→ I/1.3.6) random variables (→ I/1.2.2) where each of them is following a standard normal distribution (→ II/3.2.3):

$$X_i \sim \mathcal{N}(0,1) \quad \text{for} \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the sum of their squares follows a chi-squared distribution with k degrees of freedom:

$$Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k) \quad \text{where} \quad k > 0 \; . \tag{2}$$

The probability density function of the chi-squared distribution (→ II/3.7.3) with k degrees of freedom is

$$\chi^2(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} \, x^{k/2-1} \, e^{-x/2} \tag{3}$$

where $k > 0$ and the density is zero if $x \leq 0$.

Sources:

- Wikipedia (2020): "Chi-square distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-10-12; URL: https://en.wikipedia.org/wiki/Chi-square_distribution#Definitions.
- Robert V. Hogg, Joseph W. McKean, Allen T. Craig (2018): "The Chi-Squared-Distribution"; in: *Introduction to Mathematical Statistics*, Pearson, Boston, 2019, p. 178, eq. 3.3.7; URL: https://www.pearson.com/store/p/introduction-to-mathematical-statistics/P100000843744.

3.7.2 Special case of gamma distribution

Theorem: The chi-squared distribution (→ II/3.7.1) with k degrees of freedom is a special case of the gamma distribution (→ II/3.4.1) with shape $\frac{k}{2}$ and rate $\frac{1}{2}$:

$$X \sim \mathrm{Gam}\left(\frac{k}{2}, \frac{1}{2}\right) \Rightarrow X \sim \chi^2(k) \; . \tag{1}$$

Proof: The probability density function of the gamma distribution (→ II/3.4.7) for $x > 0$, where α is the shape parameter and β is the rate parameter, is as follows:

$$\mathrm{Gam}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, x^{\alpha-1} \, e^{-\beta x} \tag{2}$$

If we let $\alpha = k/2$ and $\beta = 1/2$, we obtain

$$\mathrm{Gam}\left(x; \frac{k}{2}, \frac{1}{2}\right) = \frac{x^{k/2-1} \, e^{-x/2}}{\Gamma(k/2) 2^{k/2}} = \frac{1}{2^{k/2}\Gamma(k/2)} \, x^{k/2-1} \, e^{-x/2} \tag{3}$$

which is equivalent to the probability density function of the chi-squared distribution (→ II/3.7.3).

■

3.7.3 Probability density function

Theorem: Let Y be a random variable (→ I/1.2.2) following a chi-squared distribution (→ II/3.7.1):

$$Y \sim \chi^2(k) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of Y is

$$f_Y(y) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\, y^{k/2-1}\, e^{-y/2} \; . \tag{2}$$

Proof: A chi-square-distributed random variable (→ II/3.7.1) with k degrees of freedom is defined as the sum of k squared standard normal random variables (→ II/3.2.3):

$$X_1, \ldots, X_k \sim \mathcal{N}(0,1) \quad \Rightarrow \quad Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k) \; . \tag{3}$$

Let $x_1, \ldots, x_k$ be values of $X_1, \ldots, X_k$ and consider $x = (x_1, \ldots, x_k)$ to be a point in k-dimensional space. Define

$$y = \sum_{i=1}^{k} x_i^2 \tag{4}$$

and let $f_Y(y)$ and $F_Y(y)$ be the probability density function (→ I/1.7.1) and cumulative distribution function (→ I/1.8.1) of Y. Because the PDF is the first derivative of the CDF (→ I/1.7.7), we can write:

$$F_Y(y) = \frac{F_Y(y)}{\mathrm{d}y}\,\mathrm{d}y = f_Y(y)\,\mathrm{d}y \; . \tag{5}$$

Then, the cumulative distribution function (→ I/1.8.1) of Y can be expressed as

$$f_Y(y)\,\mathrm{d}y = \int_V \prod_{i=1}^{k} \left(\mathcal{N}(x_i;0,1)\,\mathrm{d}x_i\right) \tag{6}$$

where $\mathcal{N}(x_i;0,1)$ is the probability density function (→ I/1.7.1) of the standard normal distribution (→ II/3.2.3) and V is the elemental shell volume at $y(x)$, which is proportional to the $(k-1)$-dimensional surface in k-space for which equation (4) is fulfilled. Using the probability density function of the normal distribution (→ II/3.2.10), equation (6) can be developed as follows:

$$\begin{split} f_Y(y)\,\mathrm{d}y &= \int_V \prod_{i=1}^{k} \left(\frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2}x_i^2\right]\,\mathrm{d}x_i \right) \\ &= \int_V \frac{\exp\left[-\frac{1}{2}\left(x_1^2 + \ldots + x_k^2\right)\right]}{(2\pi)^{k/2}}\,\mathrm{d}x_1 \ldots \mathrm{d}x_k \\ &= \frac{1}{(2\pi)^{k/2}} \int_V \exp\left[-\frac{y}{2}\right]\,\mathrm{d}x_1 \ldots \mathrm{d}x_k \; . \end{split} \tag{7}$$

Because y is constant within the set V, it can be moved out of the integral:

$$f_Y(y)\,\mathrm{d}y = \frac{\exp\left[-y/2\right]}{(2\pi)^{k/2}} \int_V \mathrm{d}x_1 \, \ldots \, \mathrm{d}x_k \; . \tag{8}$$

Now, the integral is simply the surface area of the $(k-1)$-dimensional sphere with radius $r = \sqrt{y}$, which is

$$A = 2r^{k-1} \, \frac{\pi^{k/2}}{\Gamma(k/2)} \; , \tag{9}$$

times the infinitesimal thickness of the sphere, which is

$$\frac{\mathrm{d}r}{\mathrm{d}y} = \frac{1}{2} y^{-1/2} \quad \Leftrightarrow \quad \mathrm{d}r = \frac{\mathrm{d}y}{2y^{1/2}} \; . \tag{10}$$

Substituting (9) and (10) into (8), we have:

$$\begin{split}
f_Y(y)\,\mathrm{d}y &= \frac{\exp\left[-y/2\right]}{(2\pi)^{k/2}} \cdot A \,\mathrm{d}r \\
&= \frac{\exp\left[-y/2\right]}{(2\pi)^{k/2}} \cdot 2r^{k-1} \, \frac{\pi^{k/2}}{\Gamma(k/2)} \cdot \frac{\mathrm{d}y}{2y^{1/2}} \\
&= \frac{1}{2^{k/2}\,\Gamma(k/2)} \cdot \frac{2\sqrt{y}^{k-1}}{2\sqrt{y}} \cdot \exp\left[-y/2\right] \, \mathrm{d}y \\
&= \frac{1}{2^{k/2}\,\Gamma(k/2)} \cdot y^{\frac{k}{2}-1} \cdot \exp\left[-\frac{y}{2}\right] \, \mathrm{d}y \; .
\end{split} \tag{11}$$

From this, we get the final result in (2):

$$f_Y(y) = \frac{1}{2^{k/2}\,\Gamma(k/2)} \, y^{k/2-1} \, e^{-y/2} \; . \tag{12}$$

■

Sources:

- Wikipedia (2020): "Proofs related to chi-squared distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/Proofs_related_to_chi-squared_distribution#Derivation_of_the_pdf_for_k_degrees_of_freedom.
- Wikipedia (2020): "n-sphere"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/N-sphere#Volume_and_surface_area.

3.7.4 Moments

Theorem: Let X be a random variable (→ I/1.2.2) following a chi-squared distribution (→ II/3.7.1) with k degrees of freedom:

$$X \sim \chi^2(k) \; . \tag{1}$$

Then, if $m > -k/2$, the moment $\mathrm{E}(X^m)$ exists and is equal to:

$$\mathrm{E}(X^m) = 2^m \frac{\Gamma\left(\frac{k}{2}+m\right)}{\Gamma\left(\frac{k}{2}\right)} \; . \tag{2}$$

Proof: Combining the definition of the raw moment (→ I/1.18.3) with the probability density function of the chi-squared distribution (→ II/3.7.3), we have:

$$
\begin{split}
\mathrm{E}(X^m) &= \int_0^\infty x^m \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} \, x^{k/2-1} \, e^{-x/2} \, \mathrm{d}x \\
&= \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} \int_0^\infty x^{(k/2)+m-1} \, e^{-x/2} \, \mathrm{d}x \; .
\end{split} \tag{3}
$$

Now, we substitute $u = x/2$, such that $x = 2u$. As a result, we obtain:

$$
\begin{split}
\mathrm{E}(X^m) &= \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} \int_0^\infty 2^{(k/2)+m-1} \, u^{(k/2)+m-1} \, e^{-u} \, \mathrm{d}(2u) \\
&= \frac{2^{(k/2)+m}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} \int_0^\infty u^{(k/2)+m-1} \, e^{-u} \, \mathrm{d}u \\
&= \frac{2^m}{\Gamma\left(\frac{k}{2}\right)} \int_0^\infty u^{(k/2)+m-1} \, e^{-u} \, \mathrm{d}u \; .
\end{split} \tag{4}
$$

With the definition of the gamma function as

$$
\Gamma(x) = \int_0^\infty t^{x-1} \, e^{-t} \, \mathrm{d}t, \; z > 0 \; , \tag{5}
$$

this leads to the desired result when $m > -k/2$. Observe that, if m is a nonnegative integer, then $m > -k/2$ is always true. Therefore, all moments (→ I/1.18.1) of a chi-squared distribution (→ II/3.7.1) exist and the m-th raw moment is given by the equation above.

■

Sources:
- Robert V. Hogg, Joseph W. McKean, Allen T. Craig (2018): "The 2-Distribution"; in: *Introduction to Mathematical Statistics*, Pearson, Boston, 2019, p. 179, eq. 3.3.8; URL: https://www.pearson.com/store/p/introduction-to-mathematical-statistics/P100000843744.

3.8 F-distribution

3.8.1 Definition

Definition: Let X_1 and X_2 be independent (→ I/1.3.6) random variables (→ I/1.2.2) following a chi-squared distribution (→ II/3.7.1) with d_1 and d_2 degrees of freedom, respectively:

$$
\begin{split}
X_1 &\sim \chi^2(d_1) \\
X_2 &\sim \chi^2(d_2) \; .
\end{split} \tag{1}
$$

Then, the ratio of X_1 to X_2, divided by their respective degrees of freedom, is said to be F-distributed with numerator degrees of freedom d_1 and denominator degrees of freedom d_2:

$$
Y = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2) \quad \text{where} \quad d_1, d_2 > 0 \; . \tag{2}
$$

The F-distribution is also called "Snedecor's F-distribution" or "Fisher–Snedecor distribution", after Ronald A. Fisher and George W. Snedecor.

Sources:
- Wikipedia (2021): "F-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-04-21; URL: https://en.wikipedia.org/wiki/F-distribution#Characterization.

3.8.2 Probability density function

Theorem: Let F be a random variable (→ I/1.2.2) following an F-distribution (→ II/3.8.1):

$$F \sim \mathrm{F}(u, v) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of F is

$$f_F(f) = \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right)} \cdot \left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1} \cdot \left(\frac{u}{v} f + 1\right)^{-\frac{u+v}{2}} \; . \tag{2}$$

Proof: An F-distributed random variable (→ II/3.8.1) is defined as the ratio of two chi-squared random variables (→ II/3.7.1), divided by their degrees of freedom

$$X \sim \chi^2(u), \; Y \sim \chi^2(v) \quad \Rightarrow \quad F = \frac{X/u}{Y/v} \sim \mathrm{F}(u, v) \tag{3}$$

where X and Y are independent of each other (→ I/1.3.6).
The probability density function of the chi-squared distribution (→ II/3.7.3) is

$$f_X(x) = \frac{1}{\Gamma\left(\frac{u}{2}\right) \cdot 2^{u/2}} \cdot x^{\frac{u}{2}-1} \cdot e^{-\frac{x}{2}} \; . \tag{4}$$

Define the random variables F and W as functions of X and Y

$$\begin{aligned} F &= \frac{X/u}{Y/v} \\ W &= Y \; , \end{aligned} \tag{5}$$

such that the inverse functions X and Y in terms of F and W are

$$\begin{aligned} X &= \frac{u}{v} F W \\ Y &= W \; . \end{aligned} \tag{6}$$

This implies the following Jacobian matrix and determinant:

$$\begin{aligned} J &= \begin{bmatrix} \frac{\mathrm{d}X}{\mathrm{d}F} & \frac{\mathrm{d}X}{\mathrm{d}W} \\ \frac{\mathrm{d}Y}{\mathrm{d}F} & \frac{\mathrm{d}Y}{\mathrm{d}W} \end{bmatrix} = \begin{bmatrix} \frac{u}{v} W & \frac{u}{v} F \\ 0 & 1 \end{bmatrix} \\ |J| &= \frac{u}{v} W \; . \end{aligned} \tag{7}$$

Because X and Y are independent (→ I/1.3.6), the joint density (→ I/1.5.2) of X and Y is equal to the product (→ I/1.3.8) of the marginal densities (→ I/1.5.3):

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \; . \tag{8}$$

With the probability density function of an invertible function (→ I/1.7.5), the joint density (→ I/1.5.2) of F and W can be derived as:

$$f_{F,W}(f,w) = f_{X,Y}(x,y) \cdot |J| \; . \tag{9}$$

Substituting (6) into (4), and then with (7) into (9), we get:

$$\begin{split} f_{F,W}(f,w) &= f_X\left(\frac{u}{v}fw\right) \cdot f_Y(w) \cdot |J| \\ &= \frac{1}{\Gamma\left(\frac{u}{2}\right) \cdot 2^{u/2}} \cdot \left(\frac{u}{v}fw\right)^{\frac{u}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{u}{v}fw\right)} \cdot \frac{1}{\Gamma\left(\frac{v}{2}\right) \cdot 2^{v/2}} \cdot w^{\frac{v}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \frac{u}{v}w \\ &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot w^{\frac{u+v}{2}-1} \cdot e^{-\frac{w}{2}\left(\frac{u}{v}f+1\right)} \; . \end{split} \tag{10}$$

The marginal density (→ I/1.5.3) of F can now be obtained by integrating out (→ I/1.3.3) W:

$$\begin{split} f_F(f) &= \int_0^\infty f_{F,W}(f,w) \, \mathrm{d}w \\ &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot \int_0^\infty w^{\frac{u+v}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{u}{v}f+1\right)w\right] \mathrm{d}w \\ &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot \frac{\Gamma\left(\frac{u+v}{2}\right)}{\left[\frac{1}{2}\left(\frac{u}{v}f+1\right)\right]^{(u+v)/2}} \cdot \int_0^\infty \frac{\left[\frac{1}{2}\left(\frac{u}{v}f+1\right)\right]^{(u+v)/2}}{\Gamma\left(\frac{u+v}{2}\right)} \cdot w^{\frac{u+v}{2}-1} \cdot \exp\left[-\frac{1}{2}\left(\frac{u}{v}f+1\right)w\right] \end{split} \tag{11}$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution (→ II/3.4.7) with

$$a = \frac{u+v}{2} \quad \text{and} \quad b = \frac{1}{2}\left(\frac{u}{v}f+1\right) \; , \tag{12}$$

and because a probability density function integrates to one (→ I/1.7.1), we finally have:

$$\begin{split} f_F(f) &= \frac{\left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right) \cdot 2^{(u+v)/2}} \cdot \frac{\Gamma\left(\frac{u+v}{2}\right)}{\left[\frac{1}{2}\left(\frac{u}{v}f+1\right)\right]^{(u+v)/2}} \\ &= \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right) \cdot \Gamma\left(\frac{v}{2}\right)} \cdot \left(\frac{u}{v}\right)^{\frac{u}{2}} \cdot f^{\frac{u}{2}-1} \cdot \left(\frac{u}{v}f+1\right)^{-\frac{u+v}{2}} \; . \end{split} \tag{13}$$

■

Sources:

- statisticsmatt (2018): "Statistical Distributions: Derive the F Distribution"; in: *YouTube*, retrieved on 2021-10-11; URL: https://www.youtube.com/watch?v=AmHiOKYmHkI.

3.9 Beta distribution

3.9.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to follow a beta distribution with shape parameters α and β

$$X \sim \mathrm{Bet}(\alpha, \beta) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathrm{Bet}(x; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \tag{2}$$

where $\alpha > 0$ and $\beta > 0$, and the density is zero, if $x \notin [0, 1]$.

Sources:
- Wikipedia (2020): "Beta distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Beta_distribution#Definitions.

3.9.2 Relationship to chi-squared distribution

Theorem: Let X and Y be independent (→ I/1.3.6) random variables (→ I/1.2.2) following chi-squared distributions (→ II/3.7.1):

$$X \sim \chi^2(m) \quad \text{and} \quad Y \sim \chi^2(n) \; . \tag{1}$$

Then, the quantity $X/(X+Y)$ follows a beta distribution (→ II/3.9.1):

$$\frac{X}{X+Y} \sim \mathrm{Bet}\left(\frac{m}{2}, \frac{n}{2}\right) \; . \tag{2}$$

Proof: The probability density function of the chi-squared distribution (→ II/3.7.3) is

$$X \sim \chi^2(u) \quad \Rightarrow \quad f_X(x) = \frac{1}{\Gamma\left(\frac{u}{2}\right) \cdot 2^{u/2}} \cdot x^{\frac{u}{2}-1} \cdot e^{-\frac{x}{2}} \; . \tag{3}$$

Define the random variables Z and W as functions of X and Y

$$\begin{aligned} Z &= \frac{X}{X+Y} \\ W &= Y \; , \end{aligned} \tag{4}$$

such that the inverse functions X and Y in terms of Z and W are

$$\begin{aligned} X &= \frac{ZW}{1-Z} \\ Y &= W \; . \end{aligned} \tag{5}$$

This implies the following Jacobian matrix and determinant:

$$
\begin{aligned}
J &= \begin{bmatrix} \frac{\mathrm{d}X}{\mathrm{d}Z} & \frac{\mathrm{d}X}{\mathrm{d}W} \\ \frac{\mathrm{d}Y}{\mathrm{d}Z} & \frac{\mathrm{d}Y}{\mathrm{d}W} \end{bmatrix} = \begin{bmatrix} \frac{W}{(1-Z)^2} & \frac{Z}{1-Z} \\ 0 & 1 \end{bmatrix} \\
|J| &= \frac{W}{(1-Z)^2} \; .
\end{aligned} \tag{6}
$$

Because X and Y are independent (→ I/1.3.6), the joint density (→ I/1.5.2) of X and Y is equal to the product (→ I/1.3.8) of the marginal densities (→ I/1.5.3):

$$
f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \; . \tag{7}
$$

With the probability density function of an invertible function (→ I/1.7.5), the joint density (→ I/1.5.2) of Z and W can be derived as:

$$
f_{Z,W}(z,w) = f_{X,Y}(x,y) \cdot |J| \; . \tag{8}
$$

Substituting (5) into (3), and then with (6) into (8), we get:

$$
\begin{aligned}
f_{Z,W}(z,w) &= f_X\left(\frac{zw}{1-z}\right) \cdot f_Y(w) \cdot |J| \\
&= \frac{1}{\Gamma\left(\frac{m}{2}\right) \cdot 2^{m/2}} \cdot \left(\frac{zw}{1-z}\right)^{\frac{m}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{zw}{1-z}\right)} \cdot \frac{1}{\Gamma\left(\frac{n}{2}\right) \cdot 2^{n/2}} \cdot w^{\frac{n}{2}-1} \cdot e^{-\frac{w}{2}} \cdot \frac{w}{(1-z)^2} \\
&= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) \cdot 2^{m/2}2^{n/2}} \cdot \left(\frac{z}{1-z}\right)^{\frac{m}{2}-1} \left(\frac{1}{(1-z)}\right)^{2} \cdot w^{\frac{m}{2}+\frac{n}{2}-1} e^{-\frac{1}{2}\left(\frac{zw}{1-z}+\frac{w(1-z)}{1-z}\right)} \\
&= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot w^{\frac{m+n}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{w}{1-z}\right)} \; .
\end{aligned} \tag{9}
$$

The marginal density (→ I/1.5.3) of Z can now be obtained by integrating out (→ I/1.3.3) W:

$$
\begin{aligned}
f_Z(z) &= \int_0^\infty f_{Z,W}(z,w) \, \mathrm{d}w \\
&= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot \int_0^\infty w^{\frac{m+n}{2}-1} \cdot e^{-\frac{1}{2}\left(\frac{w}{1-z}\right)} \, \mathrm{d}w \\
&= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right) \cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot \frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{1}{2(1-z)}\right)^{\frac{m+n}{2}}} \cdot \\
&\quad \int_0^\infty \frac{\left(\frac{1}{2(1-z)}\right)^{\frac{m+n}{2}}}{\Gamma\left(\frac{m+n}{2}\right)} \cdot w^{\frac{m+n}{2}-1} \cdot e^{-\frac{1}{2(1-z)} w} \, \mathrm{d}w \; .
\end{aligned} \tag{10}
$$

At this point, we can recognize that the integrand is equal to the probability density function of a gamma distribution (→ II/3.4.7) with

$$
a = \frac{m+n}{2} \quad \text{and} \quad b = \frac{1}{2(1-z)} \; , \tag{11}
$$

and because a probability density function integrates to one (→ I/1.7.1), we have:

$$
\begin{aligned}
f_Z(z) &= \frac{1}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}-1} \cdot \frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{1}{2(1-z)}\right)^{\frac{m+n}{2}}} \\
&= \frac{\Gamma\left(\frac{m+n}{2}\right)\cdot 2^{(m+n)/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\cdot 2^{(m+n)/2}} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{-\frac{m}{2}+\frac{m+n}{2}-1} \\
&= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{\frac{n}{2}-1} \; .
\end{aligned} \tag{12}
$$

With the definition of the beta function (→ II/3.9.6), this becomes

$$
f_Z(z) = \frac{1}{\mathrm{B}\left(\frac{m}{2},\frac{n}{2}\right)} \cdot z^{\frac{m}{2}-1} \cdot (1-z)^{\frac{n}{2}-1} \tag{13}
$$

which is the probability density function of the beta distribution (→ II/3.9.3) with parameters

$$
\alpha = \frac{m}{2} \quad \text{and} \quad \beta = \frac{n}{2} \; , \tag{14}
$$

such that

$$
Z \sim \mathrm{Bet}\left(\frac{m}{2},\frac{n}{2}\right) \; . \tag{15}
$$

■

Sources:

- Probability Fact (2021): "If X   chisq(m) and Y   chisq(n) are independent"; in: *Twitter*, retrieved on 2022-10-17; URL: https://twitter.com/ProbFact/status/1450492787854647300.

3.9.3 Probability density function

Theorem: Let X be a random variable (→ I/1.2.2) following a beta distribution (→ II/3.9.1):

$$
X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}
$$

Then, the probability density function (→ I/1.7.1) of X is

$$
f_X(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \; . \tag{2}
$$

Proof: This follows directly from the definition of the beta distribution (→ II/3.9.1).

■

3.9.4 Moment-generating function

Theorem: Let X be a positive random variable (→ I/1.2.2) following a beta distribution (→ II/3.4.1):

$$X \sim \mathrm{Bet}(\alpha, \beta) . \tag{1}$$

Then, the moment-generating function (→ I/1.9.5) of X is

$$M_X(t) = 1 + \sum_{n=1}^{\infty} \left(\prod_{m=0}^{n-1} \frac{\alpha+m}{\alpha+\beta+m} \right) \frac{t^n}{n!} . \tag{2}$$

Proof: The probability density function of the beta distribution (→ II/3.9.3) is

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \tag{3}$$

and the moment-generating function (→ I/1.9.5) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] . \tag{4}$$

Using the expected value for continuous random variables (→ I/1.10.1), the moment-generating function of X therefore is

$$\begin{aligned} M_X(t) &= \int_0^1 \exp[tx] \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \\ &= \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 e^{tx} \, x^{\alpha-1} \, (1-x)^{\beta-1} \, \mathrm{d}x . \end{aligned} \tag{5}$$

With the relationship between beta function and gamma function

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha) \, \Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{6}$$

and the integral representation of the confluent hypergeometric function (Kummer's function of the first kind)

$${}_1F_1(a, b, z) = \frac{\Gamma(b)}{\Gamma(a) \, \Gamma(b-a)} \int_0^1 e^{zu} \, u^{a-1} \, (1-u)^{(b-a)-1} \, \mathrm{d}u , \tag{7}$$

the moment-generating function can be written as

$$M_X(t) = {}_1F_1(\alpha, \alpha+\beta, t) . \tag{8}$$

Note that the series equation for the confluent hypergeometric function (Kummer's function of the first kind) is

$${}_1F_1(a, b, z) = \sum_{n=0}^{\infty} \frac{a^{\overline{n}}}{b^{\overline{n}}} \, \frac{z^n}{n!} \tag{9}$$

where $m^{\overline{n}}$ is the rising factorial

$$m^{\overline{n}} = \prod_{i=0}^{n-1} (m+i) \; , \tag{10}$$

so that the moment-generating function can be written as

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\alpha^{\overline{n}}}{(\alpha+\beta)^{\overline{n}}} \frac{t^n}{n!} \; . \tag{11}$$

Applying the rising factorial equation (10) and using $m^{\overline{0}} = x^0 = 0! = 1$, we finally have:

$$M_X(t) = 1 + \sum_{n=1}^{\infty} \left(\prod_{m=0}^{n-1} \frac{\alpha+m}{\alpha+\beta+m} \right) \frac{t^n}{n!} \; . \tag{12}$$

■

Sources:

- Wikipedia (2020): "Beta distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/Beta_distribution#Moment_generating_function.
- Wikipedia (2020): "Confluent hypergeometric function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/Confluent_hypergeometric_function#Kummer's_equation.

3.9.5 Cumulative distribution function

Theorem: Let X be a positive random variable (→ I/1.2.2) following a beta distribution (→ II/3.4.1):

$$X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}$$

Then, the cumulative distribution function (→ I/1.8.1) of X is

$$F_X(x) = \frac{B(x; \alpha, \beta)}{B(\alpha, \beta)} \tag{2}$$

where $B(a,b)$ is the beta function and $B(x;a,b)$ is the incomplete gamma function.

Proof: The probability density function of the beta distribution (→ II/3.9.3) is:

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \; . \tag{3}$$

Thus, the cumulative distribution function (→ I/1.8.1) is:

$$\begin{split} F_X(x) &= \int_0^x \mathrm{Bet}(z; \alpha, \beta) \, \mathrm{d}z \\ &= \int_0^x \frac{1}{\mathrm{B}(\alpha, \beta)} \, z^{\alpha-1} \, (1-z)^{\beta-1} \, \mathrm{d}z \\ &= \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^x z^{\alpha-1} \, (1-z)^{\beta-1} \, \mathrm{d}z \; . \end{split} \tag{4}$$

With the definition of the incomplete beta function

$$B(x;a,b) = \int_0^x t^{a-1} \, (1-t)^{b-1} \, \mathrm{d}t \; , \tag{5}$$

we arrive at the final result given by equation (2):

$$F_X(x) = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)} \; . \tag{6}$$

■

Sources:
- Wikipedia (2020): "Beta distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: https://en.wikipedia.org/wiki/Beta_distribution#Cumulative_distribution_function.
- Wikipedia (2020): "Beta function"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-19; URL: https://en.wikipedia.org/wiki/Beta_function#Incomplete_beta_function.

3.9.6 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following a beta distribution (→ II/3.9.1):

$$X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta} \; . \tag{2}$$

Proof: The expected value (→ I/1.10.1) is the probability-weighted average over all possible values:

$$\mathrm{E}(X) = \int_{\mathcal{X}} x \cdot f_X(x) \, \mathrm{d}x \; . \tag{3}$$

The probability density function of the beta distribution (→ II/3.9.3) is

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1}, \quad 0 \leq x \leq 1 \tag{4}$$

where the beta function is given by a ratio gamma functions:

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha+\beta)} \; . \tag{5}$$

Combining (3), (4) and (5), we have:

$$\begin{split} \mathrm{E}(X) &= \int_0^1 x \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1+\beta)} \int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1) \cdot \Gamma(\beta)} \, x^{(\alpha+1)-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \; . \end{split} \tag{6}$$

Employing the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$
\begin{split}
\mathrm{E}(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \cdot \frac{\alpha \cdot \Gamma(\alpha)}{(\alpha+\beta) \cdot \Gamma(\alpha+\beta)} \int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1) \cdot \Gamma(\beta)} \, x^{(\alpha+1)-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \\
&= \frac{\alpha}{\alpha+\beta} \int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1) \cdot \Gamma(\beta)} \, x^{(\alpha+1)-1} \, (1-x)^{\beta-1} \, \mathrm{d}x
\end{split} \tag{7}
$$

and again using the density of the beta distribution (→ II/3.9.3), we get

$$
\begin{split}
\mathrm{E}(X) &= \frac{\alpha}{\alpha+\beta} \int_0^1 \mathrm{Bet}(x; \alpha+1, \beta) \, \mathrm{d}x \\
&= \frac{\alpha}{\alpha+\beta} \; .
\end{split} \tag{8}
$$

■

Sources:
- Boer Commander (2020): "Beta Distribution Mean and Variance Proof"; in: *YouTube*, retrieved on 2021-04-29; URL: https://www.youtube.com/watch?v=3OgCcnpZtZ8.

3.9.7 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following a beta distribution (→ II/3.9.1):

$$
X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}
$$

Then, the variance (→ I/1.11.1) of X is

$$
\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1) \cdot (\alpha+\beta)^2} \; . \tag{2}
$$

Proof: The variance (→ I/1.11.1) can be expressed in terms of expected values (→ I/1.11.3) as

$$
\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; . \tag{3}
$$

The expected value of a beta random variable (→ II/3.9.6) is

$$
\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta} \; . \tag{4}
$$

The probability density function of the beta distribution (→ II/3.9.3) is

$$
f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1}, \quad 0 \leq x \leq 1 \tag{5}
$$

where the beta function is given by a ratio gamma functions:

$$
\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha+\beta)} \; . \tag{6}
$$

Therefore, the expected value of a squared beta random variable becomes

$$
\begin{split}
\mathrm{E}(X^2) &= \int_0^1 x^2 \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+2)}{\Gamma(\alpha+2+\beta)} \int_0^1 \frac{\Gamma(\alpha+2+\beta)}{\Gamma(\alpha+2)\cdot\Gamma(\beta)} \, x^{(\alpha+2)-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \; .
\end{split} \tag{7}
$$

Twice-applying the relation $\Gamma(x+1) = \Gamma(x) \cdot x$, we have

$$
\begin{split}
\mathrm{E}(X^2) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \cdot \frac{(\alpha+1)\cdot\alpha\cdot\Gamma(\alpha)}{(\alpha+\beta+1)\cdot(\alpha+\beta)\cdot\Gamma(\alpha+\beta)} \int_0^1 \frac{\Gamma(\alpha+2+\beta)}{\Gamma(\alpha+2)\cdot\Gamma(\beta)} \, x^{(\alpha+2)-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \\
&= \frac{(\alpha+1)\cdot\alpha}{(\alpha+\beta+1)\cdot(\alpha+\beta)} \int_0^1 \frac{\Gamma(\alpha+2+\beta)}{\Gamma(\alpha+2)\cdot\Gamma(\beta)} \, x^{(\alpha+2)-1} \, (1-x)^{\beta-1} \, \mathrm{d}x
\end{split} \tag{8}
$$

and again using the density of the beta distribution (→ II/3.9.3), we get

$$
\begin{split}
\mathrm{E}(X^2) &= \frac{(\alpha+1)\cdot\alpha}{(\alpha+\beta+1)\cdot(\alpha+\beta)} \int_0^1 \mathrm{Bet}(x;\alpha+2,\beta) \, \mathrm{d}x \\
&= \frac{(\alpha+1)\cdot\alpha}{(\alpha+\beta+1)\cdot(\alpha+\beta)} \; .
\end{split} \tag{9}
$$

Plugging (9) and (4) into (3), the variance of a beta random variable finally becomes

$$
\begin{split}
\mathrm{Var}(X) &= \frac{(\alpha+1)\cdot\alpha}{(\alpha+\beta+1)\cdot(\alpha+\beta)} - \left(\frac{\alpha}{\alpha+\beta} \right)^2 \\
&= \frac{(\alpha^2+\alpha)\cdot(\alpha+\beta)}{(\alpha+\beta+1)\cdot(\alpha+\beta)^2} - \frac{\alpha^2\cdot(\alpha+\beta+1)}{(\alpha+\beta+1)\cdot(\alpha+\beta)^2} \\
&= \frac{(\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta) - (\alpha^3+\alpha^2\beta+\alpha^2)}{(\alpha+\beta+1)\cdot(\alpha+\beta)^2} \\
&= \frac{\alpha\beta}{(\alpha+\beta+1)\cdot(\alpha+\beta)^2} \; .
\end{split} \tag{10}
$$

■

Sources:

- Boer Commander (2020): "Beta Distribution Mean and Variance Proof"; in: *YouTube*, retrieved on 2021-04-29; URL: https://www.youtube.com/watch?v=3OgCcnpZtZ8.

3.10 Wald distribution

3.10.1 Definition

Definition: Let X be a random variable (→ I/1.2.2). Then, X is said to follow a Wald distribution with drift rate γ and threshold α

$$
X \sim \mathrm{Wald}(\gamma, \alpha) \; , \tag{1}
$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathrm{Wald}(x;\gamma,\alpha) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha-\gamma x)^2}{2x}\right) \tag{2}$$

where $\gamma > 0$, $\alpha > 0$, and the density is zero if $x \leq 0$.

Sources:

- Anders, R., Alario, F.-X., and van Maanen, L. (2016): "The Shifted Wald Distribution for Response Time Data Analysis"; in: *Psychological Methods*, vol. 21, no. 3, pp. 309-327; URL: https://dx.doi.org/10.1037/met0000066; DOI: 10.1037/met0000066.

3.10.2 Probability density function

Theorem: Let X be a positive random variable (→ I/1.2.2) following a Wald distribution (→ II/3.10.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha-\gamma x)^2}{2x}\right) \; . \tag{2}$$

Proof: This follows directly from the definition of the Wald distribution (→ II/3.10.1).

■

3.10.3 Moment-generating function

Theorem: Let X be a positive random variable (→ I/1.2.2) following a Wald distribution (→ II/3.10.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}$$

Then, the moment-generating function (→ I/1.9.5) of X is

$$M_X(t) = \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \; . \tag{2}$$

Proof: The probability density function of the Wald distribution (→ II/3.10.2) is

$$f_X(x) = \frac{\alpha}{\sqrt{2\pi x^3}} \exp\left(-\frac{(\alpha-\gamma x)^2}{2x}\right) \tag{3}$$

and the moment-generating function (→ I/1.9.5) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] \; . \tag{4}$$

Using the definition of expected value for continuous random variables (→ I/1.10.1), the moment-generating function of X therefore is

$$
\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx} \cdot \frac{\alpha}{\sqrt{2\pi x^3}} \cdot \exp\left[-\frac{(\alpha-\gamma x)^2}{2x}\right] dx \\
&= \frac{\alpha}{\sqrt{2\pi}} \int_0^\infty x^{-3/2} \cdot \exp\left[tx - \frac{(\alpha-\gamma x)^2}{2x}\right] dx \; .
\end{aligned} \tag{5}
$$

To evaluate this integral, we will need two identities about modified Bessel functions of the second kind[1], denoted K_p. The function K_p (for $p \in \mathbb{R}$) is one of the two linearly independent solutions of the differential equation

$$
x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} - (x^2+p^2)y = 0 \; . \tag{6}
$$

The first of these identities[2] gives an explicit solution for $K_{-1/2}$:

$$
K_{-1/2}(x) = \sqrt{\frac{\pi}{2x}} e^{-x} \; . \tag{7}
$$

The second of these identities[3] gives an integral representation of K_p:

$$
K_p(\sqrt{ab}) = \frac{1}{2}\left(\frac{a}{b}\right)^{p/2} \int_0^\infty x^{p-1} \cdot \exp\left[-\frac{1}{2}\left(ax + \frac{b}{x}\right)\right] dx \; . \tag{8}
$$

Starting from (5), we can expand the binomial term and rearrange the moment generating function into the following form:

$$
\begin{aligned}
M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \int_0^\infty x^{-3/2} \cdot \exp\left[tx - \frac{\alpha^2}{2x} + \alpha\gamma - \frac{\gamma^2 x}{2}\right] dx \\
&= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \int_0^\infty x^{-3/2} \cdot \exp\left[\left(t - \frac{\gamma^2}{2}\right)x - \frac{\alpha^2}{2x}\right] dx \\
&= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \int_0^\infty x^{-3/2} \cdot \exp\left[-\frac{1}{2}\left(\gamma^2 - 2t\right)x - \frac{1}{2} \cdot \frac{\alpha^2}{x}\right] dx \; .
\end{aligned} \tag{9}
$$

The integral now has the form of the integral in (8) with $p = -1/2$, $a = \gamma^2 - 2t$, and $b = \alpha^2$. This allows us to write the moment-generating function in terms of the modified Bessel function $K_{-1/2}$:

$$
M_X(t) = \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \cdot 2\left(\frac{\gamma^2-2t}{\alpha^2}\right)^{1/4} \cdot K_{-1/2}\left(\sqrt{\alpha^2(\gamma^2-2t)}\right) . \tag{10}
$$

Combining with (7) and simplifying gives

---

[1] https://dlmf.nist.gov/10.25
[2] https://dlmf.nist.gov/10.39.2
[3] https://dlmf.nist.gov/10.32.10

$$
\begin{aligned}
M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \cdot 2 \left(\frac{\gamma^2 - 2t}{\alpha^2} \right)^{1/4} \cdot \sqrt{\frac{\pi}{2\sqrt{\alpha^2(\gamma^2 - 2t)}}} \cdot \exp\left[-\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&= \frac{\alpha}{\sqrt{2} \cdot \sqrt{\pi}} \cdot e^{\alpha\gamma} \cdot 2 \cdot \frac{(\gamma^2 - 2t)^{1/4}}{\sqrt{\alpha}} \cdot \frac{\sqrt{\pi}}{\sqrt{2} \cdot \sqrt{\alpha} \cdot (\gamma^2 - 2t)^{1/4}} \cdot \exp\left[-\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&= e^{\alpha\gamma} \cdot \exp\left[-\sqrt{\alpha^2(\gamma^2 - 2t)} \right] \\
&= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \; .
\end{aligned} \tag{11}
$$

This finishes the proof of (2).

■

Sources:

- Siegrist, K. (2020): "The Wald Distribution"; in: *Random: Probability, Mathematical Statistics, Stochastic Processes*, retrieved on 2020-09-13; URL: https://www.randomservices.org/random/special/Wald.html.
- National Institute of Standards and Technology (2020): "NIST Digital Library of Mathematical Functions", retrieved on 2020-09-13; URL: https://dlmf.nist.gov.

3.10.4 Mean

Theorem: Let X be a positive random variable (→ I/1.2.2) following a Wald distribution (→ II/3.10.1):

$$
X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}
$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$
\mathrm{E}(X) = \frac{\alpha}{\gamma} \; . \tag{2}
$$

Proof: The mean or expected value $\mathrm{E}(X)$ is the first moment (→ I/1.18.1) of X, so we can use (→ I/1.18.2) the moment-generating function of the Wald distribution (→ II/3.10.3) to calculate

$$
\mathrm{E}(X) = M_X'(0) \; . \tag{3}
$$

First we differentiate

$$
M_X(t) = \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \tag{4}
$$

with respect to t. Using the chain rule gives

$$
\begin{aligned}
M_X'(t) &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot -\frac{1}{2} \left(\alpha^2(\gamma^2 - 2t) \right)^{-1/2} \cdot -2\alpha^2 \\
&= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)} \right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2t)}} \; .
\end{aligned} \tag{5}
$$

Evaluating (5) at $t = 0$ gives the desired result:

$$
\begin{aligned}
M_X'(0) &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2-2(0))}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2-2(0))}} \\
&= \exp\left[\alpha\gamma - \sqrt{\alpha^2 \cdot \gamma^2}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2 \cdot \gamma^2}} \\
&= \exp[0] \cdot \frac{\alpha^2}{\alpha\gamma} \\
&= \frac{\alpha}{\gamma} \; .
\end{aligned} \tag{6}
$$

■

3.10.5 Variance

Theorem: Let X be a positive random variable (→ I/1.2.2) following a Wald distribution (→ II/3.10.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = \frac{\alpha}{\gamma^3} \; . \tag{2}$$

Proof: To compute the variance of X, we partition the variance into expected values (→ I/1.11.3):

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2. \tag{3}$$

We then use the moment-generating function of the Wald distribution (→ II/3.10.3) to calculate

$$\mathrm{E}(X^2) = M_X''(0) \; . \tag{4}$$

First we differentiate

$$M_X(t) = \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2-2t)}\right] \tag{5}$$

with respect to t. Using the chain rule gives

$$
\begin{aligned}
M_X'(t) &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2-2t)}\right] \cdot -\frac{1}{2}\left(\alpha^2(\gamma^2-2t)\right)^{-1/2} \cdot -2\alpha^2 \\
&= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2-2t)}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2-2t)}} \\
&= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2-2t)}\right] \cdot (\gamma^2-2t)^{-1/2} \; .
\end{aligned} \tag{6}
$$

Now we use the product rule to obtain the second derivative:

$$
\begin{split}
M_X''(t) &= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1/2} \cdot -\frac{1}{2}\left(\alpha^2(\gamma^2 - 2t)\right)^{-1/2} \cdot -2\alpha^2 \\
&+ \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}(\gamma^2 - 2t)^{-3/2} \cdot -2 \\
&= \alpha^2 \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1} \\
&+ \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-3/2} \\
&= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \left[\frac{\alpha}{\gamma^2 - 2t} + \frac{1}{\sqrt{(\gamma^2 - 2t)^3}}\right] .
\end{split} \tag{7}
$$

Applying (4) yields

$$
\begin{split}
\mathrm{E}(X^2) &= M_X''(0) \\
&= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2(0))}\right] \left[\frac{\alpha}{\gamma^2 - 2(0)} + \frac{1}{\sqrt{(\gamma^2 - 2(0))^3}}\right] \\
&= \alpha \cdot \exp\left[\alpha\gamma - \alpha\gamma\right] \cdot \left[\frac{\alpha}{\gamma^2} + \frac{1}{\gamma^3}\right] \\
&= \frac{\alpha^2}{\gamma^2} + \frac{\alpha}{\gamma^3} .
\end{split} \tag{8}
$$

Since the mean of a Wald distribution (→ II/3.10.4) is given by $\mathrm{E}(X) = \alpha/\gamma$, we can apply (3) to show

$$
\begin{split}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \\
&= \frac{\alpha^2}{\gamma^2} + \frac{\alpha}{\gamma^3} - \left(\frac{\alpha}{\gamma}\right)^2 \\
&= \frac{\alpha}{\gamma^3}
\end{split} \tag{9}
$$

which completes the proof of (2).

■

3.10.6 Skewness

Theorem: Let X be a random variable (→ I/1.2.2) following a Wald distribution (→ II/3.10.1):

$$
X \sim \mathrm{Wald}(\gamma, \alpha) . \tag{1}
$$

Then the skewness (→ I/1.12.1) of X is

$$
\mathrm{Skew}(X) = \frac{3}{\sqrt{\alpha\gamma}} . \tag{2}
$$

Proof:

To compute the skewness of X, we partition the skewness into expected values (→ I/1.12.3):

$$\mathrm{Skew}(X) = \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \; , \tag{3}$$

where μ and σ are the mean and standard deviation of X, respectively. Since X follows an Wald distribution (→ II/3.10.1), the mean (→ II/3.10.4) of X is given by

$$\mu = \mathrm{E}(X) = \frac{\alpha}{\gamma} \tag{4}$$

and the standard deviation (→ II/3.10.5) of X is given by

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{\alpha}{\gamma^3}} \; . \tag{5}$$

Substituting (4) and (5) into (3) gives:

$$\begin{aligned} \mathrm{Skew}(X) &= \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \\ &= \frac{\mathrm{E}(X^3) - 3\left(\frac{\alpha}{\gamma}\right)\left(\frac{\alpha}{\gamma^3}\right) - \left(\frac{\alpha}{\gamma}\right)^3}{\left(\sqrt{\frac{\alpha}{\gamma^3}}\right)^3} \\ &= \frac{\gamma^{9/2}}{\alpha^{3/2}} \left[\mathrm{E}(X^3) - \frac{3\alpha^2}{\gamma^4} - \frac{\alpha^3}{\gamma^3}\right] \; . \end{aligned} \tag{6}$$

Thus, the remaining work is to compute $\mathrm{E}(X^3)$. To do this, we use the moment-generating function of the Wald distribution (→ II/3.10.3) to calculate

$$\mathrm{E}(X^3) = M_X'''(0) \tag{7}$$

based on the relationship between raw moment and moment-generating function (→ I/1.18.2). First, we differentiate the moment-generating function

$$M_X(t) = \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \tag{8}$$

with respect to t. Using the chain rule, we have:

$$\begin{aligned} M_X'(t) &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}\left(\alpha^2(\gamma^2 - 2t)\right)^{-1/2} \cdot -2\alpha^2 \\ &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2t)}} \\ &= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1/2} \; . \end{aligned} \tag{9}$$

Now we use the product rule to obtain the second derivative:

$$
\begin{aligned}
M_X''(t) &= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1/2} \cdot -\frac{1}{2}\left(\alpha^2(\gamma^2 - 2t)\right)^{-1/2} \cdot -2\alpha^2 \\
&+ \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}(\gamma^2 - 2t)^{-3/2} \cdot -2 \\
&= \alpha^2 \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1} \\
&+ \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-3/2} \\
&= \frac{\alpha^2}{\gamma^2 - 2t} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] + \frac{\alpha}{(\gamma^2 - 2t)^{3/2}} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \; .
\end{aligned} \tag{10}
$$

Finally, one more application of the chain rule will give us the third derivative. To start, we will decompose the second derivative obtained in (10) as

$$
M''(t) = f(t) + g(t) \tag{11}
$$

where

$$
f(t) = \frac{\alpha^2}{\gamma^2 - 2t} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - t2)}\right] \tag{12}
$$

and

$$
g(t) = \frac{\alpha}{(\gamma^2 - 2t)^{3/2}} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \; . \tag{13}
$$

With this decomposition, $M_X'''(t) = f'(t) + g'(t)$. Applying the product rule to f gives:

$$
\begin{aligned}
f'(t) &= 2\alpha^2(\gamma^2 - 2t)^{-2} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\
&+ \alpha^2(\gamma^2 - 2t)^{-1} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}\left[\alpha^2(\gamma^2 - 2t)\right]^{-1/2} \cdot -2\alpha^2 \\
&= \frac{2\alpha^2}{(\gamma^2 - 2t)^2} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\
&+ \frac{\alpha^3}{(\gamma^2 - 2t)^{3/2}} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \; .
\end{aligned} \tag{14}
$$

Similarly, applying the product rule to g gives:

$$
\begin{aligned}
g'(t) &= -\frac{3}{2}\alpha(\gamma^2 - 2t)^{-5/2}(-2) \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\
&+ \alpha(\gamma^2 - 2t)^{-3/2} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}\left[\alpha^2(\gamma^2 - 2t)\right]^{-1/2} \cdot -2\alpha^2 \\
&= \frac{3\alpha}{(\gamma^2 - 2t)^{5/2}} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\
&+ \frac{\alpha^2}{(\gamma^2 - 2t)^2} \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \; .
\end{aligned} \tag{15}
$$

Applying (7), together with (14) and (15), yields

$$
\begin{aligned}
\mathrm{E}(X^3) &= M_X'''(0) \\
&= f'(0) + g'(0) \\
&= \left[\frac{2\alpha^2}{\gamma^4} + \frac{\alpha^3}{\gamma^3} \right] + \left[\frac{3\alpha}{\gamma^5} + \frac{\alpha^2}{\gamma^4} \right] \\
&= \frac{3\alpha^2}{\gamma^4} + \frac{\alpha^3}{\gamma^3} + \frac{3\alpha}{\gamma^5} \; .
\end{aligned} \tag{16}
$$

We now substitute (16) into (6), giving

$$
\begin{aligned}
\mathrm{Skew}(X) &= \frac{\gamma^{9/2}}{\alpha^{3/2}} \left[\mathrm{E}(X^3) - \frac{3\alpha^2}{\gamma^4} - \frac{\alpha^3}{\gamma^3} \right] \\
&= \frac{\gamma^{9/2}}{\alpha^{3/2}} \left[\frac{3\alpha^2}{\gamma^4} + \frac{\alpha^3}{\gamma^3} + \frac{3\alpha}{\gamma^5} - \frac{3\alpha^2}{\gamma^4} - \frac{\alpha^3}{\gamma^3} \right] \\
&= \frac{\gamma^{9/2}}{\alpha^{3/2}} \cdot \frac{3\alpha}{\gamma^5} \\
&= \frac{3}{\alpha^{1/2} \cdot \gamma^{1/2}} \\
&= \frac{3}{\sqrt{\alpha\gamma}} \; .
\end{aligned} \tag{17}
$$

This completes the proof of (2).

■

3.10.7 Method of moments

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed data independent and identically distributed according to a Wald distribution (→ II/3.10.1) with drift rate γ and threshold α:

$$
y_i \sim \mathrm{Wald}(\gamma, \alpha), \quad i = 1, \ldots, n \; . \tag{1}
$$

Then, the method-of-moments estimates (→ I/4.1.8) for the parameters γ and α are given by

$$
\begin{aligned}
\hat{\gamma} &= \sqrt{\frac{\bar{y}}{\bar{v}}} \\
\hat{\alpha} &= \sqrt{\frac{\bar{y}^3}{\bar{v}}}
\end{aligned} \tag{2}
$$

where $\bar{y}$ is the sample mean (→ I/1.10.2) and $\bar{v}$ is the unbiased sample variance (→ I/1.11.2):

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\
\bar{v} &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \; .
\end{aligned} \tag{3}
$$

Proof: The mean (→ II/3.10.4) and variance (→ II/3.10.5) of the Wald distribution (→ II/3.10.1) in terms of the parameters γ and α are given by

$$\begin{split} \mathrm{E}(X) &= \frac{\alpha}{\gamma} \\ \mathrm{Var}(X) &= \frac{\alpha}{\gamma^3} \; . \end{split} \tag{4}$$

Thus, matching the moments (→ I/4.1.8) requires us to solve the following system of equations for γ and α:

$$\begin{split} \bar{y} &= \frac{\alpha}{\gamma} \\ \bar{v} &= \frac{\alpha}{\gamma^3} \; . \end{split} \tag{5}$$

To this end, our first step is to express the second equation of (5) as follows:

$$\begin{split} \bar{v} &= \frac{\alpha}{\gamma^3} \\ &= \frac{\alpha}{\gamma} \cdot \gamma^{-2} \\ &= \bar{y} \cdot \gamma^{-2} \; . \end{split} \tag{6}$$

Rearranging (6) gives

$$\gamma^2 = \frac{\bar{y}}{\bar{v}} \; , \tag{7}$$

or equivalently,

$$\gamma = \sqrt{\frac{\bar{y}}{\bar{v}}} \; . \tag{8}$$

Our final step is to solve the first equation of (5) for α and substitute (8) for γ:

$$\begin{split} \alpha &= \bar{y} \cdot \gamma \\ &= \bar{y} \cdot \sqrt{\frac{\bar{y}}{\bar{v}}} \\ &= \sqrt{\bar{y}^2} \cdot \sqrt{\frac{\bar{y}}{\bar{v}}} \\ &= \sqrt{\frac{\bar{y}^3}{\bar{v}}} \; . \end{split} \tag{9}$$

Together, (8) and (9) constitute the method-of-moment estimates of γ and α.

■

3.11 ex-Gaussian distribution

3.11.1 Definition

Definition: Let A be a random variable (→ I/1.2.2) that is normally distributed (→ II/3.2.1) with mean μ and variance σ^2, and let B be a random variable that is exponentially distributed (→ II/3.5.1) with rate λ. Suppose further that A and B are independent (→ I/1.3.6). Then the sum $X = A + B$ is said to have an exponentially-modified Gaussian (i.e., ex-Gaussian) distribution, with parameters μ, σ, and λ; that is,

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \; , \tag{1}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\lambda > 0$.

Sources:
- Luce, R. D. (1986): "Response Times: Their Role in Inferring Elementary Mental Organization", 35-36; URL: https://global.oup.com/academic/product/response-times-9780195036428.

3.11.2 Probability density function

Theorem: Let X be a random variable (→ I/1.2.2) following an ex-Gaussian distribution (→ II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \; . \tag{1}$$

Then the probability density function (→ I/1.7.1) of X is

$$f_X(t) = \frac{\lambda}{\sqrt{2\pi}} \exp\left[\frac{\lambda^2\sigma^2}{2} - \lambda(t-\mu)\right] \cdot \int_{-\infty}^{\frac{t-\mu}{\sigma} - \lambda\sigma} \exp\left[-\frac{1}{2}y^2\right] dy \; . \tag{2}$$

Proof: Suppose X follows an ex-Gaussian distribution (→ II/3.11.1). Then $X = A + B$, where A and B are independent (→ I/1.3.6), A is normally distributed (→ II/3.2.1) with mean (→ II/3.2.16) μ and variance (→ II/3.2.19) σ^2, and B is exponentially distributed (→ II/3.5.1) with rate λ. Then, the probability density function (→ II/3.2.10) for A is given by

$$f_A(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \; , \tag{3}$$

and the probability density function (→ II/3.5.3) for B is given by

$$f_B(t) = \begin{cases} \lambda \exp[-\lambda t] \; , & \text{if } t \geq 0 \\ 0 \; , & \text{if } t < 0 \; . \end{cases} \tag{4}$$

Thus, the probability density function for the sum (→ I/1.7.2) $X = A + B$ is given by taking the convolution of f_A and f_B:

$$
\begin{aligned}
f_X(t) &= \int_{-\infty}^{\infty} f_A(x) f_B(t-x) dx \\
&= \int_{-\infty}^{t} f_A(x) f_B(t-x) dx + \int_{t}^{\infty} f_A(x) f_B(t-x) dx \\
&= \int_{-\infty}^{t} f_A(x) f_B(t-x) dx \; ,
\end{aligned} \tag{5}
$$

which follows from the fact that $f_B(t-x) = 0$ for $x > t$. From here, we substitute the expressions (3) and (4) for the probability density functions f_A and f_B in (5):

$$
\begin{aligned}
f_X(t) &= \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \cdot \lambda \exp[-\lambda(t-x)] dx \\
&= \frac{\lambda}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \cdot \exp[-\lambda t + \lambda x] dx \\
&= \frac{\lambda}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \cdot \exp[-\lambda t] \cdot \exp[\lambda x] dx \\
&= \frac{\lambda \exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 + \lambda x \right] dx \; .
\end{aligned} \tag{6}
$$

We can further simplify the integrand with a substitution; to this end, let

$$
y = g(x) = \frac{x-\mu}{\sigma} - \lambda\sigma \tag{7}
$$

This gives the following three identities:

$$
\frac{dy}{dx} = \frac{1}{\sigma} \; , \quad \text{or equivalently,} \quad dx = \sigma dy \; , \tag{8}
$$

$$
\frac{x-\mu}{\sigma} = y + \lambda\sigma \; , \quad \text{and} \tag{9}
$$

$$
x = y\sigma + \lambda\sigma^2 + \mu \; . \tag{10}
$$

Substituting these identities into (6) gives

$$
\begin{aligned}
f_X(t) &= \frac{\lambda \exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int_{-\infty}^{g(t)} \exp\left[-\frac{1}{2}(y+\lambda\sigma)^2 + \lambda(y\sigma + \lambda\sigma^2 + \mu)\right] \sigma dy \\
&= \frac{\lambda\sigma \exp[-\lambda t]}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}(y^2 + 2y\lambda\sigma + \lambda^2\sigma^2) + \lambda y\sigma + \lambda^2\sigma^2 + \lambda\mu\right] dy \\
&= \frac{\lambda \exp[-\lambda t]}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2 - y\lambda\sigma - \frac{\lambda^2\sigma^2}{2} + \lambda y\sigma + \lambda^2\sigma^2 + \lambda\mu\right] dy \\
&= \frac{\lambda \exp[-\lambda t]}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right] \cdot \exp\left[\frac{\lambda^2\sigma^2}{2} + \lambda\mu\right] dy \qquad (11) \\
&= \frac{\lambda \exp[-\lambda t]}{\sqrt{2\pi}} \cdot \exp\left[\frac{\lambda^2\sigma^2}{2} + \lambda\mu\right] \int_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right] \cdot dy \\
&= \frac{\lambda}{\sqrt{2\pi}} \cdot \exp\left[-\lambda t + \frac{\lambda^2\sigma^2}{2} + \lambda\mu\right] \int_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right] \cdot dy \\
&= \frac{\lambda}{\sqrt{2\pi}} \cdot \exp\left[\frac{\lambda^2\sigma^2}{2} - \lambda(t-\mu)\right] \int_{-\infty}^{\frac{x-\mu}{\sigma}+\lambda\sigma} \exp\left[-\frac{1}{2}y^2\right] \cdot dy \; .
\end{aligned}
$$

This finishes the proof of (2).

■

3.11.3 Moment-generating function

Theorem: Let X be a random variable (→ I/1.2.2) following an ex-Gaussian distribution (→ II/3.11.1):

$$
X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \; . \qquad (1)
$$

Then, the moment generating function (→ I/1.9.5) of X is

$$
M_X(t) = \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \; . \qquad (2)
$$

Proof: Suppose X follows an ex-Gaussian distribution (→ II/3.11.1). Then, $X = A + B$ where A and B are independent (→ I/1.3.6), A is normally distributed (→ II/3.2.1) with mean (→ II/3.2.16) μ and variance (→ II/3.2.19) σ^2, and B is exponentially distributed (→ II/3.5.1) with rate λ. Then the moment generating function (→ II/3.2.11) for A is given by

$$
M_A(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \qquad (3)
$$

and the moment generating function (→ II/3.5.4) for B is given by

$$
M_B(t) = \frac{\lambda}{\lambda - t} \; . \qquad (4)
$$

By definition, X is a linear combination of independent random variables A and B, so the moment generating function (→ I/1.9.8) of X is the product of $M_A(t)$ and $M_B(t)$. That is,

$$
\begin{aligned}
M_X(t) &= M_A(t) \cdot M_B(t) \\
&= \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \cdot \left(\frac{\lambda}{\lambda - t}\right) \\
&= \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \; .
\end{aligned} \tag{5}
$$

This finishes the proof of (2).

■

3.11.4 Mean

Theorem: Let X be a random variable (→ I/1.2.2) following an ex-Gaussian distribution (→ II/3.11.1):

$$
X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \; . \tag{1}
$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$
\mathrm{E}(X) = \mu + \frac{1}{\lambda} \; . \tag{2}
$$

Proof: The mean or expected value $\mathrm{E}(X)$ is the first raw moment (→ I/1.18.1) of X, so we can use (→ I/1.18.2) the moment-generating function of the ex-Gaussian distribution (→ II/3.11.3) to calculate

$$
\mathrm{E}(X) = M_X'(0) \; . \tag{3}
$$

First, we differentiate

$$
M_X(t) = \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \tag{4}
$$

with respect to t. Using the product rule and chain rule gives:

$$
\begin{aligned}
M_X'(t) &= \frac{\lambda}{(\lambda - t)^2} \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] + \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] (\mu + \sigma^2 t) \\
&= \left(\frac{\lambda}{\lambda - t}\right) \cdot \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \; .
\end{aligned} \tag{5}
$$

Evaluating (5) at $t = 0$ gives the desired result:

$$
\begin{aligned}
M_X'(0) &= \left(\frac{\lambda}{\lambda - 0}\right) \cdot \exp\left[\mu \cdot 0 + \frac{1}{2}\sigma^2 \cdot 0^2\right] \cdot \left[\frac{1}{\lambda - 0} + \mu + \sigma^2 \cdot 0\right] \\
&= 1 \cdot 1 \cdot \left[\frac{1}{\lambda} + \mu\right] \\
&= \mu + \frac{1}{\lambda} \; .
\end{aligned} \tag{6}
$$

■

3.11.5 Variance

Theorem: Let X be a random variable (→ I/1.2.2) following an ex-Gaussian distribution (→ II/3.11.1):

$$X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \; . \tag{1}$$

Then, the variance (→ I/1.11.1) of X is

$$\mathrm{Var}(X) = \sigma^2 + \frac{1}{\lambda^2} \; . \tag{2}$$

Proof: To compute the variance of X, we partition the variance into expected values (→ I/1.11.3):

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; . \tag{3}$$

We then use the moment-generating function of the ex-Gaussian distribution (→ II/3.11.3) to calculate

$$\mathrm{E}(X^2) = M_X''(0) \tag{4}$$

based on the relationship between raw moment and moment-generating function (→ I/1.18.2). First, we differentiate

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \tag{5}$$

with respect to t. Using the product rule and chain rule gives:

$$\begin{aligned}
M_X'(t) &= \frac{\lambda}{(\lambda - t)^2} \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] + \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] (\mu + \sigma^2 t) \\
&= \left(\frac{\lambda}{\lambda - t}\right) \cdot \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \\
&= M_X(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \; .
\end{aligned} \tag{6}$$

We now use the product rule to obtain the second derivative:

$$\begin{aligned}
M_X''(t) &= M_X'(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] + M_X(t) \cdot \left[\frac{1}{(\lambda - t)^2} + \sigma^2\right] \\
&\overset{(6)}{=} M_X(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right]^2 + M_X(t) \cdot \left[\frac{1}{(\lambda - t)^2} + \sigma^2\right] \\
&= M_X(t) \cdot \left[\left(\frac{1}{\lambda - t} + \mu + \sigma^2 t\right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2\right] \\
&= \left(\frac{\lambda}{\lambda - t}\right) \cdot \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \cdot \left[\left(\frac{1}{\lambda - t} + \mu + \sigma^2 t\right)^2 + \frac{1}{(\lambda - t)^2} + \sigma^2\right]
\end{aligned} \tag{7}$$

Applying (4) yields

$$
\begin{aligned}
\mathrm{E}(X^2) &= M_X''(0) \\
&= \left(\frac{\lambda}{\lambda-0}\right) \cdot \exp\left[\mu \cdot 0 + \frac{1}{2}\sigma^2 \cdot 0^2\right] \cdot \left[\left(\frac{1}{\lambda-0} + \mu + \sigma^2 \cdot 0\right)^2 + \frac{1}{(\lambda-0)^2} + \sigma^2\right] \\
&= 1 \cdot 1 \cdot \left[\left(\frac{1}{\lambda} + \mu\right)^2 + \frac{1}{\lambda^2} + \sigma^2\right] \\
&= \frac{1}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \frac{1}{\lambda^2} + \sigma^2 \\
&= \frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 \; .
\end{aligned} \tag{8}
$$

Since the mean of an ex-Gaussian distribution (→ II/3.11.4) is given by

$$
\mathrm{E}(X) = \mu + \frac{1}{\lambda} \; , \tag{9}
$$

we can apply (3) to show

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \\
&= \left[\frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2\right] - \left(\mu + \frac{1}{\lambda}\right)^2 \\
&= \frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 - \mu^2 - \frac{2\mu}{\lambda} - \frac{1}{\lambda^2} \\
&= \sigma^2 + \frac{1}{\lambda^2} \; .
\end{aligned} \tag{10}
$$

This completes the proof of (2).

■

3.11.6 Skewness

Theorem: Let X be a random variable (→ I/1.2.2) following an ex-Gaussian distribution (→ II/3.11.1):

$$
X \sim \text{ex-Gaussian}(\mu, \sigma, \lambda) \; . \tag{1}
$$

Then the skewness (→ I/1.12.1) of X is

$$
\mathrm{Skew}(X) = \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2}\right)^{\frac{3}{2}}} \; . \tag{2}
$$

Proof:
To compute the skewness of X, we partition the skewness into expected values (→ I/1.12.3):

$$
\mathrm{Skew}(X) = \frac{\mathrm{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3} \; , \tag{3}
$$

where μ and σ are the mean and standard deviation of X, respectively. To prevent confusion between the labels used for the ex-Gaussian parameters in (1) and the mean and standard deviation of X, we rewrite (3) as

$$\mathrm{Skew}(X) = \frac{\mathrm{E}(X^3) - 3 \cdot \mathrm{E}(X) \cdot \mathrm{Var}(X) - \mathrm{E}(X)^3}{\mathrm{Var}(X)^{\frac{3}{2}}} \; . \tag{4}$$

Since X follows an ex-Gaussian distribution (→ II/3.11.1), the mean (→ II/3.11.4) of X is given by

$$\mathrm{E}(X) = \mu + \frac{1}{\lambda} \tag{5}$$

and the variance (→ II/3.11.5) of X is given by

$$\mathrm{Var}(X) = \sigma^2 + \frac{1}{\lambda^2} \; . \tag{6}$$

Thus, the primary work is to compute $\mathrm{E}(X^3)$. To do this, we use the moment-generating function of the ex-Gaussian distribution (→ II/3.11.3) to calculate

$$\mathrm{E}(X^3) = M_X'''(0) \tag{7}$$

based on the relationship between raw moment and moment-generating function (→ I/1.18.2).
First, we differentiate the moment-generating function of the ex-Gaussian distribution (→ II/3.11.3)

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \tag{8}$$

with respect to t. Using the product rule and chain rule, we have:

$$\begin{aligned} M_X'(t) &= \frac{\lambda}{(\lambda-t)^2} \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] + \left(\frac{\lambda}{\lambda - t}\right) \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] (\mu + \sigma^2 t) \\ &= \left(\frac{\lambda}{\lambda - t}\right) \cdot \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right] \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \\ &= M_X(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] \; . \end{aligned} \tag{9}$$

We then use the product rule to obtain the second derivative:

$$\begin{aligned} M_X''(t) &= M_X'(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right] + M_X(t) \cdot \left[\frac{1}{(\lambda-t)^2} + \sigma^2\right] \\ &= M_X(t) \cdot \left[\frac{1}{\lambda - t} + \mu + \sigma^2 t\right]^2 + M_X(t) \cdot \left[\frac{1}{(\lambda-t)^2} + \sigma^2\right] \\ &= M_X(t) \cdot \left[\left(\frac{1}{\lambda - t} + \mu + \sigma^2 t\right)^2 + \frac{1}{(\lambda-t)^2} + \sigma^2\right] \; . \end{aligned} \tag{10}$$

Finally, we use the product rule and chain rule to obtain the third derivative:

$$M_X'''(t) = M_X'(t) \left[\left(\frac{1}{\lambda - t} + \mu + \sigma^2 t\right)^2 + \frac{1}{(\lambda-t)^2} + \sigma^2\right] + M_X(t) \left[2\left(\frac{1}{\lambda - t} + \mu + \sigma^2 t\right)\left(\frac{1}{(\lambda-t)^2} + \sigma^2\right) + \frac{}{(\lambda} \right. \tag{11}$$

Applying (7), together with (11), yields:

$$
\begin{split}
\mathrm{E}(X^3) &= M_X'''(0) \\
&= M_X'(0) \left[\left(\frac{1}{\lambda} + \mu \right)^2 + \frac{1}{\lambda^2} + \sigma^2 \right] + M_X(0) \left[2 \left(\frac{1}{\lambda} + \mu \right) \left(\frac{1}{\lambda^2} + \sigma^2 \right) + \frac{2}{\lambda^3} \right] \\
&= \left(\mu + \frac{1}{\lambda} \right) \left(\frac{1}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \frac{1}{\lambda^2} + \sigma^2 \right) + \left(\frac{2}{\lambda^3} + \frac{2\sigma^2}{\lambda} + \frac{2\mu}{\lambda^2} + 2\mu\sigma^2 + \frac{2}{\lambda^3} \right) \\
&= \left(\mu + \frac{1}{\lambda} \right) \left(\frac{2}{\lambda^2} + \frac{2\mu}{\lambda} + \mu^2 + \sigma^2 \right) + \left(\frac{4}{\lambda^3} + \frac{2\sigma^2}{\lambda} + \frac{2\mu}{\lambda^2} + 2\mu\sigma^2 \right) \\
&= \frac{2\mu}{\lambda^2} + \frac{2\mu^2}{\lambda} + \mu^3 + \mu\sigma^2 + \frac{2}{\lambda^3} + \frac{2\mu}{\lambda^2} + \frac{\mu^2}{\lambda} + \frac{\sigma^2}{\lambda} + \frac{4}{\lambda^3} + \frac{2\sigma^2}{\lambda} + \frac{2\mu}{\lambda^2} + 2\mu\sigma^2 \\
&= \frac{6\mu}{\lambda^2} + \frac{6}{\lambda^3} + \frac{3\mu^2 + 3\sigma^2}{\lambda} + 3\mu\sigma^2 + \mu^3 \; .
\end{split}
\tag{12}
$$

We now substitute (12), (5), and (6) into the numerator of (4), giving

$$
\begin{split}
\mathrm{E}(X^3) - 3 \cdot \mathrm{E}(X) \cdot \mathrm{Var}(X) - \mathrm{E}(X)^3 &= \left(\frac{6\mu}{\lambda^2} + \frac{6}{\lambda^3} + \frac{3\mu^2 + 3\sigma^2}{\lambda} + 3\mu\sigma^2 + \mu^3 \right) - 3 \left(\mu + \frac{1}{\lambda} \right) \left(\sigma^2 + \frac{1}{\lambda^2} \right) - \Big(\mu \\
&= \frac{6\mu}{\lambda^2} + \frac{6}{\lambda^3} + \frac{3\mu^2 + 3\sigma^2}{\lambda} + 3\mu\sigma^2 + \mu^3 - 3\mu\sigma^2 - \frac{3\mu}{\lambda^2} - \frac{3\sigma^2}{\lambda} - \frac{3}{\lambda^3} - \mu^3 - 3 \\
&= \frac{2}{\lambda^3} \; .
\end{split}
\tag{13}
$$

Thus, we have:

$$
\begin{split}
\mathrm{Skew}(X) &= \frac{\mathrm{E}(X^3) - 3 \cdot \mathrm{E}(X) \cdot \mathrm{Var}(X) - \mathrm{E}(X)^3}{\mathrm{Var}(X)^{\frac{3}{2}}} \\
&= \frac{\frac{2}{\lambda^3}}{\left(\sigma^2 + \frac{1}{\lambda^2} \right)^{\frac{3}{2}}} \\
&= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2} \right)^{\frac{3}{2}}} \; .
\end{split}
\tag{14}
$$

This completes the proof of (2).

■

3.11.7 Method of moments

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed data independent and identically distributed according to an ex-Gaussian distribution (→ II/3.11.1) with parameters μ, σ, and λ:

$$
y_i \sim \mathrm{ex-Gaussian}(\mu, \sigma, \lambda), \quad i = 1, \ldots, n \; . \tag{1}
$$

Then, the method-of-moments estimates (→ I/4.1.8) for the parameters μ, σ, and λ are given by

$$
\begin{aligned}
\hat{\mu} &= \bar{y} - \sqrt[3]{\frac{\bar{s} \cdot \bar{v}^{3/2}}{2}} \\
\hat{\sigma} &= \sqrt{\bar{v} \cdot \left(1 - \sqrt[3]{\frac{\bar{s}^2}{4}} \right)} \\
\hat{\lambda} &= \sqrt[3]{\frac{2}{\bar{s} \cdot \bar{v}^{3/2}}} \; ,
\end{aligned} \tag{2}
$$

where $\bar{y}$ is the sample mean (→ I/1.10.2), $\bar{v}$ is the sample variance (→ I/1.11.2) and $\bar{s}$ is the sample skewness (→ I/1.12.2)

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\bar{v} &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
\bar{s} &= \frac{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^3}{\left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{3/2}} \; .
\end{aligned} \tag{3}
$$

Proof: The mean (→ II/3.11.4), variance (→ II/3.11.5), and skewness (→ II/3.11.6) of the ex-Gaussian distribution (→ II/3.11.1) in terms of the parameters μ, σ, and λ are given by

$$
\begin{aligned}
\mathrm{E}(X) &= \mu + \frac{1}{\lambda} \\
\mathrm{Var}(X) &= \sigma^2 + \frac{1}{\lambda^2} \\
\mathrm{Skew}(X) &= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2} \right)^{3/2}} \; .
\end{aligned} \tag{4}
$$

Thus, matching the moments (→ I/4.1.8) requires us to solve the following system of equations for μ, σ, and λ:

$$
\begin{aligned}
\bar{y} &= \mu + \frac{1}{\lambda} \\
\bar{v} &= \sigma^2 + \frac{1}{\lambda^2} \\
\bar{s} &= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2} \right)^{3/2}} \; .
\end{aligned} \tag{5}
$$

To this end, our first step is to substitute the second equation of (5) into the third equation:

$$
\begin{aligned}
\bar{s} &= \frac{2}{\lambda^3 \left(\sigma^2 + \frac{1}{\lambda^2} \right)^{3/2}} \\
&= \frac{2}{\lambda^3 \cdot \bar{v}^{3/2}} \; .
\end{aligned} \tag{6}
$$

Re-expressing (6) in terms of λ^3 and taking the cube root gives:

$$\lambda = \sqrt[3]{\frac{2}{\bar{s} \cdot \bar{v}^{3/2}}} \; . \tag{7}$$

Next, we solve the first equation of (5) for μ and substitute (7):

$$\begin{aligned} \mu &= \bar{y} - \frac{1}{\lambda} \\ &= \bar{y} - \sqrt[3]{\frac{\bar{s} \cdot \bar{v}^{3/2}}{2}} \; . \end{aligned} \tag{8}$$

Finally, we solve the second equation of (5) for σ:

$$\sigma^2 = \bar{v} - \frac{1}{\lambda^2} \; . \tag{9}$$

Taking the square root gives and substituting (7) gives:

$$\begin{aligned} \sigma &= \sqrt{\bar{v} - \frac{1}{\lambda^2}} \\ &= \sqrt{\bar{v} - \left(\sqrt[3]{\frac{\bar{s} \cdot \bar{v}^{3/2}}{2}} \right)^2} \\ &= \sqrt{\bar{v} - \bar{v} \cdot \sqrt[3]{\frac{\bar{s}^2}{4}}} \\ &= \sqrt{\bar{v} \cdot \left(1 - \sqrt[3]{\frac{\bar{s}^2}{4}} \right)} \; . \end{aligned} \tag{10}$$

Together, (8), (10), and (7) constitute the method-of-moment estimates of μ, σ, and λ.

■

4 Multivariate continuous distributions

4.1 Multivariate normal distribution

4.1.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ I/1.2.3). Then, X is said to be multivariate normally distributed with mean μ and covariance Σ

$$X \sim \mathcal{N}(\mu, \Sigma) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \tag{2}$$

where μ is an $n \times 1$ real vector and Σ is an $n \times n$ positive definite matrix.

Sources:
- Koch KR (2007): "Multivariate Normal Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.1, pp. 51-53, eq. 2.195; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

4.1.2 Special case of matrix-normal distribution

Theorem: The multivariate normal distribution (→ II/4.1.1) is a special case of the matrix-normal distribution (→ II/5.1.1) with number of variables $p = 1$, i.e. random matrix (→ I/1.2.2) $X = x \in \mathbb{R}^{n \times 1}$, mean $M = \mu \in \mathbb{R}^{n \times 1}$, covariance across rows $U = \Sigma$ and covariance across columns $V = 1$.

Proof: The probability density function of the matrix-normal distribution (→ II/5.1.3) is

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \; . \tag{1}$$

Setting $p = 1$, $X = x$, $M = \mu$, $U = \Sigma$ and $V = 1$, we obtain

$$\begin{split} \mathcal{MN}(x; \mu, \Sigma, 1) &= \frac{1}{\sqrt{(2\pi)^n |1|^n |\Sigma|^1}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(1^{-1} (x-\mu)^\mathrm{T} \, \Sigma^{-1} (x-\mu) \right) \right] \\ &= \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \, \Sigma^{-1} (x-\mu) \right] \end{split} \tag{2}$$

which is equivalent to the probability density function of the multivariate normal distribution (→ II/4.1.7).

■

Sources:
- Wikipedia (2022): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-07-31; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Definition.

4.1.3 Relationship to chi-squared distribution

Theorem: Let x be an $n \times 1$ random vector (→ I/1.2.3) following a multivariate normal distribution (→ II/4.1.1) with zero mean (→ I/1.10.13) and arbitrary covariance matrix (→ II/4.1.10) Σ:

$$x \sim \mathcal{N}(0, \Sigma) . \tag{1}$$

Then, the quadratic form of x, weighted by Σ, follows a chi-squared distribution (→ II/3.7.1) with n degrees of freedom:

$$y = x^\mathrm{T} \Sigma^{-1} x \sim \chi^2(n) . \tag{2}$$

Proof: Define a new random vector (→ I/1.2.3) z as

$$z = \Sigma^{-1/2} x . \tag{3}$$

where $\Sigma^{-1/2}$ is the matrix square root of Σ. This matrix must exist, because Σ is a covariance matrix (→ I/1.13.9) and thus positive semi-definite (→ I/1.13.13). Due to the linear transformation theorem (→ II/4.1.13), z is distributed as

$$\begin{aligned} z &\sim \mathcal{N}\left(\Sigma^{-1/2} 0, \, \Sigma^{-1/2} \Sigma \, {\Sigma^{-1/2}}^\mathrm{T}\right) \\ &\sim \mathcal{N}\left(\Sigma^{-1/2} 0, \, \Sigma^{-1/2} \Sigma^{1/2} \Sigma^{1/2} \Sigma^{-1/2}\right) \\ &\sim \mathcal{N}(0, I_n) , \end{aligned} \tag{4}$$

i.e. each entry of this vector follows (→ II/4.1.14) a standard normal distribution (→ II/3.2.3):

$$z_i \sim \mathcal{N}(0,1) \quad \text{for all} \quad i = 1, \ldots, n . \tag{5}$$

We further observe that y can be represented in terms of z

$$y = x^\mathrm{T} \Sigma^{-1} x = \left(x^\mathrm{T} \Sigma^{-1/2}\right) \left(\Sigma^{-1/2} x\right) = z^\mathrm{T} z , \tag{6}$$

thus z is a sum of n squared standard normally distributed (→ II/3.2.3) random variables (→ I/1.2.2)

$$y = \sum_{i=1}^{n} z_i^2 \quad \text{where all} \quad z_i \sim \mathcal{N}(0,1) \tag{7}$$

which, by definition, is chi-squared distributed (→ II/3.7.1) with n degrees of freedom:

$$y \sim \chi^2(n) . \tag{8}$$

■

Sources:

- Koch KR (2007): "Chi-Squared Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.4.5, pp. 48-49, eq. 2.180; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

4.1.4 Bivariate normal distribution

Definition: Let X be an 2×1 random vector (→ I/1.2.3). Then, X is said to have a bivariate normal distribution, if X follows a multivariate normal distribution (→ II/4.1.1)

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{1}$$

with means (→ I/1.10.1) x_1 and x_2, variances (→ I/1.11.1) σ_1^2 and σ_2^2 and covariance (→ I/1.13.1) σ_{12}:

$$\mu = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \; . \tag{2}$$

Sources:
- Wikipedia (2023): "Multivariate normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-09-22; URL: https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case.

4.1.5 Probability density function of the bivariate normal distribution

Theorem: Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ follow a bivariate normal distribution (→ II/4.1.4):

$$X \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is:

$$f_X(x) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp\left[-\frac{1}{2}\frac{\sigma_2^2(x_1-\mu_1)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\right] \; . \tag{2}$$

Proof: The probability density function of the multivariate normal distribution (→ II/4.1.7) for an $n \times 1$ random vector (→ I/1.2.3) x is:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}\Sigma^{-1}(x-\mu)\right] \; . \tag{3}$$

Plugging in $n = 2$, $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$, we obtain:

$$
\begin{split}
f_X(x) &= \frac{1}{\sqrt{(2\pi)^2 \left| \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right|}} \cdot \exp\left[-\frac{1}{2} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^\mathrm{T} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right) \right] \\
&= \frac{1}{2\pi \left| \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2} \begin{bmatrix} (x_1-\mu_1) & (x_2-\mu_2) \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} (x_1-\mu_1) \\ (x_2-\mu_2) \end{bmatrix} \right] \; .
\end{split} \tag{4}
$$

Using the determinant of a 2×2 matrix

$$
\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = ad - bc \tag{5}
$$

and the inverse of of a 2×2 matrix

$$
\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \; , \tag{6}
$$

the probability density function (→ I/1.7.1) becomes:

$$
\begin{split}
f_X(x) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp\left[-\frac{1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} \begin{bmatrix} (x_1-\mu_1) & (x_2-\mu_2) \end{bmatrix} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix} \begin{bmatrix} (x_1-\mu_1) \\ (x_2-\mu_2) \end{bmatrix} \right] \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp\left[-\frac{1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} \begin{bmatrix} \sigma_2^2(x_1-\mu_1) - \sigma_{12}(x_2-\mu_2) & \sigma_1^2(x_2-\mu_2) - \sigma_{12}(x_1-\mu_1) \end{bmatrix} \left[(\right. \right. \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp\left[-\frac{1}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} (\sigma_2^2(x_1-\mu_1)^2 - \sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2 - \sigma_{12}(x \right. \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \cdot \exp\left[-\frac{1}{2} \frac{\sigma_2^2(x_1-\mu_1)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \right] \; .
\end{split} \tag{7}
$$

■

4.1.6 Probability density function in terms of correlation coefficient

Theorem: Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ follow a bivariate normal distribution:

$$X \sim \mathcal{N}\left(\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right) . \tag{1}$$

Then, the probability density function of X is

$$f_X(x) = \frac{1}{2\pi\,\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\right] \tag{2}$$

where ρ is the correlation (→ I/1.14.1) between X_1 and X_2.

Proof: Since X follows a special case of the multivariate normal distribution, its covariance matrix is (→ II/4.1.10)

$$\mathrm{Cov}(X) = \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \tag{3}$$

and the covariance matrix can be decomposed into correlation matrix and standard deviations (→ I/1.13.18):

$$\begin{split} \Sigma &= \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} . \end{split} \tag{4}$$

The determinant of this matrix is

$$\begin{split} |\Sigma| &= \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right| \\ &= \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right| \cdot \left| \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right| \cdot \left| \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right| \\ &= (\sigma_1\sigma_2)(1-\rho^2)(\sigma_1\sigma_2) \\ &= \sigma_1^2\sigma_2^2(1-\rho^2) \end{split} \tag{5}$$

and the inverse of this matrix is

$$
\begin{split}
\Sigma^{-1} &= \left(\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \right)^{-1} \\
&= \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}^{-1} \\
&= \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \; .
\end{split} \tag{6}
$$

The probability density function of the multivariate normal distribution (→ II/4.1.7) for an $n \times 1$ random vector (→ I/1.2.3) x is:

$$
f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \; . \tag{7}
$$

Plugging in $n=2$, μ from (1) and Σ from (5) and (6), the probability density function becomes:

$$
\begin{split}
f_X(x) &= \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2 (1-\rho^2)}} \cdot \exp\left[-\frac{1}{2} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^\mathrm{T} \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix} \right(\\
&= \frac{1}{2\pi \, \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1} & \frac{x_2-\mu_2}{\sigma_2} \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1} \\ \frac{x_2-\mu_2}{\sigma_2} \end{bmatrix} \right] \\
&= \frac{1}{2\pi \, \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)} \begin{bmatrix} \left(\frac{x_1-\mu_1}{\sigma_1} - \rho \frac{x_2-\mu_2}{\sigma_2} \right) & \left(\frac{x_2-\mu_2}{\sigma_2} - \rho \frac{x_1-\mu_1}{\sigma_1} \right) \end{bmatrix} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1} \\ \frac{x_2-\mu_2}{\sigma_2} \end{bmatrix} \right] \\
&= \frac{1}{2\pi \, \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)} \left(\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1 \sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right) \right] \; .
\end{split} \tag{8}
$$

■

Sources:

- Wikipedia (2023): "Multivariate normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-09-29; URL: https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Bivariate_case.

4.1.7 Probability density function

Theorem: Let X be a random vector (→ I/1.2.3) following a multivariate normal distribution (→ II/4.1.1):

$$
X \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}
$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}\Sigma^{-1}(x-\mu)\right] \; . \tag{2}$$

Proof: This follows directly from the definition of the multivariate normal distribution (→ II/4.1.1).

■

4.1.8 Moment-generating function

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the moment-generating function (→ I/1.9.5) of x is

$$M_x(t) = \exp\left[t^\mathrm{T}\mu + \frac{1}{2}t^\mathrm{T}\Sigma t\right] \; . \tag{2}$$

Proof: The moment-generating function of a random vector (→ I/1.9.5) X is defined as:

$$M_X(t) = \mathrm{E}\left[e^{t^\mathrm{T}X}\right], \quad t \in \mathbb{R}^n \; . \tag{3}$$

Applying the law of the unconscious statistician (→ I/1.10.12), we have:

$$M_x(t) = \int_{\mathcal{X}} e^{t^\mathrm{T}x} \cdot f_X(x) \, \mathrm{d}x \; . \tag{4}$$

With the probability density function of the multivariate normal distribution (→ II/4.1.7), we have:

$$M_x(t) = \int_{\mathbb{R}^n} \exp\left[t^\mathrm{T}x\right] \cdot \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}\Sigma^{-1}(x-\mu)\right] \mathrm{d}x \; . \tag{5}$$

Now we summarize the two exponential functions inside the integral:

$$\begin{aligned}
M_x(t) &= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}\Sigma^{-1}(x-\mu) + t^\mathrm{T}x\right] \mathrm{d}x \\
&= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}\left(x^\mathrm{T}\Sigma^{-1}x - 2\mu^\mathrm{T}\Sigma^{-1}x + \mu^\mathrm{T}\Sigma^{-1}\mu - 2t^\mathrm{T}x\right)\right] \mathrm{d}x \\
&= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}\left(x^\mathrm{T}\Sigma^{-1}x - 2(\mu+\Sigma t)^\mathrm{T}\Sigma^{-1}x + \mu^\mathrm{T}\Sigma^{-1}\mu\right)\right] \mathrm{d}x \\
&= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}\left((x-[\mu+\Sigma t])^\mathrm{T}\Sigma^{-1}(x-[\mu+\Sigma t]) - 2t^\mathrm{T}\mu - t^\mathrm{T}\Sigma t\right)\right] \mathrm{d}x \\
&= \exp\left[t^\mathrm{T}\mu + t^\mathrm{T}\Sigma t\right] \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-[\mu+\Sigma t])^\mathrm{T}\Sigma^{-1}(x-[\mu+\Sigma t])\right] \mathrm{d}x \; .
\end{aligned} \tag{6}$$

The integrand is equal to the probability density function of a multivariate normal distribution (→ II/4.1.7):

$$M_x(t) = \exp\left[t^\mathrm{T}\mu + t^\mathrm{T}\Sigma t\right] \int_{\mathbb{R}^n} \mathcal{N}(x; \mu + \Sigma t, \Sigma) \, \mathrm{d}x \; . \tag{7}$$

Because the entire probability density integrates to one (→ I/1.7.1), we finally have:

$$M_x(t) = \exp\left[t^\mathrm{T}\mu + t^\mathrm{T}\Sigma t\right] \; . \tag{8}$$

■

4.1.9 Mean

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of x is

$$\mathrm{E}(x) = \mu \; . \tag{2}$$

Proof:
1) First, consider a set of independent (→ I/1.3.6) and standard normally (→ II/3.2.3) distributed random variables (→ I/1.2.2):

$$z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n \; . \tag{3}$$

Then, these variables together form a multivariate normally (→ II/4.1.16) distributed random vector (→ I/1.2.3):

$$z \sim \mathcal{N}(0_n, I_n) \; . \tag{4}$$

By definition, the expected value of a random vector is equal to the vector of all expected values (→ I/1.10.13):

$$\mathrm{E}(z) = \mathrm{E}\left(\begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \right) = \begin{bmatrix} \mathrm{E}(z_1) \\ \vdots \\ \mathrm{E}(z_n) \end{bmatrix} \; . \tag{5}$$

Because the expected value of all its entries is zero (→ II/3.2.16), the expected value of the random vector is

$$\mathrm{E}(z) = \begin{bmatrix} \mathrm{E}(z_1) \\ \vdots \\ \mathrm{E}(z_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = 0_n \; . \tag{6}$$

2) Next, consider an $n \times n$ matrix A solving the equation $AA^\mathrm{T} = \Sigma$. Such a matrix exists, because Σ is defined to be positive definite (→ II/4.1.1). Then, x can be represented as a linear transformation of (→ II/4.1.13) z:

$$x = Az + \mu \sim \mathcal{N}(A0_n + \mu, AI_nA^\mathrm{T}) = \mathcal{N}(\mu, \Sigma) \; . \tag{7}$$

Thus, the expected value (→ I/1.10.1) of x can be written as:

$$\mathrm{E}(x) = \mathrm{E}(Az + \mu) \; . \tag{8}$$

With the linearity of the expected value (→ I/1.10.5), this becomes:

$$\begin{split} \mathrm{E}(x) &= \mathrm{E}(Az + \mu) \\ &= \mathrm{E}(Az) + \mathrm{E}(\mu) \\ &= A\,\mathrm{E}(z) + \mu \\ &\overset{(6)}{=} A\,0_n + \mu \\ &= \mu \; . \end{split} \tag{9}$$

■

Sources:

- Taboga, Marco (2021): "Multivariate normal distribution"; in: *Lectures on probability theory and mathematical statistics*, retrieved on 2022-09-15; URL: https://www.statlect.com/probability-distributions/multivariate-normal-distribution.

4.1.10 Covariance

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the covariance matrix (→ I/1.13.9) of x is

$$\mathrm{Cov}(x) = \Sigma \; . \tag{2}$$

Proof:
1) First, consider a set of independent (→ I/1.3.6) and standard normally (→ II/3.2.3) distributed random variables (→ I/1.2.2):

$$z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n \; . \tag{3}$$

Then, these variables together form a multivariate normally (→ II/4.1.16) distributed random vector (→ I/1.2.3):

$$z \sim \mathcal{N}(0_n, I_n) \; . \tag{4}$$

Because the covariance is zero for independent random variables (→ I/1.13.6), we have

$$\mathrm{Cov}(z_i, z_j) = 0 \quad \text{for all} \quad i \neq j \; . \tag{5}$$

Moreover, as the variance of all entries of the vector is one (→ II/3.2.19), we have

$$\mathrm{Var}(z_i) = 1 \quad \text{for all} \quad i = 1, \ldots, n \; . \tag{6}$$

Taking (5) and (6) together, the covariance matrix (→ I/1.13.9) of z is

$$\mathrm{Cov}(z) = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = I_n \; . \tag{7}$$

2) Next, consider an $n \times n$ matrix A solving the equation $AA^\mathrm{T} = \Sigma$. Such a matrix exists, because Σ is defined to be positive definite (→ II/4.1.1). Then, x can be represented as a linear transformation of (→ II/4.1.13) z:

$$x = Az + \mu \sim \mathcal{N}(A0_n + \mu, AI_nA^\mathrm{T}) = \mathcal{N}(\mu, \Sigma) \; . \tag{8}$$

Thus, the covariance (→ I/1.13.1) of x can be written as:

$$\mathrm{Cov}(x) = \mathrm{Cov}(Az + \mu) \; . \tag{9}$$

With the invariance of the covariance matrix under addition (→ I/1.13.14)

$$\mathrm{Cov}(x + a) = \mathrm{Cov}(x) \tag{10}$$

and the scaling of the covariance matrix upon multiplication (→ I/1.13.15)

$$\mathrm{Cov}(Ax) = A\mathrm{Cov}(x)A^\mathrm{T} \; , \tag{11}$$

this becomes:

$$\begin{split} \mathrm{Cov}(x) &= \mathrm{Cov}(Az + \mu) \\ &\overset{(10)}{=} \mathrm{Cov}(Az) \\ &\overset{(11)}{=} A\,\mathrm{Cov}(z)A^\mathrm{T} \\ &\overset{(7)}{=} AI_nA^\mathrm{T} \\ &= AA^\mathrm{T} \\ &= \Sigma \; . \end{split} \tag{12}$$

■

Sources:

- Rosenfeld, Meni (2016): "Deriving the Covariance of Multivariate Gaussian"; in: *StackExchange Mathematics*, retrieved on 2022-09-15; URL: https://math.stackexchange.com/questions/1905977/deriving-the-covariance-of-multivariate-gaussian.

4.1.11 Differential entropy

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the differential entropy (→ I/2.2.1) of x in nats is

$$\mathrm{h}(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln|\Sigma| + \frac{1}{2} n \; . \tag{2}$$

Proof: The differential entropy (→ I/2.2.1) of a random variable is defined as

$$\mathrm{h}(X) = -\int_{\mathcal{X}} p(x) \, \log_b p(x) \, \mathrm{d}x \; . \tag{3}$$

To measure $h(X)$ in nats, we set $b = e$, such that (→ I/1.10.1)

$$\mathrm{h}(X) = -\mathrm{E}\left[\ln p(x)\right] \; . \tag{4}$$

With the probability density function of the multivariate normal distribution (→ II/4.1.7), the differential entropy of x is:

$$\begin{split} \mathrm{h}(x) &= -\mathrm{E}\left[\ln \left(\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \right) \right] \\ &= -\mathrm{E}\left[-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \\ &= \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\,\mathrm{E}\left[(x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \; . \end{split} \tag{5}$$

The last term can be evaluted as

$$\begin{split} \mathrm{E}\left[(x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] &= \mathrm{E}\left[\mathrm{tr}\left((x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right) \right] \\ &= \mathrm{E}\left[\mathrm{tr}\left(\Sigma^{-1} (x-\mu)(x-\mu)^\mathrm{T} \right) \right] \\ &= \mathrm{tr}\left(\Sigma^{-1} \mathrm{E}\left[(x-\mu)(x-\mu)^\mathrm{T} \right] \right) \\ &= \mathrm{tr}\left(\Sigma^{-1} \Sigma \right) \\ &= \mathrm{tr}\left(I_n \right) \\ &= n \; , \end{split} \tag{6}$$

such that the differential entropy is

$$\mathrm{h}(x) = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\,n \; . \tag{7}$$

■

Sources:

- Kiuhnm (2018): "Entropy of the multivariate Gaussian"; in: *StackExchange Mathematics*, retrieved on 2020-05-14; URL: https://math.stackexchange.com/questions/2029707/entropy-of-the-multivariate-ga

4.1.12 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (→ I/1.2.3). Assume two multivariate normal distributions (→ II/4.1.1) P and Q specifying the probability distribution of x as

$$\begin{split} P: \; x &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ Q: \; x &\sim \mathcal{N}(\mu_2, \Sigma_2) \; . \end{split} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[(\mu_2-\mu_1)^\mathrm{T}\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \; . \tag{2}$$

Proof: The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P\,||\,Q] = \int_\mathcal{X} p(x)\,\ln\frac{p(x)}{q(x)}\,\mathrm{d}x \tag{3}$$

which, applied to the multivariate normal distributions (→ II/4.1.1) in (1), yields

$$\begin{split}\mathrm{KL}[P\,||\,Q] &= \int_{\mathbb{R}^n} \mathcal{N}(x;\mu_1,\Sigma_1)\,\ln\frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)}\,\mathrm{d}x \\ &= \left\langle \ln\frac{\mathcal{N}(x;\mu_1,\Sigma_1)}{\mathcal{N}(x;\mu_2,\Sigma_2)} \right\rangle_{p(x)} \; .\end{split} \tag{4}$$

Using the probability density function of the multivariate normal distribution (→ II/4.1.7), this becomes:

$$\begin{split}\mathrm{KL}[P\,||\,Q] &= \left\langle \ln\frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_1|}}\cdot\exp\left[-\frac{1}{2}(x-\mu_1)^\mathrm{T}\Sigma_1^{-1}(x-\mu_1)\right]}{\frac{1}{\sqrt{(2\pi)^n|\Sigma_2|}}\cdot\exp\left[-\frac{1}{2}(x-\mu_2)^\mathrm{T}\Sigma_2^{-1}(x-\mu_2)\right]} \right\rangle_{p(x)} \\ &= \left\langle \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^\mathrm{T}\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\mathrm{T}\Sigma_2^{-1}(x-\mu_2) \right\rangle_{p(x)} \\ &= \frac{1}{2}\left\langle \ln\frac{|\Sigma_2|}{|\Sigma_1|} - (x-\mu_1)^\mathrm{T}\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^\mathrm{T}\Sigma_2^{-1}(x-\mu_2) \right\rangle_{p(x)} \; .\end{split} \tag{5}$$

Now, using the fact that $x = \mathrm{tr}(x)$, if a is scalar, and the trace property $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, we have:

$$\begin{split}\mathrm{KL}[P\,||\,Q] &= \frac{1}{2}\left\langle \ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^\mathrm{T}\right] + \mathrm{tr}\left[\Sigma_2^{-1}(x-\mu_2)(x-\mu_2)^\mathrm{T}\right] \right\rangle_{p(x)} \\ &= \frac{1}{2}\left\langle \ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^\mathrm{T}\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(xx^\mathrm{T} - 2\mu_2x^\mathrm{T} + \mu_2\mu_2^\mathrm{T}\right)\right] \right\rangle_{p(x)} \; .\end{split} \tag{6}$$

Because trace function and expected value are both linear operators (→ I/1.10.8), the expectation can be moved inside the trace:

$$\begin{split}\mathrm{KL}[P\,||\,Q] &= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}\left\langle(x-\mu_1)(x-\mu_1)^\mathrm{T}\right\rangle_{p(x)}\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left\langle xx^\mathrm{T} - 2\mu_2x^\mathrm{T} + \mu_2\mu_2^\mathrm{T}\right\rangle_{p(x)}\right] \right) \\ &= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}\left\langle(x-\mu_1)(x-\mu_1)^\mathrm{T}\right\rangle_{p(x)}\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(\left\langle xx^\mathrm{T}\right\rangle_{p(x)} - \left\langle 2\mu_2x^\mathrm{T}\right\rangle_{p(x)} + \left\langle \mu_2\mu_2^\mathrm{T}\right\rangle_{p(x)}\right)\right] \right)\end{split} \tag{7}$$

Using the expectation of a linear form for the multivariate normal distribution (→ II/4.1.13)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle Ax \rangle = A\mu \tag{8}$$

and the expectation of a quadratic form for the multivariate normal distribution (→ I/1.10.9)

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \left\langle x^{\mathrm{T}} A x \right\rangle = \mu^{\mathrm{T}} A \mu + \mathrm{tr}(A\Sigma) \;, \tag{9}$$

the Kullback-Leibler divergence from (7) becomes:

$$\begin{split}
\mathrm{KL}[P \,||\, Q] &= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}\Sigma_1\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(\Sigma_1 + \mu_1\mu_1^{\mathrm{T}} - 2\mu_2\mu_1^{\mathrm{T}} + \mu_2\mu_2^{\mathrm{T}}\right)\right] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[I_n\right] + \mathrm{tr}\left[\Sigma_2^{-1}\Sigma_1\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(\mu_1\mu_1^{\mathrm{T}} - 2\mu_2\mu_1^{\mathrm{T}} + \mu_2\mu_2^{\mathrm{T}}\right)\right] \right) \\
&= \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}\left[\Sigma_2^{-1}\Sigma_1\right] + \mathrm{tr}\left[\mu_1^{\mathrm{T}}\Sigma_2^{-1}\mu_1 - 2\mu_1^{\mathrm{T}}\Sigma_2^{-1}\mu_2 + \mu_2^{\mathrm{T}}\Sigma_2^{-1}\mu_2\right] \right) \\
&= \frac{1}{2} \left[\ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}\left[\Sigma_2^{-1}\Sigma_1\right] + (\mu_2 - \mu_1)^{\mathrm{T}}\Sigma_2^{-1}(\mu_2 - \mu_1) \right] \;.
\end{split} \tag{10}$$

Finally, rearranging the terms, we get:

$$\mathrm{KL}[P \,||\, Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^{\mathrm{T}}\Sigma_2^{-1}(\mu_2 - \mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \;. \tag{11}$$

■

Sources:

- Duchi, John (2014): "Derivations for Linear Algebra and Optimization"; in: *University of California, Berkeley*; URL: http://www.eecs.berkeley.edu/~jduchi/projects/general_notes.pdf.

4.1.13 Linear transformation

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \;. \tag{1}$$

Then, any linear transformation of x is also multivariate normally distributed:

$$y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^{\mathrm{T}}) \;. \tag{2}$$

Proof: The moment-generating function of a random vector (→ I/1.9.5) x is

$$M_x(t) = \mathbb{E}\left(\exp\left[t^{\mathrm{T}} x\right]\right) \tag{3}$$

and therefore the moment-generating function of the random vector y is given by

$$\begin{split}
M_y(t) &\overset{(2)}{=} \mathbb{E}\left(\exp\left[t^{\mathrm{T}}(Ax + b)\right]\right) \\
&= \mathbb{E}\left(\exp\left[t^{\mathrm{T}} A x\right] \cdot \exp\left[t^{\mathrm{T}} b\right]\right) \\
&= \exp\left[t^{\mathrm{T}} b\right] \cdot \mathbb{E}\left(\exp\left[t^{\mathrm{T}} A x\right]\right) \\
&\overset{(3)}{=} \exp\left[t^{\mathrm{T}} b\right] \cdot M_x(A^{\mathrm{T}} t) \;.
\end{split} \tag{4}$$

The moment-generating function of the multivariate normal distribution (→ II/4.1.8) is

$$M_x(t) = \exp\left[t^\mathrm{T}\mu + \frac{1}{2}t^\mathrm{T}\Sigma t\right] \tag{5}$$

and therefore the moment-generating function of the random vector y becomes

$$\begin{aligned} M_y(t) &\overset{(4)}{=} \exp\left[t^\mathrm{T}b\right] \cdot M_x(A^\mathrm{T}t) \\ &\overset{(5)}{=} \exp\left[t^\mathrm{T}b\right] \cdot \exp\left[t^\mathrm{T}A\mu + \frac{1}{2}t^\mathrm{T}A\Sigma A^\mathrm{T}t\right] \\ &= \exp\left[t^\mathrm{T}\left(A\mu + b\right) + \frac{1}{2}t^\mathrm{T}A\Sigma A^\mathrm{T}t\right] \; . \end{aligned} \tag{6}$$

Because moment-generating function and probability density function of a random variable are equivalent, this demonstrates that y is following a multivariate normal distribution with mean $A\mu+b$ and covariance $A\Sigma A^\mathrm{T}$.

■

Sources:
- Taboga, Marco (2010): "Linear combinations of normal random variables"; in: *Lectures on probability and statistics*, retrieved on 2019-08-27; URL: https://www.statlect.com/probability-distributions/normal-distribution-linear-combinations.

4.1.14 Marginal distributions

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the marginal distribution (→ I/1.5.3) of any subset vector x_s is also a multivariate normal distribution

$$x_s \sim \mathcal{N}(\mu_s, \Sigma_s) \tag{2}$$

where μ_s drops the irrelevant variables (the ones not in the subset, i.e. marginalized out) from the mean vector μ and Σ_s drops the corresponding rows and columns from the covariance matrix Σ.

Proof: Define an $m \times n$ subset matrix S such that $s_{ij} = 1$, if the j-th element in x_s corresponds to the i-th element in x, and $s_{ij} = 0$ otherwise. Then,

$$x_s = Sx \tag{3}$$

and we can apply the linear transformation theorem (→ II/4.1.13) to give

$$x_s \sim \mathcal{N}(S\mu, S\Sigma S^\mathrm{T}) \; . \tag{4}$$

Finally, we see that $S\mu = \mu_s$ and $S\Sigma S^\mathrm{T} = \Sigma_s$.

■

4.1.15 Conditional distributions

Theorem: Let x follow a multivariate normal distribution (→ II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma) . \tag{1}$$

Then, the conditional distribution (→ I/1.5.4) of any subset vector x_1, given the complement vector x_2, is also a multivariate normal distribution

$$x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \tag{2}$$

where the conditional mean (→ I/1.10.1) and covariance (→ I/1.13.1) are

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned} \tag{3}$$

with block-wise mean and covariance defined as

$$\begin{aligned} \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} . \end{aligned} \tag{4}$$

Proof: Without loss of generality, we assume that, in parallel to (4),

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{5}$$

where x_1 is an $n_1 \times 1$ vector, x_2 is an $n_2 \times 1$ vector and x is an $n_1 + n_2 = n \times 1$ vector.

By construction, the joint distribution (→ I/1.5.2) of x_1 and x_2 is:

$$x_1, x_2 \sim \mathcal{N}(\mu, \Sigma) . \tag{6}$$

Moreover, the marginal distribution (→ I/1.5.3) of x_2 follows from (→ II/4.1.14) (1) and (4) as

$$x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) . \tag{7}$$

According to the law of conditional probability (→ I/1.3.4), it holds that

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \tag{8}$$

Applying (6) and (7) to (8), we have:

$$p(x_1|x_2) = \frac{\mathcal{N}(x; \mu, \Sigma)}{\mathcal{N}(x_2; \mu_2, \Sigma_{22})} . \tag{9}$$

Using the probability density function of the multivariate normal distribution (→ II/4.1.7), this becomes:

$$
\begin{split}
p(x_1|x_2) &= \frac{1/\sqrt{(2\pi)^n|\Sigma|} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}\Sigma^{-1}(x-\mu)\right]}{1/\sqrt{(2\pi)^{n_2}|\Sigma_{22}|} \cdot \exp\left[-\frac{1}{2}(x_2-\mu_2)^\mathrm{T}\Sigma_{22}^{-1}(x_2-\mu_2)\right]} \\
&= \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}\Sigma^{-1}(x-\mu) + \frac{1}{2}(x_2-\mu_2)^\mathrm{T}\Sigma_{22}^{-1}(x_2-\mu_2)\right] \; .
\end{split} \tag{10}
$$

Writing the inverse of Σ as

$$
\Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \tag{11}
$$

and applying (4) to (10), we get:

$$
\begin{split}
p(x_1|x_2) = &\frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
&\exp\left[-\frac{1}{2}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^\mathrm{T} \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) \right. \\
&\left. \quad +\frac{1}{2} \left(x_2-\mu_2\right)^\mathrm{T} \Sigma_{22}^{-1} \left(x_2-\mu_2\right) \right] \; .
\end{split} \tag{12}
$$

Multiplying out within the exponent of (12), we have

$$
\begin{split}
p(x_1|x_2) = &\frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
&\exp\left[-\frac{1}{2}\left((x_1-\mu_1)^\mathrm{T}\Sigma^{11}(x_1-\mu_1) + 2(x_1-\mu_1)^\mathrm{T}\Sigma^{12}(x_2-\mu_2) + (x_2-\mu_2)^\mathrm{T}\Sigma^{22}(x_2-\mu_2)\right) \right. \\
&\left. \quad +\frac{1}{2}(x_2-\mu_2)^\mathrm{T}\Sigma_{22}^{-1}(x_2-\mu_2)\right]
\end{split} \tag{13}
$$

where we have used the fact that ${\Sigma^{21}}^\mathrm{T} = \Sigma^{12}$, because Σ^{-1} is a symmetric matrix.

The inverse of a block matrix is

$$
\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \; , \tag{14}
$$

thus the inverse of Σ in (11) is

$$
\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix} . \tag{15}
$$

Plugging this into (13), we have:

$$
\begin{split}
p(x_1|x_2) = & \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
& \exp\left[-\frac{1}{2} \left((x_1-\mu_1)^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1-\mu_1) \; - \right.\right. \\
& \qquad 2(x_1-\mu_1)^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2-\mu_2) + \\
& \qquad \left. (x_2-\mu_2)^\mathrm{T} \left[\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \right] (x_2-\mu_2) \right) \\
& \quad \left. + \frac{1}{2} \left((x_2-\mu_2)^\mathrm{T} \Sigma_{22}^{-1} (x_2-\mu_2) \right) \right] \; .
\end{split} \tag{16}
$$

Eliminating some terms, we have:

$$
\begin{split}
p(x_1|x_2) = & \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\
& \exp\left[-\frac{1}{2} \left((x_1-\mu_1)^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} (x_1-\mu_1) \; - \right.\right. \\
& \qquad 2(x_1-\mu_1)^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \Sigma_{12}\Sigma_{22}^{-1} (x_2-\mu_2) + \\
& \qquad \left.\left. (x_2-\mu_2)^\mathrm{T} \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} (x_2-\mu_2) \right) \right] \; .
\end{split} \tag{17}
$$

Rearranging the terms, we have

$$
\begin{split}
p(x_1|x_2) = & \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2} \cdot \right. \\
& \left. \left[(x_1-\mu_1) - \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \right]^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[(x_1-\mu_1) - \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \right] \right] \\
= & \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2} \cdot \right. \\
& \left. \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \right) \right]^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \right) \right] \right]
\end{split} \tag{18}
$$

where we have used the fact that $\Sigma_{21} = \Sigma_{12}^\mathrm{T}$, because Σ is a covariance matrix (→ I/1.13.9).

The determinant of a block matrix is

$$
\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| \cdot |A - BD^{-1}C| \; , \tag{19}
$$

such that we have for Σ that

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{22}| \cdot |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}| \; . \tag{20}$$

With this and $n - n_2 = n_1$, we finally arrive at

$$\begin{split} p(x_1|x_2) = & \frac{1}{\sqrt{(2\pi)^{n_1}|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|}} \cdot \exp\left[-\frac{1}{2} \cdot \right. \\ & \left. \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)\right)\right]^\mathrm{T} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)\right)\right]\right] \end{split} \tag{21}$$

which is the probability density function of a multivariate normal distribution (→ II/4.1.7)

$$p(x_1|x_2) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{1|2}) \tag{22}$$

with the mean (→ I/1.10.1) $\mu_{1|2}$ and covariance (→ I/1.13.1) $\Sigma_{1|2}$ given by (3).

■

Sources:

- Wang, Ruye (2006): "Marginal and conditional distributions of multivariate normal distribution"; in: *Computer Image Processing and Analysis*; URL: http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html.
- Wikipedia (2020): "Multivariate normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-20; URL: https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions.

4.1.16 Conditions for independence

Theorem: Let x be an $n \times 1$ random vector (→ I/1.2.3) following a multivariate normal distribution (→ II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the components of x are statistically independent (→ I/1.3.6), if and only if the covariance matrix (→ I/1.13.9) is a diagonal matrix:

$$p(x) = p(x_1) \cdot \ldots \cdot p(x_n) \quad \Leftrightarrow \quad \Sigma = \mathrm{diag}\left(\left[\sigma_1^2, \ldots, \sigma_n^2\right]\right) \; . \tag{2}$$

Proof: The marginal distribution of one entry from a multivariate normal random vector is a univariate normal distribution (→ II/4.1.14) where mean (→ I/1.10.1) and variance (→ I/1.11.1) are equal to the corresponding entries of the mean vector and covariance matrix:

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad x_i \sim \mathcal{N}(\mu_i, \sigma_{ii}^2) \; . \tag{3}$$

The probability density function of the multivariate normal distribution (→ II/4.1.7) is

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] \tag{4}$$

and the probability density function of the univariate normal distribution (→ II/3.2.10) is

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left[-\frac{1}{2} \left(\frac{x_i-\mu_i}{\sigma_i} \right)^2 \right] \; . \tag{5}$$

1) Let

$$p(x) = p(x_1) \cdot \ldots \cdot p(x_n) \; . \tag{6}$$

Then, we have

$$\begin{split} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] &\overset{(4),(5)}{=} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left[-\frac{1}{2} \left(\frac{x_i-\mu_i}{\sigma_i} \right)^2 \right] \\ \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) \right] &= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \cdot \exp\left[-\frac{1}{2} \sum_{i=1}^n (x_i-\mu_i) \frac{1}{\sigma_i^2} (x_i-\mu_i) \right] \\ -\frac{1}{2} \log|\Sigma| - \frac{1}{2} (x-\mu)^\mathrm{T} \Sigma^{-1} (x-\mu) &= -\frac{1}{2} \sum_{i=1}^n \log\sigma_i^2 - \frac{1}{2} \sum_{i=1}^n (x_i-\mu_i) \frac{1}{\sigma_i^2} (x_i-\mu_i) \end{split} \tag{7}$$

which is only fulfilled by a diagonal covariance matrix

$$\Sigma = \mathrm{diag}\left(\left[\sigma_1^2, \ldots, \sigma_n^2 \right] \right) \; , \tag{8}$$

because the determinant of a diagonal matrix is a product

$$|\mathrm{diag}\left([a_1, \ldots, a_n] \right)| = \prod_{i=1}^n a_i \; , \tag{9}$$

the inverse of a diagonal matrix is a diagonal matrix

$$\mathrm{diag}\left([a_1, \ldots, a_n] \right)^{-1} = \mathrm{diag}\left([1/a_1, \ldots, 1/a_n] \right) \tag{10}$$

and the squared form with a diagonal matrix is

$$x^\mathrm{T} \mathrm{diag}\left([a_1, \ldots, a_n] \right) x = \sum_{i=1}^n a_i x_i^2 \; . \tag{11}$$

2) Let

$$\Sigma = \mathrm{diag}\left(\left[\sigma_1^2, \ldots, \sigma_n^2 \right] \right) \; . \tag{12}$$

Then, we have

$$
\begin{aligned}
p(x) &\overset{(4)}{=} \frac{1}{\sqrt{(2\pi)^n |\mathrm{diag}\left([\sigma_1^2, \ldots, \sigma_n^2]\right)|}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T} \mathrm{diag}\left(\left[\sigma_1^2, \ldots, \sigma_n^2\right]\right)^{-1} (x-\mu)\right] \\
&= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T} \mathrm{diag}\left(\left[1/\sigma_1^2, \ldots, 1/\sigma_n^2\right]\right) (x-\mu)\right] \\
&= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^n \sigma_i^2}} \cdot \exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\mu_i)^2}{\sigma_i^2}\right] \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]
\end{aligned} \tag{13}
$$

which implies that

$$
p(x) = p(x_1) \cdot \ldots \cdot p(x_n) \; . \tag{14}
$$

■

4.1.17 Independence of products

Theorem: Let X be an $n \times 1$ random vector (→ I/1.2.3) following a multivariate normal distribution (→ II/4.1.1):

$$
X \sim \mathcal{N}(\mu, \Sigma) \tag{1}
$$

and consider two matrices $A \in \mathbb{R}^{k \times n}$ and $B \in \mathbb{R}^{l \times n}$. Then, AX and BX are independent (→ I/1.3.6), if and only if the cross-matrix product, weighted with the covariance matrix (→ II/4.1.10) is equal to the zero matrix:

$$
AX \quad \text{and} \quad BX \quad \text{ind.} \quad \Leftrightarrow \quad A\Sigma B^\mathrm{T} = 0_{kl} \; . \tag{2}
$$

Proof: Define a new random vector (→ I/1.2.3) C as

$$
C = \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{(k+l) \times n} \; . \tag{3}
$$

Then, due to the linear transformation theorem (→ II/4.1.13), we have

$$
CX = \begin{bmatrix} AX \\ BX \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} A\mu \\ B\mu \end{bmatrix}, C\Sigma C^\mathrm{T}\right) \tag{4}
$$

with the combined covariance matrix (→ I/1.13.9)

$$
C\Sigma C^\mathrm{T} = \begin{bmatrix} A\Sigma A^\mathrm{T} & A\Sigma B^\mathrm{T} \\ B\Sigma A^\mathrm{T} & B\Sigma B^\mathrm{T} \end{bmatrix} \; . \tag{5}
$$

We know that the necessary and sufficient condition for two components of a multivariate normal random vector to be independent is that their entries in the covariance matrix are zero (→ II/4.1.16). Thus, AX and BX are independent (→ I/1.3.6), if and only if

$$A\Sigma B^{\mathrm{T}} = (B\Sigma A^{\mathrm{T}})^{\mathrm{T}} = 0_{kl} \tag{6}$$

where 0_{kl} is the $k \times l$ zero matrix. This proves the result in (2).

■

Sources:
- jld (2018): "Understanding t-test for linear regression"; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: https://stats.stackexchange.com/a/344008.

4.2 Multivariate t-distribution

4.2.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ I/1.2.3). Then, X is said to follow a multivariate t-distribution with mean μ, scale matrix Σ and degrees of freedom ν

$$X \sim t(\mu, \Sigma, \nu) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$t(x; \mu, \Sigma, \nu) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \, \frac{\Gamma([\nu+n]/2)}{\Gamma(\nu/2)} \, \left[1 + \frac{1}{\nu}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu) \right]^{-(\nu+n)/2} \tag{2}$$

where μ is an $n \times 1$ real vector, Σ is an $n \times n$ positive definite matrix and $\nu > 0$.

Sources:
- Koch KR (2007): "Multivariate t-Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.2, pp. 53-55; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

4.2.2 Probability density function

Theorem: Let X be a random vector (→ I/1.2.3) following a multivariate t-distribution (→ II/4.2.1):

$$X \sim t(\mu, \Sigma, \nu) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \sqrt{\frac{1}{(\nu\pi)^n |\Sigma|}} \, \frac{\Gamma([\nu+n]/2)}{\Gamma(\nu/2)} \, \left[1 + \frac{1}{\nu}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu) \right]^{-(\nu+n)/2} \; . \tag{2}$$

Proof: This follows directly from the definition of the multivariate t-distribution (→ II/4.2.1).

■

4.2.3 Relationship to F-distribution

Theorem: Let X be a $n \times 1$ random vector (→ I/1.2.3) following a multivariate t-distribution (→ II/4.2.1) with mean μ, scale matrix Σ and degrees of freedom ν:

$$X \sim t(\mu, \Sigma, \nu) \; . \tag{1}$$

Then, the centered, weighted and standardized quadratic form of X follows an F-distribution (→ II/3.8.1) with degrees of freedom n and ν:

$$(X-\mu)^\mathrm{T} \, \Sigma^{-1} (X-\mu)/n \sim F(n, \nu) \; . \tag{2}$$

Proof: The linear transformation theorem for the multivariate t-distribution states

$$x \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad y = Ax + b \sim t(A\mu + b, A \Sigma A^\mathrm{T}, \nu) \tag{3}$$

where x is an $n \times 1$ random vector (→ I/1.2.3) following a multivariate t-distribution (→ II/4.2.1), A is an $m \times n$ matrix and b is an $m \times 1$ vector. Define the following quantities

$$\begin{split} Y &= \Sigma^{-1/2}(X-\mu) = \Sigma^{-1/2}X - \Sigma^{-1/2}\mu \\ Z &= Y^\mathrm{T} Y/n = (X-\mu)^\mathrm{T} \, \Sigma^{-1}(X-\mu)/n \end{split} \tag{4}$$

where $\Sigma^{-1/2}$ is a matrix square root of the inverse of Σ. Then, applying (3) to (4) with (1), one obtains the distribution of Y as

$$\begin{split} Y &\sim t(\Sigma^{-1/2}\mu - \Sigma^{-1/2}\mu, \Sigma^{-1/2}\Sigma \, \Sigma^{-1/2}, \nu) \\ &= t(0_n, \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2}\Sigma^{-1/2}, \nu) \\ &= t(0_n, I_n, \nu) \; , \end{split} \tag{5}$$

i.e. the marginal distributions (→ I/1.5.3) of the individual entries of Y are univariate t-distributions (→ II/3.3.1) with ν degrees of freedom:

$$Y_i \sim t(\nu), \; i = 1, \ldots, n \; . \tag{6}$$

Note that, when X follows a t-distribution with n degrees of freedom, this is equivalent to (→ II/3.3.1) an expression of X in terms of a standard normal (→ II/3.2.3) random variable Z and a chi-squared (→ II/3.7.1) random variable V:

$$X \sim t(n) \quad \Leftrightarrow \quad X = \frac{Z}{\sqrt{V/n}} \quad \text{with independent} \quad Z \sim \mathcal{N}(0,1) \quad \text{and} \quad V \sim \chi^2(n) \; . \tag{7}$$

With that, Z from (4) can be rewritten as follows:

$$
\begin{aligned}
Z &\overset{(4)}{=} Y^{\mathrm{T}} Y/n \\
&= \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \\
&\overset{(7)}{=} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Z_i}{\sqrt{V/\nu}} \right)^2 \\
&= \frac{\left(\sum_{i=1}^{n} Z_i^2 \right)/n}{V/\nu} \; .
\end{aligned} \tag{8}
$$

Because by definition, the sum of squared standard normal random variables follows a chi-squared distribution (→ II/3.7.1)

$$
X_i \sim \mathcal{N}(0,1), \; i = 1, \ldots, n \quad \Rightarrow \quad \sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \; , \tag{9}
$$

the quantity Z becomes a ratio of the following form

$$
Z = \frac{W/n}{V/\nu} \quad \text{with} \quad W \sim \chi^2(n) \quad \text{and} \quad V \sim \chi^2(\nu) \; , \tag{10}
$$

such that Z, by definition, follows an F-distribution (→ II/3.8.1):

$$
Z = \frac{W/n}{V/\nu} \sim F(n, \nu) \; . \tag{11}
$$

■

Sources:

- Lin, Pi-Erh (1972): "Some Characterizations of the Multivariate t Distribution"; in: *Journal of Multivariate Analysis*, vol. 2, pp. 339-344, Lemma 2; URL: https://core.ac.uk/download/pdf/81139018.pdf; DOI: 10.1016/0047-259X(72)90021-8.
- Nadarajah, Saralees; Kotz, Samuel (2005): "Mathematical Properties of the Multivariate t Distribution"; in: *Acta Applicandae Mathematicae*, vol. 89, pp. 53-84, page 56; URL: https://link.springer.com/content/pdf/10.1007/s10440-005-9003-4.pdf; DOI: 10.1007/s10440-005-9003-4.

4.3 Normal-gamma distribution

4.3.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ I/1.2.3) and let Y be a positive random variable (→ I/1.2.2). Then, X and Y are said to follow a normal-gamma distribution

$$
X, Y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; , \tag{1}
$$

if the distribution of X conditional on Y is a multivariate normal distribution (→ II/4.1.1) with mean vector μ and covariance matrix $(y\Lambda)^{-1}$ and Y follows a gamma distribution (→ II/3.4.1) with shape parameter a and rate parameter b:

$$
\begin{aligned}
X|Y &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) \\
Y &\sim \mathrm{Gam}(a, b) \; .
\end{aligned} \tag{2}
$$

The $n \times n$ matrix Λ is referred to as the precision matrix (→ I/1.13.19) of the normal-gamma distribution.

Sources:
- Koch KR (2007): "Normal-Gamma Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.3, pp. 55-56, eq. 2.212; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

4.3.2 Special case of normal-Wishart distribution

Theorem: The normal-gamma distribution (→ II/4.3.1) is a special case of the normal-Wishart distribution (→ II/5.3.1) where the number of columns of the random matrices (→ I/1.2.4) is $p = 1$.

Proof: Let X be an $n \times p$ real matrix and let Y be a $p \times p$ positive-definite symmetric matrix, such that X and Y jointly follow a normal-Wishart distribution (→ II/5.3.1):

$$
X, Y \sim \mathrm{NW}(M, U, V, \nu) \; . \tag{1}
$$

Then, X and Y are described by the probability density function (→ II/5.3.2)

$$
\begin{split}
p(X, Y) = \; & \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^\nu}} \cdot \frac{\sqrt{2^{-\nu p}}}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot |Y|^{(\nu+n-p-1)/2} \cdot \\
& \exp\left[-\frac{1}{2} \mathrm{tr}\left(Y \left[(X-M)^\mathrm{T} \, U^{-1} (X-M) + V^{-1} \right] \right) \right]
\end{split} \tag{2}
$$

where $|A|$ is a matrix determinant, A^{-1} is a matrix inverse and $\Gamma_p(x)$ is the multivariate gamma function of order p. If $p = 1$, then $\Gamma_p(x) = \Gamma(x)$ is the ordinary gamma function, $x = X$ is a column vector and $y = Y$ is a real number. Thus, the probability density function (→ I/1.7.1) of x and y can be developed as

$$
\begin{aligned}
p(x,y) &= \frac{1}{\sqrt{(2\pi)^n |U||V|^\nu}} \cdot \frac{\sqrt{2^{-\nu}}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot y^{(\nu+n-2)/2} \cdot \\
&\quad \exp\left[-\frac{1}{2}\mathrm{tr}\left(y\left[(x-M)^\mathrm{T} U^{-1}(x-M) + V^{-1}\right]\right)\right] \\
&= \sqrt{\frac{|U^{-1}|}{(2\pi)^n}} \cdot \frac{\sqrt{(2|V|)^{-\nu}}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot y^{\frac{\nu}{2}+\frac{n}{2}-1} \cdot \\
&\quad \exp\left[-\frac{1}{2}\left(y\left[(x-M)^\mathrm{T} U^{-1}(x-M) + 2\,(2V)^{-1}\right]\right)\right] \\
&= \sqrt{\frac{|U^{-1}|}{(2\pi)^n}} \cdot \frac{\left(\frac{1}{2|V|}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} \cdot y^{\frac{\nu}{2}+\frac{n}{2}-1} \cdot \\
&\quad \exp\left[-\frac{y}{2}\left((x-M)^\mathrm{T} U^{-1}(x-M) + 2\left(\frac{1}{2V}\right)\right)\right]
\end{aligned} \tag{3}
$$

In the matrix-normal distribution (→ II/5.1.1), we have $M \in \mathbb{R}^{n \times p}$, $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{p \times p}$ and $\nu \in \mathbb{R}$. Thus, with $p = 1$, M becomes a column vector and V becomes a real number, such that $V = |V| = 1/V^{-1}$. Finally, substituting $\mu = M$, $\Lambda = U^{-1}$, $a = \frac{\nu}{2}$ and $b = \frac{1}{2V}$, we get

$$
p(x,y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp\left[-\frac{y}{2}\left((x-\mu)^\mathrm{T}\Lambda(x-\mu) + 2b\right)\right] \tag{4}
$$

which is the probability density function of the normal-gamma distribution (→ II/4.3.3).

■

4.3.3 Probability density function

Theorem: Let x and y follow a normal-gamma distribution (→ II/4.3.1):

$$
x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}
$$

Then, the joint probability (→ I/1.3.2) density function (→ I/1.7.1) of x and y is

$$
p(x,y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp\left[-\frac{y}{2}\left((x-\mu)^\mathrm{T}\Lambda(x-\mu) + 2b\right)\right] \; . \tag{2}
$$

Proof: The normal-gamma distribution (→ II/4.3.1) is defined as X conditional on Y following a multivariate distribution (→ II/4.1.1) and Y following a gamma distribution (→ II/3.4.1):

$$
\begin{aligned}
X|Y &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) \\
Y &\sim \mathrm{Gam}(a, b) \; .
\end{aligned} \tag{3}
$$

Thus, using the probability density function of the multivariate normal distribution (→ II/4.1.7) and the probability density function of the gamma distribution (→ II/3.4.7), we have the following probabilities:

$$
\begin{aligned}
p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\
&= \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \\
p(y) &= \mathrm{Gam}(y; a, b) \\
&= \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right] \; .
\end{aligned} \tag{4}
$$

The law of conditional probability (→ I/1.3.4) implies that

$$
p(x,y) = p(x|y) \, p(y) \; , \tag{5}
$$

such that the normal-gamma density function becomes:

$$
p(x,y) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right] \; . \tag{6}
$$

Using the relation $|yA| = y^n|A|$ for an $n \times n$ matrix A and rearranging the terms, we have:

$$
p(x,y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp\left[-\frac{y}{2}\left((x-\mu)^\mathrm{T}\Lambda(x-\mu) + 2b\right)\right] \; . \tag{7}
$$

■

Sources:

- Koch KR (2007): "Normal-Gamma Distribution"; in: *Introduction to Bayesian Statistics*, ch. 2.5.3, pp. 55-56, eq. 2.212; URL: https://www.springer.com/gp/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

4.3.4 Mean

Theorem: Let $x \in \mathbb{R}^n$ and $y > 0$ follow a normal-gamma distribution (→ II/4.3.1):

$$
x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}
$$

Then, the expected value (→ I/1.10.1) of x and y is

$$
\mathrm{E}[(x,y)] = \left(\mu, \frac{a}{b}\right) \; . \tag{2}
$$

Proof: Consider the random vector (→ I/1.2.3)

$$
\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ y \end{bmatrix} \; . \tag{3}
$$

According to the expected value of a random vector (→ I/1.10.13), its expected value is

$$\mathrm{E}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}(x_1) \\ \vdots \\ \mathrm{E}(x_n) \\ \mathrm{E}(y) \end{bmatrix} = \begin{bmatrix} \mathrm{E}(x) \\ \mathrm{E}(y) \end{bmatrix} . \tag{4}$$

When x and y are jointly normal-gamma distributed, then (→ II/4.3.1) by definition x follows a multivariate normal distribution (→ II/4.1.1) conditional on y and y follows a univariate gamma distribution (→ II/3.4.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \quad \Leftrightarrow \quad x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) \quad \wedge \quad y \sim \mathrm{Gam}(a, b) . \tag{5}$$

Thus, with the expected value of the multivariate normal distribution (→ II/4.1.9) and the law of conditional probability (→ I/1.3.4), $\mathrm{E}(x)$ becomes

$$\begin{aligned} \mathrm{E}(x) &= \iint x \cdot p(x, y) \, \mathrm{d}x \, \mathrm{d}y \\ &= \iint x \cdot p(x|y) \cdot p(y) \, \mathrm{d}x \, \mathrm{d}y \\ &= \int p(y) \int x \cdot p(x|y) \, \mathrm{d}x \, \mathrm{d}y \\ &= \int p(y) \, \langle x \rangle_{\mathcal{N}(\mu, (y\Lambda)^{-1})} \, \mathrm{d}y \\ &= \int p(y) \cdot \mu \, \mathrm{d}y \\ &= \mu \int p(y) \, \mathrm{d}y \\ &= \mu , \end{aligned} \tag{6}$$

and with the expected value of the gamma distribution (→ II/3.4.11), $\mathrm{E}(y)$ becomes

$$\begin{aligned} \mathrm{E}(y) &= \int y \cdot p(y) \, \mathrm{d}y \\ &= \langle y \rangle_{\mathrm{Gam}(a,b)} \\ &= \frac{a}{b} . \end{aligned} \tag{7}$$

Thus, the expectation of the random vector in equations (3) and (4) is

$$\mathrm{E}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \mu \\ a/b \end{bmatrix} , \tag{8}$$

as indicated by equation (2).

■

4.3.5 Covariance

Theorem: Let $x \in \mathbb{R}^n$ and $y > 0$ follow a normal-gamma distribution (→ II/4.3.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then,
1) the covariance (→ I/1.13.1) of x, conditional (→ I/1.5.4) on y is

$$\mathrm{Cov}(x|y) = \frac{1}{y}\Lambda^{-1} \; ; \tag{2}$$

2) the covariance (→ I/1.13.1) of x, unconditional (→ I/1.5.3) on y is

$$\mathrm{Cov}(x) = \frac{b}{a-1}\Lambda^{-1} \; ; \tag{3}$$

3) the variance (→ I/1.11.1) of y is

$$\mathrm{Var}(y) = \frac{a}{b^2} \; . \tag{4}$$

Proof:
1) According to the definition of the normal-gamma distribution (→ II/4.3.1), the distribution of x given y is a multivariate normal distribution (→ II/4.1.1):

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) \; . \tag{5}$$

The covariance of the multivariate normal distribution (→ II/4.1.10) is

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \mathrm{Cov}(x) = \Sigma \; , \tag{6}$$

such that we have:

$$\mathrm{Cov}(x|y) = (y\Lambda)^{-1} = \frac{1}{y}\Lambda^{-1} \; . \tag{7}$$

2) The marginal distribution of the normal-gamma distribution (→ II/4.3.8) with respect to x is a multivariate t-distribution (→ II/4.2.1):

$$x \sim t\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \; . \tag{8}$$

The covariance of the multivariate t-distribution is

$$x \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad \mathrm{Cov}(x) = \frac{\nu}{\nu-2}\Sigma \; , \tag{9}$$

such that we have:

$$\mathrm{Cov}(x) = \frac{2a}{2a-2}\left(\frac{a}{b}\Lambda\right)^{-1} = \frac{a}{a-1}\,\frac{b}{a}\,\Lambda^{-1} = \frac{b}{a-1}\Lambda^{-1} \; . \tag{10}$$

3) The marginal distribution of the normal-gamma distribution (→ II/4.3.8) with respect to y is a univariate gamma distribution (→ II/3.4.1):

$$y \sim \mathrm{Gam}(a, b) \; . \tag{11}$$

The variance of the gamma distribution (→ II/3.4.12) is

$$x \sim \mathrm{Gam}(a,b) \quad \Rightarrow \quad \mathrm{Var}(x) = \frac{a}{b^2} \; , \tag{12}$$

such that we have:

$$\mathrm{Var}(y) = \frac{a}{b^2} \; . \tag{13}$$

■

4.3.6 Differential entropy

Theorem: Let x be an $n \times 1$ random vector (→ I/1.2.3) and let y be a positive random variable (→ I/1.2.2). Assume that x and y are jointly normal-gamma distributed:

$$(x,y) \sim \mathrm{NG}(\mu, \Lambda^{-1}, a, b) \tag{1}$$

Then, the differential entropy (→ I/2.2.1) of x in nats is

$$\begin{split} \mathrm{h}(x,y) &= \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Lambda| + \frac{1}{2}n \\ &+ a + \ln\Gamma(a) - \frac{n-2+2a}{2}\psi(a) + \frac{n-2}{2}\ln b \; . \end{split} \tag{2}$$

Proof: The probabibility density function of the normal-gamma distribution (→ II/4.3.3) is

$$p(x,y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \; . \tag{3}$$

The differential entropy of the multivariate normal distribution (→ II/4.1.11) is

$$\mathrm{h}(x) = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}n \tag{4}$$

and the differential entropy of the univariate gamma distribution (→ II/3.4.15) is

$$\mathrm{h}(y) = a + \ln\Gamma(a) + (1-a)\cdot\psi(a) - \ln b \tag{5}$$

where $\Gamma(x)$ is the gamma function and $\psi(x)$ is the digamma function.

The differential entropy of a continuous random variable (→ I/2.2.1) in nats is given by

$$\mathrm{h}(Z) = -\int_{\mathcal{Z}} p(z)\ln p(z)\,\mathrm{d}z \tag{6}$$

which, applied to the normal-gamma distribution (→ II/4.3.1) over x and y, yields

$$\mathrm{h}(x,y) = -\int_0^\infty \int_{\mathbb{R}^n} p(x,y)\,\ln p(x,y)\,\mathrm{d}x\,\mathrm{d}y \; . \tag{7}$$

Using the law of conditional probability (→ I/1.3.4), this can be evaluated as follows:

$$
\begin{aligned}
\mathrm{h}(x,y) &= -\int_0^\infty \int_{\mathbb{R}^n} p(x|y)\,p(y)\,\ln p(x|y)\,p(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= -\int_0^\infty \int_{\mathbb{R}^n} p(x|y)\,p(y)\,\ln p(x|y)\,\mathrm{d}x\,\mathrm{d}y - \int_0^\infty \int_{\mathbb{R}^n} p(x|y)\,p(y)\,\ln p(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^n} p(x|y)\,\ln p(x|y)\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty p(y)\,\ln p(y) \int_{\mathbb{R}^n} p(x|y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left\langle \mathrm{h}(x|y) \right\rangle_{p(y)} + \mathrm{h}(y) \; .
\end{aligned} \tag{8}
$$

In other words, the differential entropy of the normal-gamma distribution over x and y is equal to the sum of a multivariate normal entropy regarding x conditional on y, expected over y, and a univariate gamma entropy regarding y.

From equations (3) and (4), the first term becomes

$$
\begin{aligned}
\left\langle \mathrm{h}(x|y) \right\rangle_{p(y)} &= \left\langle \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|(y\Lambda)^{-1}| + \frac{1}{2}n \right\rangle_{p(y)} \\
&= \left\langle \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|(y\Lambda)| + \frac{1}{2}n \right\rangle_{p(y)} \\
&= \left\langle \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(y^n|\Lambda|) + \frac{1}{2}n \right\rangle_{p(y)} \\
&= \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Lambda| + \frac{1}{2}n - \left\langle \frac{n}{2}\ln y \right\rangle_{p(y)}
\end{aligned} \tag{9}
$$

and using the relation (→ II/3.4.13) $y \sim \mathrm{Gam}(a,b) \Rightarrow \langle \ln y \rangle = \psi(a) - \ln(b)$, we have

$$
\left\langle \mathrm{h}(x|y) \right\rangle_{p(y)} = \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Lambda| + \frac{1}{2}n - \frac{n}{2}\psi(a) + \frac{n}{2}\ln b \; . \tag{10}
$$

By plugging (10) and (5) into (8), one arrives at the differential entropy given by (2).

■

4.3.7 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (→ I/1.2.3) and let y be a positive random variable (→ I/1.2.2). Assume two normal-gamma distributions (→ II/4.3.1) P and Q specifying the joint distribution of x and y as

$$
\begin{aligned}
P &: \; (x,y) \sim \mathrm{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\
Q &: \; (x,y) \sim \mathrm{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) \; .
\end{aligned} \tag{1}
$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] = &\frac{1}{2}\frac{a_1}{b_1}\left[(\mu_2-\mu_1)^\mathrm{T}\Lambda_2(\mu_2-\mu_1)\right] + \frac{1}{2}\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \\
&+ a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \; .
\end{aligned} \tag{2}
$$

Proof: The probabibility density function of the normal-gamma distribution (→ II/4.3.3) is

$$p(x,y) = p(x|y) \cdot p(y) = \mathcal{N}(x;\mu,(y\Lambda)^{-1}) \cdot \mathrm{Gam}(y;a,b) \; . \tag{3}$$

The Kullback-Leibler divergence of the multivariate normal distribution (→ II/4.1.12) is

$$\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[(\mu_2-\mu_1)^\mathrm{T}\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \tag{4}$$

and the Kullback-Leibler divergence of the univariate gamma distribution (→ II/3.4.16) is

$$\mathrm{KL}[P\,||\,Q] = a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \tag{5}$$

where $\Gamma(x)$ is the gamma function and $\psi(x)$ is the digamma function.

The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P\,||\,Q] = \int_{\mathcal{Z}} p(z)\,\ln\frac{p(z)}{q(z)}\,\mathrm{d}z \tag{6}$$

which, applied to the normal-gamma distribution (→ II/4.3.1) over x and y, yields

$$\mathrm{KL}[P\,||\,Q] = \int_0^\infty \int_{\mathbb{R}^n} p(x,y)\,\ln\frac{p(x,y)}{q(x,y)}\,\mathrm{d}x\,\mathrm{d}y \; . \tag{7}$$

Using the law of conditional probability (→ I/1.3.4), this can be evaluated as follows:

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \int_0^\infty \int_{\mathbb{R}^n} p(x|y)\,p(y)\,\ln\frac{p(x|y)\,p(y)}{q(x|y)\,q(y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty \int_{\mathbb{R}^n} p(x|y)\,p(y)\,\ln\frac{p(x|y)}{q(x|y)}\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty \int_{\mathbb{R}^n} p(x|y)\,p(y)\,\ln\frac{p(y)}{q(y)}\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^n} p(x|y)\,\ln\frac{p(x|y)}{q(x|y)}\,\mathrm{d}x\,\mathrm{d}y + \int_0^\infty p(y)\,\ln\frac{p(y)}{q(y)} \int_{\mathbb{R}^n} p(x|y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} + \mathrm{KL}[p(y)\,||\,q(y)] \; .
\end{split} \tag{8}$$

In other words, the KL divergence between two normal-gamma distributions over x and y is equal to the sum of a multivariate normal KL divergence regarding x conditional on y, expected over y, and a univariate gamma KL divergence regarding y.

From equations (3) and (4), the first term becomes

$$\begin{split}
&\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} \\
&= \left\langle \frac{1}{2}\left[(\mu_2-\mu_1)^\mathrm{T}(y\Lambda_2)(\mu_2-\mu_1) + \mathrm{tr}\left((y\Lambda_2)(y\Lambda_1)^{-1}\right) - \ln\frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - n\right] \right\rangle_{p(y)} \\
&= \left\langle \frac{y}{2}(\mu_2-\mu_1)^\mathrm{T}\Lambda_2(\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \right\rangle_{p(y)}
\end{split} \tag{9}$$

and using the relation (→ II/3.4.11) $y \sim \mathrm{Gam}(a,b) \Rightarrow \langle y \rangle = a/b$, we have

$$\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \rangle_{p(y)} = \frac{1}{2}\frac{a_1}{b_1}(\mu_2-\mu_1)^\mathrm{T}\Lambda_2(\mu_2-\mu_1) + \frac{1}{2}\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \; . \tag{10}$$

By plugging (10) and (5) into (8), one arrives at the KL divergence given by (2).

■

Sources:
- Soch J, Allefeld A (2016): "Kullback-Leibler Divergence for the Normal-Gamma Distribution"; in: *arXiv math.ST*, 1611.01437; URL: https://arxiv.org/abs/1611.01437.

4.3.8 Marginal distributions

Theorem: Let x and y follow a normal-gamma distribution (→ II/4.3.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then, the marginal distribution (→ I/1.5.3) of y is a gamma distribution (→ II/3.4.1)

$$y \sim \mathrm{Gam}(a, b) \tag{2}$$

and the marginal distribution (→ I/1.5.3) of x is a multivariate t-distribution (→ II/4.2.1)

$$x \sim t\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \; . \tag{3}$$

Proof: The probability density function of the normal-gamma distribution (→ II/4.3.3) is given by

$$\begin{aligned} p(x,y) &= p(x|y) \cdot p(y) \\ p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\ p(y) &= \mathrm{Gam}(y; a, b) \; . \end{aligned} \tag{4}$$

Using the law of marginal probability (→ I/1.3.3), the marginal distribution of y can be derived as

$$\begin{aligned} p(y) &= \int p(x,y) \, \mathrm{d}x \\ &= \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \, \mathrm{Gam}(y; a, b) \, \mathrm{d}x \\ &= \mathrm{Gam}(y; a, b) \int \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \, \mathrm{d}x \\ &= \mathrm{Gam}(y; a, b) \end{aligned} \tag{5}$$

which is the probability density function of the gamma distribution (→ II/3.4.7) with shape parameter a and rate parameter b.

Using the law of marginal probability (→ I/1.3.3), the marginal distribution of x can be derived as

$$
\begin{aligned}
p(x) &= \int p(x,y)\,\mathrm{d}y \\
&= \int \mathcal{N}(x;\mu,(y\Lambda)^{-1})\,\mathrm{Gam}(y;a,b)\,\mathrm{d}y \\
&= \int \sqrt{\frac{|y\Lambda|}{(2\pi)^n}}\,\exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right]\cdot\frac{b^a}{\Gamma(a)}\,y^{a-1}\exp[-by]\,\mathrm{d}y \\
&= \int \sqrt{\frac{y^n|\Lambda|}{(2\pi)^n}}\,\exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right]\cdot\frac{b^a}{\Gamma(a)}\,y^{a-1}\exp[-by]\,\mathrm{d}y \\
&= \int \sqrt{\frac{|\Lambda|}{(2\pi)^n}}\cdot\frac{b^a}{\Gamma(a)}\cdot y^{a+\frac{n}{2}-1}\cdot\exp\left[-\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)y\right]\mathrm{d}y \\
&= \int \sqrt{\frac{|\Lambda|}{(2\pi)^n}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}\cdot\mathrm{Gam}\left(y;a+\frac{n}{2},b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\mathrm{d}y \\
&= \sqrt{\frac{|\Lambda|}{(2\pi)^n}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}\int\mathrm{Gam}\left(y;a+\frac{n}{2},b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\mathrm{d}y \\
&= \sqrt{\frac{|\Lambda|}{(2\pi)^n}}\cdot\frac{b^a}{\Gamma(a)}\cdot\frac{\Gamma\left(a+\frac{n}{2}\right)}{\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}} \\
&= \frac{\sqrt{|\Lambda|}}{(2\pi)^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot b^a\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\left(a+\frac{n}{2}\right)} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(\frac{1}{b}\right)^{-a}\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-a}\cdot 2^{-\frac{n}{2}}\cdot\left(b+\frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2b}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-a}\cdot\left(2b+(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{|\Lambda|}}{\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(\frac{1}{2a}\right)^{-a}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot\left(\frac{b}{a}\right)^{-\frac{n}{2}}\cdot\Big(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{-a}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot\left(2a+(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{n}{2}} \\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{-a}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot(2a)^{-a}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-a}\cdot(2a)^{-\frac{n}{2}}\cdot\Big(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)\Big(\\
&= \frac{\sqrt{\left(\frac{a}{b}\right)^n|\Lambda|}}{(2a)^{\frac{n}{2}}\,\pi^{\frac{n}{2}}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}} \\
&= \sqrt{\frac{\left|\frac{a}{b}\Lambda\right|}{(2a\,\pi)^n}}\cdot\frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)}\cdot\left(1+\frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}}
\end{aligned}
\tag{6}
$$

which is the probability density function of a multivariate t-distribution (→ II/4.2.2) with mean vector μ, shape matrix $\left(\frac{a}{b}\Lambda\right)^{-1}$ and $2a$ degrees of freedom.

■

4.3.9 Conditional distributions

Theorem: Let x and y follow a normal-gamma distribution (→ II/4.3.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then,
1) the conditional distribution (→ I/1.5.4) of x given y is a multivariate normal distribution (→ II/4.1.1)

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) \; ; \tag{2}$$

2) the conditional distribution (→ I/1.5.4) of a subset vector x_1, given the complement vector x_2 and y, is also a multivariate normal distribution (→ II/4.1.1)

$$x_1|x_2, y \sim \mathcal{N}(\mu_{1|2}(y), \Sigma_{1|2}(y)) \tag{3}$$

with the conditional mean (→ I/1.10.1) and covariance (→ I/1.13.1)

$$\begin{aligned} \mu_{1|2}(y) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2}(y) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} \end{aligned} \tag{4}$$

where μ_1, μ_2 and Σ_{11}, Σ_{12}, Σ_{22}, Σ_{21} are block-wise components (→ II/4.1.15) of μ and $\Sigma(y) = (y\Lambda)^{-1}$;
3) the conditional distribution (→ I/1.5.4) of y given x is a gamma distribution (→ II/3.4.1)

$$y|x \sim \mathrm{Gam}\left(a + \frac{n}{2}, b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right) \tag{5}$$

where n is the dimensionality of x.

Proof:
1) This follows from the definition of the normal-gamma distribution (→ II/4.3.1):

$$\begin{aligned} p(x, y) &= p(x|y) \cdot p(y) \\ &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \; . \end{aligned} \tag{6}$$

2) This follows from (2) and the conditional distributions of the multivariate normal distribution (→ II/4.1.15):

$$\begin{aligned} x &\sim \mathcal{N}(\mu, \Sigma) \\ \Rightarrow x_1|x_2 &\sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \; . \end{aligned} \tag{7}$$

3) The conditional density of y given x follows from Bayes' theorem (→ I/5.3.1) as

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)} \; . \tag{8}$$

The conditional distribution (→ I/1.5.4) of x given y is a multivariate normal distribution (→ II/4.3.3)

$$p(x|y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \; , \tag{9}$$

the marginal distribution (→ I/1.5.3) of y is a gamma distribution (→ II/4.3.8)

$$p(y) = \mathrm{Gam}(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right] \tag{10}$$

and the marginal distribution (→ I/1.5.3) of x is a multivariate t-distribution (→ II/4.3.8)

$$\begin{split} p(x) &= t\left(x; \mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \\ &= \sqrt{\frac{\left|\frac{a}{b}\,\Lambda\right|}{(2a\,\pi)^n}} \cdot \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)} \cdot \left(1 + \frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}} \\ &= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a+\frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left(b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\left(a+\frac{n}{2}\right)} \; . \end{split} \tag{11}$$

Plugging (9), (10) and (11) into (8), we obtain

$$\begin{split} p(y|x) &= \frac{\sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right]}{\sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a+\frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left(b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\left(a+\frac{n}{2}\right)}} \\ &= y^{\frac{n}{2}} \cdot \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \cdot y^{a-1} \cdot \exp\left[-by\right] \cdot \frac{1}{\Gamma\left(a+\frac{n}{2}\right)} \cdot \left(b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}} \\ &= \frac{\left(b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{a+\frac{n}{2}}}{\Gamma\left(a+\frac{n}{2}\right)} \cdot y^{a+\frac{n}{2}-1} \cdot \exp\left[-\left(b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)\right] \end{split} \tag{12}$$

which is the probability density function of a gamma distribution (→ II/3.4.7) with shape and rate parameters

$$a + \frac{n}{2} \quad \text{and} \quad b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu) \; , \tag{13}$$

such that

$$p(y|x) = \mathrm{Gam}\left(y; a + \frac{n}{2}, b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right) \; . \tag{14}$$

■

4.3.10 Drawing samples

Theorem: Let $Z_1 \in \mathbb{R}^n$ be a random vector (→ I/1.2.3) with all entries independently following a standard normal distribution (→ II/3.2.3) and let $Z_2 \in \mathbb{R}$ be a random variable (→ I/1.2.2) following a standard gamma distribution (→ II/3.4.3) with shape a. Moreover, let $A \in \mathbb{R}^{n \times n}$ be a matrix, such that $AA^\mathrm{T} = \Lambda^{-1}$.
Then, $X = \mu + AZ_1/\sqrt{Z_2/b}$ and $Y = Z_2/b$ jointly follow a normal-gamma distribution (→ II/4.3.1) with mean vector (→ I/1.10.13) μ, precision matrix (→ I/1.13.19) Λ, shape parameter a and rate parameter b:

$$\left(X = \mu + AZ_1/\sqrt{Z_2/b}, \; Y = Z_2/b \right) \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Proof: If all entries of Z_1 are independent and standard normally distributed (→ II/3.2.3)

$$z_{1i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \quad \text{for all} \quad i = 1, \ldots, n \; , \tag{2}$$

this implies a multivariate normal distribution with diagonal covariance matrix (→ II/4.1.16):

$$Z_1 \sim \mathcal{N}\left(0_n, I_n\right) \tag{3}$$

where 0_n is an $n \times 1$ matrix of zeros and I_n is the $n \times n$ identity matrix.
If the distribution of Z_2 is a standard gamma distribution (→ II/3.4.3)

$$Z_2 \sim \mathrm{Gam}(a, 1) \; , \tag{4}$$

then due to the relationship between gamma and standard gamma distribution distribution (→ II/3.4.4), we have:

$$Y = \frac{Z_2}{b} \sim \mathrm{Gam}(a, b) \; . \tag{5}$$

Moreover, using the linear transformation theorem for the multivariate normal distribution (→ II/4.1.13), it follows that:

$$\begin{split} Z_1 &\sim \mathcal{N}\left(0_n, I_n\right) \\ X = \mu + \frac{1}{\sqrt{Z_2/b}} A Z_1 &\sim \mathcal{N}\left(\mu + \frac{1}{\sqrt{Z_2/b}} A \, 0_n, \left(\frac{1}{\sqrt{Z_2/b}} A \right) I_n \left(\frac{1}{\sqrt{Z_2/b}} A \right)^\mathrm{T} \right) \\ X &\sim \mathcal{N}\left(\mu + 0_n, \left(\frac{1}{\sqrt{Y}} \right)^2 AA^\mathrm{T} \right) \\ X &\sim \mathcal{N}\left(\mu, (Y\Lambda)^{-1} \right) \; . \end{split} \tag{6}$$

Thus, Y follows a gamma distribution (→ II/3.4.1) and the distribution of X conditional on Y is a multivariate normal distribution (→ II/4.1.1):

$$\begin{split} X|Y &\sim \mathcal{N}(\mu, (Y\Lambda)^{-1}) \\ Y &\sim \mathrm{Gam}(a, b) \; . \end{split} \tag{7}$$

This means that, by definition (→ II/4.3.1), X and Y jointly follow a normal-gamma distribution (→ II/4.3.1):

$$X, Y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; , \tag{8}$$

Thus, given Z_1 defined by (2) and Z_2 defined by (4), X and Y defined by (1) are a sample from $\mathrm{NG}(\mu, \Lambda, a, b)$.

■

Sources:
- Wikipedia (2022): "Normal-gamma distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-22; URL: https://en.wikipedia.org/wiki/Normal-gamma_distribution#Generating_normal-gamma_random_variates.

4.4 Dirichlet distribution

4.4.1 Definition

Definition: Let X be a $k \times 1$ random vector (→ I/1.2.3). Then, X is said to follow a Dirichlet distribution with concentration parameters $\alpha = [\alpha_1, \ldots, \alpha_k]$

$$X \sim \mathrm{Dir}(\alpha) \; , \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathrm{Dir}(x; \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} {x_i}^{\alpha_i - 1} \tag{2}$$

where $\alpha_i > 0$ for all $i = 1, \ldots, k$, and the density is zero, if $x_i \notin [0, 1]$ for any $i = 1, \ldots, k$ or $\sum_{i=1}^{k} x_i \neq 1$.

Sources:
- Wikipedia (2020): "Dirichlet distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-05-10; URL: https://en.wikipedia.org/wiki/Dirichlet_distribution#Probability_density_function.

4.4.2 Probability density function

Theorem: Let X be a random vector (→ I/1.2.3) following a Dirichlet distribution (→ II/4.4.1):

$$X \sim \mathrm{Dir}(\alpha) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is

$$f_X(x) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} {x_i}^{\alpha_i - 1} \; . \tag{2}$$

Proof: This follows directly from the definition of the Dirichlet distribution (→ II/4.4.1).

■

4.4.3 Kullback-Leibler divergence

Theorem: Let x be an $k \times 1$ random vector (→ I/1.2.3). Assume two Dirichlet distributions (→ II/4.4.1) P and Q specifying the probability distribution of x as

$$
\begin{split}
P &: \; x \sim \mathrm{Dir}(\alpha_1) \\
Q &: \; x \sim \mathrm{Dir}(\alpha_2) \; .
\end{split} \tag{1}
$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$
\mathrm{KL}[P\,||\,Q] = \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \left[\psi(\alpha_{1i}) - \psi\left(\sum_{i=1}^k \alpha_{1i} \right) \right] \; . \tag{2}
$$

Proof: The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$
\mathrm{KL}[P\,||\,Q] = \int_{\mathcal{X}} p(x) \, \ln \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{3}
$$

which, applied to the Dirichlet distributions (→ II/4.1.1) in (1), yields

$$
\begin{split}
\mathrm{KL}[P\,||\,Q] &= \int_{\mathcal{X}^k} \mathrm{Dir}(x;\alpha_1) \, \ln \frac{\mathrm{Dir}(x;\alpha_1)}{\mathrm{Dir}(x;\alpha_2)} \, \mathrm{d}x \\
&= \left\langle \ln \frac{\mathrm{Dir}(x;\alpha_1)}{\mathrm{Dir}(x;\alpha_2)} \right\rangle_{p(x)}
\end{split} \tag{4}
$$

where $\mathcal{X}^k$ is the set $\left\{ x \in \mathbb{R}^k \mid \sum_{i=1}^k x_i = 1, \; 0 \leq x_i \leq 1, \; i = 1, \ldots, k \right\}$.
Using the probability density function of the Dirichlet distribution (→ II/4.4.2), this becomes:

$$
\begin{split}
\mathrm{KL}[P\,||\,Q] &= \left\langle \ln \frac{\frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\prod_{i=1}^k \Gamma(\alpha_{1i})} \prod_{i=1}^k {x_i}^{\alpha_{1i}-1}}{\frac{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)}{\prod_{i=1}^k \Gamma(\alpha_{2i})} \prod_{i=1}^k {x_i}^{\alpha_{2i}-1}} \right\rangle_{p(x)} \\
&= \left\langle \ln \left(\frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} \cdot \frac{\prod_{i=1}^k \Gamma(\alpha_{2i})}{\prod_{i=1}^k \Gamma(\alpha_{1i})} \cdot \prod_{i=1}^k {x_i}^{\alpha_{1i}-\alpha_{2i}} \right) \right\rangle_{p(x)} \\
&= \left\langle \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \cdot \ln(x_i) \right\rangle_{p(x)} \\
&= \ln \frac{\Gamma\left(\sum_{i=1}^k \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^k \alpha_{2i}\right)} + \sum_{i=1}^k \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^k (\alpha_{1i} - \alpha_{2i}) \cdot \left\langle \ln x_i \right\rangle_{p(x)} \; .
\end{split} \tag{5}
$$

Using the expected value of a logarithmized Dirichlet variate

$$x \sim \mathrm{Dir}(\alpha) \quad \Rightarrow \quad \langle \ln x_i \rangle = \psi(\alpha_i) - \psi\left(\sum_{i=1}^{k} \alpha_i \right) \; , \tag{6}$$

the Kullback-Leibler divergence from (5) becomes:

$$\mathrm{KL}[P\,||\,Q] = \ln \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_{1i}\right)}{\Gamma\left(\sum_{i=1}^{k} \alpha_{2i}\right)} + \sum_{i=1}^{k} \ln \frac{\Gamma(\alpha_{2i})}{\Gamma(\alpha_{1i})} + \sum_{i=1}^{k} (\alpha_{1i} - \alpha_{2i}) \cdot \left[\psi(\alpha_{1i}) - \psi\left(\sum_{i=1}^{k} \alpha_{1i} \right) \right] \tag{7}$$

■

Sources:
- Penny, William D. (2001): "KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities"; in: *University College, London*, p. 2, eqs. 8-9; URL: https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps.

4.4.4 Exceedance probabilities

Theorem: Let $r = [r_1, \ldots, r_k]$ be a random vector (→ I/1.2.3) following a Dirichlet distribution (→ II/4.4.1) with concentration parameters $\alpha = [\alpha_1, \ldots, \alpha_k]$:

$$r \sim \mathrm{Dir}(\alpha) \; . \tag{1}$$

1) If $k = 2$, then the exceedance probability (→ I/1.3.10) for r_1 is

$$\varphi_1 = 1 - \frac{\mathrm{B}\left(\frac{1}{2}; \alpha_1, \alpha_2\right)}{\mathrm{B}(\alpha_1, \alpha_2)} \tag{2}$$

where $\mathrm{B}(x, y)$ is the beta function and $\mathrm{B}(x; a, b)$ is the incomplete beta function.

2) If $k > 2$, then the exceedance probability (→ I/1.3.10) for r_i is

$$\varphi_i = \int_0^\infty \prod_{j \neq i} \left(\frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \right) \frac{q_i^{\alpha_i - 1} \exp[-q_i]}{\Gamma(\alpha_i)} \, \mathrm{d}q_i \; . \tag{3}$$

where $\Gamma(x)$ is the gamma function and $\gamma(s, x)$ is the lowerr incomplete gamma function.

Proof: In the context of the Dirichlet distribution (→ II/4.4.1), the exceedance probability (→ I/1.3.10) for a particular r_i is defined as:

$$\begin{aligned} \varphi_i &= p\Big(\forall j \in \Big\{1, \ldots, k \,\Big|\, j \neq i\Big\} : \, r_i > r_j | \alpha\Big) \\ &= p\Big(\bigwedge_{j \neq i} r_i > r_j \,\Big|\, \alpha\Big) \; . \end{aligned} \tag{4}$$

The probability density function of the Dirichlet distribution (→ II/4.4.2) is given by:

$$\mathrm{Dir}(r; \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} {r_i}^{\alpha_i - 1} \; . \tag{5}$$

Note that the probability density function is only calculated, if

$$r_i \in [0,1] \quad \text{for} \quad i = 1, \ldots, k \quad \text{and} \quad \sum_{i=1}^{k} r_i = 1 \; , \tag{6}$$

and defined to be zero otherwise (→ II/4.4.1).

1) If $k = 2$, the probability density function of the Dirichlet distribution (→ II/4.4.2) reduces to

$$p(r) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)}\, r_1^{\alpha_1 - 1}\, r_2^{\alpha_2 - 1} \tag{7}$$

which is equivalent to the probability density function of the beta distribution (→ II/3.9.3)

$$p(r_1) = \frac{r_1^{\alpha_1 - 1}\,(1 - r_1)^{\alpha_2 - 1}}{\mathrm{B}(\alpha_1, \alpha_2)} \tag{8}$$

with the beta function given by

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x + y)} \; . \tag{9}$$

With (6), the exceedance probability for this bivariate case simplifies to

$$\varphi_1 = p(r_1 > r_2) = p(r_1 > 1 - r_1) = p(r_1 > 1/2) = \int_{\frac{1}{2}}^{1} p(r_1)\, \mathrm{d}r_1 \; . \tag{10}$$

Using the cumulative distribution function of the beta distribution (→ II/3.9.5), it evaluates to

$$\varphi_1 = 1 - \int_0^{\frac{1}{2}} p(r_1)\, \mathrm{d}r_1 = 1 - \frac{\mathrm{B}\left(\frac{1}{2}; \alpha_1, \alpha_2\right)}{\mathrm{B}(\alpha_1, \alpha_2)} \tag{11}$$

with the incomplete beta function

$$\mathrm{B}(x; a, b) = \int_0^x x^{a-1}\,(1 - x)^{b-1}\, \mathrm{d}x \; . \tag{12}$$

2) If $k > 2$, there is no similarly simple expression, because in general

$$\varphi_i = p(r_i = \max(r)) > p(r_i > 1/2) \quad \text{for} \quad i = 1, \ldots, k \; , \tag{13}$$

i.e. exceedance probabilities cannot be evaluated using a simple threshold on r_i, because r_i might be the maximal element in r without being larger than 1/2. Instead, we make use of the relationship between the Dirichlet and the gamma distribution which states that

$$\begin{gathered} Y_1 \sim \mathrm{Gam}(\alpha_1, \beta), \; \ldots, \; Y_k \sim \mathrm{Gam}(\alpha_k, \beta), \; Y_s = \sum_{i=1}^{k} Y_j \\ \Rightarrow X = (X_1, \ldots, X_k) = \left(\frac{Y_1}{Y_s}, \ldots, \frac{Y_k}{Y_s}\right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k) \; . \end{gathered} \tag{14}$$

The probability density function of the gamma distribution (→ II/3.4.7) is given by

$$\mathrm{Gam}(x; a, b) = \frac{b^a}{\Gamma(a)} \, x^{a-1} \exp[-bx] \quad \text{for} \quad x > 0 \; . \tag{15}$$

Consider the gamma random variables (→ II/3.4.1)

$$q_1 \sim \mathrm{Gam}(\alpha_1, 1), \; \ldots, \; q_k \sim \mathrm{Gam}(\alpha_k, 1), \; q_s = \sum_{j=1}^{k} q_j \tag{16}$$

and the Dirichlet random vector (→ II/4.4.1)

$$r = (r_1, \ldots, r_k) = \left(\frac{q_1}{q_s}, \ldots, \frac{q_k}{q_s} \right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k) \; . \tag{17}$$

Obviously, it holds that

$$r_i > r_j \;\Leftrightarrow\; q_i > q_j \quad \text{for} \quad i, j = 1, \ldots, k \quad \text{with} \quad j \neq i \; . \tag{18}$$

Therefore, consider the probability that q_i is larger than q_j, given q_i is known. This probability is equal to the probability that q_j is smaller than q_i, given q_i is known

$$p(q_i > q_j | q_i) = p(q_j < q_i | q_i) \tag{19}$$

which can be expressed in terms of the cumulative distribution function of the gamma distribution (→ II/3.4.9) as

$$p(q_j < q_i | q_i) = \int_0^{q_i} \mathrm{Gam}(q_j; \alpha_j, 1) \, \mathrm{d}q_j = \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \tag{20}$$

where $\Gamma(x)$ is the gamma function and $\gamma(s, x)$ is the lower incomplete gamma function. Since the gamma variates are independent of each other, these probabilties factorize:

$$p(\forall_{j \neq i} \, [q_i > q_j] \, | q_i) = \prod_{j \neq i} p(q_i > q_j | q_i) = \prod_{j \neq i} \frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \; . \tag{21}$$

In order to obtain the exceedance probability φ_i, the dependency on q_i in this probability still has to be removed. From equations (4) and (18), it follows that

$$\varphi_i = p(\forall_{j \neq i} \, [r_i > r_j]) = p(\forall_{j \neq i} \, [q_i > q_j]) \; . \tag{22}$$

Using the law of marginal probability (→ I/1.3.3), we have

$$\varphi_i = \int_0^{\infty} p(\forall_{j \neq i} \, [q_i > q_j] \, | q_i) \, p(q_i) \, \mathrm{d}q_i \; . \tag{23}$$

With (21) and (16), this becomes

$$\varphi_i = \int_0^{\infty} \prod_{j \neq i} \left(p(q_i > q_j | q_i) \right) \cdot \mathrm{Gam}(q_i; \alpha_i, 1) \, \mathrm{d}q_i \; . \tag{24}$$

And with (20) and (15), it becomes

$$\varphi_i = \int_0^{\infty} \prod_{j \neq i} \left(\frac{\gamma(\alpha_j, q_i)}{\Gamma(\alpha_j)} \right) \cdot \frac{q_i^{\alpha_i - 1} \exp[-q_i]}{\Gamma(\alpha_i)} \, \mathrm{d}q_i \; . \tag{25}$$

In other words, the exceedance probability (→ I/1.3.10) for one element from a Dirichlet-distributed (→ II/4.4.1) random vector (→ I/1.2.3) is an integral from zero to infinity where the first term in the integrand conforms to a product of gamma (→ II/3.4.1) cumulative distribution functions (→ I/1.8.1) and the second term is a gamma (→ II/3.4.1) probability density function (→ I/1.7.1).

■

Sources:

- Soch J, Allefeld C (2016): "Exceedance Probabilities for the Dirichlet Distribution"; in: *arXiv stat.AP*, 1611.01439; URL: https://arxiv.org/abs/1611.01439.

5 Matrix-variate continuous distributions

5.1 Matrix-normal distribution

5.1.1 Definition

Definition: Let X be an $n \times p$ random matrix (→ I/1.2.4). Then, X is said to be matrix-normally distributed with mean M, covariance (→ I/1.13.9) across rows U and covariance (→ I/1.13.9) across columns V

$$X \sim \mathcal{MN}(M, U, V) \;, \tag{1}$$

if and only if its probability density function (→ I/1.7.1) is given by

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \tag{2}$$

where M is an $n \times p$ real matrix, U is an $n \times n$ positive definite matrix and V is a $p \times p$ positive definite matrix.

Sources:
- Wikipedia (2020): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-27; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Definition.

5.1.2 Equivalence to multivariate normal distribution

Theorem: The matrix X is matrix-normally distributed (→ II/5.1.1)

$$X \sim \mathcal{MN}(M, U, V) \;, \tag{1}$$

if and only if $\mathrm{vec}(X)$ is multivariate normally distributed (→ II/4.1.1)

$$\mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U) \tag{2}$$

where $\mathrm{vec}(X)$ is the vectorization operator and $\otimes$ is the Kronecker product.

Proof: The probability density function of the matrix-normal distribution (→ II/5.1.3) with $n \times p$ mean M, $n \times n$ covariance across rows U and $p \times p$ covariance across columns V is

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M) \right) \right] \;. \tag{3}$$

Using the trace property $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left((X-M)^\mathrm{T} \, U^{-1} (X-M) \, V^{-1} \right) \right] \;. \tag{4}$$

Using the trace-vectorization relation $\mathrm{tr}(A^\mathrm{T} B) = \mathrm{vec}(A)^\mathrm{T} \, \mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{vec}(X-M)^\mathrm{T} \, \mathrm{vec}\left(U^{-1} (X-M) \, V^{-1} \right) \right] \;. \tag{5}$$

Using the vectorization-Kronecker relation $\mathrm{vec}(ABC) = \left(C^\mathrm{T} \otimes A\right) \mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{vec}(X-M)^\mathrm{T} \left(V^{-1} \otimes U^{-1}\right) \mathrm{vec}(X-M)\right] \; . \tag{6}$$

Using the Kronecker product property $(A^{-1} \otimes B^{-1}) = (A \otimes B)^{-1}$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{vec}(X-M)^\mathrm{T} \left(V \otimes U\right)^{-1} \mathrm{vec}(X-M)\right] \; . \tag{7}$$

Using the vectorization property $\mathrm{vec}(A+B) = \mathrm{vec}(A) + \mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]^\mathrm{T} \left(V \otimes U\right)^{-1} \left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]\right] \; . \tag{8}$$

Using the Kronecker-determinant relation $|A \otimes B| = |A|^m |B|^n$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V \otimes U|}} \cdot \exp\left[-\frac{1}{2}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]^\mathrm{T} \left(V \otimes U\right)^{-1} \left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]\right] \; . \tag{9}$$

This is the probability density function of the multivariate normal distribution (→ II/4.1.7) with the $np \times 1$ mean vector $\mathrm{vec}(M)$ and the $np \times np$ covariance matrix $V \otimes U$:

$$\mathcal{MN}(X; M, U, V) = \mathcal{N}(\mathrm{vec}(X); \mathrm{vec}(M), V \otimes U) \; . \tag{10}$$

By showing that the probability density functions (→ I/1.7.1) are identical, it is proven that the associated probability distributions (→ I/1.5.1) are equivalent.

■

Sources:
- Wikipedia (2020): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-20; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Proof.

5.1.3 Probability density function

Theorem: Let X be a random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, the probability density function (→ I/1.7.1) of X is

$$f(X) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1} (X-M)^\mathrm{T} \, U^{-1} (X-M)\right)\right] \; . \tag{2}$$

Proof: This follows directly from the definition of the matrix-normal distribution (→ II/5.1.1).

■

5.1.4 Mean

Theorem: Let X be a random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, the mean or expected value (→ I/1.10.1) of X is

$$\mathrm{E}(X) = M \; . \tag{2}$$

Proof: When X follows a matrix-normal distribution (→ II/5.1.1), its vectorized version follows a multivariate normal distribution (→ II/5.1.2)

$$\mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U) \tag{3}$$

and the expected value of this multivariate normal distribution is (→ II/4.1.9)

$$\mathrm{E}[\mathrm{vec}(X)] = \mathrm{vec}(M) \; . \tag{4}$$

Since the expected value of a random matrix is calculated element-wise (→ I/1.10.14), we can invert the vectorization operator to get:

$$\mathrm{E}[X] = M \; . \tag{5}$$

■

Sources:
- Wikipedia (2022): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-15; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Expected_values.

5.1.5 Covariance

Theorem: Let X be an $n \times p$ random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then,
1) the covariance matrix (→ I/1.13.9) of each row of X is a scalar multiple of V

$$\mathrm{Cov}(x_{i,\bullet}^\mathrm{T}) \propto V \quad \text{for all} \quad i = 1, \ldots, n \; ; \tag{2}$$

2) the covariance matrix (→ I/1.13.9) of each column of X is a scalar multiple of U

$$\mathrm{Cov}(x_{\bullet,j}) \propto U \quad \text{for all} \quad i = 1, \ldots, p \; . \tag{3}$$

Proof:
1) The marginal distribution (→ I/1.5.3) of a given row of X is a multivariate normal distribution (→ II/5.1.10)

$$x_{i,\bullet}^\mathrm{T} \sim \mathcal{N}(m_{i,\bullet}^\mathrm{T}, u_{ii} V) \; , \tag{4}$$

and the covariance of this multivariate normal distribution (→ II/4.1.10) is

$$\mathrm{Cov}(x_{i,\bullet}^\mathrm{T}) = u_{ii} V \propto V \; . \tag{5}$$

2) The marginal distribution (→ I/1.5.3) of a given column of X is a multivariate normal distribution (→ II/5.1.10)

$$x_{\bullet,j} \sim \mathcal{N}(m_{\bullet,j}, v_{jj} U) \; , \tag{6}$$

and the covariance of this multivariate normal distribution (→ II/4.1.10) is

$$\mathrm{Cov}(x_{\bullet,j}) = v_{jj} U \propto U \; . \tag{7}$$

■

Sources:
- Wikipedia (2022): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-09-15; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Expected_values.

5.1.6 Differential entropy

Theorem: Let X be an $n \times p$ random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1)

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, the differential entropy (→ I/2.2.1) of X in nats is

$$\mathrm{h}(X) = \frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln|V| + \frac{p}{2} \ln|U| + \frac{np}{2} \; . \tag{2}$$

Proof: The matrix-normal distribution is equivalent to the multivariate normal distribution (→ II/5.1.2),

$$X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U) \; , \tag{3}$$

and the differential entropy for the multivariate normal distribution (→ II/4.1.11) in nats is

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \mathrm{h}(X) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln|\Sigma| + \frac{1}{2} n \tag{4}$$

where X is an $n \times 1$ random vector (→ I/1.2.3).
Thus, we can plug the distribution parameters from (1) into the differential entropy in (4) using the relationship given by (3)

$$\mathrm{h}(X) = \frac{np}{2} \ln(2\pi) + \frac{1}{2} \ln|V \otimes U| + \frac{1}{2} np \; . \tag{5}$$

Using the Kronecker product property

$$|A \otimes B| = |A|^m \, |B|^n \quad \text{where} \quad A \in \mathbb{R}^{n \times n} \quad \text{and} \quad B \in \mathbb{R}^{m \times m} \; , \tag{6}$$

the differential entropy from (5) becomes:

$$
\begin{aligned}
\mathrm{h}(X) &= \frac{np}{2}\ln(2\pi) + \frac{1}{2}\ln\left(|V|^n|U|^p\right) + \frac{1}{2}np \\
&= \frac{np}{2}\ln(2\pi) + \frac{n}{2}\ln|V| + \frac{p}{2}\ln|U| + \frac{np}{2} \; .
\end{aligned} \tag{7}
$$

■

5.1.7 Kullback-Leibler divergence

Theorem: Let X be an $n \times p$ random matrix (→ I/1.2.4). Assume two matrix-normal distributions (→ II/5.1.1) P and Q specifying the probability distribution of X as

$$
\begin{aligned}
P &: \; X \sim \mathcal{MN}(M_1, U_1, V_1) \\
Q &: \; X \sim \mathcal{MN}(M_2, U_2, V_2) \; .
\end{aligned} \tag{1}
$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] = \frac{1}{2} \Big[& \mathrm{vec}(M_2 - M_1)^\mathrm{T} \mathrm{vec}\left(U_2^{-1}(M_2 - M_1)V_2^{-1}\right) \\
& + \mathrm{tr}\left((V_2^{-1}V_1) \otimes (U_2^{-1}U_1)\right) - n\ln\frac{|V_1|}{|V_2|} - p\ln\frac{|U_1|}{|U_2|} - np \Big] \; .
\end{aligned} \tag{2}
$$

Proof: The matrix-normal distribution is equivalent to the multivariate normal distribution (→ II/5.1.2),

$$
X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U) \; , \tag{3}
$$

and the Kullback-Leibler divergence for the multivariate normal distribution (→ II/4.1.12) is

$$
\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \tag{4}
$$

where X is an $n \times 1$ random vector (→ I/1.2.3).
Thus, we can plug the distribution parameters from (1) into the KL divergence in (4) using the relationship given by (3)

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] = \frac{1}{2} \Big[& (\mathrm{vec}(M_2) - \mathrm{vec}(M_1))^T (V_2 \otimes U_2)^{-1} (\mathrm{vec}(M_2) - \mathrm{vec}(M_1)) \\
& + \mathrm{tr}\left((V_2 \otimes U_2)^{-1}(V_1 \otimes U_1)\right) - \ln\frac{|V_1 \otimes U_1|}{|V_2 \otimes U_2|} - np \Big] \; .
\end{aligned} \tag{5}
$$

Using the vectorization operator and Kronecker product properties

$$
\mathrm{vec}(A) + \mathrm{vec}(B) = \mathrm{vec}(A + B) \tag{6}
$$

$$
(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \tag{7}
$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{8}$$

$$|A \otimes B| = |A|^m \, |B|^n \quad \text{where} \quad A \in \mathbb{R}^{n \times n} \quad \text{and} \quad B \in \mathbb{R}^{m \times m} \; , \tag{9}$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{split} \mathrm{KL}[P \, || \, Q] = \frac{1}{2} \left[\mathrm{vec}(M_2 - M_1)^\mathrm{T} \, (V_2^{-1} \otimes U_2^{-1}) \, \mathrm{vec}(M_2 - M_1) \right. \\ \left. + \; \mathrm{tr} \left((V_2^{-1} V_1) \otimes (U_2^{-1} U_1) \right) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right] \; . \end{split} \tag{10}$$

Using the relationship between Kronecker product and vectorization operator

$$(C^\mathrm{T} \otimes A) \, \mathrm{vec}(B) = \mathrm{vec}(ABC) \; , \tag{11}$$

we finally have:

$$\begin{split} \mathrm{KL}[P \, || \, Q] = \frac{1}{2} \left[\mathrm{vec}(M_2 - M_1)^\mathrm{T} \mathrm{vec} \left(U_2^{-1} (M_2 - M_1) V_2^{-1} \right) \right. \\ \left. + \; \mathrm{tr} \left((V_2^{-1} V_1) \otimes (U_2^{-1} U_1) \right) - n \ln \frac{|V_1|}{|V_2|} - p \ln \frac{|U_1|}{|U_2|} - np \right] \; . \end{split} \tag{12}$$

■

5.1.8 Transposition

Theorem: Let X be a random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, the transpose of X also has a matrix-normal distribution:

$$X^\mathrm{T} \sim \mathcal{MN}(M^\mathrm{T}, V, U) \; . \tag{2}$$

Proof: The probability density function of the matrix-normal distribution (→ II/5.1.3) is:

$$f(X) = \mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr} \left(V^{-1} (X - M)^\mathrm{T} \, U^{-1} (X - M) \right) \right] \; . \tag{3}$$

Define $Y = X^\mathrm{T}$. Then, $X = Y^\mathrm{T}$ and we can substitute:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr} \left(V^{-1} (Y^\mathrm{T} - M)^\mathrm{T} \, U^{-1} (Y^\mathrm{T} - M) \right) \right] \; . \tag{4}$$

Using $(A + B)^\mathrm{T} = (A^\mathrm{T} + B^\mathrm{T})$, we have:

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr} \left(V^{-1} (Y - M^\mathrm{T}) \, U^{-1} (Y - M^\mathrm{T})^\mathrm{T} \right) \right] \; . \tag{5}$$

Using $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$, we obtain

$$f(Y) = \frac{1}{\sqrt{(2\pi)^{np}|V|^n|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(U^{-1}(Y-M^\mathrm{T})^\mathrm{T}\,V^{-1}(Y-M^\mathrm{T})\right)\right] \tag{6}$$

which is the probability density function of a matrix-normal distribution (→ II/5.1.3) with mean M^T, covariance across rows V and covariance across columns U.

■

5.1.9 Linear transformation

Theorem: Let X be an $n \times p$ random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, a linear transformation of X is also matrix-normally distributed

$$Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^\mathrm{T}, B^\mathrm{T}VB) \tag{2}$$

where A us ab $r \times n$ matrix of full rank $r \leq b$ and B is a $p \times s$ matrix of full rank $s \leq p$ and C is an $r \times s$ matrix.

Proof: The matrix-normal distribution is equivalent to the multivariate normal distribution (→ II/5.1.2),

$$X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U) \; , \tag{3}$$

and the linear transformation theorem for the multivariate normal distribution (→ II/4.1.13) states:

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathrm{T}) \; . \tag{4}$$

The vectorization of $Y = AXB + C$ is

$$\begin{aligned} \mathrm{vec}(Y) &= \mathrm{vec}(AXB + C) \\ &= \mathrm{vec}(AXB) + \mathrm{vec}(C) \\ &= (B^\mathrm{T} \otimes A)\mathrm{vec}(X) + \mathrm{vec}(C) \end{aligned} \tag{5}$$

and the Kronecker product obeys

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \; . \tag{6}$$

Using (3) and (4), we have

$$\begin{aligned} \mathrm{vec}(Y) &\sim \mathcal{N}((B^\mathrm{T} \otimes A)\mathrm{vec}(M) + \mathrm{vec}(C), (B^\mathrm{T} \otimes A)(V \otimes U)(B^\mathrm{T} \otimes A)^\mathrm{T}) \\ &= \mathcal{N}(\mathrm{vec}(AMB) + \mathrm{vec}(C), (B^\mathrm{T}V \otimes AU)(B^\mathrm{T} \otimes A)^\mathrm{T}) \\ &= \mathcal{N}(\mathrm{vec}(AMB + C), B^\mathrm{T}VB \otimes AUA^\mathrm{T}) \; . \end{aligned} \tag{7}$$

Using (3), we finally have:

$$Y \sim \mathcal{MN}(AMB + C, AUA^\mathrm{T}, B^\mathrm{T}VB) \; . \tag{8}$$

■

5.1.10 Marginal distributions

Theorem: Let X be an $n \times p$ random matrix (→ I/1.2.4) following a matrix-normal distribution (→ II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then,
1) the marginal distribution (→ I/1.5.3) of any subset matrix $X_{I,J}$, obtained by dropping some rows and/or columns from X, is also a matrix-normal distribution (→ II/5.1.1)

$$X_{I,J} \sim \mathcal{MN}(M_{I,J}, U_{I,I}, V_{J,J}) \tag{2}$$

where $I \subseteq \{1, \ldots, n\}$ is an (ordered) subset of all row indices and $J \subseteq \{1, \ldots, p\}$ is an (ordered) subset of all column indices, such that $M_{I,J}$ is the matrix dropping the irrelevant rows and columns (the ones not in the subset, i.e. marginalized out) from the mean matrix M; $U_{I,I}$ is the matrix dropping rows not in I from U; and $V_{J,J}$ is the matrix dropping columns not in J from V;
2) the marginal distribution (→ I/1.5.3) of each row vector is a multivariate normal distribution (→ II/4.1.1)

$$x_{i,\bullet}^{\mathrm{T}} \sim \mathcal{N}(m_{i,\bullet}^{\mathrm{T}}, u_{ii} V) \tag{3}$$

where $m_{i,\bullet}$ is the i-th row of M and u_{ii} is the i-th diagonal entry of U;
3) the marginal distribution (→ I/1.5.3) of each column vector is a multivariate normal distribution (→ II/4.1.1)

$$x_{\bullet,j} \sim \mathcal{N}(m_{\bullet,j}, v_{jj} U) \tag{4}$$

where $m_{\bullet,j}$ is the j-th column of M and v_{jj} is the j-th diagonal entry of V; and
4) the marginal distribution (→ I/1.5.3) of one element of X is a univariate normal distribution (→ II/3.2.1)

$$x_{ij} \sim \mathcal{N}(m_{ij}, u_{ii} v_{jj}) \tag{5}$$

where m_{ij} is the (i,j)-th entry of M.

Proof:
1) Define a selector matrix A, such that $a_{ij} = 1$, if the i-th row in the subset matrix should be the j-th row from the original matrix (and $a_{ij} = 0$ otherwise)

$$A \in \mathbb{R}^{|I| \times n}, \quad \text{s.t.} \quad a_{ij} = \begin{cases} 1 \; , & \text{if } I_i = j \\ 0 \; , & \text{otherwise} \end{cases} \tag{6}$$

and define a selector matrix B, such that $b_{ij} = 1$, if the j-th column in the subset matrix should be the i-th column from the original matrix (and $b_{ij} = 0$ otherwise)

$$B \in \mathbb{R}^{p \times |J|}, \quad \text{s.t.} \quad b_{ij} = \begin{cases} 1 \; , & \text{if } J_j = i \\ 0 \; , & \text{otherwise} \; . \end{cases} \tag{7}$$

Then, $X_{I,J}$ can be expressed as

$$X_{I,J} = AXB \tag{8}$$

and we can apply the linear transformation theorem (→ II/5.1.9) to give

$$X_{I,J} \sim \mathcal{MN}(AMB, AUA^{\mathrm{T}}, B^{\mathrm{T}}VB) \; . \tag{9}$$

Finally, we see that $AMB = M_{I,J}$, $AUA^{\mathrm{T}} = U_{I,I}$ and $B^{\mathrm{T}}VB = V_{J,J}$.
2) This is a special case of 1). Setting A to the i-th elementary row vector in n dimensions and B to the $p \times p$ identity matrix

$$A = e_i, \; B = I_p \; , \tag{10}$$

the i-th row of X can be expressed as

$$\begin{split} x_{i,\bullet} &= AXB = e_i X I_p = e_i X \\ &\overset{(9)}{\sim} \mathcal{MN}(m_{i,\bullet}, u_{ii}, V) \; . \end{split} \tag{11}$$

Thus, the transpose of the row vector is distributed as (→ II/5.1.8)

$$x_{i,\bullet}^{\mathrm{T}} \sim \mathcal{MN}(m_{i,\bullet}^{\mathrm{T}}, V, u_{ii}) \tag{12}$$

which is equivalent to a multivariate normal distribution (→ II/5.1.2):

$$x_{i,\bullet}^{\mathrm{T}} \sim \mathcal{N}(m_{i,\bullet}^{\mathrm{T}}, u_{ii}V) \; . \tag{13}$$

3) This is a special case of 1). Setting A to the $n \times n$ identity matrix and B to the j-th elementary row vector in p dimensions

$$A = I_n, \; B = e_j^{\mathrm{T}} \; , \tag{14}$$

the j-th column of X can be expressed as

$$\begin{split} x_{\bullet,j} &= AXB = I_n X e_j^{\mathrm{T}} = X e_j^{\mathrm{T}} \\ &\overset{(9)}{\sim} \mathcal{MN}(m_{\bullet,j}, U, v_{jj}) \end{split} \tag{15}$$

which is equivalent to a multivariate normal distribution (→ II/5.1.2):

$$x_{\bullet,j} \sim \mathcal{N}(m_{\bullet,j}, v_{jj}U) \; . \tag{16}$$

4) This is a special case of 2) and 3). Setting A to the i-th elementary row vector in n dimensions and B to the j-th elementary row vector in p dimensions

$$A = e_i, \; B = e_j^{\mathrm{T}} \; , \tag{17}$$

the (i,j)-th entry of X can be expressed as

$$\begin{split} x_{ij} &= AXB = e_i X e_j^{\mathrm{T}} \\ &\overset{(9)}{\sim} \mathcal{MN}(m_{ij}, u_{ii}, v_{jj}) \; . \end{split} \tag{18}$$

As x_{ij} is a scalar, this is equivalent to a univariate normal distribution (→ II/3.2.1) as a special case (→ II/3.2.2) of the matrix-normal distribution (→ II/4.1.2):

$$x_{ij} \sim \mathcal{N}(m_{ij}, u_{ii}v_{jj}) \; . \tag{19}$$

■

5.1.11 Drawing samples

Theorem: Let $X \in \mathbb{R}^{n \times p}$ be a random matrix (→ I/1.2.4) with all entries independently following a standard normal distribution (→ II/3.2.3). Moreover, let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{p \times p}$, such that $AA^\mathrm{T} = U$ and $B^\mathrm{T} B = V$.
Then, $Y = M + AXB$ follows a matrix-normal distribution (→ II/5.1.1) with mean (→ I/1.10.14) M, covariance (→ I/1.13.9) across rows U and covariance (→ I/1.13.9) across columns V:

$$Y = M + AXB \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Proof: If all entries of X are independent and standard normally distributed (→ II/3.2.3)

$$x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \quad \text{for all} \quad i = 1, \ldots, n \quad \text{and} \quad j = 1, \ldots, p \; , \tag{2}$$

this implies a multivariate normal distribution with diagonal covariance matrix (→ II/4.1.16):

$$\begin{aligned} \mathrm{vec}(X) &\sim \mathcal{N}\left(\mathrm{vec}(0_{np}), I_{np}\right) \\ &\sim \mathcal{N}\left(\mathrm{vec}(0_{np}), I_p \otimes I_n\right) \end{aligned} \tag{3}$$

where 0_{np} is an $n \times p$ matrix of zeros and I_n is the $n \times n$ identity matrix.
Due to the relationship between multivariate and matrix-normal distribution (→ II/5.1.2), we have:

$$X \sim \mathcal{MN}(0_{np}, I_n, I_p) \; . \tag{4}$$

Thus, with the linear transformation theorem for the matrix-normal distribution (→ II/5.1.9), it follows that

$$\begin{aligned} Y = M + AXB &\sim \mathcal{MN}\left(M + A0_{np}B, AI_nA^\mathrm{T}, B^\mathrm{T}I_pB\right) \\ &\sim \mathcal{MN}\left(M, AA^\mathrm{T}, B^\mathrm{T}B\right) \\ &\sim \mathcal{MN}\left(M, U, V\right) \; . \end{aligned} \tag{5}$$

Thus, given X defined by (2), Y defined by (1) is a sample from $\mathcal{MN}(M, U, V)$.

■

Sources:

- Wikipedia (2021): "Matrix normal distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-07; URL: https://en.wikipedia.org/wiki/Matrix_normal_distribution#Drawing_values_from_the_distribution.

5.2 Wishart distribution

5.2.1 Definition

Definition: Let X be an $n \times p$ matrix following a matrix-normal distribution (→ II/5.1.1) with mean zero, independence across rows and covariance across columns V:

$$X \sim \mathcal{MN}(0, I_n, V) \; . \tag{1}$$

Define the scatter matrix S as the product of the transpose of X with itself:

$$S = X^T X = \sum_{i=1}^{n} x_i^\mathrm{T} x_i \; . \tag{2}$$

Then, the matrix S is said to follow a Wishart distribution with scale matrix V and degrees of freedom n

$$S \sim \mathcal{W}(V, n) \tag{3}$$

where $n > p - 1$ and V is a positive definite symmetric covariance matrix.

Sources:
- Wikipedia (2020): "Wishart distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Wishart_distribution#Definition.

5.2.2 Kullback-Leibler divergence

Theorem: Let S be a $p \times p$ random matrix (→ I/1.2.4). Assume two Wishart distributions (→ II/5.2.1) P and Q specifying the probability distribution of S as

$$\begin{split} P &: \; S \sim \mathcal{W}(V_1, n_1) \\ Q &: \; S \sim \mathcal{W}(V_2, n_2) \; . \end{split} \tag{1}$$

Then, the Kullback-Leibler divergence (→ I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P \, || \, Q] = \frac{1}{2} \left[n_2 \left(\ln |V_2| - \ln |V_1| \right) + n_1 \mathrm{tr}(V_2^{-1} V_1) + 2 \ln \frac{\Gamma_p \left(\frac{n_2}{2} \right)}{\Gamma_p \left(\frac{n_1}{2} \right)} + (n_1 - n_2) \psi_p \left(\frac{n_1}{2} \right) - n_1 p \right] \tag{2}$$

where $\Gamma_p(x)$ is the multivariate gamma function

$$\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{k} \Gamma \left(x - \frac{j-1}{2} \right) \tag{3}$$

and $\psi_p(x)$ is the multivariate digamma function

$$\psi_p(x) = \frac{\mathrm{d} \ln \Gamma_p(x)}{\mathrm{d}x} = \sum_{j=1}^{k} \psi \left(x - \frac{j-1}{2} \right) \; . \tag{4}$$

Proof: The KL divergence for a continuous random variable (→ I/2.5.1) is given by

$$\mathrm{KL}[P\,||\,Q] = \int_{\mathcal{X}} p(x)\, \ln \frac{p(x)}{q(x)}\, \mathrm{d}x \tag{5}$$

which, applied to the Wishart distributions (→ II/5.2.1) in (1), yields

$$\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \int_{\mathcal{S}^p} \mathcal{W}(S; V_1, n_1)\, \ln \frac{\mathcal{W}(S; V_1, n_1)}{\mathcal{W}(S; V_2, n_2)}\, \mathrm{d}S \\
&= \left\langle \ln \frac{\mathcal{W}(S; \alpha_1)}{\mathcal{W}(S; \alpha_1)} \right\rangle_{p(S)}
\end{aligned} \tag{6}$$

where $\mathcal{S}^p$ is the set of all positive-definite symmetric $p \times p$ matrices.
Using the probability density function of the Wishart distribution, this becomes:

$$\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{2^{n_1 p} |V_1|^{n_1}} \Gamma_p\left(\frac{n_1}{2}\right)} \cdot |S|^{(n_1-p-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V_1^{-1} S\right)\right]}{\frac{1}{\sqrt{2^{n_2 p} |V_2|^{n_2}} \Gamma_p\left(\frac{n_2}{2}\right)} \cdot |S|^{(n_2-p-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V_2^{-1} S\right)\right]} \right\rangle_{p(S)} \\
&= \left\langle \ln \left(\sqrt{2^{(n_2-n_1)p} \cdot \frac{|V_2|^{n_2}}{|V_1|^{n_1}}} \cdot \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \cdot |S|^{(n_1-n_2)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V_1^{-1} S\right) + \frac{1}{2}\mathrm{tr}\left(V_2^{-1} S\right) \right] \right) \right\rangle_{p(S)} \\
&= \left\langle \frac{(n_2-n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \right. \\
&+ \left. \frac{n_1-n_2}{2} \ln |S| - \frac{1}{2}\mathrm{tr}\left(V_1^{-1} S\right) + \frac{1}{2}\mathrm{tr}\left(V_2^{-1} S\right) \right\rangle_{p(S)} \\
&= \frac{(n_2-n_1)p}{2} \ln 2 + \frac{n_2}{2} \ln |V_2| - \frac{n_1}{2} \ln |V_1| + \ln \frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \\
&+ \frac{n_1-n_2}{2} \left\langle \ln |S| \right\rangle_{p(S)} - \frac{1}{2} \left\langle \mathrm{tr}\left(V_1^{-1} S\right) \right\rangle_{p(S)} + \frac{1}{2} \left\langle \mathrm{tr}\left(V_2^{-1} S\right) \right\rangle_{p(S)} \; .
\end{aligned} \tag{7}$$

Using the expected value of a Wishart random matrix

$$S \sim \mathcal{W}(V, n) \quad \Rightarrow \quad \langle S \rangle = nV \; , \tag{8}$$

such that the expected value of the matrix trace (→ I/1.10.8) becomes

$$\langle \mathrm{tr}(AS) \rangle = \mathrm{tr}\left(\langle AS \rangle\right) = \mathrm{tr}\left(A \langle S \rangle\right) = \mathrm{tr}\left(A \cdot (nV)\right) = n \cdot \mathrm{tr}(AV) \; , \tag{9}$$

and the expected value of a Wishart log-determinant

$$S \sim \mathcal{W}(V, n) \quad \Rightarrow \quad \langle \ln |S| \rangle = \psi_p\left(\frac{n}{2}\right) + p \cdot \ln 2 + \ln |V| \; , \tag{10}$$

the Kullback-Leibler divergence from (7) becomes:

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \frac{(n_2-n_1)p}{2}\ln 2 + \frac{n_2}{2}\ln|V_2| - \frac{n_1}{2}\ln|V_1| + \ln\frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} \\
&+ \frac{n_1-n_2}{2}\left[\psi_p\left(\frac{n_1}{2}\right) + p\cdot\ln 2 + \ln|V_1|\right] - \frac{n_1}{2}\mathrm{tr}\left(V_1^{-1}V_1\right) + \frac{n_1}{2}\mathrm{tr}\left(V_2^{-1}V_1\right) \\
&= \frac{n_2}{2}\left(\ln|V_2| - \ln|V_1|\right) + \ln\frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} + \frac{n_1-n_2}{2}\psi_p\left(\frac{n_1}{2}\right) - \frac{n_1}{2}\mathrm{tr}\left(I_p\right) + \frac{n_1}{2}\mathrm{tr}\left(V_2^{-1}V_1\right) \\
&= \frac{1}{2}\left[n_2\left(\ln|V_2| - \ln|V_1|\right) + n_1\mathrm{tr}(V_2^{-1}V_1) + 2\ln\frac{\Gamma_p\left(\frac{n_2}{2}\right)}{\Gamma_p\left(\frac{n_1}{2}\right)} + (n_1-n_2)\psi_p\left(\frac{n_1}{2}\right) - n_1 p\right] \; .
\end{aligned} \tag{11}
$$

■

Sources:

- Penny, William D. (2001): "KL-Divergences of Normal, Gamma, Dirichlet and Wishart densities"; in: *University College, London*, pp. 2-3, eqs. 13/15; URL: https://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps.
- Wikipedia (2021): "Wishart distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-12-02; URL: https://en.wikipedia.org/wiki/Wishart_distribution#KL-divergence.

5.3 Normal-Wishart distribution

5.3.1 Definition

Definition: Let X be an $n \times p$ random matrix (→ I/1.2.4) and let Y be a $p \times p$ positive-definite symmetric matrix. Then, X and Y are said to follow a normal-Wishart distribution

$$
X, Y \sim \mathrm{NW}(M, U, V, \nu) \; , \tag{1}
$$

if the distribution of X conditional on Y is a matrix-normal distribution (→ II/5.1.1) with mean M, covariance across rows U, covariance across columns Y^{-1} and Y follows a Wishart distribution (→ II/5.2.1) with scale matrix V and degrees of freedom ν:

$$
\begin{aligned}
X|Y &\sim \mathcal{MN}(M, U, Y^{-1}) \\
Y &\sim \mathcal{W}(V, \nu) \; .
\end{aligned} \tag{2}
$$

The $p \times p$ matrix Y can be seen as the precision matrix (→ I/1.13.19) across the columns of the $n \times p$ matrix X.

5.3.2 Probability density function

Theorem: Let X and Y follow a normal-Wishart distribution (→ II/5.3.1):

$$
X, Y \sim \mathrm{NW}(M, U, V, \nu) \; . \tag{1}
$$

Then, the joint probability (→ I/1.3.2) density function (→ I/1.7.1) of X and Y is

$$
\begin{aligned}
p(X,Y) = \; & \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^\nu}} \cdot \frac{\sqrt{2^{-\nu p}}}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot |Y|^{(\nu+n-p-1)/2} \cdot \\
& \exp\left[-\frac{1}{2}\mathrm{tr}\left(Y\left[(X-M)^\mathrm{T} U^{-1} (X-M) + V^{-1}\right]\right)\right] \; .
\end{aligned} \tag{2}
$$

Proof: The normal-Wishart distribution (→ II/5.3.1) is defined as X conditional on Y following a matrix-normal distribution (→ II/5.1.1) and Y following a Wishart distribution (→ II/5.2.1):

$$
\begin{aligned}
X|Y &\sim \mathcal{MN}(M, U, Y^{-1}) \\
Y &\sim \mathcal{W}(V, \nu) \; .
\end{aligned} \tag{3}
$$

Thus, using the probability density function of the matrix-normal distribution (→ II/5.1.3) and the probability density function of the Wishart distribution, we have the following probabilities:

$$
\begin{aligned}
p(X|Y) &= \mathcal{MN}(X; M, U, Y^{-1}) \\
&= \sqrt{\frac{|Y|^n}{(2\pi)^{np}|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(Y(X-M)^\mathrm{T} U^{-1} (X-M)\right)\right] \\
p(Y) &= \mathcal{W}(Y; V, \nu) \\
&= \frac{1}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{2^{\nu p}|V|^\nu}} \cdot |Y|^{(\nu-p-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1}Y\right)\right] \; .
\end{aligned} \tag{4}
$$

The law of conditional probability (→ I/1.3.4) implies that

$$
p(X,Y) = p(X|Y) \, p(Y) \; , \tag{5}
$$

such that the normal-Wishart density function becomes:

$$
\begin{aligned}
p(X,Y) &= \mathcal{MN}(X; M, U, Y^{-1}) \cdot \mathcal{W}(Y; V, \nu) \\
&= \sqrt{\frac{|Y|^n}{(2\pi)^{np}|U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(Y(X-M)^\mathrm{T} U^{-1} (X-M)\right)\right] \cdot \\
&\quad\;\; \frac{1}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{2^{\nu p}|V|^\nu}} \cdot |Y|^{(\nu-p-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(V^{-1}Y\right)\right] \\
&= \frac{1}{\sqrt{(2\pi)^{np} |U|^p |V|^\nu}} \cdot \frac{\sqrt{2^{-\nu p}}}{\Gamma_p\left(\frac{\nu}{2}\right)} \cdot |Y|^{(\nu+n-p-1)/2} \cdot \\
&\quad\;\; \exp\left[-\frac{1}{2}\mathrm{tr}\left(Y\left[(X-M)^\mathrm{T} U^{-1} (X-M) + V^{-1}\right]\right)\right] \; .
\end{aligned} \tag{6}
$$

■

5.3.3 Mean

Theorem: Let $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{p \times p}$ follow a normal-Wishart distribution (→ II/5.3.1):

$$X, Y \sim \mathrm{NW}(M, U, V, \nu) \; . \tag{1}$$

Then, the expected value (→ I/1.10.1) of X and Y is

$$\mathrm{E}[(X, Y)] = (M, \nu V) \; . \tag{2}$$

Proof: Consider the random matrix (→ I/1.2.4)

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \\ y_{11} & \ldots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{p1} & \ldots & y_{pp} \end{bmatrix} \; . \tag{3}$$

According to the expected value of a random matrix (→ I/1.10.14), its expected value is

$$\mathrm{E}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \right) = \begin{bmatrix} \mathrm{E}(x_{11}) & \ldots & \mathrm{E}(x_{1p}) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(x_{n1}) & \ldots & \mathrm{E}(x_{np}) \\ \mathrm{E}(y_{11}) & \ldots & \mathrm{E}(y_{1p}) \\ \vdots & \ddots & \vdots \\ \mathrm{E}(y_{p1}) & \ldots & \mathrm{E}(y_{pp}) \end{bmatrix} = \begin{bmatrix} \mathrm{E}(X) \\ \mathrm{E}(Y) \end{bmatrix} \; . \tag{4}$$

When X and Y are jointly normal-Wishart distributed, then (→ II/5.3.1) by definition X follows a matrix-normal distribution (→ II/5.1.1) conditional on Y and Y follows a Wishart distribution (→ II/5.2.1):

$$X, Y \sim \mathrm{NW}(M, U, V, \nu) \quad \Leftrightarrow \quad X|Y \sim \mathcal{MN}(M, U, Y^{-1}) \quad \wedge \quad Y \sim \mathcal{W}(V, \nu) \; . \tag{5}$$

Thus, with the expected value of the matrix-variate normal distribution (→ II/5.1.4) and the law of conditional probability (→ I/1.3.4), $\mathrm{E}(X)$ becomes

$$
\begin{aligned}
\mathrm{E}(X) &= \iint X \cdot p(X,Y) \, \mathrm{d}X \, \mathrm{d}Y \\
&= \iint X \cdot p(X|Y) \cdot p(Y) \, \mathrm{d}X \, \mathrm{d}Y \\
&= \int p(Y) \int X \cdot p(X|Y) \, \mathrm{d}X \, \mathrm{d}Y \\
&= \int p(Y) \, \langle X \rangle_{\mathcal{MN}(M,U,Y^{-1})} \, \mathrm{d}Y \\
&= \int p(Y) \cdot M \, \mathrm{d}Y \\
&= M \int p(Y) \, \mathrm{d}Y \\
&= M \; ,
\end{aligned} \tag{6}
$$

and with the expected value of the Wishart distribution, $\mathrm{E}(Y)$ becomes

$$
\begin{aligned}
\mathrm{E}(Y) &= \int Y \cdot p(Y) \, \mathrm{d}Y \\
&= \langle Y \rangle_{\mathcal{W}(V,\nu)} \\
&= \nu V \; .
\end{aligned} \tag{7}
$$

Thus, the expectation of the random matrix in equations (3) and (4) is

$$
\mathrm{E}\left(\left[\begin{array}{c} X \\ Y \end{array} \right] \right) = \left[\begin{array}{c} M \\ \nu V \end{array} \right] \; , \tag{8}
$$

as indicated by equation (2).

■

Chapter III

Statistical Models

1 Univariate normal data

1.1 Univariate Gaussian

1.1.1 Definition

Definition: A univariate Gaussian data set is given by a set of real numbers $y = \{y_1, \ldots, y_n\}$, independent and identically distributed according to a normal distribution (→ II/3.2.1) with unknown mean μ and unknown variance σ^2:

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Sources:
- Bishop, Christopher M. (2006): "Example: The univariate Gaussian"; in: *Pattern Recognition for Machine Learning*, ch. 10.1.3, p. 470, eq. 10.21; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.1.2 Maximum likelihood estimation

Theorem: Let there be a univariate Gaussian data set (→ III/1.1.1) $y = \{y_1, \ldots, y_n\}$:

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the maximum likelihood estimates (→ I/4.1.3) for mean μ and variance σ^2 are given by

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \end{aligned} \tag{2}$$

Proof: The likelihood function (→ I/5.1.2) for each observation is given by the probability density function of the normal distribution (→ II/3.2.10)

$$p(y_i|\mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \tag{3}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\mu, \sigma^2) = \prod_{i=1}^{n} p(y_i|\mu) = \sqrt{\frac{1}{(2\pi\sigma^2)^n}} \cdot \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \; . \tag{4}$$

This can be developed into

$$
\begin{aligned}
p(y|\mu,\sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i^2 - 2y_i\mu + \mu^2}{\sigma^2}\right)\right] \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot \exp\left[-\frac{1}{2\sigma^2}\left(y^\mathrm{T}y - 2n\bar{y}\mu + n\mu^2\right)\right]
\end{aligned} \tag{5}
$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of data points and $y^\mathrm{T}y = \sum_{i=1}^{n} y_i^2$ is the sum of squared data points. Thus, the log-likelihood function (→ I/4.1.2) is

$$
\mathrm{LL}(\mu,\sigma^2) = \log p(y|\mu,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y^\mathrm{T}y - 2n\bar{y}\mu + n\mu^2\right) \; . \tag{6}
$$

The derivative of the log-likelihood function (6) with respect to μ is

$$
\frac{\mathrm{dLL}(\mu,\sigma^2)}{\mathrm{d}\mu} = \frac{n\bar{y}}{\sigma^2} - \frac{n\mu}{\sigma^2} = \frac{n}{\sigma^2}(\bar{y}-\mu) \tag{7}
$$

and setting this derivative to zero gives the MLE for μ:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\mu},\sigma^2)}{\mathrm{d}\mu} &= 0 \\
0 &= \frac{n}{\sigma^2}(\bar{y}-\hat{\mu}) \\
0 &= \bar{y}-\hat{\mu} \\
\hat{\mu} &= \bar{y} \\
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} y_i \; .
\end{aligned} \tag{8}
$$

The derivative of the log-likelihood function (6) at $\hat{\mu}$ with respect to σ^2 is

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\mu},\sigma^2)}{\mathrm{d}\sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(y^\mathrm{T}y - 2n\bar{y}\hat{\mu} + n\hat{\mu}^2\right) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\hat{\mu} + \hat{\mu}^2\right) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\hat{\mu})^2
\end{aligned} \tag{9}
$$

and setting this derivative to zero gives the MLE for σ^2:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\mu},\hat{\sigma}^2)}{\mathrm{d}\sigma^2} &= 0 \\
0 &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \\
\frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \\
\frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{n}{2\hat{\sigma}^2} &= \frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2
\end{aligned}
\tag{10}
$$

Together, (8) and (10) constitute the MLE for the univariate Gaussian.

■

Sources:

- Bishop CM (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, pp. 93-94, eqs. 2.121, 2.122; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.1.3 One-sample t-test

Theorem: Let

$$
y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}
$$

be a univariate Gaussian data set (→ III/1.1.1) with unknown mean μ and unknown variance σ^2. Then, the test statistic (→ I/4.3.5)

$$
t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \tag{2}
$$

with sample mean (→ I/1.10.2) $\bar{y}$ and sample variance (→ I/1.11.2) s^2 follows a Student's t-distribution (→ II/3.3.1) with $n-1$ degrees of freedom

$$
t \sim \mathrm{t}(n-1) \tag{3}
$$

under the null hypothesis (→ I/4.3.2)

$$
H_0: \; \mu = \mu_0 \; . \tag{4}
$$

Proof: The sample mean (→ I/1.10.2) is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{5}$$

and the sample variance (→ I/1.11.2) is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \tag{6}$$

Using the linear combination formula for normal random variables (→ II/3.2.26), the sample mean follows a normal distribution (→ II/3.2.1) with the following parameters:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\frac{1}{n} n\mu, \left(\frac{1}{n}\right)^2 n\sigma^2 \right) = \mathcal{N}\left(\mu, \sigma^2/n\right) \; . \tag{7}$$

Again employing the linear combination theorem and applying the null hypothesis from (4), the distribution of $Z = \sqrt{n}(\bar{y} - \mu_0)/\sigma$ becomes standard normal (→ II/3.2.3)

$$Z = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma} \sim \mathcal{N}\left(\frac{\sqrt{n}}{\sigma}(\mu - \mu_0), \left(\frac{\sqrt{n}}{\sigma}\right)^2 \frac{\sigma^2}{n} \right) \overset{H_0}{=} \mathcal{N}(0, 1) \; . \tag{8}$$

Because sample variances calculated from independent normal random variables follow a chi-squared distribution (→ II/3.2.7), the distribution of $V = (n-1)\, s^2/\sigma^2$ is

$$V = \frac{(n-1)\, s^2}{\sigma^2} \sim \chi^2(n-1) \; . \tag{9}$$

Finally, since the ratio of a standard normal random variable and the square root of a chi-squared random variable follows a t-distribution (→ II/3.3.1), the distribution of the test statistic (→ I/4.3.5) is given by

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{Z}{\sqrt{V/(n-1)}} \sim \mathrm{t}(n-1) \; . \tag{10}$$

This means that the null hypothesis (→ I/4.3.2) can be rejected when t is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from the Student's t-distribution (→ II/3.3.1) with $n-1$ degrees of freedom using a significance level (→ I/4.3.8) α.

■

Sources:
- Wikipedia (2021): "Student's t-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: https://en.wikipedia.org/wiki/Student%27s_t-distribution#Derivation.

1.1.4 Two-sample t-test

Theorem: Let

$$\begin{aligned} y_{1i} &\sim \mathcal{N}(\mu_1, \sigma^2), \quad i = 1, \ldots, n_1 \\ y_{2i} &\sim \mathcal{N}(\mu_2, \sigma^2), \quad i = 1, \ldots, n_2 \end{aligned} \tag{1}$$

be two univariate Gaussian data sets (→ III/1.1.1) representing two groups of unequal size n_1 and n_2 with unknown means μ_1 and μ_2 and equal unknown variance σ^2. Then, the test statistic (→ I/4.3.5)

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

with sample means (→ I/1.10.2) $\bar{y}_1$ and $\bar{y}_2$ and pooled standard deviation s_p follows a Student's t-distribution (→ II/3.3.1) with $n_1 + n_2 - 2$ degrees of freedom

$$t \sim \mathrm{t}(n_1 + n_2 - 2) \tag{3}$$

under the null hypothesis (→ I/4.3.2)

$$H_0: \; \mu_1 - \mu_2 = \mu_\Delta \; . \tag{4}$$

Proof: The sample means (→ I/1.10.2) are given by

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \end{aligned} \tag{5}$$

and the pooled standard deviation is given by

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \tag{6}$$

with the sample variances (→ I/1.11.2)

$$\begin{aligned} s_1^2 &= \frac{1}{n_1-1} \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 \\ s_2^2 &= \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2 \; . \end{aligned} \tag{7}$$

Using the linear combination formula for normal random variables (→ II/3.2.26), the sample means follows normal distributions (→ II/3.2.1) with the following parameters:

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \sim \mathcal{N}\left(\frac{1}{n_1} n_1 \mu_1, \left(\frac{1}{n_1}\right)^2 n_1 \sigma^2 \right) = \mathcal{N}\left(\mu_1, \sigma^2/n_1\right) \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \sim \mathcal{N}\left(\frac{1}{n_2} n_2 \mu_2, \left(\frac{1}{n_2}\right)^2 n_2 \sigma^2 \right) = \mathcal{N}\left(\mu_2, \sigma^2/n_2\right) \; . \end{aligned} \tag{8}$$

Again employing the linear combination theorem and applying the null hypothesis from (4), the distribution of $Z = ((\bar{y}\_1 - \bar{y}\_2) - \mu\_\Delta)/(\sigma\sqrt{1/n\_1 + 1/n\_2})$ becomes standard normal (→ II/3.2.3)

$$Z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}\left(\frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \left(\frac{1}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)^2 \left(\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} \right) \right) \overset{H_0}{=} \mathcal{N}(0, 1) \; . \tag{9}$$

Because sample variances calculated from independent normal random variables follow a chi-squared distribution (→ II/3.2.7), the distribution of $V = (n_1 + n_2 - 2)\, s_p^2/\sigma^2$ is

$$V = \frac{(n_1 + n_2 - 2)\, s_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) \; . \tag{10}$$

Finally, since the ratio of a standard normal random variable and the square root of a chi-squared random variable follows a t-distribution (→ II/3.3.1), the distribution of the test statistic (→ I/4.3.5) is given by

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} \sim \mathrm{t}(n_1 + n_2 - 2) \; . \tag{11}$$

This means that the null hypothesis (→ I/4.3.2) can be rejected when t is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from the Student's t-distribution (→ II/3.3.1) with $n_1 + n_2 - 2$ degrees of freedom using a significance level (→ I/4.3.8) α.

■

Sources:

- Wikipedia (2021): "Student's t-distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: https://en.wikipedia.org/wiki/Student%27s_t-distribution#Derivation.
- Wikipedia (2021): "Student's t-test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: https://en.wikipedia.org/wiki/Student%27s_t-test#Equal_or_unequal_sample_sizes,_similar_variances_(1/2_%3C_sX1/sX2_%3C_2).

1.1.5 Paired t-test

Theorem: Let y_{i1} and y_{i2} with $i = 1, \ldots, n$ be paired observations, such that

$$y_{i1} \sim \mathcal{N}(y_{i2} + \mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

is a univariate Gaussian data set (→ III/1.1.1) with unknown shift μ and unknown variance σ^2. Then, the test statistic (→ I/4.3.5)

$$t = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}} \quad \text{where} \quad d_i = y_{i1} - y_{i2} \tag{2}$$

with sample mean (→ I/1.10.2) $\bar{d}$ and sample variance (→ I/1.11.2) s_d^2 follows a Student's t-distribution (→ II/3.3.1) with $n - 1$ degrees of freedom

$$t \sim \mathrm{t}(n-1) \tag{3}$$

under the null hypothesis (→ I/4.3.2)

$$H_0: \; \mu = \mu_0 \; . \tag{4}$$

Proof: Define the pair-wise difference $d_i = y_{i1} - y_{i2}$ which is, according to the linearity of the expected value (→ I/1.10.5) and the invariance of the variance under addition (→ I/1.11.6), distributed as

$$d_i = y_{i1} - y_{i2} \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{5}$$

Therefore, $d_1, \ldots, d_n$ satisfy the conditions of the one-sample t-test (→ III/1.1.3) which results in the test statistic given by (2).

■

Sources:
- Wikipedia (2021): "Student's t-test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-12; URL: https://en.wikipedia.org/wiki/Student%27s_t-test#Dependent_t-test_for_paired_samples.

1.1.6 F-test for equality of variances

Theorem: Let

$$\begin{aligned} y_{1i} &\sim \mathcal{N}(\mu_1, \sigma_1^2), \quad i = 1, \ldots, n_1 \\ y_{2i} &\sim \mathcal{N}(\mu_2, \sigma_2^2), \quad i = 1, \ldots, n_2 \end{aligned} \tag{1}$$

be two univariate Gaussian data sets (→ III/1.1.1) representing two groups of unequal size n_1 and n_2 with unknown means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2. Then, the test statistic (→ I/4.3.5)

$$F = \frac{s_1^2}{s_1^2} = \frac{\frac{1}{n_1-1} \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2}{\frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2} \tag{2}$$

with sample means (→ I/1.10.2) $\bar{y}_1$ and $\bar{y}_2$ and samle variances (→ I/1.11.2) s_1^2 and s_2^2 follows an F-distribution (→ II/3.8.1) with numerator degrees of freedom $n_1 - 1$ and denominator degrees of freedom $n_2 - 1$

$$F \sim \mathrm{F}(n_1 - 1, n_2 - 1) \tag{3}$$

under the null hypothesis (→ I/4.3.2) that the two variances (→ II/3.2.1) are equal:

$$H_0: \; \sigma_1^2 = \sigma_2^2 \; . \tag{4}$$

Proof: We know that, for a sample of normal random variables, the sample variance is following a chi-squared distribution (→ II/3.2.7):

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \; i = 1, \ldots, n \quad \Rightarrow \quad V = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2(n-1) \; . \tag{5}$$

Thus, we have:

$$\begin{aligned} V_1 &= (n_1 - 1)\frac{s_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{and} \\ V_2 &= (n_2 - 1)\frac{s_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1) \; . \end{aligned} \tag{6}$$

Moreover, by definition, the ratio of two chi-squared random variables, divided by their degrees of freedom, is following an F-distribution (→ II/3.8.1):

$$X_1 \sim \chi^2(d_1),\; X_2 \sim \chi^2(d_2) \quad \Rightarrow \quad Y = \frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2) \; . \tag{7}$$

Thus, we have:

$$\begin{split} F &= \frac{V_1/(n_1-1)}{V_2/(n_2-1)} \\ &= \frac{(n_1-1)\frac{s_1^2}{\sigma_1^2} \Big/ (n_1-1)}{(n_2-1)\frac{s_2^2}{\sigma_2^2} \Big/ (n_2-1)} \\ &= \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \\ &\overset{H_0}{=} \frac{s_1^2}{s_2^2} \; . \end{split} \tag{8}$$

This means that the null hypothesis (→ I/4.3.2) of equal variances can be rejected when F is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from the F-distribution (→ II/**??**) with degrees of freedom $n_1 - 1$ and $n_2 - 1$ using a significance level (→ I/4.3.8) α.

■

Sources:

- Wikipedia (2024): "F-test of equality of variances"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-07-05; URL: https://en.wikipedia.org/wiki/F-test_of_equality_of_variances#The_test.

1.1.7 Conjugate prior distribution

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \; , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.1.1) with unknown mean μ and unknown variance σ^2. Then, the conjugate prior (→ I/5.2.5) for this model is a normal-gamma distribution (→ II/4.3.1)

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau\lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{2}$$

where $\tau = 1/\sigma^2$ is the inverse variance or precision.

Proof: By definition, a conjugate prior (→ I/5.2.5) is a prior distribution (→ I/5.1.3) that, when combined with the likelihood function (→ I/5.1.2), leads to a posterior distribution (→ I/5.1.7) that belongs to the same family of probability distributions (→ I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model model parameters in the same way, i.e. the model parameters appear in the same functional form in both.
Equation (1) implies the following likelihood function (→ I/5.1.2)

$$
\begin{split}
p(y|\mu, \sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]
\end{split} \tag{3}
$$

which, for mathematical convenience, can also be parametrized as

$$
\begin{split}
p(y|\mu, \tau) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \tau^{-1}) \\
&= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left[-\frac{\tau}{2}(y_i-\mu)^2\right] \\
&= \left(\sqrt{\frac{\tau}{2\pi}}\right)^n \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right]
\end{split} \tag{4}
$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

Separating constant and variable terms, we have:

$$
p(y|\mu, \tau) = \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \; . \tag{5}
$$

Expanding the product in the exponent, we have

$$
\begin{split}
p(y|\mu, \tau) &= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}\left(y_i^2 - 2\mu y_i + \mu^2\right)\right] \\
&= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right)\right] \\
&= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T}y - 2\mu n\bar{y} + n\mu^2\right)\right] \\
&= \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau n}{2}\left(\frac{1}{n}y^\mathrm{T}y - 2\mu\bar{y} + \mu^2\right)\right]
\end{split} \tag{6}
$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of data points and $y^\mathrm{T}y = \sum_{i=1}^{n} y_i^2$ is the sum of squared data points. Completing the square over μ, finally gives

$$
p(y|\mu, \tau) = \sqrt{\frac{1}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau n}{2}\left((\mu-\bar{y})^2 - \bar{y}^2 + \frac{1}{n}y^\mathrm{T}y\right)\right] \tag{7}
$$

In other words, the likelihood function (→ I/5.1.2) is proportional to a power of τ times an exponential of τ and an exponential of a squared form of μ, weighted by τ:

$$p(y|\mu,\tau) \propto \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} y - n\bar{y}^2\right)\right] \cdot \exp\left[-\frac{\tau n}{2}\left(\mu - \bar{y}\right)^2\right] \; . \tag{8}$$

The same is true for a normal-gamma distribution (→ II/4.3.1) over μ and τ

$$p(\mu,\tau) = \mathcal{N}(\mu; \mu_0, (\tau\lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{9}$$

the probability density function of which (→ II/4.3.3)

$$p(\mu,\tau) = \sqrt{\frac{\tau\lambda_0}{2\pi}} \cdot \exp\left[-\frac{\tau\lambda_0}{2}\left(\mu - \mu_0\right)^2\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \tag{10}$$

exhibits the same proportionality

$$p(\mu,\tau) \propto \tau^{a_0+1/2-1} \cdot \exp[-\tau b_0] \cdot \exp\left[-\frac{\tau\lambda_0}{2}\left(\mu - \mu_0\right)^2\right] \tag{11}$$

and is therefore conjugate relative to the likelihood.

■

Sources:
- Bishop CM (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, pp. 97-102, eq. 2.154; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.1.8 Posterior distribution

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.1.1) with unknown mean μ and unknown variance σ^2. Moreover, assume a normal-gamma prior distribution (→ III/1.1.7) over the model parameters μ and $\tau = 1/\sigma^2$:

$$p(\mu,\tau) = \mathcal{N}(\mu; \mu_0, (\tau\lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a normal-gamma distribution (→ II/4.3.1)

$$p(\mu,\tau|y) = \mathcal{N}(\mu; \mu_n, (\tau\lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned}
\mu_n &= \frac{\lambda_0\mu_0 + n\bar{y}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) \; .
\end{aligned} \tag{4}$$

Proof: According to Bayes' theorem (→ I/5.3.1), the posterior distribution (→ I/5.1.7) is given by

$$p(\mu, \tau|y) = \frac{p(y|\mu, \tau)\, p(\mu, \tau)}{p(y)} \; . \tag{5}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (→ I/5.1.9) to the numerator:

$$p(\mu, \tau|y) \propto p(y|\mu, \tau)\, p(\mu, \tau) = p(y, \mu, \tau) \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$\begin{split} p(y|\mu, \sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \end{split} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{split} p(y|\mu, \tau) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \tau^{-1}) \\ &= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left[-\frac{\tau}{2} (y_i - \mu)^2 \right] \\ &= \left(\sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp\left[-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \end{split} \tag{8}$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

Combining the likelihood function (→ I/5.1.2) (8) with the prior distribution (→ I/5.1.3) (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{split} p(y, \mu, \tau) &= p(y|\mu, \tau)\, p(\mu, \tau) \\ &= \left(\sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp\left[-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] \cdot \\ &\quad \sqrt{\frac{\tau\lambda_0}{2\pi}} \cdot \exp\left[-\frac{\tau\lambda_0}{2} (\mu - \mu_0)^2 \right] \cdot \\ &\quad \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0-1} \exp[-b_0\tau] \; . \end{split} \tag{9}$$

Collecting identical variables gives:

$$\begin{split} p(y,\mu,\tau) = &\sqrt{\frac{\tau^{n+1}\lambda_0}{(2\pi)^{n+1}}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau}{2} \left(\sum_{i=1}^{n} (y_i-\mu)^2 + \lambda_0 \left(\mu-\mu_0\right)^2 \right) \right] \; . \end{split} \tag{10}$$

Expanding the products in the exponent (→ III/1.1.7) gives

$$\begin{split} p(y,\mu,\tau) = &\sqrt{\frac{\tau^{n+1}\lambda_0}{(2\pi)^{n+1}}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau}{2} \left((y^\mathrm{T} y - 2\mu n \bar{y} + n\mu^2) + \lambda_0 (\mu^2 - 2\mu\mu_0 + \mu_0^2) \right) \right] \end{split} \tag{11}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $y^\mathrm{T} y = \sum_{i=1}^{n} y_i^2$, such that

$$\begin{split} p(y,\mu,\tau) = &\sqrt{\frac{\tau^{n+1}\lambda_0}{(2\pi)^{n+1}}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau}{2} \left(\mu^2 (\lambda_0+n) - 2\mu(\lambda_0\mu_0 + n\bar{y}) + (y^\mathrm{T} y + \lambda_0 \mu_0^2) \right) \right] \end{split} \tag{12}$$

Completing the square over μ, we finally have

$$\begin{split} p(y,\mu,\tau) = &\sqrt{\frac{\tau^{n+1}\lambda_0}{(2\pi)^{n+1}}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau \lambda_n}{2} \left(\mu-\mu_n\right)^2 - \frac{\tau}{2} \left(y^\mathrm{T} y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2 \right) \right] \end{split} \tag{13}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{split} \mu_n &= \frac{\lambda_0 \mu_0 + n\bar{y}}{\lambda_0 + n} \\ \lambda_n &= \lambda_0 + n \; . \end{split} \tag{14}$$

Ergo, the joint likelihood is proportional to

$$p(y,\mu,\tau) \propto \tau^{1/2} \cdot \exp\left[-\frac{\tau \lambda_n}{2} \left(\mu-\mu_n\right)^2 \right] \cdot \tau^{a_n-1} \cdot \exp\left[-b_n \tau \right] \tag{15}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{split} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} \left(y^\mathrm{T} y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2 \right) \; . \end{split} \tag{16}$$

From the term in (13), we can isolate the posterior distribution over μ given τ:

$$p(\mu|\tau, y) = \mathcal{N}(\mu; \mu_n, (\tau\lambda_n)^{-1}) \; . \tag{17}$$

From the remaining term, we can isolate the posterior distribution over τ:

$$p(\tau|y) = \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{18}$$

Together, (17) and (18) constitute the joint (→ I/1.3.2) posterior distribution (→ I/5.1.7) of μ and τ.

■

Sources:

- Bishop CM (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, pp. 97-102, eq. 2.154; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.1.9 Log model evidence

Theorem: Let

$$m: \; y = \{y_1, \ldots, y_n\} \; , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.1.1) with unknown mean μ and unknown variance σ^2. Moreover, assume a normal-gamma prior distribution (→ III/1.1.7) over the model parameters μ and $\tau = 1/\sigma^2$:

$$p(\mu, \tau) = \mathcal{N}(\mu; \mu_0, (\tau\lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the log model evidence (→ IV/??) for this model is

$$\log p(y|m) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\frac{\lambda_0}{\lambda_n} + \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n \tag{3}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned}
\mu_n &= \frac{\lambda_0\mu_0 + n\bar{y}}{\lambda_0 + n} \\
\lambda_n &= \lambda_0 + n \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2) \; .
\end{aligned} \tag{4}$$

Proof: According to the law of marginal probability (→ I/1.3.3), the model evidence (→ I/5.1.11) for this model is:

$$p(y|m) = \iint p(y|\mu, \tau)\, p(\mu, \tau) \, \mathrm{d}\mu \, \mathrm{d}\tau \; . \tag{5}$$

According to the law of conditional probability (→ I/1.3.4), the integrand is equivalent to the joint likelihood (→ I/5.1.5):

$$p(y|m) = \iint p(y,\mu,\tau) \, \mathrm{d}\mu \, \mathrm{d}\tau \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$\begin{split} p(y|\mu,\sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i;\mu,\sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right] \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \end{split} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{split} p(y|\mu,\tau) &= \prod_{i=1}^{n} \mathcal{N}(y_i;\mu,\tau^{-1}) \\ &= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left[-\frac{\tau}{2}(y_i-\mu)^2\right] \\ &= \left(\sqrt{\frac{\tau}{2\pi}}\right)^n \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \end{split} \tag{8}$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

When deriving the posterior distribution (→ III/1.1.8) $p(\mu,\tau|y)$, the joint likelihood $p(y,\mu,\tau)$ is obtained as

$$\begin{split} p(y,\mu,\tau) = & \sqrt{\frac{\tau^{n+1}\lambda_0}{(2\pi)^{n+1}}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ & \exp\left[-\frac{\tau\lambda_n}{2}(\mu-\mu_n)^2 - \frac{\tau}{2}\left(y^\mathrm{T}y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2\right)\right] \; . \end{split} \tag{9}$$

Using the probability density function of the normal distribution (→ II/3.2.10), we can rewrite this as

$$\begin{split} p(y,\mu,\tau) = & \sqrt{\frac{\tau^n}{(2\pi)^n}} \sqrt{\frac{\tau\lambda_0}{2\pi}} \sqrt{\frac{2\pi}{\tau\lambda_n}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ & \mathcal{N}(\mu;\mu_n,(\tau\lambda_n)^{-1}) \, \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T}y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2\right)\right] \; . \end{split} \tag{10}$$

Now, μ can be integrated out easily:

$$\begin{split} \int p(y,\mu,\tau) \, \mathrm{d}\mu = & \sqrt{\frac{1}{(2\pi)^n}} \sqrt{\frac{\lambda_0}{\lambda_n}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0+n/2-1} \exp[-b_0\tau] \cdot \\ & \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T}y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2\right)\right] \; . \end{split} \tag{11}$$

Using the probability density function of the gamma distribution (→ II/3.4.7), we can rewrite this as

$$\int p(y,\mu,\tau)\,\mathrm{d}\mu = \sqrt{\frac{1}{(2\pi)^n}}\sqrt{\frac{\lambda_0}{\lambda_n}}\,\frac{{b_0}^{a_0}}{\Gamma(a_0)}\,\frac{\Gamma(a_n)}{{b_n}^{a_n}}\,\mathrm{Gam}(\tau;a_n,b_n)\;. \tag{12}$$

Finally, τ can also be integrated out:

$$\iint p(y,\mu,\tau)\,\mathrm{d}\mu\,\mathrm{d}\tau = \sqrt{\frac{1}{(2\pi)^n}}\sqrt{\frac{\lambda_0}{\lambda_n}}\,\frac{\Gamma(a_n)}{\Gamma(a_0)}\,\frac{{b_0}^{a_0}}{{b_n}^{a_n}}\;. \tag{13}$$

Thus, the log model evidence (→ IV/**??**) of this model is given by

$$\log p(y|m) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\frac{\lambda_0}{\lambda_n} + \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n\;. \tag{14}$$

■

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.23, eq. 3.118; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.1.10 Accuracy and complexity

Theorem: Let

$$m:\; y = \{y_1,\ldots,y_n\}\,,\quad y_i \sim \mathcal{N}(\mu,\sigma^2),\quad i=1,\ldots,n \tag{1}$$

be a univariate Gaussian data set (→ III/1.1.1) with unknown mean μ and unknown variance σ^2. Moreover, assume a normal-gamma prior distribution (→ III/1.1.7) over the model parameters μ and $\tau = 1/\sigma^2$:

$$p(\mu,\tau) = \mathcal{N}(\mu;\mu_0,(\tau\lambda_0)^{-1})\cdot\mathrm{Gam}(\tau;a_0,b_0)\;. \tag{2}$$

Then, accuracy and complexity (→ IV/**??**) of this model are

$$\begin{aligned}
\mathrm{Acc}(m) &= -\frac{1}{2}\frac{a_n}{b_n}\left(y^\mathrm{T}y - 2n\bar{y}\mu_n + n\mu_n^2\right) - \frac{1}{2}n\lambda_n^{-1} + \frac{n}{2}\left(\psi(a_n) - \log(b_n)\right) - \frac{n}{2}\log(2\pi) \\
\mathrm{Com}(m) &= \frac{1}{2}\frac{a_n}{b_n}\left[\lambda_0(\mu_0-\mu_n)^2 - 2(b_n-b_0)\right] + \frac{1}{2}\frac{\lambda_0}{\lambda_n} - \frac{1}{2}\log\frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&\quad + a_0\cdot\log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n-a_0)\cdot\psi(a_n)
\end{aligned} \tag{3}$$

where μ_n and λ_n as well as a_n and b_n are the posterior hyperparameters for the univariate Gaussian (→ III/1.1.8) and $\bar{y}$ is the sample mean (→ I/1.10.2).

Proof: Model accuracy and complexity are defined as (→ IV/**??**)

$$\begin{aligned}
\mathrm{LME}(m) &= \mathrm{Acc}(m) - \mathrm{Com}(m) \\
\mathrm{Acc}(m) &= \langle\log p(y|\mu,\tau,m)\rangle_{p(\mu,\tau|y,m)} \\
\mathrm{Com}(m) &= \mathrm{KL}\left[p(\mu,\tau|y,m)\,||\,p(\mu,\tau|m)\right]\;.
\end{aligned} \tag{4}$$

The accuracy term is the expectation (→ I/1.10.1) of the log-likelihood function (→ I/4.1.2) $\log p(y|\mu,\tau)$ with respect to the posterior distribution (→ I/5.1.7) $p(\mu,\tau|y)$. With the log-likelihood function for the univariate Gaussian (→ III/1.1.2) and the posterior distribution for the univariate Gaussian (→ III/1.1.8), the model accuracy of m evaluates to:

$$\begin{aligned}
\mathrm{Acc}(m) &= \left\langle \log p(y|\mu,\tau) \right\rangle_{p(\mu,\tau|y)} \\
&= \left\langle \left\langle \log p(y|\mu,\tau) \right\rangle_{p(\mu|\tau,y)} \right\rangle_{p(\tau|y)} \\
&= \left\langle \left\langle \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{\tau}{2}\left(y^\mathrm{T} y - 2n\bar{y}\mu + n\mu^2\right) \right\rangle_{\mathcal{N}(\mu_n,(\tau\lambda_n)^{-1})} \right\rangle_{\mathrm{Gam}(a_n,b_n)} \\
&= \left\langle \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{\tau}{2}\left(y^\mathrm{T} y - 2n\bar{y}\mu_n + n\mu_n^2\right) - \frac{1}{2}n\lambda_n^{-1} \right\rangle_{\mathrm{Gam}(a_n,b_n)} \\
&= \frac{n}{2}\left(\psi(a_n) - \log(b_n)\right) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\frac{a_n}{b_n}\left(y^\mathrm{T} y - 2n\bar{y}\mu_n + n\mu_n^2\right) - \frac{1}{2}n\lambda_n^{-1} \\
&= -\frac{1}{2}\frac{a_n}{b_n}\left(y^\mathrm{T} y - 2n\bar{y}\mu_n + n\mu_n^2\right) - \frac{1}{2}n\lambda_n^{-1} + \frac{n}{2}\left(\psi(a_n) - \log(b_n)\right) - \frac{n}{2}\log(2\pi)
\end{aligned} \tag{5}$$

The complexity penalty is the Kullback-Leibler divergence (→ I/2.5.1) of the posterior distribution (→ I/5.1.7) $p(\mu,\tau|y)$ from the prior distribution (→ I/5.1.3) $p(\mu,\tau)$. With the prior distribution (→ III/1.1.7) given by (2), the posterior distribution for the univariate Gaussian (→ III/1.1.8) and the Kullback-Leibler divergence of the normal-gamma distribution (→ II/4.3.7), the model complexity of m evaluates to:

$$\begin{aligned}
\mathrm{Com}(m) &= \mathrm{KL}\left[p(\mu,\tau|y)\,||\,p(\mu,\tau)\right] \\
&= \mathrm{KL}\left[\mathrm{NG}(\mu_n,\lambda_n^{-1},a_n,b_n)\,||\,\mathrm{NG}(\mu_0,\lambda_0^{-1},a_0,b_0)\right] \\
&= \frac{1}{2}\frac{a_n}{b_n}\left[\lambda_0(\mu_0-\mu_n)^2\right] + \frac{1}{2}\frac{\lambda_0}{\lambda_n} - \frac{1}{2}\log\frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&+ a_0 \cdot \log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0)\cdot\psi(a_n) - (b_n - b_0)\cdot\frac{a_n}{b_n} \\
&= \frac{1}{2}\frac{a_n}{b_n}\left[\lambda_0(\mu_0-\mu_n)^2 - 2(b_n - b_0)\right] + \frac{1}{2}\frac{\lambda_0}{\lambda_n} - \frac{1}{2}\log\frac{\lambda_0}{\lambda_n} - \frac{1}{2} \\
&+ a_0 \cdot \log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n - a_0)\cdot\psi(a_n) \; .
\end{aligned} \tag{6}$$

A control calculation confirms that

$$\mathrm{Acc}(m) - \mathrm{Com}(m) = \mathrm{LME}(m) \tag{7}$$

where $\mathrm{LME}(m)$ is the log model evidence for the univariate Gaussian (→ III/1.1.9).

■

1.2 Univariate Gaussian with known variance

1.2.1 Definition

Definition: A univariate Gaussian data set with known variance is given by a set of real numbers $y = \{y_1, \ldots, y_n\}$, independent and identically distributed according to a normal distribution (→

II/3.2.1) with unknown mean μ and known variance σ^2:

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Sources:

- Bishop, Christopher M. (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, p. 97, eq. 2.137; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.2.2 Maximum likelihood estimation

Theorem: Let there be univariate Gaussian data with known variance (→ III/1.2.1) $y = \{y_1, \ldots, y_n\}$:

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the maximum likelihood estimate (→ I/4.1.3) for the mean μ is given by

$$\hat{\mu} = \bar{y} \tag{2}$$

where $\bar{y}$ is the sample mean (→ I/1.10.2)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{3}$$

Proof: The likelihood function (→ I/5.1.2) for each observation is given by the probability density function of the normal distribution (→ II/3.2.10)

$$p(y_i|\mu) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right] \tag{4}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\mu) = \prod_{i=1}^{n} p(y_i|\mu) = \sqrt{\frac{1}{(2\pi\sigma^2)^n}} \cdot \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2\right] \; . \tag{5}$$

This can be developed into

$$\begin{aligned} p(y|\mu) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i^2 - 2y_i\mu + \mu^2}{\sigma^2}\right)\right] \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot \exp\left[-\frac{1}{2\sigma^2}\left(y^\mathrm{T} y - 2n\bar{y}\mu + n\mu^2\right)\right] \end{aligned} \tag{6}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of data points and $y^\mathrm{T} y = \sum_{i=1}^{n} y_i^2$ is the sum of squared data points. Thus, the log-likelihood function (→ I/4.1.2) is

$$\mathrm{LL}(\mu) = \log p(y|\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y^\mathrm{T} y - 2n\bar{y}\mu + n\mu^2\right) \; . \tag{7}$$

The derivatives of the log-likelihood with respect to μ are

$$\begin{split}
\frac{\mathrm{dLL}(\mu)}{\mathrm{d}\mu} &= \frac{n\bar{y}}{\sigma^2} - \frac{n\mu}{\sigma^2} = \frac{n}{\sigma^2}(\bar{y}-\mu) \\
\frac{\mathrm{d}^2\mathrm{LL}(\mu)}{\mathrm{d}\mu^2} &= -\frac{n}{\sigma^2} \; .
\end{split} \tag{8}$$

Setting the first derivative to zero, we obtain:

$$\begin{split}
\frac{\mathrm{dLL}(\hat{\mu})}{\mathrm{d}\mu} &= 0 \\
0 &= \frac{n}{\sigma^2}(\bar{y}-\hat{\mu}) \\
0 &= \bar{y}-\hat{\mu} \\
\hat{\mu} &= \bar{y}
\end{split} \tag{9}$$

Plugging this value into the second derivative, we confirm:

$$\frac{\mathrm{d}^2\mathrm{LL}(\hat{\mu})}{\mathrm{d}\mu^2} = -\frac{n}{\sigma^2} < 0 \; . \tag{10}$$

This demonstrates that the estimate $\hat{\mu} = \bar{y}$ maximizes the likelihood $p(y|\mu)$.

■

Sources:

- Bishop, Christopher M. (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, p. 98, eq. 2.143; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.2.3 One-sample z-test

Theorem: Let

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Then, the test statistic (→ I/4.3.5)

$$z = \sqrt{n}\,\frac{\bar{y}-\mu_0}{\sigma} \tag{2}$$

with sample mean (→ I/1.10.2) $\bar{y}$ follows a standard normal distribution (→ II/3.2.3)

$$z \sim \mathcal{N}(0, 1) \tag{3}$$

under the null hypothesis (→ I/4.3.2)

$$H_0 : \; \mu = \mu_0 \; . \tag{4}$$

Proof: The sample mean (→ I/1.10.2) is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{5}$$

Using the linear combination formula for normal random variables (→ II/3.2.26), the sample mean follows a normal distribution (→ II/3.2.1) with the following parameters:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\frac{1}{n} n\mu, \left(\frac{1}{n}\right)^2 n\sigma^2 \right) = \mathcal{N}\left(\mu, \sigma^2/n\right) \; . \tag{6}$$

Again employing the linear combination theorem, the distribution of $z = \sqrt{n/\sigma^2}(\bar{y} - \mu_0)$ becomes

$$z = \sqrt{\frac{n}{\sigma^2}}(\bar{y} - \mu_0) \sim \mathcal{N}\left(\sqrt{\frac{n}{\sigma^2}}(\mu - \mu_0), \left(\sqrt{\frac{n}{\sigma^2}}\right)^2 \frac{\sigma^2}{n} \right) = \mathcal{N}\left(\sqrt{n}\, \frac{\mu - \mu_0}{\sigma}, 1 \right) \; , \tag{7}$$

such that, under the null hypothesis in (4), we have:

$$z \sim \mathcal{N}(0,1), \quad \text{if } \mu = \mu_0 \; . \tag{8}$$

This means that the null hypothesis (→ I/4.3.2) can be rejected when z is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from the standard normal distribution (→ II/3.2.3) using a significance level (→ I/4.3.8) α.

■

Sources:
- Wikipedia (2021): “Z-test”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-24; URL: https://en.wikipedia.org/wiki/Z-test#Use_in_location_testing.
- Wikipedia (2021): “Gauß-Test”; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2021-03-24; URL: https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Einstichproben-Gau%C3%9F-Test.

1.2.4 Two-sample z-test

Theorem: Let

$$\begin{aligned} y_{1i} &\sim \mathcal{N}(\mu_1, \sigma_1^2), \quad i = 1, \ldots, n_1 \\ y_{2i} &\sim \mathcal{N}(\mu_2, \sigma_2^2), \quad i = 1, \ldots, n_2 \end{aligned} \tag{1}$$

be two univariate Gaussian data sets (→ III/1.1.1) representing two groups of unequal size n_1 and n_2 with unknown means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2. Then, the test statistic (→ I/4.3.5)

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2}$$

with sample means (→ I/1.10.2) $\bar{y}_1$ and $\bar{y}_2$ follows a standard normal distribution (→ II/3.2.3)

$$z \sim \mathcal{N}(0,1) \tag{3}$$

under the null hypothesis (→ I/4.3.2)

$$H_0: \; \mu_1 - \mu_2 = \mu_\Delta \; . \tag{4}$$

Proof: The sample means (→ I/1.10.2) are given by

$$\begin{split} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \; . \end{split} \tag{5}$$

Using the linear combination formula for normal random variables (→ II/3.2.26), the sample means follows normal distributions (→ II/3.2.1) with the following parameters:

$$\begin{split} \bar{y}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1i} \sim \mathcal{N}\left(\frac{1}{n_1} n_1 \mu_1, \left(\frac{1}{n_1} \right)^2 n_1 \sigma^2 \right) = \mathcal{N}\left(\mu_1, \sigma_1^2/n_1 \right) \\ \bar{y}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} y_{2i} \sim \mathcal{N}\left(\frac{1}{n_2} n_2 \mu_2, \left(\frac{1}{n_2} \right)^2 n_2 \sigma^2 \right) = \mathcal{N}\left(\mu_2, \sigma_2^2/n_2 \right) \; . \end{split} \tag{6}$$

Again employing the linear combination theorem, the distribution of $z = [(\bar{y}_1 - \bar{y}_2) - \mu_\Delta]/\sigma_\Delta$ becomes

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - \mu_\Delta}{\sigma_\Delta} \sim \mathcal{N}\left(\frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma_\Delta}, \left(\frac{1}{\sigma_\Delta} \right)^2 \sigma_\Delta^2 \right) = \mathcal{N}\left(\frac{(\mu_1 - \mu_2) - \mu_\Delta}{\sigma_\Delta}, 1 \right) \tag{7}$$

where σ_Δ is the pooled standard deviation

$$\sigma_\Delta = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \; , \tag{8}$$

such that, under the null hypothesis in (4), we have:

$$z \sim \mathcal{N}(0,1), \quad \text{if} \; \mu_\Delta = \mu_1 - \mu_2 \; . \tag{9}$$

This means that the null hypothesis (→ I/4.3.2) can be rejected when z is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from the standard normal distribution (→ II/3.2.3) using a significance level (→ I/4.3.8) α.

■

Sources:

- Wikipedia (2021): "Z-test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-24; URL: https://en.wikipedia.org/wiki/Z-test#Use_in_location_testing.
- Wikipedia (2021): "Gauß-Test"; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2021-03-24; URL: https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Zweistichproben-Gau%C3%9F-Test_f%C3%BCr_unabh%C3%A4ngige_Stichproben.

1.2.5 Paired z-test

Theorem: Let y_{i1} and y_{i2} with $i = 1, \ldots, n$ be paired observations, such that

$$y_{i1} \sim \mathcal{N}(y_{i2} + \mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

is a univariate Gaussian data set (→ III/1.2.1) with unknown shift μ and known variance σ^2. Then, the test statistic (→ I/4.3.5)

$$z = \sqrt{n}\,\frac{\bar{d} - \mu_0}{\sigma} \quad \text{where} \quad d_i = y_{i1} - y_{i2} \tag{2}$$

with sample mean (→ I/1.10.2) $\bar{d}$ follows a standard normal distribution (→ II/3.2.3)

$$z \sim \mathcal{N}(0, 1) \tag{3}$$

under the null hypothesis (→ I/4.3.2)

$$H_0: \; \mu = \mu_0 \; . \tag{4}$$

Proof: Define the pair-wise difference $d_i = y_{i1} - y_{i2}$ which is, according to the linearity of the expected value (→ I/1.10.5) and the invariance of the variance under addition (→ I/1.11.6), distributed as

$$d_i = y_{i1} - y_{i2} \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{5}$$

Therefore, $d_1, \ldots, d_n$ satisfy the conditions of the one-sample z-test (→ III/1.2.3) which results in the test statistic given by (2).

■

Sources:
- Wikipedia (2021): "Z-test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-03-24; URL: https://en.wikipedia.org/wiki/Z-test#Use_in_location_testing.
- Wikipedia (2021): "Gauß-Test"; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2021-03-24; URL: https://de.wikipedia.org/wiki/Gau%C3%9F-Test#Zweistichproben-Gau%C3%9F-Test_f%C3%BCr_abh%C3%A4ngige_(verbundene)_Stichproben.

1.2.6 Conjugate prior distribution

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \,, \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Then, the conjugate prior (→ I/5.2.5) for this model is a normal distribution (→ II/3.2.1)

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \tag{2}$$

with prior (→ I/5.1.3) mean (→ I/1.10.1) μ_0 and prior (→ I/5.1.3) precision (→ I/1.11.12) λ_0.

Proof: By definition, a conjugate prior (→ I/5.2.5) is a prior distribution (→ I/5.1.3) that, when combined with the likelihood function (→ I/5.1.2), leads to a posterior distribution (→ I/5.1.7) that belongs to the same family of probability distributions (→ I/1.5.1). This is fulfilled when the prior

density and the likelihood function are proportional to the model model parameters in the same way, i.e. the model parameters appear in the same functional form in both.
Equation (1) implies the following likelihood function (→ I/5.1.2)

$$
\begin{aligned}
p(y|\mu) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right] \\
&= \left(\sqrt{\frac{1}{2\pi\sigma^2}} \right)^n \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]
\end{aligned} \tag{3}
$$

which, for mathematical convenience, can also be parametrized as

$$
\begin{aligned}
p(y|\mu) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \tau^{-1}) \\
&= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left[-\frac{\tau}{2} (y_i - \mu)^2 \right] \\
&= \left(\sqrt{\frac{\tau}{2\pi}} \right)^n \cdot \exp\left[-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]
\end{aligned} \tag{4}
$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

Expanding the product in the exponent, we have

$$
\begin{aligned}
p(y|\mu) &= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp\left[-\frac{\tau}{2} \sum_{i=1}^{n} \left(y_i^2 - 2\mu y_i + \mu^2 \right) \right] \\
&= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp\left[-\frac{\tau}{2} \left(\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right] \\
&= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp\left[-\frac{\tau}{2} \left(y^\mathrm{T} y - 2\mu n \bar{y} + n\mu^2 \right) \right] \\
&= \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp\left[-\frac{\tau n}{2} \left(\frac{1}{n} y^\mathrm{T} y - 2\mu \bar{y} + \mu^2 \right) \right]
\end{aligned} \tag{5}
$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is the mean of data points and $y^\mathrm{T} y = \sum_{i=1}^{n} y_i^2$ is the sum of squared data points.
Completing the square over μ, finally gives

$$
p(y|\mu) = \left(\frac{\tau}{2\pi} \right)^{n/2} \cdot \exp\left[-\frac{\tau n}{2} \left((\mu - \bar{y})^2 - \bar{y}^2 + \frac{1}{n} y^\mathrm{T} y \right) \right] \tag{6}
$$

In other words, the likelihood function (→ I/5.1.2) is proportional to an exponential of a squared form of μ, weighted by some constant:

$$p(y|\mu) \propto \exp\left[-\frac{\tau n}{2}(\mu - \bar{y})^2\right] \; . \tag{7}$$

The same is true for a normal distribution (→ II/3.2.1) over μ

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \tag{8}$$

the probability density function of which (→ II/3.2.10)

$$p(\mu) = \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp\left[-\frac{\lambda_0}{2}(\mu - \mu_0)^2\right] \tag{9}$$

exhibits the same proportionality

$$p(\mu) \propto \exp\left[-\frac{\lambda_0}{2}(\mu - \mu_0)^2\right] \tag{10}$$

and is therefore conjugate relative to the likelihood.

■

Sources:
- Bishop, Christopher M. (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, pp. 97-98, eq. 2.138; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

1.2.7 Posterior distribution

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume a normal distribution (→ III/1.2.6) over the model parameter μ:

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a normal distribution (→ II/3.2.1)

$$p(\mu|y) = \mathcal{N}(\mu; \mu_n, \lambda_n^{-1}) \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \tag{4}$$

with the sample mean (→ I/1.10.2) $\bar{y}$ and the inverse variance or precision (→ I/1.11.12) $\tau = 1/\sigma^2$.

Proof: According to Bayes' theorem (→ I/5.3.1), the posterior distribution (→ I/5.1.7) is given by

$$p(\mu|y) = \frac{p(y|\mu)\, p(\mu)}{p(y)} \; . \tag{5}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (→ I/5.1.9) to the numerator:

$$p(\mu|y) \propto p(y|\mu)\, p(\mu) = p(y, \mu) \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$\begin{aligned} p(y|\mu) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right] \\ &= \left(\sqrt{\frac{1}{2\pi\sigma^2}}\right)^n \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \end{aligned} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$\begin{aligned} p(y|\mu) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \tau^{-1}) \\ &= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left[-\frac{\tau}{2}(y_i-\mu)^2\right] \\ &= \left(\sqrt{\frac{\tau}{2\pi}}\right)^n \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \end{aligned} \tag{8}$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

Combining the likelihood function (→ I/5.1.2) (8) with the prior distribution (→ I/5.1.3) (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{aligned} p(y, \mu) &= p(y|\mu)\, p(\mu) \\ &= \left(\sqrt{\frac{\tau}{2\pi}}\right)^n \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp\left[-\frac{\lambda_0}{2}(\mu-\mu_0)^2\right] \; . \end{aligned} \tag{9}$$

Rearranging the terms, we then have:

$$p(y, \mu) = \left(\frac{\tau}{2\pi}\right)^{n/2} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{\lambda_0}{2}(\mu-\mu_0)^2\right] \; . \tag{10}$$

Expanding the products in the exponent (→ III/1.2.6) gives

$$
\begin{aligned}
p(y,\mu) &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \left(\frac{\lambda_0}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n} \tau(y_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right)\right] \\
&= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \left(\frac{\lambda_0}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n} \tau(y_i^2-2y_i\mu+\mu^2) + \lambda_0(\mu^2-2\mu\mu_0+\mu_0^2)\right)\right] \\
&= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \left(\frac{\lambda_0}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\left(\tau(y^\mathrm{T}y-2n\bar{y}\mu+n\mu^2) + \lambda_0(\mu^2-2\mu\mu_0+\mu_0^2)\right)\right] \\
&= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \left(\frac{\lambda_0}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}\left(\mu^2(\tau n+\lambda_0) - 2\mu(\tau n\bar{y}+\lambda_0\mu_0) + (\tau y^\mathrm{T}y+\lambda_0\mu_0^2)\right)\right]
\end{aligned} \tag{11}
$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $y^\mathrm{T}y = \sum_{i=1}^{n} y_i^2$. Completing the square in μ then yields

$$
p(y,\mu) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \left(\frac{\lambda_0}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left[-\frac{\lambda_n}{2}(\mu-\mu_n)^2 + f_n\right] \tag{12}
$$

with the posterior hyperparameters (→ I/5.1.7)

$$
\begin{aligned}
\mu_n &= \frac{\lambda_0\mu_0 + \tau n\bar{y}}{\lambda_0 + \tau n} \\
\lambda_n &= \lambda_0 + \tau n
\end{aligned} \tag{13}
$$

and the remaining independent term

$$
f_n = -\frac{1}{2}\left(\tau y^\mathrm{T}y + \lambda_0\mu_0^2 - \lambda_n\mu_n^2\right) \; . \tag{14}
$$

Ergo, the joint likelihood in (12) is proportional to

$$
p(y,\mu) \propto \exp\left[-\frac{\lambda_n}{2}(\mu-\mu_n)^2\right] \; , \tag{15}
$$

such that the posterior distribution over μ is given by

$$
p(\mu|y) = \mathcal{N}(\mu; \mu_n, \lambda_n^{-1}) \; . \tag{16}
$$

with the posterior hyperparameters given in (13).

■

Sources:

- Bishop, Christopher M. (2006): "Bayesian inference for the Gaussian"; in: *Pattern Recognition for Machine Learning*, ch. 2.3.6, p. 98, eqs. 2.139-2.142; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springe 20%202006.pdf.

1.2.8 Log model evidence

Theorem: Let

$$m: \; y = \{y_1, \ldots, y_n\} \; , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume a normal distribution (→ III/1.2.6) over the model parameter μ:

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \lambda_0^{-1}) \; . \tag{2}$$

Then, the log model evidence (→ IV/??) for this model is

$$\log p(y|m) = \frac{n}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{\lambda_0}{\lambda_n}\right) - \frac{1}{2}\left(\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2\right) \; . \tag{3}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \tag{4}$$

with the sample mean (→ I/1.10.2) $\bar{y}$ and the inverse variance or precision (→ I/1.11.12) $\tau = 1/\sigma^2$.

Proof: According to the law of marginal probability (→ I/1.3.3), the model evidence (→ I/5.1.11) for this model is:

$$p(y|m) = \int p(y|\mu) \, p(\mu) \, \mathrm{d}\mu \; . \tag{5}$$

According to the law of conditional probability (→ I/1.3.4), the integrand is equivalent to the joint likelihood (→ I/5.1.5):

$$p(y|m) = \int p(y, \mu) \, \mathrm{d}\mu \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$\begin{aligned} p(y|\mu, \sigma^2) &= \prod_{i=1}^{n} \mathcal{N}(y_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right] \\ &= \left(\sqrt{\frac{1}{2\pi\sigma^2}}\right)^n \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right] \end{aligned} \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$
\begin{aligned}
p(y|\mu,\tau) &= \prod_{i=1}^{n} \mathcal{N}(y_i;\mu,\tau^{-1}) \\
&= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \cdot \exp\left[-\frac{\tau}{2}\left(y_i-\mu\right)^2\right] \\
&= \left(\sqrt{\frac{\tau}{2\pi}}\right)^n \cdot \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right]
\end{aligned} \tag{8}
$$

using the inverse variance or precision $\tau = 1/\sigma^2$.

When deriving the posterior distribution (→ III/1.2.7) $p(\mu|y)$, the joint likelihood $p(y,\mu)$ is obtained as

$$
p(y,\mu) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \exp\left[-\frac{\lambda_n}{2}(\mu-\mu_n)^2 - \frac{1}{2}\left(\tau y^\mathrm{T} y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2\right)\right] \; . \tag{9}
$$

Using the probability density function of the normal distribution (→ II/3.2.10), we can rewrite this as

$$
p(y,\mu) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{2\pi}} \cdot \sqrt{\frac{2\pi}{\lambda_n}} \cdot \mathcal{N}(\mu;\lambda_n^{-1}) \cdot \exp\left[-\frac{1}{2}\left(\tau y^\mathrm{T} y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2\right)\right] \; . \tag{10}
$$

Now, μ can be integrated out using the properties of the probability density function (→ I/1.7.1):

$$
p(y|m) = \int p(y,\mu)\,\mathrm{d}\mu = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \cdot \sqrt{\frac{\lambda_0}{\lambda_n}} \cdot \exp\left[-\frac{1}{2}\left(\tau y^\mathrm{T} y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2\right)\right] \; . \tag{11}
$$

Thus, the log model evidence (→ IV/**??**) of this model is given by

$$
\log p(y|m) = \frac{n}{2}\log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2}\log\left(\frac{\lambda_0}{\lambda_n}\right) - \frac{1}{2}\left(\tau y^T y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2\right) \; . \tag{12}
$$

■

1.2.9 Accuracy and complexity

Theorem: Let

$$
y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}
$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume a statistical model (→ I/5.1.4) imposing a normal distribution (→ III/1.2.6) as the prior distribution (→ I/5.1.3) on the model parameter μ:

$$
m: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \; . \tag{2}
$$

Then, accuracy and complexity (→ IV/**??**) of this model are

$$
\begin{aligned}
\mathrm{Acc}(m) &= \frac{n}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2}\left[\tau y^\mathrm{T} y - 2\,\tau n \bar{y} \mu_n + \tau n \mu_n^2 + \frac{\tau n}{\lambda_n}\right] \\
\mathrm{Com}(m) &= \frac{1}{2}\left[\frac{\lambda_0}{\lambda_n} + \lambda_0 (\mu_0 - \mu_n)^2 - 1 + \log\left(\frac{\lambda_0}{\lambda_n}\right)\right]
\end{aligned} \tag{3}
$$

where μ_n and λ_n are the posterior hyperparameters for the univariate Gaussian with known variance (→ III/1.2.7), $\tau = 1/\sigma^2$ is the inverse variance or precision (→ I/1.11.12) and $\bar{y}$ is the sample mean (→ I/1.10.2).

Proof: Model accuracy and complexity are defined as (→ IV/**??**)

$$
\begin{aligned}
\mathrm{LME}(m) &= \mathrm{Acc}(m) - \mathrm{Com}(m) \\
\mathrm{Acc}(m) &= \langle \log p(y|\mu,m) \rangle_{p(\mu|y,m)} \\
\mathrm{Com}(m) &= \mathrm{KL}\left[p(\mu|y,m)\,||\,p(\mu|m)\right] \; .
\end{aligned} \tag{4}
$$

The accuracy term is the expectation (→ I/1.10.1) of the log-likelihood function (→ I/4.1.2) $\log p(y|\mu)$ with respect to the posterior distribution (→ I/5.1.7) $p(\mu|y)$. With the log-likelihood function for the univariate Gaussian with known variance (→ III/1.2.2) and the posterior distribution for the univariate Gaussian with known variance (→ III/1.2.7), the model accuracy of m evaluates to:

$$
\begin{aligned}
\mathrm{Acc}(m) &= \langle \log p(y|\mu) \rangle_{p(\mu|y)} \\
&= \left\langle \frac{n}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{\tau}{2}\left(y^\mathrm{T} y - 2n\bar{y}\mu + n\mu^2\right) \right\rangle_{\mathcal{N}(\mu_n,\lambda_n^{-1})} \\
&= \frac{n}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2}\left[\tau y^\mathrm{T} y - 2\tau n \bar{y} \mu_n + \tau n \mu_n^2 + \frac{\tau n}{\lambda_n}\right] \; .
\end{aligned} \tag{5}
$$

The complexity penalty is the Kullback-Leibler divergence (→ I/2.5.1) of the posterior distribution (→ I/5.1.7) $p(\mu|y)$ from the prior distribution (→ I/5.1.3) $p(\mu)$. With the prior distribution (→ III/1.2.6) given by (2), the posterior distribution for the univariate Gaussian with known variance (→ III/1.2.7) and the Kullback-Leibler divergence of the normal distribution (→ II/3.2.24), the model complexity of m evaluates to:

$$
\begin{aligned}
\mathrm{Com}(m) &= \mathrm{KL}\left[p(\mu|y)\,||\,p(\mu)\right] \\
&= \mathrm{KL}\left[\mathcal{N}(\mu_n,\lambda_n^{-1})\,||\,\mathcal{N}(\mu_0,\lambda_0^{-1})\right] \\
&= \frac{1}{2}\left[\frac{\lambda_0}{\lambda_n} + \lambda_0 (\mu_0 - \mu_n)^2 - 1 + \log\left(\frac{\lambda_0}{\lambda_n}\right)\right] \; .
\end{aligned} \tag{6}
$$

A control calculation confirms that

$$
\mathrm{Acc}(m) - \mathrm{Com}(m) = \mathrm{LME}(m) \tag{7}
$$

where $\mathrm{LME}(m)$ is the log model evidence for the univariate Gaussian with known variance (→ III/1.2.8).

■

1.2.10 Log Bayes factor

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \; , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume two statistical models (→ I/5.1.4), one assuming that μ is zero (null model (→ I/4.3.2)), the other imposing a normal distribution (→ III/1.2.6) as the prior distribution (→ I/5.1.3) on the model parameter μ (alternative (→ I/4.3.3)):

$$\begin{aligned} m_0 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu = 0 \\ m_1 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \; . \end{aligned} \tag{2}$$

Then, the log Bayes factor (→ IV/**??**) in favor of m_1 against m_0 is

$$\mathrm{LBF}_{10} = \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left(\lambda_0 \mu_0^2 - \lambda_n \mu_n^2 \right) \tag{3}$$

where μ_n and λ_n are the posterior hyperparameters for the univariate Gaussian with known variance (→ III/1.2.7) which are functions of the inverse variance or precision (→ I/1.11.12) $\tau = 1/\sigma^2$ and the sample mean (→ I/1.10.2) $\bar{y}$.

Proof: The log Bayes factor is equal to the difference of two log model evidences (→ IV/**??**):

$$\mathrm{LBF}_{12} = \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \tag{4}$$

The LME of the alternative m_1 is equal to the log model evidence for the univariate Gaussian with known variance (→ III/1.2.8):

$$\mathrm{LME}(m_1) = \log p(y|m_1) = \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left(\tau y^\mathrm{T} y + \lambda_0 \mu_0^2 - \lambda_n \mu_n^2 \right) \; . \tag{5}$$

Because the null model m_0 has no free parameter, its log model evidence (→ IV/**??**) (logarithmized marginal likelihood (→ I/5.1.11)) is equal to the log-likelihood function for the univariate Gaussian with known variance (→ III/1.2.2) at the value $\mu = 0$:

$$\mathrm{LME}(m_0) = \log p(y|\mu = 0) = \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) - \frac{1}{2} \left(\tau y^\mathrm{T} y \right) \; . \tag{6}$$

Subtracting the two LMEs from each other, the LBF emerges as

$$\mathrm{LBF}_{10} = \mathrm{LME}(m_1) - \mathrm{LME}(m_0) = \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left(\lambda_0 \mu_0^2 - \lambda_n \mu_n^2 \right) \tag{7}$$

where the posterior hyperparameters (→ I/5.1.7) are given by (→ III/1.2.7)

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \tag{8}$$

with the sample mean (→ I/1.10.2) $\bar{y}$ and the inverse variance or precision (→ I/1.11.12) $\tau = 1/\sigma^2$.

■

1.2.11 Expectation of log Bayes factor

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume two statistical models (→ I/5.1.4), one assuming that μ is zero (null model (→ I/4.3.2)), the other imposing a normal distribution (→ III/1.2.6) as the prior distribution (→ I/5.1.3) on the model parameter μ (alternative (→ I/4.3.3)):

$$\begin{aligned} m_0 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu = 0 \\ m_1 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \; . \end{aligned} \tag{2}$$

Then, under the null hypothesis (→ I/4.3.2) that m_0 generated the data, the expectation (→ I/1.10.1) of the log Bayes factor (→ IV/**??**) in favor of m_1 with $\mu_0 = 0$ against m_0 is

$$\langle \mathrm{LBF}_{10} \rangle = \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left(\frac{\lambda_n - \lambda_0}{\lambda_n} \right) \tag{3}$$

where λ_n is the posterior precision for the univariate Gaussian with known variance (→ III/1.2.7).

Proof: The log Bayes factor for the univariate Gaussian with known variance (→ III/1.2.10) is

$$\mathrm{LBF}_{10} = \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left(\lambda_0 \mu_0^2 - \lambda_n \mu_n^2 \right) \tag{4}$$

where the posterior hyperparameters (→ I/5.1.7) are given by (→ III/1.2.7)

$$\begin{aligned} \mu_n &= \frac{\lambda_0 \mu_0 + \tau n \bar{y}}{\lambda_0 + \tau n} \\ \lambda_n &= \lambda_0 + \tau n \end{aligned} \tag{5}$$

with the sample mean (→ I/1.10.2) $\bar{y}$ and the inverse variance or precision (→ I/1.11.12) $\tau = 1/\sigma^2$. Plugging μ_n from (5) into (4), we obtain:

$$\begin{aligned} \mathrm{LBF}_{10} &= \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left(\lambda_0 \mu_0^2 - \lambda_n \, \frac{(\lambda_0 \mu_0 + \tau n \bar{y})^2}{\lambda_n^2} \right) \\ &= \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) - \frac{1}{2} \left(\lambda_0 \mu_0^2 - \frac{1}{\lambda_n} (\lambda_0^2 \mu_0^2 - 2 \tau n \lambda_0 \mu_0 \bar{y} + \tau^2 (n \bar{y})^2) \right) \end{aligned} \tag{6}$$

Because m_1 uses a zero-mean prior distribution (→ I/5.1.3) with prior mean (→ I/1.10.1) $\mu_0 = 0$ per construction, the log Bayes factor simplifies to:

$$\mathrm{LBF}_{10} = \frac{1}{2} \log \left(\frac{\lambda_0}{\lambda_n} \right) + \frac{1}{2} \left(\frac{\tau^2 (n \bar{y})^2}{\lambda_n} \right) \; . \tag{7}$$

From (1), we know that the data are distributed as $y_i \sim \mathcal{N}(\mu, \sigma^2)$, such that we can derive the expectation (→ I/1.10.1) of $(n\bar{y})^2$ as follows:

$$
\begin{aligned}
\left\langle (n\bar{y})^2 \right\rangle = \left\langle \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \right\rangle &= \left\langle n y_i^2 + (n^2-n)[y_i y_j]_{i \neq j} \right\rangle \\
&= n(\mu^2 + \sigma^2) + (n^2 - n)\mu^2 \\
&= n^2 \mu^2 + n\sigma^2 \; .
\end{aligned} \tag{8}
$$

Applying this expected value (→ I/1.10.1) to (7), the expected LBF emerges as:

$$
\begin{aligned}
\langle \mathrm{LBF}_{10} \rangle &= \frac{1}{2} \log\left(\frac{\lambda_0}{\lambda_n}\right) + \frac{1}{2}\left(\frac{\tau^2(n^2\mu^2 + n\sigma^2)}{\lambda_n}\right) \\
&= \frac{1}{2} \log\left(\frac{\lambda_0}{\lambda_n}\right) + \frac{1}{2}\left(\frac{(\tau n \mu)^2 + \tau n}{\lambda_n}\right)
\end{aligned} \tag{9}
$$

Under the null hypothesis (→ I/4.3.2) that m_0 generated the data, the unknown mean is $\mu = 0$, such that the log Bayes factor further simplifies to:

$$
\langle \mathrm{LBF}_{10} \rangle = \frac{1}{2} \log\left(\frac{\lambda_0}{\lambda_n}\right) + \frac{1}{2}\left(\frac{\tau n}{\lambda_n}\right) \; . \tag{10}
$$

Finally, plugging λ_n from (5) into (10), we obtain:

$$
\langle \mathrm{LBF}_{10} \rangle = \frac{1}{2} \log\left(\frac{\lambda_0}{\lambda_n}\right) + \frac{1}{2}\left(\frac{\lambda_n - \lambda_0}{\lambda_n}\right) \; . \tag{11}
$$

■

1.2.12 Cross-validated log model evidence

Theorem: Let

$$
y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}
$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume two statistical models (→ I/5.1.4), one assuming that μ is zero (null model (→ I/4.3.2)), the other imposing a normal distribution (→ III/1.2.6) as the prior distribution (→ I/5.1.3) on the model parameter μ (alternative (→ I/4.3.3)):

$$
\begin{aligned}
m_0 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu = 0 \\
m_1 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \; .
\end{aligned} \tag{2}
$$

Then, the cross-validated log model evidences (→ IV/??) of m_0 and m_1 are

$$
\begin{aligned}
\mathrm{cvLME}(m_0) &= \frac{n}{2} \log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2}\left(\tau y^\mathrm{T} y\right) \\
\mathrm{cvLME}(m_1) &= \frac{n}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2}\left[y^\mathrm{T} y + \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)}\right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right]
\end{aligned} \tag{3}
$$

where $\bar{y}$ is the sample mean (→ I/1.10.2), $\tau = 1/\sigma^2$ is the inverse variance or precision (→ I/1.11.12), $y_1^{(i)}$ are the training data in the i-th cross-validation fold and S is the number of data subsets (→ IV/??).

Proof: For evaluation of the cross-validated log model evidences (→ IV/??) (cvLME), we assume that n data points are divided into $S \mid n$ data subsets without remainder. Then, the number of training data points n_1 and test data points n_2 are given by

$$\begin{split} n &= n_1 + n_2 \\ n_1 &= \frac{S-1}{S} n \\ n_2 &= \frac{1}{S} n \; , \end{split} \tag{4}$$

such that training data y_1 and test data y_2 in the i-th cross-validation fold are

$$\begin{split} y &= \left\lbrace y_1, \ldots, y_n \right\rbrace \\ y_1^{(i)} &= \left\lbrace x \in y \mid x \notin y_2^{(i)} \right\rbrace = y \backslash y_2^{(i)} \\ y_2^{(i)} &= \left\lbrace y_{(i-1) \cdot n_2 + 1}, \ldots, y_{i \cdot n_2} \right\rbrace \; . \end{split} \tag{5}$$

First, we consider the null model m_0 assuming $\mu = 0$. Because this model has no free parameter, nothing is estimated from the training data and the assumed parameter value is applied to the test data. Consequently, the out-of-sample log model evidence (oosLME) is equal to the log-likelihood function (→ III/1.2.2) of the test data at $\mu = 0$:

$$\mathrm{oosLME}_i(m_0) = \log p\left(y_2^{(i)} \middle| \mu = 0 \right) = \frac{n_2}{2} \log\left(\frac{\tau}{2\pi} \right) - \frac{1}{2} \left[\tau y_2^{(i)\mathrm{T}} y_2^{(i)} \right] \; . \tag{6}$$

By definition, the cross-validated log model evidence is the sum of out-of-sample log model evidences (→ IV/??) over cross-validation folds, such that the cvLME of m_0 is:

$$\begin{split} \mathrm{cvLME}(m_0) &= \sum_{i=1}^{S} \mathrm{oosLME}_i(m_0) \\ &= \sum_{i=1}^{S} \left(\frac{n_2}{2} \log\left(\frac{\tau}{2\pi} \right) - \frac{1}{2} \left[\tau y_2^{(i)\mathrm{T}} y_2^{(i)} \right] \right) \\ &= \frac{n}{2} \log\left(\frac{\tau}{2\pi} \right) - \frac{1}{2} \left[\tau y^\mathrm{T} y \right] \; . \end{split} \tag{7}$$

Next, we have a look at the alternative m_1 assuming $\mu \neq 0$. First, the training data $y_1^{(i)}$ are analyzed using a non-informative prior distribution (→ I/5.2.3) and applying the posterior distribution for the univariate Gaussian with known variance (→ III/1.2.7):

$$
\begin{aligned}
\mu_0^{(1)} &= 0 \\
\lambda_0^{(1)} &= 0 \\
\mu_n^{(1)} &= \frac{\tau n_1 \bar{y}_1^{(i)} + \lambda_0^{(1)} \mu_0^{(1)}}{\tau n_1 + \lambda_0^{(1)}} = \bar{y}_1^{(i)} \\
\lambda_n^{(1)} &= \tau n_1 + \lambda_0^{(1)} = \tau n_1 \; .
\end{aligned} \tag{8}
$$

This results in a posterior characterized by $\mu_n^{(1)}$ and $\lambda_n^{(1)}$. Then, the test data $y_2^{(i)}$ are analyzed using this posterior as an informative prior distribution (→ I/5.2.3), again applying the posterior distribution for the univariate Gaussian with known variance (→ III/1.2.7):

$$
\begin{aligned}
\mu_0^{(2)} &= \mu_n^{(1)} = \bar{y}_1^{(i)} \\
\lambda_0^{(2)} &= \lambda_n^{(1)} = \tau n_1 \\
\mu_n^{(2)} &= \frac{\tau n_2 \bar{y}_2^{(i)} + \lambda_0^{(2)} \mu_0^{(2)}}{\tau n_2 + \lambda_0^{(2)}} = \bar{y} \\
\lambda_n^{(2)} &= \tau n_2 + \lambda_0^{(2)} = \tau n \; .
\end{aligned} \tag{9}
$$

In the test data, we now have a prior characterized by $\mu_0^{(2)}/\lambda_0^{(2)}$ and a posterior characterized $\mu_n^{(2)}/\lambda_n^{(2)}$. Applying the log model evidence for the univariate Gaussian with known variance (→ III/1.2.8), the out-of-sample log model evidence (oosLME) therefore follows as

$$
\begin{aligned}
\mathrm{oosLME}_i(m_1) &= \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{\lambda_0^{(2)}}{\lambda_n^{(2)}}\right) - \frac{1}{2}\left[\tau {y_2^{(i)}}^\mathrm{T} y_2^{(i)} + \lambda_0^{(2)} {\mu_0^{(2)}}^2 - \lambda_n^{(2)} {\mu_n^{(2)}}^2\right] \\
&= \frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{n_1}{n}\right) - \frac{1}{2}\left[\tau {y_2^{(i)}}^\mathrm{T} y_2^{(i)} + \frac{\tau}{n_1}\left(n_1 \bar{y}_1^{(i)}\right)^2 - \frac{\tau}{n}(n\bar{y})^2\right] \; .
\end{aligned} \tag{10}
$$

Again, because the cross-validated log model evidence is the sum of out-of-sample log model evidences (→ IV/**??**) over cross-validation folds, the cvLME of m_1 becomes:

$$
\begin{aligned}
\mathrm{cvLME}(m_1) &= \sum_{i=1}^{S} \mathrm{oosLME}_i(m_1) \\
&= \sum_{i=1}^{S} \left(\frac{n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{1}{2} \log\left(\frac{n_1}{n}\right) - \frac{1}{2}\left[\tau {y_2^{(i)}}^\mathrm{T} y_2^{(i)} + \frac{\tau}{n_1}\left(n_1 \bar{y}_1^{(i)}\right)^2 - \frac{\tau}{n}(n\bar{y})^2\right]\right) \\
&= \frac{S \cdot n_2}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{S}{2} \log\left(\frac{n_1}{n}\right) - \frac{\tau}{2} \sum_{i=1}^{S} \left[{y_2^{(i)}}^\mathrm{T} y_2^{(i)} + \frac{\left(n_1 \bar{y}_1^{(i)}\right)^2}{n_1} - \frac{(n\bar{y})^2}{n}\right] \\
&= \frac{n}{2} \log\left(\frac{\tau}{2\pi}\right) + \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2}\left[y^\mathrm{T} y + \sum_{i=1}^{S}\left(\frac{\left(n_1 \bar{y}_1^{(i)}\right)^2}{n_1} - \frac{(n\bar{y})^2}{n}\right)\right] \; .
\end{aligned} \tag{11}
$$

Together, (7) and (11) conform to the results given in (3).

■

1.2.13 Cross-validated log Bayes factor

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume two statistical models (→ I/5.1.4), one assuming that μ is zero (null model (→ I/4.3.2)), the other imposing a normal distribution (→ III/1.2.6) as the prior distribution (→ I/5.1.3) on the model parameter μ (alternative (→ I/4.3.3)):

$$\begin{aligned} m_0 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu = 0 \\ m_1 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \; . \end{aligned} \tag{2}$$

Then, the cross-validated (→ IV/??) log Bayes factor (→ IV/??) in favor of m_1 against m_0 is

$$\mathrm{cvLBF}_{10} = \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \tag{3}$$

where $\bar{y}$ is the sample mean (→ I/1.10.2), $\tau = 1/\sigma^2$ is the inverse variance or precision (→ I/1.11.12), $y_1^{(i)}$ are the training data in the i-th cross-validation fold and S is the number of data subsets (→ IV/??).

Proof: The relationship between log Bayes factor and log model evidences (→ IV/??) also holds for cross-validated log bayes factor (→ IV/??) (cvLBF) and cross-validated log model evidences (→ IV/??) (cvLME):

$$\mathrm{cvLBF}_{12} = \mathrm{cvLME}(m_1) - \mathrm{cvLME}(m_2) \; . \tag{4}$$

The cross-validated log model evidences (→ IV/??) of m_0 and m_1 are given by (→ III/1.2.12)

$$\begin{aligned} \mathrm{cvLME}(m_0) &= \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) - \frac{1}{2} \left(\tau y^\mathrm{T} y \right) \\ \mathrm{cvLME}(m_1) &= \frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \left[y^\mathrm{T} y + \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right] \; . \end{aligned} \tag{5}$$

Subtracting the two cvLMEs from each other, the cvLBF emerges as

$$\begin{aligned} \mathrm{cvLBF}_{10} &= \mathrm{cvLME}(m_1) - \mathrm{LME}(m_0) \\ &= \left(\frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) + \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \left[y^\mathrm{T} y + \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right] \right) \\ &- \left(\frac{n}{2} \log \left(\frac{\tau}{2\pi} \right) - \frac{1}{2} \left(\tau y^\mathrm{T} y \right) \right) \\ &= \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \; . \end{aligned} \tag{6}$$

■

1.2.14 Expectation of cross-validated log Bayes factor

Theorem: Let

$$y = \{y_1, \ldots, y_n\} \, , \quad y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

be a univariate Gaussian data set (→ III/1.2.1) with unknown mean μ and known variance σ^2. Moreover, assume two statistical models (→ I/5.1.4), one assuming that μ is zero (null model (→ I/4.3.2)), the other imposing a normal distribution (→ III/1.2.6) as the prior distribution (→ I/5.1.3) on the model parameter μ (alternative (→ I/4.3.3)):

$$\begin{aligned} m_0 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu = 0 \\ m_1 &: \; y_i \sim \mathcal{N}(\mu, \sigma^2), \; \mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \; . \end{aligned} \tag{2}$$

Then, the expectation (→ I/1.10.1) of the cross-validated (→ IV/??) log Bayes factor (→ IV/??) (cvLBF) in favor of m_1 against m_0 is

$$\langle \mathrm{cvLBF}_{10} \rangle = \frac{S}{2} \log \left(\frac{S-1}{S} \right) + \frac{1}{2} \left[\tau n \mu^2 \right] \tag{3}$$

where $\tau = 1/\sigma^2$ is the inverse variance or precision (→ I/1.11.12) and S is the number of data subsets (→ IV/??).

Proof: The cross-validated log Bayes factor for the univariate Gaussian with known variance (→ III/1.2.13) is

$$\mathrm{cvLBF}_{10} = \frac{S}{2} \log \left(\frac{S-1}{S} \right) - \frac{\tau}{2} \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)} \right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \tag{4}$$

From (1), we know that the data are distributed as $y_i \sim \mathcal{N}(\mu, \sigma^2)$, such that we can derive the expectation (→ I/1.10.1) of $(n\bar{y})^2$ and $\left(n_1 \bar{y}_1^{(i)} \right)^2$ as follows:

$$\begin{aligned} \left\langle (n\bar{y})^2 \right\rangle = \left\langle \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \right\rangle &= \left\langle n y_i^2 + (n^2 - n)[y_i y_j]_{i \neq j} \right\rangle \\ &= n(\mu^2 + \sigma^2) + (n^2 - n)\mu^2 \\ &= n^2 \mu^2 + n\sigma^2 \; . \end{aligned} \tag{5}$$

Applying this expected value (→ I/1.10.1) to (4), the expected cvLBF emerges as:

$$
\begin{aligned}
\langle \mathrm{cvLBF}_{10} \rangle &= \left\langle \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2} \sum_{i=1}^{S} \left(\frac{\left(n_1 \bar{y}_1^{(i)}\right)^2}{n_1} - \frac{(n\bar{y})^2}{n} \right) \right\rangle \\
&= \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2} \sum_{i=1}^{S} \left(\frac{\left\langle \left(n_1 \bar{y}_1^{(i)}\right)^2 \right\rangle}{n_1} - \frac{\langle (n\bar{y})^2 \rangle}{n} \right) \\
&\overset{(5)}{=} \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2} \sum_{i=1}^{S} \left(\frac{n_1^2\mu^2 + n_1\sigma^2}{n_1} - \frac{n^2\mu^2 + n\sigma^2}{n} \right) \\
&= \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2} \sum_{i=1}^{S} \left([n_1\mu^2 + \sigma^2] - [n\mu^2 + \sigma^2] \right) \\
&= \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2} \sum_{i=1}^{S} (n_1 - n)\mu^2
\end{aligned} \tag{6}
$$

Because it holds that (→ III/1.2.12) $n_1 + n_2 = n$ and $n_2 = n/S$, we finally have:

$$
\begin{aligned}
\langle \mathrm{cvLBF}_{10} \rangle &= \frac{S}{2} \log\left(\frac{S-1}{S}\right) - \frac{\tau}{2} \sum_{i=1}^{S} (-n_2)\mu^2 \\
&= \frac{S}{2} \log\left(\frac{S-1}{S}\right) + \frac{1}{2} \left[\tau n \mu^2\right] \; .
\end{aligned} \tag{7}
$$

■

1.3 Analysis of variance

1.3.1 One-way ANOVA

Definition: Consider measurements $y_{ij} \in \mathbb{R}$ from distinct objects $j = 1, \ldots, n_i$ in separate groups $i = 1, \ldots, k$.
Then, in one-way analysis of variance (ANOVA), these measurements are assumed to come from normal distributions (→ II/3.2.1)

$$
y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \quad \text{for all} \quad i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, n_i \tag{1}
$$

where
- μ_i is the expected value (→ I/1.10.1) in group i and
- σ^2 is the common variance (→ I/1.11.1) across groups.

Alternatively, the model may be written as

$$
\begin{aligned}
y_{ij} &= \mu_i + \varepsilon_{ij} \\
\varepsilon_{ij} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned} \tag{2}
$$

where ε_{ij} is the error term (→ III/1.4.1) belonging to observation j in category i and ε_{ij} are the independent and identically distributed.

Sources:
- Bortz, Jürgen (1977): "Einfaktorielle Varianzanalyse"; in: *Lehrbuch der Statistik. Für Sozialwissenschaftler*, ch. 12.1, pp. 528ff.; URL: https://books.google.de/books?id=lNCyBgAAQBAJ.
- Denziloe (2018): "Derive the distribution of the ANOVA F-statistic under the alternative hypothesis"; in: *StackExchange CrossValidated*, retrieved on 2022-11-06; URL: https://stats.stackexchange.com/questions/355594/derive-the-distribution-of-the-anova-f-statistic-under-the-alternative-hypothesi.

1.3.2 Treatment sum of squares

Definition: Let there be an analysis of variance (ANOVA) model with one (→ III/1.3.1), two (→ III/1.3.8) or multiple factors influencing the measured data y (here, using the reparametrized version (→ III/1.3.7) of one-way ANOVA (→ III/1.3.1)):

$$y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{1}$$

Then, the treatment sum of squares is defined as the explained sum of squares (→ III/1.5.8) (ESS) for each main effect, i.e. as the sum of squared deviations of the average for each level of the factor, from the average across all observations:

$$\mathrm{SS}_{\mathrm{treat}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \; . \tag{2}$$

Here, $\bar{y}\_i$ is the mean for the i-th level of the factor (out of k levels), computed from n_i values y_ij, and $\bar{y}$ is the mean across all values y_ij.

Sources:
- Wikipedia (2022): "Analysis of variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: https://en.wikipedia.org/wiki/Analysis_of_variance#Partitioning_of_the_sum_of_squares.

1.3.3 Ordinary least squares for one-way ANOVA

Theorem: Given the one-way analysis of variance (→ III/1.3.1) assumption

$$y_{ij} = \mu_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, k, \; j = 1, \ldots, n_i \; , \tag{1}$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) are given by

$$\hat{\mu}_i = \bar{y}_i \tag{2}$$

where $\bar{y}_i$ is the sample mean (→ I/1.10.2) of all observations in group (→ III/1.3.1) i:

$$\hat{\mu}_i = \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \; . \tag{3}$$

Proof: The residual sum of squares (→ III/1.5.9) for this model is

$$\mathrm{RSS}(\mu) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \tag{4}$$

and the derivatives of RSS with respect to μ_i are

$$\begin{aligned}
\frac{\mathrm{dRSS}(\mu)}{\mathrm{d}\mu_i} &= \sum_{j=1}^{n_i} \frac{\mathrm{d}}{\mathrm{d}\mu_i}(y_{ij} - \mu_i)^2 \\
&= \sum_{j=1}^{n_i} 2(y_{ij} - \mu_i)(-1) \\
&= 2\sum_{j=1}^{n_i}(\mu_i - y_{ij}) \\
&= 2n_i\mu_i - 2\sum_{j=1}^{n_i} y_{ij} \quad \text{for} \quad i = 1, \ldots, k \; .
\end{aligned} \tag{5}$$

Setting these derivatives to zero, we obtain the estimates of μ_i:

$$\begin{aligned}
0 &= 2n_i\hat{\mu}_i - 2\sum_{j=1}^{n_i} y_{ij} \\
\hat{\mu}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad \text{for} \quad i = 1, \ldots, k \; .
\end{aligned} \tag{6}$$

■

1.3.4 Sums of squares in one-way ANOVA

Theorem: Given one-way analysis of variance (→ III/1.3.1),

$$y_{ij} = \mu_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

sums of squares can be partitioned as follows

$$\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_{\mathrm{treat}} + \mathrm{SS}_{\mathrm{res}} \tag{2}$$

where $\mathrm{SS}_{\mathrm{tot}}$ is the total sum of squares (→ III/1.5.7), $\mathrm{SS}_{\mathrm{treat}}$ is the treatment sum of squares (→ III/1.3.2) (equivalent to explained sum of squares (→ III/1.5.8)) and $\mathrm{SS}_{\mathrm{res}}$ is the residual sum of squares (→ III/1.5.9).

Proof: The total sum of squares (→ III/1.5.7) for one-way ANOVA (→ III/1.3.1) is given by

$$\mathrm{SS}_{\mathrm{tot}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 \tag{3}$$

where $\bar{y}$ is the mean across all values y_{ij}. This can be rewritten as

$$
\begin{split}
\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \right]^2 \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + 2 (y_{ij} - \bar{y}_i) (\bar{y}_i - \bar{y}) \right] \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{k} (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) \; .
\end{split} \tag{4}
$$

Note that the following sum is zero

$$
\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \bar{y}_i = \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \; , \tag{5}
$$

so that the sum in (4) reduces to

$$
\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \; . \tag{6}
$$

With the treatment sum of squares (→ III/1.3.2) for one-way ANOVA (→ III/1.3.1)

$$
\mathrm{SS}_\mathrm{treat} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \tag{7}
$$

and the residual sum of squares (→ III/1.5.9) for one-way ANOVA (→ III/1.3.1)

$$
\mathrm{SS}_\mathrm{res} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \; , \tag{8}
$$

we finally have:

$$
\mathrm{SS}_\mathrm{tot} = \mathrm{SS}_\mathrm{treat} + \mathrm{SS}_\mathrm{res} \; . \tag{9}
$$

■

Sources:

- Wikipedia (2022): "Analysis of variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: https://en.wikipedia.org/wiki/Analysis_of_variance#Partitioning_of_the_sum_of_squares.

1.3.5 F-test for main effect in one-way ANOVA

Theorem: Assume the one-way analysis of variance (→ III/1.3.1) model

$$
y_{ij} = \mu_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, k, \; j = 1, \ldots, n_i \; . \tag{1}
$$

Then, the test statistic (→ I/4.3.5)

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2} \tag{2}$$

follows an F-distribution (→ II/3.8.1)

$$F \sim \mathrm{F}(k-1, n-k) \tag{3}$$

under the null hypothesis (→ I/4.3.2)

$$\begin{aligned} &H_0 : \mu_1 = \ldots = \mu_k \\ &H_1 : \mu_i \neq \mu_j \quad \text{for at least one} \quad i, j \in \{1, \ldots, k\}, \; i \neq j \; . \end{aligned} \tag{4}$$

Proof: Denote sample sizes as

$$\begin{aligned} n_i &- \text{number of samples in category } i \\ n &= \sum_{i=1}^{k} n_i j \end{aligned} \tag{5}$$

and denote sample means as

$$\begin{aligned} \bar{y}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \\ \bar{y} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \; . \end{aligned} \tag{6}$$

Let μ be the common mean (→ I/1.10.1) according to H_0 given by (4), i.e. $\mu_1 = \ldots = \mu_k = \mu$. Under this null hypothesis, we have:

$$y_{ij} \sim \mathcal{N}(\mu, \sigma^2) \quad \text{for all} \quad i = 1, \ldots, k, \; j = 1, \ldots, n_i \; . \tag{7}$$

Thus, the random variable (→ I/1.2.2) $U_{ij} = (y_{ij} - \mu)/\sigma$ follows a standard normal distribution (→ II/3.2.4)

$$U_{ij} = \frac{y_{ij} - \mu}{\sigma} \sim \mathcal{N}(0, 1) \; . \tag{8}$$

Now consider the following sum:

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{n_i} U_{ij}^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\frac{y_{ij} - \mu}{\sigma}\right)^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) + (\bar{y} - \mu)\right)^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + (\bar{y} - \mu)^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) + 2(y_{ij} - \bar{y}_i)(\bar{y} - \mu) + 2(\bar{y}_i - \right. \end{aligned} \tag{9}$$

Because the following sum over j is zero for all i

$$\begin{aligned}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) &= \sum_{j=1}^{n_i} y_{ij} - n_i \bar{y}_i \\ &= \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \\ &= 0, \; i = 1, \ldots, k\end{aligned} \tag{10}$$

and the following sum over i and j is also zero

$$\begin{aligned}\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}) &= \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}) \\ &= \sum_{i=1}^{k} n_i \bar{y}_i - \bar{y} \sum_{i=1}^{k} n_i \\ &= \sum_{i=1}^{k} n_i \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - n \cdot \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \\ &= 0 \; ,\end{aligned} \tag{11}$$

non-square products in (9) disappear and the sum reduces to

$$\begin{aligned}\sum_{i=1}^{k} \sum_{j=1}^{n_i} U_{ij}^2 &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[\left(\frac{y_{ij} - \bar{y}_i}{\sigma} \right)^2 + \left(\frac{\bar{y}_i - \bar{y}}{\sigma} \right)^2 + \left(\frac{\bar{y} - \mu}{\sigma} \right)^2 \right] \\ &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - \bar{y}_i}{\sigma} \right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{\bar{y}_i - \bar{y}}{\sigma} \right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{\bar{y} - \mu}{\sigma} \right)^2 \; .\end{aligned} \tag{12}$$

Cochran's theorem states that, if a sum of squared standard normal (→ II/3.2.3) random variables (→ I/1.2.2) can be written as a sum of squared forms

$$\begin{aligned}\sum_{i=1}^{n} U_i^2 = \sum_{j=1}^{m} Q_j \quad &\text{where} \quad Q_j = U^\mathrm{T} B^{(j)} U \\ &\text{with} \quad \sum_{j=1}^{m} B^{(j)} = I_n \\ &\text{and} \quad r_j = \mathrm{rank}(B^{(j)}) \; ,\end{aligned} \tag{13}$$

then the terms Q_j are independent (→ I/1.3.6) and each term Q_j follows a chi-squared distribution (→ II/3.7.1) with r_j degrees of freedom:

$$Q_j \sim \chi^2(r_j), \; j = 1, \ldots, m \; . \tag{14}$$

Let U be the $n \times 1$ column vector of all observations

$$U = \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix} \tag{15}$$

where the group-wise $n_i \times 1$ column vectors are

$$u_1 = \begin{bmatrix} (y_{1,1} - \mu)/\sigma \\ \vdots \\ (y_{1,n_1} - \mu)/\sigma \end{bmatrix}, \quad \ldots, \quad u_k = \begin{bmatrix} (y_{k,1} - \mu)/\sigma \\ \vdots \\ (y_{k,n_k} - \mu)/\sigma \end{bmatrix} . \tag{16}$$

Then, we observe that the sum in (12) can be represented in the form of (13) using the matrices

$$\begin{aligned} B^{(1)} &= I_n - \mathrm{diag}\left(\frac{1}{n_1} J_{n_1}, \; \ldots, \; \frac{1}{n_k} J_{n_k} \right) \\ B^{(2)} &= \mathrm{diag}\left(\frac{1}{n_1} J_{n_1}, \; \ldots, \; \frac{1}{n_k} J_{n_k} \right) - \frac{1}{n} J_n \\ B^{(3)} &= \frac{1}{n} J_n \end{aligned} \tag{17}$$

where J_n is an $n \times n$ matrix of ones and $\mathrm{diag}\left(A_1, \ldots, A_n\right)$ denotes a block-diagonal matrix composed of $A_1, \ldots, A_n$. We observe that those matrices satisfy

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} U_{ij}^2 = Q_1 + Q_2 + Q_3 = U^\mathrm{T} B^{(1)} U + U^\mathrm{T} B^{(2)} U + U^\mathrm{T} B^{(3)} U \tag{18}$$

as well as

$$B^{(1)} + B^{(2)} + B^{(3)} = I_n \tag{19}$$

and their ranks are:

$$\begin{aligned} \mathrm{rank}\left(B^{(1)}\right) &= n - k \\ \mathrm{rank}\left(B^{(2)}\right) &= k - 1 \\ \mathrm{rank}\left(B^{(3)}\right) &= 1 \; . \end{aligned} \tag{20}$$

Let's write down the explained sum of squares (→ III/1.5.8) and the residual sum of squares (→ III/1.5.9) for one-way analysis of variance (→ III/1.3.1) as

$$\begin{aligned} \mathrm{ESS} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{y}_i - \bar{y}\right)^2 \\ \mathrm{RSS} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 \; . \end{aligned} \tag{21}$$

Then, using (12), (13), (14), (17) and (20), we find that

$$
\begin{aligned}
\frac{\mathrm{ESS}}{\sigma^2} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{\bar{y}_i - \bar{y}}{\sigma} \right)^2 = Q_2 = U^\mathrm{T} B^{(2)} U \sim \chi^2(k-1) \\
\frac{\mathrm{RSS}}{\sigma^2} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - \bar{y}_i}{\sigma} \right)^2 = Q_1 = U^\mathrm{T} B^{(1)} U \sim \chi^2(n-k) \; .
\end{aligned} \tag{22}
$$

Because ESS/σ^2 and RSS/σ^2 are also independent by (14), the F-statistic from (2) is equal to the ratio of two independent chi-squared distributed (→ II/3.7.1) random variables (→ I/1.2.2) divided by their degrees of freedom

$$
\begin{aligned}
F &= \frac{(\mathrm{ESS}/\sigma^2)/(k-1)}{(\mathrm{RSS}/\sigma^2)/(n-k)} \\
&= \frac{\mathrm{ESS}/(k-1)}{\mathrm{RSS}/(n-k)} \\
&= \frac{\frac{1}{k-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \\
&= \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}
\end{aligned} \tag{23}
$$

which, by definition of the F-distribution (→ II/3.8.1), is distributed as

$$
F \sim \mathrm{F}(k-1, n-k) \tag{24}
$$

under the null hypothesis (→ I/4.3.2) for the main effect.

■

Sources:

- Denziloe (2018): "Derive the distribution of the ANOVA F-statistic under the alternative hypothesis"; in: *StackExchange CrossValidated*, retrieved on 2022-11-06; URL: https://stats.stackexchange.com/questions/355594/derive-the-distribution-of-the-anova-f-statistic-under-the-alternative-hypothesi.

1.3.6 F-statistic in terms of OLS estimates

Theorem: Given the one-way analysis of variance (→ III/1.3.1) assumption

$$
y_{ij} = \mu_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}
$$

1) the F-statistic for the main effect (→ III/1.3.5) can be expressed in terms of ordinary least squares parameter estimates (→ III/1.3.3) as

$$
F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\hat{\mu}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2} \tag{2}
$$

2) or, when using the reparametrized version of one-way ANOVA (→ III/1.3.7), the F-statistic can be expressed as

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\delta}_i)^2} \; . \tag{3}$$

Proof: The F-statistic for the main effect in one-way ANOVA (→ III/1.3.5) is given in terms of the sample means (→ I/1.10.2) as

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \tag{4}$$

where $\bar{y}\_i$ is the average of all values y_ij from category i and $\bar{y}$ is the grand mean of all values y_ij from all categories $i = 1, \ldots, k$.
1) The ordinary least squares estimates for one-way ANOVA (→ III/1.3.3) are

$$\hat{\mu}_i = \bar{y}_i \; , \tag{5}$$

such that

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\hat{\mu}_i - \bar{y})^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2} \; . \tag{6}$$

2) The OLS estimates for reparametrized one-way ANOVA (→ III/1.3.7) are

$$\begin{aligned} \hat{\mu} &= \bar{y} \\ \hat{\delta}_i &= \bar{y}_i - \bar{y} \; , \end{aligned} \tag{7}$$

such that

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\delta}_i)^2} \; . \tag{8}$$

■

1.3.7 Reparametrization of one-way ANOVA

Theorem: The one-way analysis of variance (→ III/1.3.1) model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

can be rewritten using paraneters μ and δ_i instead of μ_i

$$y_{ij} = \mu + \delta_i + \varepsilon_{ij}, \; \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{2}$$

with the constraint

$$\sum_{i=1}^{k} \frac{n_i}{n} \delta_i = 0 \; , \tag{3}$$

in which case
1) the model parameters are related to each other as

$$\delta_i = \mu_i - \mu, \; i = 1, \ldots, k \; ; \tag{4}$$

2) the ordinary least squares estimates (→ III/1.3.3) are given by

$$\hat{\delta}_i = \bar{y}_i - \bar{y} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \; ; \tag{5}$$

3) the following sum of squares (→ III/1.3.4) is chi-square distributed (→ II/3.7.1)

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\hat{\delta}_i - \delta_i \right)^2 \sim \chi^2(k-1) \; ; \tag{6}$$

4) and the following test statistic (→ I/4.3.5) is F-distributed (→ II/3.8.1)

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \sim \mathrm{F}(k-1, n-k) \tag{7}$$

under the null hypothesis for the main effect (→ III/1.3.5)

$$H_0: \; \delta_1 = \ldots = \delta_k = 0 \; . \tag{8}$$

Proof:
1) Equating (1) with (2), we get:

$$\begin{split} y_{ij} = \mu + \delta_i + \varepsilon_{ij} &= \mu_i + \varepsilon_{ij} = y_{ij} \\ \mu + \delta_i &= \mu_i \\ \delta_i &= \mu_i - \mu \; . \end{split} \tag{9}$$

2) Equation (2) is a special case of the two-way analysis of variance (→ III/1.3.8) with (i) just one factor A and (ii) no interaction term. Thus, OLS estimates are identical to that of two-way ANOVA (→ III/1.3.10), i.e. given by

$$\begin{split} \hat{\mu} = \bar{y}_{\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \\ \hat{\delta}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \; . \end{split} \tag{10}$$

3) Let $U_{ij} = (y_{ij} - \mu - \delta_i)/\sigma$, such that (→ II/3.2.4) $U_{ij} \sim \mathcal{N}(0,1)$ and consider the sum of all squared random variables (→ I/1.2.2) U_{ij}:

$$\begin{split} \sum_{i=1}^{k} \sum_{j=1}^{n_i} U_{ij}^2 &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - \mu - \delta_i}{\sigma} \right)^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i) + ([\bar{y}_i - \bar{y}] - \delta_i) + (\bar{y} - \mu) \right]^2 \; . \end{split} \tag{11}$$

This square of sums, using a number of intermediate steps, can be developed (→ III/1.3.5) into a sum of squares:

$$
\begin{split}
\sum_{i=1}^{k} \sum_{j=1}^{n_i} U_{ij}^2 &= \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_i)^2 + ([\bar{y}_i - \bar{y}] - \delta_i)^2 + (\bar{y} - \mu)^2 \right] \\
&= \frac{1}{\sigma^2} \left[\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} ([\bar{y}_i - \bar{y}] - \delta_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y} - \mu)^2 \right] \; .
\end{split} \tag{12}
$$

To this sum, Cochran's theorem for one-way analysis of variance can be applied (→ III/1.3.5), yielding the distributions:

$$
\begin{split}
\frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 &\sim \chi^2(n-k) \\
\frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} ([\bar{y}_i - \bar{y}] - \delta_i)^2 \overset{(10)}{=} \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2 &\sim \chi^2(k-1) \; .
\end{split} \tag{13}
$$

4) The ratio of two chi-square distributed (→ II/3.7.1) random variables (→ I/1.2.2), divided by their degrees of freedom, is defined to be F-distributed (→ II/3.8.1), so that

$$
\begin{split}
F &= \frac{\left(\frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2 \right) / (k-1)}{\left(\frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \right) / (n-k)} \\
&= \frac{\frac{1}{k-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\hat{\delta}_i - \delta_i)^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \\
&= \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i (\hat{\delta}_i - \delta_i)^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} \\
&\overset{(8)}{=} \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i \hat{\delta}_i^2}{\frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}
\end{split} \tag{14}
$$

follows the F-distribution

$$
F \sim \mathrm{F}(k-1, n-k) \tag{15}
$$

under the null hypothesis.

■

Sources:

- Wikipedia (2022): "Analysis of variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: https://en.wikipedia.org/wiki/Analysis_of_variance#For_a_single_factor.

1.3.8 Two-way ANOVA

Definition: Let there be two factors A and B with levels $i = 1, \ldots, a$ and $j = 1, \ldots, b$ that are used to group measurements $y_{ijk} \in \mathbb{R}$ from distinct objects $k = 1, \ldots, n_{ij}$ into $a \cdot b$ categories $(i, j) \in \{1, \ldots, a\} \times \{1, \ldots, b\}$.
Then, in two-way analysis of variance (ANOVA), these measurements are assumed to come from normal distributions (→ II/3.2.1)

$$y_{ijk} \sim \mathcal{N}(\mu_{ij}, \sigma^2) \quad \text{for all} \quad i = 1, \ldots, a, \quad j = 1, \ldots, b, \quad \text{and} \quad k = 1, \ldots, n_{ij} \tag{1}$$

with

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{2}$$

where
- μ is called the "grand mean";
- α_i is the additive "main effect" of the i-th level of factor A;
- β_j is the additive "main effect" of the j-th level of factor B;
- γ_{ij} is the non-additive "interaction effect" of category (i, j);
- μ_{ij} is the expected value (→ I/1.10.1) in category (i, j); and
- σ^2 is common variance (→ I/1.11.1) across all categories.

Alternatively, the model may be written as

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{3}$$

where ε_{ijk} is the error term (→ III/1.4.1) corresponding to observation k belonging to the i-th level of A and the j-th level of B.
As the two-way ANOVA model is underdetermined, the parameters of the model are additionally subject to the constraints

$$\begin{aligned} \sum_{i=1}^{a} w_{ij} \alpha_i &= 0 \quad \text{for all} \quad j = 1, \ldots, b \\ \sum_{j=1}^{b} w_{ij} \beta_j &= 0 \quad \text{for all} \quad i = 1, \ldots, a \\ \sum_{i=1}^{a} w_{ij} \gamma_{ij} &= 0 \quad \text{for all} \quad j = 1, \ldots, b \\ \sum_{j=1}^{b} w_{ij} \gamma_{ij} &= 0 \quad \text{for all} \quad i = 1, \ldots, a \end{aligned} \tag{4}$$

where the weights are $w_{ij} = n_{ij}/n$ and the total sample size is $n = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}$.

Sources:

- Bortz, Jürgen (1977): "Zwei- und mehrfaktorielle Varianzanalyse"; in: *Lehrbuch der Statistik. Für Sozialwissenschaftler*, ch. 12.2, pp. 538ff.; URL: https://books.google.de/books?id=lNCyBgAAQBAJ.

- ttd (2021): "Proof on SSAB/s2 chi2(I-1)(J-1) under the null hypothesis HAB: dij=0 for i=1,...,I and j=1,...,J"; in: *StackExchange CrossValidated*, retrieved on 2022-11-06; URL: https://stats.stackexchange.com/questions/545807/proof-on-ss-ab-sigma2-sim-chi2-i-1j-1-under-the-null-hypothesis.

1.3.9 Interaction sum of squares

Definition: Let there be an analysis of variance (ANOVA) model with two (→ III/1.3.8) or more factors influencing the measured data y (here, using the standard formulation (→ III/1.3.11) of two-way ANOVA (→ III/1.3.8)):

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \; \varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{1}$$

Then, the interaction sum of squares is defined as the explained sum of squares (→ III/1.5.8) (ESS) for each interaction, i.e. as the sum of squared deviations of the average for each cell from the average across all observations, controlling for the treatment sums of squares (→ III/1.3.2) of the corresponding factors:

$$\begin{split} \mathrm{SS}_{\mathrm{A}\times\mathrm{B}} &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet}] - [\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - [\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}])^2 \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 \; . \end{split} \tag{2}$$

Here, $\bar{y}\_ij\bullet$ is the mean for the (i,j)-th cell (out of $a \times b$ cells), computed from n_ij values y_ijk, $\bar{y}\_i\bullet\bullet$ and $\bar{y}\_\bullet j\bullet$ are the level means for the two factors and and $\bar{y}\_\bullet\bullet\bullet$ is the mean across all values y_ijk.

Sources:

- Nandy, Siddhartha (2018): "Two-Way Analysis of Variance"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf.

1.3.10 Ordinary least squares for two-way ANOVA

Theorem: Given the two-way analysis of variance (→ III/1.3.8) assumption

$$\begin{split} &y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ &\varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, a, \; j = 1, \ldots, b, \; k = 1, \ldots, n_{ij} \; , \end{split} \tag{1}$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) and satisfying the constraints for the model parameters (→ III/1.3.8) are given by

$$\begin{split} \hat{\mu} &= \bar{y}_{\bullet\bullet\bullet} \\ \hat{\alpha}_i &= \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \\ \hat{\beta}_j &= \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet} \\ \hat{\gamma}_{ij} &= \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet} \end{split} \tag{2}$$

where $\bar{y}_{\bullet\bullet\bullet}$, $\bar{y}_{i\bullet\bullet}$, $\bar{y}_{\bullet j\bullet}$ and $\bar{y}_{ij\bullet}$ are the following sample means (→ I/1.10.2):

$$
\begin{aligned}
\bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{i\bullet\bullet} &= \frac{1}{n_{i\bullet}} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{\bullet j\bullet} &= \frac{1}{n_{\bullet j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{ij\bullet} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}
\end{aligned} \tag{3}
$$

with the sample size numbers

$$
\begin{aligned}
&n_{ij} - \text{number of samples in category } (i,j) \\
&n_{i\bullet} = \sum_{j=1}^{b} n_{ij} \\
&n_{\bullet j} = \sum_{i=1}^{a} n_{ij} \\
&n = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} \; .
\end{aligned} \tag{4}
$$

Proof: In two-way ANOVA, model parameters are subject to the constraints (→ III/1.3.8)

$$
\begin{aligned}
\sum_{i=1}^{a} w_{ij} \alpha_i &= 0 \quad \text{for all} \quad j = 1, \ldots, b \\
\sum_{j=1}^{b} w_{ij} \beta_j &= 0 \quad \text{for all} \quad i = 1, \ldots, a \\
\sum_{i=1}^{a} w_{ij} \gamma_{ij} &= 0 \quad \text{for all} \quad j = 1, \ldots, b \\
\sum_{j=1}^{b} w_{ij} \gamma_{ij} &= 0 \quad \text{for all} \quad i = 1, \ldots, a
\end{aligned} \tag{5}
$$

where $w_{ij} = n_{ij}/n$. The residual sum of squares (→ III/1.5.9) for this model is

$$
\mathrm{RSS}(\mu, \alpha, \beta, \gamma) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} \varepsilon_{ijk}^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ij} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \tag{6}
$$

and the derivatives of RSS with respect to μ, α, β and γ are

$$\begin{aligned}
\frac{\mathrm{dRSS}}{\mathrm{d}\mu} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \frac{\mathrm{d}}{\mathrm{d}\mu}(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} -2(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij}) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(2n_{ij}\mu+2n_{ij}(\alpha_i+\beta_j+\gamma_{ij})-2\sum_{k=1}^{n_{ij}} y_{ijk}\right) \\
&= 2n\mu+2\left(\sum_{i=1}^{a} n_{i\bullet}\alpha_i+\sum_{j=1}^{b} n_{\bullet j}\beta_j+\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij}\right)-2\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\frac{\mathrm{dRSS}}{\mathrm{d}\alpha_i} &= \sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \frac{\mathrm{d}}{\mathrm{d}\alpha_i}(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2 \\
&= \sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} -2(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij}) \\
&= 2n_{i\bullet}\mu+2n_{i\bullet}\alpha_i+2\left(\sum_{j=1}^{b} n_{ij}\beta_j+\sum_{j=1}^{b} n_{ij}\gamma_{ij}\right)-2\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{\mathrm{dRSS}}{\mathrm{d}\beta_j} &= \sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} \frac{\mathrm{d}}{\mathrm{d}\beta_j}(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2 \\
&= \sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} -2(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij}) \\
&= 2n_{\bullet j}\mu+2n_{\bullet j}\beta_j+2\left(\sum_{i=1}^{a} n_{ij}\alpha_i+\sum_{i=1}^{a} n_{ij}\gamma_{ij}\right)-2\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\mathrm{dRSS}}{\mathrm{d}\gamma_{ij}} &= \sum_{k=1}^{n_{ij}} \frac{\mathrm{d}}{\mathrm{d}\gamma_{ij}}(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2 \\
&= \sum_{k=1}^{n_{ij}} -2(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij}) \\
&= 2n_{ij}(\mu+\alpha_i+\beta_j+\gamma_{ij})-2\sum_{k=1}^{n_{ij}} y_{ijk}\ .
\end{aligned} \tag{10}$$

Setting these derivatives to zero, we obtain the estimates of μ_i, α_i, β_j and γ_{ij}:

$$
\begin{aligned}
0 &= 2n\hat{\mu} + 2\left(\sum_{i=1}^{a} n_{i\bullet}\alpha_i + \sum_{j=1}^{b} n_{\bullet j}\beta_j + \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij}\right) - 2\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \\
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{i=1}^{a}\frac{n_{i\bullet}}{n}\alpha_i - \sum_{j=1}^{b}\frac{n_{\bullet j}}{n}\beta_j - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{n_{ij}}{n}\gamma_{ij} \\
&\overset{(4)}{=} \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{j=1}^{b}\sum_{i=1}^{a}\frac{n_{ij}}{n}\alpha_i - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{n_{ij}}{n}\beta_j - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{n_{ij}}{n}\gamma_{ij} \\
&\overset{(5)}{=} \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \\
&\overset{(3)}{=} \bar{y}_{\bullet\bullet\bullet}
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
0 &= 2n_{i\bullet}\hat{\mu} + 2n_{i\bullet}\hat{\alpha}_i + 2\left(\sum_{j=1}^{b} n_{ij}\beta_j + \sum_{j=1}^{b} n_{ij}\gamma_{ij}\right) - 2\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \\
\hat{\alpha}_i &= \frac{1}{n_{i\bullet}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \sum_{j=1}^{b}\frac{n_{ij}}{n_{i\bullet}}\beta_j - \sum_{j=1}^{b}\frac{n_{ij}}{n_{i\bullet}}\gamma_{ij} \\
&= \frac{1}{n_{i\bullet}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \frac{n}{n_{i\bullet}}\sum_{j=1}^{b}\frac{n_{ij}}{n}\beta_j - \frac{n}{n_{i\bullet}}\sum_{j=1}^{b}\frac{n_{ij}}{n}\gamma_{ij} \\
&\overset{(5)}{=} \frac{1}{n_{i\bullet}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \\
&\overset{(3)}{=} \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
0 &= 2n_{\bullet j}\hat{\mu} + 2n_{\bullet j}\hat{\beta}_j + 2\left(\sum_{i=1}^{a} n_{ij}\alpha_i + \sum_{i=1}^{a} n_{ij}\gamma_{ij}\right) - 2\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk} \\
\hat{\beta}_j &= \frac{1}{n_{\bullet j}}\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \sum_{i=1}^{a}\frac{n_{ij}}{n_{\bullet j}}\alpha_i - \sum_{i=1}^{a}\frac{n_{ij}}{n_{\bullet j}}\gamma_{ij} \\
&= \frac{1}{n_{\bullet j}}\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\mu} - \frac{n}{n_{\bullet j}}\sum_{i=1}^{a}\frac{n_{ij}}{n}\alpha_i - \frac{n}{n_{\bullet j}}\sum_{i=1}^{a}\frac{n_{ij}}{n}\gamma_{ij} \\
&\overset{(5)}{=} \frac{1}{n_{\bullet j}}\sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \\
&\overset{(3)}{=} \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
0 &= 2n_{ij}(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma_{ij}}) - 2\sum_{k=1}^{n_{ij}} y_{ijk} \\
\hat{\gamma_{ij}} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu} \\
&= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n_{i\bullet}} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} - \frac{1}{n_{\bullet j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} y_{ijk} + \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
&\overset{(3)}{=} \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet} \; .
\end{aligned}
\tag{14}
$$

■

Sources:
- Olbricht, Gayla R. (2011): "Two-Way ANOVA: Interaction"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Spring 2011, Lect. 27; URL: https://www.stat.purdue.edu/~ghobbs/STAT_512/Lecture_Notes/ANOVA/Topic_27.pdf.

1.3.11 Sums of squares in two-way ANOVA

Theorem: Given two-way analysis of variance (→ III/1.3.8),

$$
y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \; \varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}
$$

sums of squares can be partitioned as follows

$$
\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A \times B} + \mathrm{SS}_{\mathrm{res}} \tag{2}
$$

where SS_tot is the total sum of squares (→ III/1.5.7), SS_A, SS_B and SS_$A \times B$ are treatment (→ III/1.3.2) and interaction sum of squares (→ III/1.3.9) (summing into the explained sum of squares (→ III/1.5.8)) and SS_res is the residual sum of squares (→ III/1.5.9).

Proof: The total sum of squares (→ III/1.5.7) for two-way ANOVA (→ III/1.3.8) is given by

$$
\mathrm{SS}_{\mathrm{tot}} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 \tag{3}
$$

where $\bar{y}$_$\bullet\bullet\bullet$ is the mean across all values y_ijk. This can be rewritten as

$$
\begin{split}
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} [&(y_{ijk} - \bar{y}_{ij\bullet}) + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + \\
&(\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})]^2
\end{split}
\tag{4}
$$

It can be shown (→ III/1.3.12) that the following sums are all zero:

$$
\begin{aligned}
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet}) = 0 \\
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) = 0 \\
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) = 0 \\
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}) = 0 \; .
\end{aligned} \tag{5}
$$

This means that the sum in (4) reduces to

$$
\begin{split}
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} \left[(y_{ijk} - \bar{y}_{ij\bullet})^2 + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \right. \\
&\qquad \left. (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 \right] \\
&= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 + \\
&\quad \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 \; .
\end{split} \tag{6}
$$

With the treatment sums of squares (→ III/1.3.2)

$$
\begin{split}
\mathrm{SS}_A &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 \\
\mathrm{SS}_B &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 \; ,
\end{split} \tag{7}
$$

the interaction sum of squares (→ III/1.3.9)

$$
\mathrm{SS}_{\mathrm{A\times B}} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 \tag{8}
$$

and the residual sum of squares (→ III/1.5.9) for two-way ANOVA (→ III/1.3.8)

$$
\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 \; , \tag{9}
$$

we finally have:

$$
\mathrm{SS_{tot}} = \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A \times B} + \mathrm{SS_{res}} \; . \tag{10}
$$

■

Sources:

- Nandy, Siddhartha (2018): "Two-Way Analysis of Variance"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf.
- Wikipedia (2022): "Analysis of variance"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-15; URL: https://en.wikipedia.org/wiki/Analysis_of_variance#Partitioning_of_the_sum_of_squares.

1.3.12 Cochran's theorem for two-way ANOVA

Theorem: Assume the two-way analysis of variance (→ III/1.3.8) model

$$
\begin{split}
y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\
\varepsilon_{ijk} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, a, \; j = 1, \ldots, b, \; k = 1, \ldots, n_{ij}
\end{split} \tag{1}
$$

under the well-known constraints for the model parameters (→ III/1.3.8)

$$
\begin{split}
\sum_{i=1}^{a} \frac{n_{ij}}{n} \alpha_i = 0 \quad &\text{for all} \quad j = 1, \ldots, b \\
\sum_{j=1}^{b} \frac{n_{ij}}{n} \beta_j = 0 \quad &\text{for all} \quad i = 1, \ldots, a \\
\sum_{i=1}^{a} \frac{n_{ij}}{n} \gamma_{ij} = 0 \quad &\text{for all} \quad j = 1, \ldots, b \\
\sum_{j=1}^{b} \frac{n_{ij}}{n} \gamma_{ij} = 0 \quad &\text{for all} \quad i = 1, \ldots, a \; .
\end{split} \tag{2}
$$

Then, the following sums of squares (→ III/1.3.11) are chi-square distributed (→ II/3.7.1)

$$
\begin{split}
\frac{1}{\sigma^2} n (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 = \frac{\mathrm{SS}_M}{\sigma^2} &\sim \chi^2(1) \\
\frac{1}{\sigma^2} \sum_{i=1}^{a} n_{i\bullet} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 = \frac{\mathrm{SS}_A}{\sigma^2} &\sim \chi^2(a-1) \\
\frac{1}{\sigma^2} \sum_{j=1}^{b} n_{\bullet j} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 = \frac{\mathrm{SS}_B}{\sigma^2} &\sim \chi^2(b-1) \\
\frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 = \frac{\mathrm{SS}_{A \times B}}{\sigma^2} &\sim \chi^2\left((a-1)(b-1)\right) \\
\frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 = \frac{\mathrm{SS_{res}}}{\sigma^2} &\sim \chi^2(n-ab) \; .
\end{split} \tag{3}
$$

Proof: Denote sample sizes as

$$
\begin{aligned}
n_{ij} &- \text{number of samples in category } (i,j) \\
n_{i\bullet} &= \sum_{j=1}^{b} n_{ij} \\
n_{\bullet j} &= \sum_{i=1}^{a} n_{ij} \\
n &= \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}
\end{aligned} \tag{4}
$$

and denote sample means as

$$
\begin{aligned}
\bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{i\bullet\bullet} &= \frac{1}{n_{i\bullet}} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{\bullet j\bullet} &= \frac{1}{n_{\bullet j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{ij\bullet} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \; .
\end{aligned} \tag{5}
$$

According to the model given by (1), the observations are distributed as:

$$
y_{ijk} \sim \mathcal{N}(\mu + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2) \quad \text{for all} \quad i, j, k \; . \tag{6}
$$

Thus, the random variable (→ I/1.2.2) $U_{ijk} = (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})/\sigma$ follows a standard normal distribution (→ II/3.2.4)

$$
U_{ijk} = \frac{y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}}{\sigma} \sim \mathcal{N}(0,1) \; . \tag{7}
$$

Now consider the following sum

$$
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} U_{ijk}^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} \left(\frac{y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij}}{\sigma} \right)^2 \tag{8}
$$

which can be rewritten as follows:

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} U_{ijk}^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} [(y_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})- \\
&\qquad [\bar{y}_{\bullet\bullet\bullet}+(\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet})+(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet})+(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})]+ \\
&\qquad [\bar{y}_{\bullet\bullet\bullet}+(\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet})+(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet})+(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})]]^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} [(y_{ijk}-[\bar{y}_{\bullet\bullet\bullet}+(\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet})+(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet})+(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})])+ \\
&\qquad (\bar{y}_{\bullet\bullet\bullet}-\mu)+([\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\alpha_i)+([\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\beta_j)+ \\
&\qquad ([\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}]-\gamma_{ij})]^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} [(y_{ijk}-\bar{y}_{ij\bullet})+(\bar{y}_{\bullet\bullet\bullet}-\mu)+([\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\alpha_i)+ \\
&\qquad ([\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\beta_j)+([\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}]-\gamma_{ij})]^2\ .
\end{aligned} \tag{9}$$

Note that the following sums are all zero:

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet}) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left[\sum_{k=1}^{n_{ij}} y_{ijk}-n_{ij}\cdot\bar{y}_{ij\bullet}\right] \\
&\overset{(5)}{=} \sum_{i=1}^{a}\sum_{j=1}^{b}\left[\sum_{k=1}^{n_{ij}} y_{ijk}-n_{ij}\cdot\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk}\right] \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b} 0 = 0
\end{aligned} \tag{10}$$

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\alpha_i) &= \sum_{i=1}^{a} n_{i\bullet}\cdot(\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}-\alpha_i) \\
&= \sum_{i=1}^{a} n_{i\bullet}\cdot\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}\sum_{i=1}^{a} n_{i\bullet}-\sum_{i=1}^{a} n_{i\bullet}\alpha_i \\
&\overset{(5)}{=} \sum_{i=1}^{a} n_{i\bullet}\cdot\frac{1}{n_{i\bullet}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}-n\cdot\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk}-\sum_{i=1}^{a} n_{i\bullet}\alpha_i \\
&= -\sum_{i=1}^{a} n_{i\bullet}\alpha_i \overset{(4)}{=} -n\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{n_{ij}}{n}\alpha_i \overset{(2)}{=} 0
\end{aligned} \tag{11}$$

$$
\begin{split}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j) &= \sum_{j=1}^{b} n_{\bullet j} \cdot (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet} - \beta_j) \\
&= \sum_{j=1}^{b} n_{\bullet j} \cdot \bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet} \sum_{j=1}^{b} n_{\bullet j} - \sum_{j=1}^{b} n_{\bullet j}\beta_j \\
&\overset{(5)}{=} \sum_{j=1}^{b} n_{\bullet j} \cdot \frac{1}{n_{\bullet j}} \sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk} - n \cdot \frac{1}{n} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{j=1}^{b} n_{\bullet j}\beta_j \\
&= -\sum_{j=1}^{b} n_{\bullet j}\beta_j \overset{(4)}{=} -n \sum_{j=1}^{b}\sum_{i=1}^{a} \frac{n_{ij}}{n}\beta_j \overset{(2)}{=} 0
\end{split}
\tag{12}
$$

$$
\begin{split}
&\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j \bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij}) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \left[(\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet}) - (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) - (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet}) - \gamma_{ij}\right] \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet} - \gamma_{ij}) - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet}) \\
&\overset{(12)}{=} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet} - \gamma_{ij}) - \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) \\
&\overset{(11)}{=} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet} - \gamma_{ij}) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\bar{y}_{ij\bullet} - \bar{y}_{\bullet\bullet\bullet} \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} - \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij} \\
&\overset{(5)}{=} \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} \cdot \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} - n \cdot \frac{1}{n} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} - \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij} \\
&= -\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\gamma_{ij} = -\frac{1}{n} \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{n_{ij}}{n}\gamma_{ij} \overset{(2)}{=} 0 \; .
\end{split}
\tag{13}
$$

Note further that $\bar{y}_{\bullet\bullet\bullet}$ and μ are not dependent on i, j and k:

$$
\bar{y}_{\bullet\bullet\bullet} = \text{const.} \quad \text{and} \quad \mu = \text{const.}
\tag{14}
$$

Thus, all the non-square products in (9) disappear and the sum reduces to

$$
\begin{split}
\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} U_{ijk}^2 = \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} &\left[(y_{ijk} - \bar{y}_{ij\bullet})^2 + (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 + ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 + \right. \\
&\left. ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 + ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \right] \\
= \frac{1}{\sigma^2} &\left[\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 + \right. \\
&\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} ([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} ([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 + \\
&\left. \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \right] \; .
\end{split}
\tag{15}
$$

Cochran's theorem states that, if a sum of squared standard normal (→ II/3.2.3) random variables (→ I/1.2.2) can be written as a sum of squared forms

$$
\begin{split}
\sum_{i=1}^{n} U_i^2 = \sum_{j=1}^{m} Q_j \quad &\text{where} \quad Q_j = U^\mathrm{T} B^{(j)} U \\
&\text{with} \quad \sum_{j=1}^{m} B^{(j)} = I_n \\
&\text{and} \quad r_j = \mathrm{rank}(B^{(j)}) \; ,
\end{split}
\tag{16}
$$

then the terms Q_j are independent (→ I/1.3.6) and each term Q_j follows a chi-squared distribution (→ II/3.7.1) with r_j degrees of freedom:

$$
Q_j \sim \chi^2(r_j), \; j = 1, \ldots, m \; . \tag{17}
$$

First, we define the $n \times 1$ vector U:

$$
U = \begin{bmatrix} u_{1\bullet} \\ \vdots \\ u_{a\bullet} \end{bmatrix} \quad \text{where} \quad u_{i\bullet} = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{ib} \end{bmatrix} \quad \text{where} \quad u_{ij} = \begin{bmatrix} (y_{i,j,1} - \mu - \alpha_i - \beta_j - \gamma_{ij})/\sigma \\ \vdots \\ (y_{i,j,n_{ij}} - \mu - \alpha_i - \beta_j - \gamma_{ij})/\sigma \end{bmatrix} \; . \tag{18}
$$

Next, we specify the $n \times n$ matrices B

$$
\begin{aligned}
B^{(1)} &= I_n - \mathrm{diag}\left[\mathrm{diag}\left(\frac{1}{n_{11}}J_{n_{11}},\,\ldots,\,\frac{1}{n_{1b}}J_{n_{1b}}\right),\,\ldots,\,\mathrm{diag}\left(\frac{1}{n_{a1}}J_{n_{a1}},\,\ldots,\,\frac{1}{n_{ab}}J_{n_{ab}}\right)\right] \\
B^{(2)} &= \frac{1}{n}J_n \\
B^{(3)} &= \mathrm{diag}\left(\frac{1}{n_{1\bullet}}J_{n_{1\bullet}},\,\ldots,\,\frac{1}{n_{a\bullet}}J_{n_{a\bullet}}\right) - \frac{1}{n}J_n \\
B^{(4)} &= M_B - \frac{1}{n}J_n \\
B^{(5)} &= \mathrm{diag}\left[\mathrm{diag}\left(\frac{1}{n_{11}}J_{n_{11}},\,\ldots,\,\frac{1}{n_{1b}}J_{n_{1b}}\right),\,\ldots,\,\mathrm{diag}\left(\frac{1}{n_{a1}}J_{n_{a1}},\,\ldots,\,\frac{1}{n_{ab}}J_{n_{ab}}\right)\right] \\
&\quad - \mathrm{diag}\left(\frac{1}{n_{1\bullet}}J_{n_{1\bullet}},\,\ldots,\,\frac{1}{n_{a\bullet}}J_{n_{a\bullet}}\right) - M_B + \frac{1}{n}J_n
\end{aligned} \tag{19}
$$

with the factor B matrix M_B given by

$$
M_B = \begin{bmatrix}
\mathrm{diag}\left(\frac{1}{n_{\bullet 1}}J_{n_{11},n_{11}},\,\ldots,\,\frac{1}{n_{\bullet b}}J_{n_{1b},n_{1b}}\right) & \cdots & \mathrm{diag}\left(\frac{1}{n_{\bullet 1}}J_{n_{11},n_{a1}},\,\ldots,\,\frac{1}{n_{\bullet b}}J_{n_{1b},n_{ab}}\right) \\
\vdots & \ddots & \vdots \\
\mathrm{diag}\left(\frac{1}{n_{\bullet 1}}J_{n_{a1},n_{11}},\,\ldots,\,\frac{1}{n_{\bullet b}}J_{n_{ab},n_{1b}}\right) & \cdots & \mathrm{diag}\left(\frac{1}{n_{\bullet 1}}J_{n_{a1},n_{a1}},\,\ldots,\,\frac{1}{n_{\bullet b}}J_{n_{ab},n_{ab}}\right)
\end{bmatrix} . \tag{20}
$$

where J_n is an $n \times n$ matrix of ones, $J_{n,m}$ is an $n \times m$ matrix of ones and $\mathrm{diag}\,(A_1, \ldots, A_n)$ denotes a block-diagonal matrix composed of $A_1, \ldots, A_n$. We observe that those matrices satisfy

$$
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} U_{ijk}^2 = \sum_{l=1}^{5} Q_l = \sum_{l=1}^{5} U^{\mathrm{T}} B^{(l)} U \tag{21}
$$

as well as

$$
\sum_{l=1}^{5} B^{(l)} = I_n \tag{22}
$$

and their ranks are

$$
\begin{aligned}
\mathrm{rank}\left(B^{(1)}\right) &= n - ab \\
\mathrm{rank}\left(B^{(2)}\right) &= 1 \\
\mathrm{rank}\left(B^{(3)}\right) &= a - 1 \\
\mathrm{rank}\left(B^{(4)}\right) &= b - 1 \\
\mathrm{rank}\left(B^{(5)}\right) &= (a-1)(b-1) \; .
\end{aligned} \tag{23}
$$

Thus, the conditions for applying Cochran's theorem given by (16) are fulfilled and we can use (15), (17), (19) and (23) to conclude that

$$
\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{\bullet\bullet\bullet}-\mu)^2 &= Q_2 = U^\mathrm{T} B^{(2)} U \sim \chi^2(1) \\
\frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\alpha_i)^2 &= Q_3 = U^\mathrm{T} B^{(3)} U \sim \chi^2(a-1) \\
\frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\beta_j)^2 &= Q_4 = U^\mathrm{T} B^{(4)} U \sim \chi^2(b-1) \\
\frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}]-\gamma_{ij})^2 &= Q_5 = U^\mathrm{T} B^{(5)} U \sim \chi^2\left((a-1)(b-1)\right) \\
\frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2 &= Q_1 = U^\mathrm{T} B^{(1)} U \sim \chi^2(n-ab) \; .
\end{aligned} \tag{24}
$$

Finally, we identify the terms Q with sums of squares in two-way ANOVA (→ III/1.3.11) and simplify them to reach the expressions given by (3):

$$
\begin{aligned}
\frac{\mathrm{SS}_M}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{y}_{\bullet\bullet\bullet}-\mu)^2 = \frac{1}{\sigma^2} n(\bar{y}_{\bullet\bullet\bullet}-\mu)^2 \\
\frac{\mathrm{SS}_A}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\alpha_i)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{a} n_{i\bullet}([\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\alpha_i)^2 \\
\frac{\mathrm{SS}_B}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\beta_j)^2 = \frac{1}{\sigma^2}\sum_{j=1}^{b} n_{\bullet j}([\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet}]-\beta_j)^2 \\
\frac{\mathrm{SS}_{A\times B}}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}]-\gamma_{ij})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}([\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}]-\gamma_{ij})^2 \\
\frac{\mathrm{SS}_\mathrm{res}}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2 \; .
\end{aligned} \tag{25}
$$

■

Sources:

- Nandy, Siddhartha (2018): "Two-Way Analysis of Variance"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf.

1.3.13 F-test for main effect in two-way ANOVA

Theorem: Assume the two-way analysis of variance (→ III/1.3.8) model

$$
\begin{aligned}
&y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\
&\varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2), \; i=1,\ldots,a, \; j=1,\ldots,b, \; k=1,\ldots,n_{ij} \; .
\end{aligned} \tag{1}
$$

Then, the test statistic (→ I/4.3.5)

$$F_A = \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_{i\bullet}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2} \tag{2}$$

follows an F-distribution (→ II/3.8.1)

$$F_A \sim \mathrm{F}(a-1, n-ab) \tag{3}$$

under the null hypothesis (→ I/4.3.2) for the main effect (→ III/1.3.8) of factor A

$$\begin{split} H_0 &: \alpha_1 = \ldots = \alpha_a = 0 \\ H_1 &: \alpha_i \neq 0 \quad \text{for at least one} \quad i \in \{1, \ldots, a\} \end{split} \tag{4}$$

and the test statistic (→ I/4.3.5)

$$F_B = \frac{\frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2} \tag{5}$$

follows an F-distribution (→ II/3.8.1)

$$F_B \sim \mathrm{F}(b-1, n-ab) \tag{6}$$

under the null hypothesis (→ I/4.3.2) for the main effect (→ III/1.3.8) of factor B

$$\begin{split} H_0 &: \beta_1 = \ldots = \beta_b = 0 \\ H_1 &: \beta_j \neq 0 \quad \text{for at least one} \quad j \in \{1, \ldots, b\} \; . \end{split} \tag{7}$$

Proof: Applying Cochran's theorem for two-analysis of variance (→ III/1.3.12), we find that the following squared sums

$$\begin{split} \frac{\mathrm{SS}_A}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{a} n_{i\bullet}([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2 \\ \frac{\mathrm{SS}_B}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 = \frac{1}{\sigma^2}\sum_{j=1}^{b} n_{\bullet j}([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2 \\ \frac{\mathrm{SS}_\mathrm{res}}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2 \end{split} \tag{8}$$

are independent (→ I/1.3.6) and chi-squared distributed (→ II/3.7.1):

$$\begin{split} \frac{\mathrm{SS}_A}{\sigma^2} &\sim \chi^2(a-1) \\ \frac{\mathrm{SS}_B}{\sigma^2} &\sim \chi^2(b-1) \\ \frac{\mathrm{SS}_\mathrm{res}}{\sigma^2} &\sim \chi^2(n-ab) \; . \end{split} \tag{9}$$

1) Thus, the F-statistic from (2) is equal to the ratio of two independent (→ I/1.3.6) chi-squared distributed (→ II/3.7.1) random variables (→ I/1.2.2) divided by their degrees of freedom

$$\begin{aligned} F_A &= \frac{(\mathrm{SS}_A/\sigma^2)/(a-1)}{(\mathrm{SS}_\mathrm{res}/\sigma^2)/(n-ab)} \\ &= \frac{\mathrm{SS}_A/(a-1)}{\mathrm{SS}_\mathrm{res}/(n-ab)} \\ &\overset{(8)}{=} \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_{i\bullet}([\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \alpha_i)^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2} \\ &\overset{(4)}{=} \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_{i\bullet}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2} \end{aligned} \tag{10}$$

which, by definition of the F-distribution (→ II/3.8.1), is distributed as

$$F_A \sim \mathrm{F}(a-1, n-ab) \tag{11}$$

under the null hypothesis (→ I/4.3.2) for main effect of A.
2) Similarly, the F-statistic from (5) is equal to the ratio of two independent chi-squared distributed (→ II/3.7.1) random variables (→ I/1.2.2) divided by their degrees of freedom

$$\begin{aligned} F_B &= \frac{(\mathrm{SS}_B/\sigma^2)/(b-1)}{(\mathrm{SS}_\mathrm{res}/\sigma^2)/(n-ab)} \\ &= \frac{\mathrm{SS}_B/(b-1)}{\mathrm{SS}_\mathrm{res}/(n-ab)} \\ &\overset{(8)}{=} \frac{\frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}([\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}] - \beta_j)^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2} \\ &\overset{(7)}{=} \frac{\frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\bullet})^2} \end{aligned} \tag{12}$$

which, by definition of the F-distribution (→ II/3.8.1), is distributed as

$$F_B \sim \mathrm{F}(b-1, n-ab) \tag{13}$$

under the null hypothesis (→ I/4.3.2) for main effect of B.

■

Sources:

- ttd (2021): "Proof on SSAB/s2 chi2(I-1)(J-1) under the null hypothesis HAB: dij=0 for i=1,...,I and j=1,...,J"; in: *StackExchange CrossValidated*, retrieved on 2022-11-10; URL: https://stats.stackexchange.com/questions/545807/proof-on-ss-ab-sigma2-sim-chi2-i-1j-1-under-the-null-hypothesis.
- JohnK (2014): "In a two-way ANOVA, how can the F-statistic for one factor have a central distribution if the null is false for the other factor?"; in: *StackExchange CrossValidated*, retrieved on 2022-11-10; URL: https://stats.stackexchange.com/questions/124166/in-a-two-way-anova-how-can-the-f-

1.3.14 F-test for interaction in two-way ANOVA

Theorem: Assume the two-way analysis of variance (→ III/1.3.8) model

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, a, \; j = 1, \ldots, b, \; k = 1, \ldots, n_{ij} \; . \end{aligned} \tag{1}$$

Then, the test statistic (→ I/4.3.5)

$$F_{A \times B} = \frac{\frac{1}{(a-1)(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \tag{2}$$

follows an F-distribution (→ II/3.8.1)

$$F_{A \times B} \sim \mathrm{F}\left((a-1)(b-1), n-ab\right) \tag{3}$$

under the null hypothesis (→ I/4.3.2) for the interaction effect (→ III/1.3.8) of factors A and B

$$\begin{aligned} &H_0 : \gamma_{11} = \ldots = \gamma_{ab} = 0 \\ &H_1 : \gamma_{ij} \neq 0 \quad \text{for at least one} \quad (i,j) \in \{1, \ldots, a\} \times \{1, \ldots, b\} \; . \end{aligned} \tag{4}$$

Proof: Applying Cochran's theorem for two-analysis of variance (→ III/1.3.12), we find that the following squared sums

$$\begin{aligned} \frac{\mathrm{SS}_{A \times B}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} ([\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}] - \gamma_{ij})^2 \\ \frac{\mathrm{SS}_{\mathrm{res}}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 \end{aligned} \tag{5}$$

are independent (→ I/1.3.6) and chi-squared distributed (→ II/3.7.1):

$$\begin{aligned} \frac{\mathrm{SS}_{A \times B}}{\sigma^2} &\sim \chi^2\left((a-1)(b-1)\right) \\ \frac{\mathrm{SS}_{\mathrm{res}}}{\sigma^2} &\sim \chi^2(n-ab) \; . \end{aligned} \tag{6}$$

Thus, the F-statistic from (2) is equal to the ratio of two independent (→ I/1.3.6) chi-squared distributed (→ II/3.7.1) random variables (→ I/1.2.2) divided by their degrees of freedom

$$
\begin{aligned}
F_{A\times B} &= \frac{(\mathrm{SS}_{A\times B}/\sigma^2)/\left((a-1)(b-1)\right)}{(\mathrm{SS}_\mathrm{res}/\sigma^2)/(n-ab)} \\
&= \frac{\mathrm{SS}_{A\times B}/\left((a-1)(b-1)\right)}{\mathrm{SS}_\mathrm{res}/(n-ab)} \\
&\overset{(5)}{=} \frac{\frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}([\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet}]-\gamma_{ij})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2} \\
&\overset{(3)}{=} \frac{\frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2}
\end{aligned} \tag{7}
$$

which, by definition of the F-distribution (→ II/3.8.1), is distributed as

$$
F_{A\times B} \sim \mathrm{F}\left((a-1)(b-1), n-ab\right) \tag{8}
$$

under the null hypothesis (→ I/4.3.2) for an interaction of A and B.

■

Sources:

- Nandy, Siddhartha (2018): "Two-Way Analysis of Variance"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf.
- ttd (2021): "Proof on SSAB/s2 chi2(I-1)(J-1) under the null hypothesis HAB: dij=0 for i=1,...,I and j=1,...,J"; in: *StackExchange CrossValidated*, retrieved on 2022-11-10; URL: https://stats.stackexchange.com/questions/545807/proof-on-ss-ab-sigma2-sim-chi2-i-1j-1-under-the-null-hypothesis.

1.3.15 F-test for grand mean in two-way ANOVA

Theorem: Assume the two-way analysis of variance (→ III/1.3.8) model

$$
\begin{aligned}
y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\
\varepsilon_{ijk} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2), \; i=1,\ldots,a, \; j=1,\ldots,b, \; k=1,\ldots,n_{ij} \; .
\end{aligned} \tag{1}
$$

Then, the test statistic (→ I/4.3.5)

$$
F_M = \frac{n(\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2} \tag{2}
$$

follows an F-distribution (→ II/3.8.1)

$$
F_M \sim \mathrm{F}\left(1, n-ab\right) \tag{3}
$$

under the null hypothesis (→ I/4.3.2) for the grand mean (→ III/1.3.8)

$$
\begin{aligned}
H_0 &: \mu = 0 \\
H_1 &: \mu \neq 0 \; .
\end{aligned} \tag{4}
$$

Proof: Applying Cochran's theorem for two-analysis of variance (→ III/1.3.12), we find that the following squared sums

$$\begin{aligned}
\frac{\mathrm{SS}_M}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 = \frac{1}{\sigma^2} n (\bar{y}_{\bullet\bullet\bullet} - \mu)^2 \\
\frac{\mathrm{SS}_\mathrm{res}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2 = \frac{1}{\sigma^2} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2
\end{aligned} \tag{5}$$

are independent (→ I/1.3.6) and chi-squared distributed (→ II/3.7.1):

$$\begin{aligned}
\frac{\mathrm{SS}_M}{\sigma^2} &\sim \chi^2(1) \\
\frac{\mathrm{SS}_\mathrm{res}}{\sigma^2} &\sim \chi^2(n - ab) \; .
\end{aligned} \tag{6}$$

Thus, the F-statistic from (2) is equal to the ratio of two independent (→ I/1.3.6) chi-squared distributed (→ II/3.7.1) random variables (→ I/1.2.2) divided by their degrees of freedom

$$\begin{aligned}
F_M &= \frac{(\mathrm{SS}_M/\sigma^2)/(1)}{(\mathrm{SS}_\mathrm{res}/\sigma^2)/(n-ab)} \\
&= \frac{\mathrm{SS}_M/(1)}{\mathrm{SS}_\mathrm{res}/(n-ab)} \\
&\overset{(5)}{=} \frac{n(\bar{y}_{\bullet\bullet\bullet} - \mu)^2}{\frac{1}{n-ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2} \\
&\overset{(4)}{=} \frac{n(\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij\bullet})^2}
\end{aligned} \tag{7}$$

which, by definition of the F-distribution (→ II/3.8.1), is distributed as

$$F_M \sim \mathrm{F}(1, n-ab) \tag{8}$$

under the null hypothesis (→ I/4.3.2) for the grand mean.

■

Sources:

- Nandy, Siddhartha (2018): "Two-Way Analysis of Variance"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Summer 2018, Ch. 19; URL: https://www.stat.purdue.edu/~snandy/stat512/topic7.pdf.
- Olbricht, Gayla R. (2011): "Two-Way ANOVA: Interaction"; in: *Stat 512: Applied Regression Analysis*, Purdue University, Spring 2011, Lect. 27; URL: https://www.stat.purdue.edu/~ghobbs/STAT_512/Lecture_Notes/ANOVA/Topic_27.pdf.

1.3.16 F-statistics in terms of OLS estimates

Theorem: Given the two-way analysis of variance (→ III/1.3.8) assumption

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \; \varepsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the F-statistics for the grand mean (→ III/1.3.15), the main effects (→ III/1.3.13) and the interaction (→ III/1.3.14) can be expressed in terms of ordinary least squares parameter estimates (→ III/1.3.10) as

$$\begin{aligned}
F_M &= \frac{n\hat{\mu}^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \\
F_A &= \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_{i\bullet}\hat{\alpha}_i^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \\
F_B &= \frac{\frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}\hat{\beta}_j^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \\
F_{A\times B} &= \frac{\frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\hat{\gamma}_{ij}^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2}
\end{aligned} \tag{2}$$

where the predicted values $\hat{y}_{ijk}$ are given by

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} \; . \tag{3}$$

Theorem: The F-statistics for the grand mean (→ III/1.3.15), the main effects (→ III/1.3.13) and the interaction (→ III/1.3.14) in two-way ANOVA (→ III/1.3.8) are calculated as

$$\begin{aligned}
F_M &= \frac{n(\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2} \\
F_A &= \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_{i\bullet}(\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2} \\
F_B &= \frac{\frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2} \\
F_{A\times B} &= \frac{\frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y}_{\bullet\bullet\bullet})^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\bar{y}_{ij\bullet})^2}
\end{aligned} \tag{4}$$

and the ordinary least squares estimates for two-way ANOVA (→ III/1.3.10) are

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{\bullet\bullet\bullet} \\
\hat{\alpha}_i &= \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet} \\
\hat{\beta}_j &= \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet} \\
\hat{\gamma}_{ij} &= \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}
\end{aligned} \tag{5}$$

where the the sample means (→ I/1.10.2) are given by

$$
\begin{aligned}
\bar{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{i\bullet\bullet} &= \frac{1}{n_{i\bullet}} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{\bullet j\bullet} &= \frac{1}{n_{\bullet j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} y_{ijk} \\
\bar{y}_{ij\bullet} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} \; .
\end{aligned} \tag{6}
$$

We first note that the predicted values can be evaluated as

$$
\begin{split}
\hat{y}_{ijk} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} \\
&= \bar{y}_{\bullet\bullet\bullet} + (\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}) + (\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}) \\
&= \bar{y}_{i\bullet\bullet} + \bar{y}_{\bullet j\bullet} + \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} \\
&= \bar{y}_{ij\bullet} \; .
\end{split} \tag{7}
$$

Substituting this (7) and the OLS estimates (5) into the F-formulas (4), we obtain:

$$
\begin{aligned}
F_M &= \frac{n\hat{\mu}^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \\
F_A &= \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_{i\bullet}\hat{\alpha}_i^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \\
F_B &= \frac{\frac{1}{b-1}\sum_{j=1}^{b} n_{\bullet j}\hat{\beta}_j^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \\
F_{A\times B} &= \frac{\frac{1}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\hat{\gamma}_{ij}^2}{\frac{1}{n-ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk}-\hat{y}_{ijk})^2} \; .
\end{aligned} \tag{8}
$$

■

1.4 Simple linear regression

1.4.1 Definition

Definition: Let y and x be two $n \times 1$ vectors.
Then, a statement asserting a linear relationship between x and y

$$
y = \beta_0 + \beta_1 x + \varepsilon \; , \tag{1}
$$

together with a statement asserting a normal distribution (→ II/4.1.1) for ε

$$
\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{2}
$$

is called a univariate simple regression model or simply, "simple linear regression".

- y is called “dependent variable”, “measured data” or “signal”;
- x is called “independent variable”, “predictor” or “covariate”;
- V is called “covariance matrix” or “covariance structure”;
- β_1 is called “slope of the regression line (→ III/1.4.10)”;
- β_0 is called “intercept of the regression line (→ III/1.4.10)”;
- ε is called “noise”, “errors” or “error terms”;
- σ^2 is called “noise variance” or “error variance”;
- n is the number of observations.

When the covariance structure V is equal to the $n \times n$ identity matrix, this is called simple linear regression with independent and identically distributed (i.i.d.) observations:

$$V = I_n \quad \Rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \Rightarrow \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{3}$$

In this case, the linear regression model can also be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \; . \tag{4}$$

Otherwise, it is called simple linear regression with correlated observations.

Sources:

- Wikipedia (2021): “Simple linear regression”; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line.

1.4.2 Special case of multiple linear regression

Theorem: Simple linear regression (→ III/1.4.1) is a special case of multiple linear regression (→ III/1.5.1) with design matrix X and regression coefficients β

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{1}$$

where 1_n is an $n \times 1$ vector of ones, x is the $n \times 1$ single predictor variable, β_0 is the intercept and β_1 is the slope of the regression line (→ III/1.4.10).

Proof: Without loss of generality, consider the simple linear regression case with uncorrelated errors (→ III/1.4.1):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; . \tag{2}$$

In matrix notation and using the multivariate normal distribution (→ II/4.1.1), this can also be written as

$$\begin{aligned} y &= \beta_0 1_n + \beta_1 x + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, I_n) \\ y &= \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, I_n) \; . \end{aligned} \tag{3}$$

Comparing with the multiple linear regression equations for uncorrelated errors (→ III/1.5.1), we finally note:

$$y = X\beta + \varepsilon \quad \text{with} \quad X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \; . \tag{4}$$

In the case of correlated observations (→ III/1.4.1), the error distribution changes to (→ III/1.5.1):

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{5}$$

■

1.4.3 Ordinary least squares

Theorem: Given a simple linear regression model (→ III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) are given by

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \end{aligned} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2), s_x^2 is the sample variance (→ I/1.11.2) of x and s_{xy} is the sample covariance (→ I/1.13.2) between x and y.

Proof: The residual sum of squares (→ III/1.5.9) is defined as

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \; . \tag{3}$$

The derivatives of $\mathrm{RSS}(\beta_0, \beta_1)$ with respect to β_0 and β_1 are

$$\begin{aligned} \frac{\mathrm{dRSS}(\beta_0, \beta_1)}{\mathrm{d}\beta_0} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-1) \\ &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\mathrm{dRSS}(\beta_0, \beta_1)}{\mathrm{d}\beta_1} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \\ &= -2 \sum_{i=1}^n (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \end{aligned} \tag{4}$$

and setting these derivatives to zero

$$
\begin{aligned}
0 &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
0 &= -2 \sum_{i=1}^{n} (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)
\end{aligned} \tag{5}
$$

yields the following equations:

$$
\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_0 \cdot n &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i \; .
\end{aligned} \tag{6}
$$

From the first equation, we can derive the estimate for the intercept:

$$
\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i \\
&= \bar{y} - \hat{\beta}_1 \bar{x} \; .
\end{aligned} \tag{7}
$$

From the second equation, we can derive the estimate for the slope:

$$
\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \sum_{i=1}^{n} x_i &\overset{(7)}{=} \sum_{i=1}^{n} x_i y_i \\
\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) &= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \; .
\end{aligned} \tag{8}
$$

Note that the numerator can be rewritten as

$$
\begin{aligned}
\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x} \bar{y} \\
&= \sum_{i=1}^{n} \left(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y} \right) \\
&= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\end{aligned} \tag{9}
$$

and that the denominator can be rewritten as

$$
\begin{aligned}
\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2 \right) \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2 \; .
\end{aligned} \tag{10}
$$

With (9) and (10), the estimate from (8) can be simplified as follows:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \frac{s_{xy}}{s_x^2} \; .
\end{aligned} \tag{11}
$$

Together, (7) and (11) constitute the ordinary least squares parameter estimates for simple linear regression.

■

Sources:

- Penny, William (2006): "Linear regression"; in: *Mathematics for Brain Imaging*, ch. 1.2.2, pp. 14-16, eqs. 1.24/1.25; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Wikipedia (2021): "Proofs involving ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Derivation_of_simple_linear_regression_estimators.

1.4.4 Ordinary least squares

Theorem: Given a simple linear regression model (→ III/1.4.1) with independent observations

$$
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}
$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) are given by

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2}
\end{aligned} \tag{2}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2), s_x^2 is the sample variance (→ I/1.11.2) of x and s_{xy} is the sample covariance (→ I/1.13.2) between x and y.

Proof: Simple linear regression is a special case of multiple linear regression (→ III/1.4.2) with

$$
X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{3}
$$

and ordinary least squares estimates (→ III/1.5.3) are given by

$$
\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{4}
$$

Writing out equation (4), we have

$$
\begin{aligned}
\hat{\beta} &= \left(\begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} y \\
&= \left(\begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^\mathrm{T} x \end{bmatrix} \right)^{-1} \begin{bmatrix} n\bar{y} \\ x^\mathrm{T} y \end{bmatrix} \\
&= \frac{1}{n x^\mathrm{T} x - (n\bar{x})^2} \begin{bmatrix} x^\mathrm{T} x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ x^\mathrm{T} y \end{bmatrix} \\
&= \frac{1}{n x^\mathrm{T} x - (n\bar{x})^2} \begin{bmatrix} n\bar{y}\, x^\mathrm{T} x - n\bar{x}\, x^\mathrm{T} y \\ n\, x^\mathrm{T} y - (n\bar{x})(n\bar{y}) \end{bmatrix} \; .
\end{aligned} \tag{5}
$$

Thus, the second entry of $\hat{\beta}$ is equal to (→ III/1.4.3):

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{n\, x^\mathrm{T} y - (n\bar{x})(n\bar{y})}{n x^\mathrm{T} x - (n\bar{x})^2} \\
&= \frac{x^\mathrm{T} y - n\bar{x}\bar{y}}{x^\mathrm{T} x - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{s_{xy}}{s_x^2} \; .
\end{aligned} \tag{6}
$$

Moreover, the first entry of $\hat{\beta}$ is equal to:

$$
\begin{aligned}
\hat{\beta}_0 &= \frac{n\bar{y}\,x^\mathrm{T}x - n\bar{x}\,x^\mathrm{T}y}{nx^\mathrm{T}x - (n\bar{x})^2} \\
&= \frac{\bar{y}\,x^\mathrm{T}x - \bar{x}\,x^\mathrm{T}y}{x^\mathrm{T}x - n\bar{x}^2} \\
&= \frac{\bar{y}\,x^\mathrm{T}x - \bar{x}\,x^\mathrm{T}y + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y}}{x^\mathrm{T}x - n\bar{x}^2} \\
&= \frac{\bar{y}(x^\mathrm{T}x - n\bar{x}^2) - \bar{x}(x^\mathrm{T}y - n\bar{x}\bar{y})}{x^\mathrm{T}x - n\bar{x}^2} \\
&= \frac{\bar{y}(x^\mathrm{T}x - n\bar{x}^2)}{x^\mathrm{T}x - n\bar{x}^2} - \frac{\bar{x}(x^\mathrm{T}y - n\bar{x}\bar{y})}{x^\mathrm{T}x - n\bar{x}^2} \\
&= \bar{y} - \bar{x}\,\frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\
&= \bar{y} - \hat{\beta}_1 \bar{x} \; .
\end{aligned} \tag{7}
$$

■

1.4.5 Expectation of estimates

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$
y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}
$$

and consider estimation using ordinary least squares (→ III/1.4.3). Then, the expected values (→ I/1.10.1) of the estimated parameters are

$$
\begin{aligned}
\mathrm{E}(\hat{\beta}_0) &= \beta_0 \\
\mathrm{E}(\hat{\beta}_1) &= \beta_1
\end{aligned} \tag{2}
$$

which means that the ordinary least squares solution (→ III/1.4.3) produces unbiased estimators.

Proof: According to the simple linear regression model in (1), the expectation of a single data point is

$$
\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i \; . \tag{3}
$$

The ordinary least squares estimates for simple linear regression (→ III/1.4.3) are given by

$$
\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^n x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \; .
\end{aligned} \tag{4}
$$

If we define the following quantity

$$
c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \; , \tag{5}
$$

we note that

$$\begin{aligned}\sum_{i=1}^{n} c_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{n\bar{x} - n\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0 \; ,\end{aligned} \tag{6}$$

and

$$\begin{aligned}\sum_{i=1}^{n} c_i x_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\left(x_i^2 - \bar{x}x_i\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1 \; .\end{aligned} \tag{7}$$

With (5), the estimate for the slope from (4) becomes

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \sum_{i=1}^{n} c_i (y_i - \bar{y}) \\ &= \sum_{i=1}^{n} c_i y_i - \bar{y} \sum_{i=1}^{n} c_i\end{aligned} \tag{8}$$

and with (3), (6) and (7), its expectation becomes:

$$\begin{aligned}\mathrm{E}(\hat{\beta}_1) &= \mathrm{E}\left(\sum_{i=1}^{n} c_i y_i - \bar{y} \sum_{i=1}^{n} c_i\right) \\ &= \sum_{i=1}^{n} c_i \mathrm{E}(y_i) - \bar{y} \sum_{i=1}^{n} c_i \\ &= \beta_1 \sum_{i=1}^{n} c_i x_i + \beta_0 \sum_{i=1}^{n} c_i - \bar{y} \sum_{i=1}^{n} c_i \\ &= \beta_1 \; .\end{aligned} \tag{9}$$

Finally, with (3) and (9), the expectation of the intercept estimate from (4) becomes

$$
\begin{aligned}
\mathrm{E}(\hat{\beta}_0) &= \mathrm{E}\left(\frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i \right) \\
&= \frac{1}{n} \sum_{i=1}^n \mathrm{E}(y_i) - \mathrm{E}(\hat{\beta}_1) \cdot \bar{x} \\
&= \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \cdot \bar{x} \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0 \; .
\end{aligned} \tag{10}
$$

■

Sources:

- Penny, William (2006): "Finding the uncertainty in estimating the slope"; in: *Mathematics for Brain Imaging*, ch. 1.2.4, pp. 18-20, eq. 1.37; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Wikipedia (2021): "Proofs involving ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F'%22%60UNIQ--postMath-00000037-QINU%60%22'%7F.

1.4.6 Variance of estimates

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$
y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}
$$

and consider estimation using ordinary least squares (→ III/1.4.3). Then, the variances (→ I/1.11.1) of the estimated parameters are

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \frac{x^\mathrm{T} x}{n} \cdot \frac{\sigma^2}{(n-1) s_x^2} \\
\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{(n-1) s_x^2}
\end{aligned} \tag{2}
$$

where s_x^2 is the sample variance (→ I/1.11.2) of x and $x^\mathrm{T} x$ is the sum of squared values of the covariate.

Proof: According to the simple linear regression model in (1), the variance of a single data point is

$$
\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2 \; . \tag{3}
$$

The ordinary least squares estimates for simple linear regression (→ III/1.4.3) are given by

$$
\begin{aligned}
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \; .
\end{aligned} \tag{4}
$$

If we define the following quantity

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \; , \tag{5}$$

we note that

$$\begin{split}
\sum_{i=1}^n c_i^2 &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \; .
\end{split} \tag{6}$$

With (5), the estimate for the slope from (4) becomes

$$\begin{split}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sum_{i=1}^n c_i (y_i - \bar{y}) \\
&= \sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i
\end{split} \tag{7}$$

and with (3) and (6) as well as invariance (→ I/1.11.6), scaling (→ I/1.11.7) and additivity (→ I/1.11.10) of the variance, the variance of $\hat{\beta}_1$ is:

$$\begin{split}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\sum_{i=1}^n c_i y_i - \bar{y} \sum_{i=1}^n c_i \right) \\
&= \mathrm{Var}\left(\sum_{i=1}^n c_i y_i \right) \\
&= \sum_{i=1}^n c_i^2 \mathrm{Var}(y_i) \\
&= \sigma^2 \sum_{i=1}^n c_i^2 \\
&= \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{(n-1)\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{(n-1) s_x^2} \; .
\end{split} \tag{8}$$

Finally, with (3) and (8), the variance of the intercept estimate from (4) becomes:

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i \right) \\
&= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i \right) + \mathrm{Var}\left(\hat{\beta}_1 \cdot \bar{x} \right) \\
&= \left(\frac{1}{n} \right)^2 \sum_{i=1}^n \mathrm{Var}(y_i) + \bar{x}^2 \cdot \mathrm{Var}(\hat{\beta}_1) \\
&= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 + \bar{x}^2 \frac{\sigma^2}{(n-1) s_x^2} \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{(n-1) s_x^2} \; .
\end{aligned} \tag{9}
$$

Applying the formula for the sample variance (→ I/1.11.2) s_x^2, we finally get:

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n \left(x_i^2 - 2 \bar{x} x_i + \bar{x}^2 \right) + \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 - 2 \bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \bar{x}^2 \right) + \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n x_i^2 - 2 \bar{x}^2 + 2 \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{\frac{1}{n} \sum_{i=1}^n x_i^2}{(n-1) \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \frac{x^\mathrm{T} x}{n} \cdot \frac{\sigma^2}{(n-1) s_x^2} \; .
\end{aligned} \tag{10}
$$

■

Sources:

- Penny, William (2006): "Finding the uncertainty in estimating the slope"; in: *Mathematics for Brain Imaging*, ch. 1.2.4, pp. 18-20, eq. 1.37; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Wikipedia (2021): "Proofs involving ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F'%22%60UNIQ--postMath-00000037-QINU%60%22'%7F.

1.4.7 Distribution of estimates

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}$$

and consider estimation using ordinary least squares (→ III/1.4.3). Then, the estimated parameters are normally distributed (→ II/4.1.1) as

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \, \frac{\sigma^2}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \tag{2}$$

where $\bar{x}$ is the sample mean (→ I/1.10.2) and s_x^2 is the sample variance (→ I/1.11.2) of x.

Proof: Simple linear regression is a special case of multiple linear regression (→ III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \, , \tag{3}$$

such that (1) can also be written as

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \tag{4}$$

and ordinary least sqaures estimates (→ III/1.5.3) are given by

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{5}$$

From (4) and the linear transformation theorem for the multivariate normal distribution (→ II/4.1.13), it follows that

$$y \sim \mathcal{N}\left(X\beta, \, \sigma^2 I_n \right) \; . \tag{6}$$

From (5), in combination with (6) and the transformation theorem (→ II/4.1.13), it follows that

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}\left((X^\mathrm{T} X)^{-1} X^\mathrm{T} X\beta, \, \sigma^2 (X^\mathrm{T} X)^{-1} X^\mathrm{T} I_n X (X^\mathrm{T} X)^{-1} \right) \\ &\sim \mathcal{N}\left(\beta, \, \sigma^2 (X^\mathrm{T} X)^{-1} \right) \; . \end{aligned} \tag{7}$$

Applying (3), the covariance matrix (→ II/4.1.1) can be further developed as follows:

$$\begin{aligned} \sigma^2 (X^\mathrm{T} X)^{-1} &= \sigma^2 \left(\begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \\ &= \sigma^2 \left(\begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^\mathrm{T} x \end{bmatrix} \right)^{-1} \\ &= \frac{\sigma^2}{n x^\mathrm{T} x - (n\bar{x})^2} \begin{bmatrix} x^\mathrm{T} x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\ &= \frac{\sigma^2}{x^\mathrm{T} x - n\bar{x}^2} \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \; . \end{aligned} \tag{8}$$

Note that the denominator in the first factor is equal to

$$
\begin{aligned}
x^{\mathrm{T}}x - n\bar{x}^2 &= x^{\mathrm{T}}x - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \sum_{i=1}^{n} \left(x_i^2 - \bar{x}\right)^2 \\
&= (n-1)\, s_x^2 \; .
\end{aligned} \tag{9}
$$

Thus, combining (7), (8) and (9), we have

$$
\hat{\beta} \sim \mathcal{N}\left(\beta,\, \frac{\sigma^2}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^{\mathrm{T}}x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\right) \tag{10}
$$

which is equivalent to equation (2).

■

Sources:

- Wikipedia (2021): "Proofs involving ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-09; URL: https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Unbiasedness_and_variance_of_%7F'%22%60UNIQ--postMath-00000037-QINU%60%22'%7F.

1.4.8 Correlation of estimates

Theorem: In simple linear regression (→ III/1.4.1), when the independent variable x is mean-centered (→ I/1.10.1), the ordinary least squares (→ III/1.4.3) estimates for slope and intercept are uncorrelated (→ I/1.14.1).

Proof: The parameter estimates for simple linear regression are bivariate normally distributed under ordinary least squares (→ III/1.4.7):

$$
\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},\, \frac{\sigma^2}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^{\mathrm{T}}x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\right) \tag{1}
$$

Because the covariance matrix (→ I/1.13.9) of the multivariate normal distribution (→ II/4.1.1) contains the pairwise covariances of the random variables (→ I/1.2.2), we can deduce that the covariance (→ I/1.13.1) of $\hat{\beta}_0$ and $\hat{\beta}_1$ is:

$$
\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2\, \bar{x}}{(n-1)\, s_x^2} \tag{2}
$$

where σ^2 is the noise variance (→ III/1.4.1), s_x^2 is the sample variance (→ I/1.11.2) of x and n is the number of observations. When x is mean-centered, we have $\bar{x} = 0$, such that:

$$\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = 0 \; . \tag{3}$$

Because correlation is equal to covariance divided by standard deviations (→ I/1.14.1), we can conclude that the correlation of $\hat{\beta}_0$ and $\hat{\beta}_1$ is also zero:

$$\mathrm{Corr}\left(\hat{\beta}_0, \hat{\beta}_1\right) = 0 \; . \tag{4}$$

■

1.4.9 Effects of mean-centering

Theorem: In simple linear regression (→ III/1.4.1), when the dependent variable y and/or the independent variable x are mean-centered (→ I/1.10.1), the ordinary least squares (→ III/1.4.3) estimate for the intercept changes, but that of the slope does not.

Proof:
1) Under unaltered y and x, ordinary least squares estimates for simple linear regression (→ III/1.4.3) are

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \end{aligned} \tag{1}$$

with sample means (→ I/1.10.2) $\bar{x}$ and $\bar{y}$, sample variance (→ I/1.11.2) s_x^2 and sample covariance (→ I/1.13.2) s_{xy}, such that β_0 estimates "the mean y at $x = 0$".

2) Let $\tilde{x}$ be the mean-centered covariate vector (→ III/1.4.1):

$$\tilde{x}_i = x_i - \bar{x} \quad \Rightarrow \quad \bar{\tilde{x}} = 0 \; . \tag{2}$$

Under this condition, the parameter estimates become

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{\tilde{x}} \\ &= \bar{y} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})(y_i - \bar{y})}{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \end{aligned} \tag{3}$$

and we can see that $\hat{\beta}_1(\tilde{x}, y) = \hat{\beta}_1(x, y)$, but $\hat{\beta}_0(\tilde{x}, y) \neq \hat{\beta}_0(x, y)$, specifically β_0 now estimates "the mean y at the mean x".

3) Let $\tilde{y}$ be the mean-centered data vector (→ III/1.4.1):

$$\tilde{y}_i = y_i - \bar{y} \quad \Rightarrow \quad \bar{\tilde{y}} = 0 \; . \tag{4}$$

Under this condition, the parameter estimates become

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{\tilde{y}} - \hat{\beta}_1 \bar{x} \\
&= -\hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\tilde{y}_i - \bar{\tilde{y}})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}
\end{aligned} \tag{5}
$$

and we can see that $\hat{\beta}_1(x, \tilde{y}) = \hat{\beta}_1(x, y)$, but $\hat{\beta}_0(x, \tilde{y}) \neq \hat{\beta}_0(x, y)$, specifically β_0 now estimates "the mean x, multiplied with the negative slope".

4) Finally, consider mean-centering both x and y::

$$
\begin{aligned}
\tilde{x}_i = x_i - \bar{x} \quad &\Rightarrow \quad \bar{\tilde{x}} = 0 \\
\tilde{y}_i = y_i - \bar{y} \quad &\Rightarrow \quad \bar{\tilde{y}} = 0 \; .
\end{aligned} \tag{6}
$$

Under this condition, the parameter estimates become

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{\tilde{y}} - \hat{\beta}_1 \bar{\tilde{x}} \\
&= 0 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})}{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}
\end{aligned} \tag{7}
$$

and we can see that $\hat{\beta}_1(\tilde{x}, \tilde{y}) = \hat{\beta}_1(x, y)$, but $\hat{\beta}_0(\tilde{x}, \tilde{y}) \neq \hat{\beta}_0(x, y)$, specifically β_0 is now forced to become zero.

■

1.4.10 Regression line

Definition: Let there be a simple linear regression with independent observations (→ III/1.4.1) using dependent variable y and independent variable x:

$$
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \; . \tag{1}
$$

Then, given some parameters $\beta_0, \beta_1 \in \mathbb{R}$, the set

$$
L(\beta_0, \beta_1) = \left\{ (x, y) \in \mathbb{R}^2 \mid y = \beta_0 + \beta_1 x \right\} \tag{2}
$$

is called a "regression line" and the set

$$
L(\hat{\beta}_0, \hat{\beta}_1) \tag{3}
$$

is called the "fitted regression line", with estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1$, e.g. obtained via ordinary least squares (→ III/1.4.3).

Sources:

- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line.

1.4.11 Regression line includes center of mass

Theorem: In simple linear regression (→ III/1.4.1), the regression line (→ III/1.4.10) estimated using ordinary least squares (→ III/1.4.3) includes the point $M(\bar{x}, \bar{y})$.

Proof: The fitted regression line (→ III/1.4.10) is described by the equation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{where} \quad x, y \in \mathbb{R} \; . \tag{1}$$

Plugging in the coordinates of M and the ordinary least squares estimate of the intercept (→ III/1.4.3), we obtain

$$\begin{aligned} \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ \bar{y} &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} \\ \bar{y} &= \bar{y} \; . \end{aligned} \tag{2}$$

which is a true statement. Thus, the regression line (→ III/1.4.10) goes through the center of mass point $(\bar{x}, \bar{y})$, if the model (→ III/1.4.1) includes an intercept term β_0.

■

Sources:
- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties.

1.4.12 Projection of data point to regression line

Theorem: Consider simple linear regression (→ III/1.4.1) and an estimated regression line (→ III/1.4.10) specified by

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{where} \quad x, y \in \mathbb{R} \; . \tag{1}$$

For any given data point $O(x_o|y_o)$, the point on the regression line $P(x_p|y_p)$ that is closest to this data point is given by:

$$P\left(w \mid \hat{\beta}_0 + \hat{\beta}_1 w\right) \quad \text{with} \quad w = \frac{x_0 + (y_o - \hat{\beta}_0)\hat{\beta}_1}{1 + \hat{\beta}_1^2} \; . \tag{2}$$

Proof: The intersection point of the regression line (→ III/1.4.10) with the y-axis is

$$S(0|\hat{\beta}_0) \; . \tag{3}$$

Let a be a vector describing the direction of the regression line, let b be the vector pointing from S to O and let p be the vector pointing from S to P.
Because $\hat{\beta}_1$ is the slope of the regression line, we have

$$a = \begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix} \; . \tag{4}$$

Moreover, with the points O and S, we have

$$b = \begin{pmatrix} x_o \\ y_o \end{pmatrix} - \begin{pmatrix} 0 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} x_o \\ y_o - \hat{\beta}_0 \end{pmatrix} \; . \tag{5}$$

Because P is located on the regression line, p is collinear with a and thus a scalar multiple of this vector:

$$p = w \cdot a \; . \tag{6}$$

Moreover, as P is the point on the regression line which is closest to O, this means that the vector $b - p$ is orthogonal to a, such that the inner product of these two vectors is equal to zero:

$$a^\mathrm{T}(b - p) = 0 \; . \tag{7}$$

Rearranging this equation gives

$$\begin{split}
a^\mathrm{T}(b - p) &= 0 \\
a^\mathrm{T}(b - w \cdot a) &= 0 \\
a^\mathrm{T} b - w \cdot a^\mathrm{T} a &= 0 \\
w \cdot a^\mathrm{T} a &= a^\mathrm{T} b \\
w &= \frac{a^\mathrm{T} b}{a^\mathrm{T} a} \; .
\end{split} \tag{8}$$

With (4) and (5), w can be calculated as

$$\begin{split}
w &= \frac{a^\mathrm{T} b}{a^\mathrm{T} a} \\
w &= \frac{\begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix}^\mathrm{T} \begin{pmatrix} x_o \\ y_o - \hat{\beta}_0 \end{pmatrix}}{\begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix}^\mathrm{T} \begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix}} \\
w &= \frac{x_0 + (y_o - \hat{\beta}_0)\hat{\beta}_1}{1 + \hat{\beta}_1^2}
\end{split} \tag{9}$$

Finally, with the point S (3) and the vector p (6), the coordinates of P are obtained as

$$\begin{pmatrix} x_p \\ y_p \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\beta}_0 \end{pmatrix} + w \cdot \begin{pmatrix} 1 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} w \\ \hat{\beta}_0 + \hat{\beta}_1 w \end{pmatrix} \; . \tag{10}$$

Together, (10) and (9) constitute the proof of equation (2).

■

Sources:
- Penny, William (2006): "Projections"; in: *Mathematics for Brain Imaging*, ch. 1.4.10, pp. 34-35, eqs. 1.87/1.88; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.

1.4.13 Sums of squares

Theorem: Under ordinary least squares (→ III/1.4.3) for simple linear regression (→ III/1.4.1), total (→ III/1.5.7), explained (→ III/1.5.8) and residual (→ III/1.5.9) sums of squares are given by

$$
\begin{aligned}
\mathrm{TSS} &= (n-1)\, s_y^2 \\
\mathrm{ESS} &= (n-1)\, \frac{s_{xy}^2}{s_x^2} \\
\mathrm{RSS} &= (n-1) \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right)
\end{aligned} \tag{1}
$$

where s_x^2 and s_y^2 are the sample variances (→ I/1.11.2) of x and y and s_{xy} is the sample covariance (→ I/1.13.2) between x and y.

Proof: The ordinary least squares parameter estimates (→ III/1.4.3) are given by

$$
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \; . \tag{2}
$$

1) The total sum of squares (→ III/1.5.7) is defined as

$$
\mathrm{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2 \tag{3}
$$

which can be reformulated as follows:

$$
\begin{aligned}
\mathrm{TSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
&= (n-1) \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\
&= (n-1) s_y^2 \; .
\end{aligned} \tag{4}
$$

2) The explained sum of squares (→ III/1.5.8) is defined as

$$
\mathrm{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{5}
$$

which, with the OLS parameter estimates, becomes:

$$
\begin{aligned}
\mathrm{ESS} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\
&\overset{(2)}{=} \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\
&\overset{(2)}{=} \sum_{i=1}^{n} \left(\frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right)^2 \\
&= \left(\frac{s_{xy}}{s_x^2} \right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \left(\frac{s_{xy}}{s_x^2} \right)^2 (n-1) s_x^2 \\
&= (n-1) \, \frac{s_{xy}^2}{s_x^2} \; .
\end{aligned} \tag{6}
$$

3) The residual sum of squares (→ III/1.5.9) is defined as

$$
\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{7}
$$

which, with the OLS parameter estimates, becomes:

$$
\begin{aligned}
\mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&\overset{(2)}{=} \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\
&= \sum_{i=1}^{n} \left((y_i - \bar{y})^2 - 2 \hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 (x_i - \bar{x})^2 \right) \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2 \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= (n-1)\, s_y^2 - 2(n-1)\, \hat{\beta}_1\, s_{xy} + (n-1)\, \hat{\beta}_1^2\, s_x^2 \\
&\overset{(2)}{=} (n-1)\, s_y^2 - 2(n-1) \left(\frac{s_{xy}}{s_x^2} \right) s_{xy} + (n-1) \left(\frac{s_{xy}}{s_x^2} \right)^2 s_x^2 \\
&= (n-1)\, s_y^2 - (n-1)\, \frac{s_{xy}^2}{s_x^2} \\
&= (n-1) \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right) \; .
\end{aligned} \tag{8}
$$

■

1.4.14 Partition of sums of squares

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$
y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}
$$

where β_0 and β_1 are intercept and slope parameter (→ III/1.4.1), respectively. Then, it holds that

$$
\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS} \tag{2}
$$

where TSS is the total sum of squares (→ III/1.5.7), ESS is the explained sum of squares (→ III/1.5.8) and RSS is the residual sum of squares (→ III/1.5.9).

Proof: For simple linear regression, total, explained and residual sum squares are given by (→ III/1.4.13)

$$
\begin{aligned}
\mathrm{TSS} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
\mathrm{ESS} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\
\mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned} \tag{3}
$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated regression coefficients obtained via ordinary least squares (→ III/1.4.3)

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2}
\end{aligned} \tag{4}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2) of x and y, s_{xy} is the unbiased sample covariance (→ I/1.13.2) of x and y and s_x^2 is the unbiased sample variance (→ I/1.13.2) of x:

$$
\begin{aligned}
s_{xy} &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
s_x^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \; .
\end{aligned} \tag{5}
$$

With that in mind, we start working out the total sum of squares:

$$
\begin{aligned}
\mathrm{TSS} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2 \\
&= \sum_{i=1}^{n} \left((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2 \right) \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&\overset{(3)}{=} \mathrm{ESS} + \mathrm{RSS} + \sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \; .
\end{aligned} \tag{6}
$$

Thus, what remains to be shown is that the following sum is zero:

$$\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \tag{7}$$

Using the expression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for the fitted signal values (→ III/1.4.10), we proceed as follows:

$$\begin{split}
\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) \\
&\overset{(4)}{=} \sum_{i=1}^{n} 2(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y}) \\
&= \sum_{i=1}^{n} 2 \left((y_i - \bar{y}) - (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}) \right) (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}) \\
&= 2 \sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right) \hat{\beta}_1 (x_i - \bar{x}) \\
&= 2 \sum_{i=1}^{n} \left((y_i - \hat{y}_i) \hat{\beta}_1 (x_i - \bar{x}) - \hat{\beta}_1 (x_i - \bar{x}) \hat{\beta}_1 (x_i - \bar{x}) \right) \\
&= 2 \left[\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i - \bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) \right] \; .
\end{split} \tag{8}$$

Next, we recognize the sample covariance and sample variance terms from (5):

$$\begin{split}
\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2 \left[\hat{\beta}_1 (n-1) s_{xy} - \hat{\beta}_1^2 (n-1) s_x^2 \right] \\
&= 2(n-1) \left[\hat{\beta}_1 s_{xy} - \hat{\beta}_1^2 s_x^2 \right] \; .
\end{split} \tag{9}$$

Now, we can apply to functional form of the estimate $\hat{\beta}_1$ from (4) to get:

$$\begin{split}
\sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= 2(n-1) \left[\left(\frac{s_{xy}}{s_x^2} \right) s_{xy} - \left(\frac{s_{xy}}{s_x^2} \right)^2 s_x^2 \right] \\
&= 2(n-1) \left[\frac{s_{xy}^2}{s_x^2} - \frac{s_{xy}^2}{s_x^2} \right] \\
&= 2(n-1) \cdot 0 \\
&= 0 \; .
\end{split} \tag{10}$$

Plugging the result from (10) into (6), we finally get:

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS} \; . \tag{11}$$

■

Sources:

- Ostwald, Dirk (2023): "Korrelation"; in: *Allgemeines Lineares Modell*, Einheit (2), Folien 19-23; URL: https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/2_Korrelation.pdf.
- Manabu, Hayashi (2021): "TSS = RSS + ESS | Simple Linear Regression"; in: *YouTube*, retrieved on 2024-07-12; URL: https://www.youtube.com/watch?v=N7pHym1L9b0.

1.4.15 Transformation matrices

Theorem: Under ordinary least squares (→ III/1.4.3) for simple linear regression (→ III/1.4.1), estimation (→ III/1.5.11), projection (→ III/1.5.12) and residual-forming (→ III/1.5.13) matrices are given by

$$
\begin{split}
E &= \frac{1}{(n-1)\,s_x^2} \begin{bmatrix} (x^\mathrm{T}x/n)\,1_n^\mathrm{T} - \bar{x}\,x^\mathrm{T} \\ -\bar{x}\,1_n^\mathrm{T} + x^\mathrm{T} \end{bmatrix} \\
P &= \frac{1}{(n-1)\,s_x^2} \begin{bmatrix} (x^\mathrm{T}x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^\mathrm{T}x/n) - \bar{x}(x_1+x_n) + x_1 x_n \\ \vdots & \ddots & \vdots \\ (x^\mathrm{T}x/n) - \bar{x}(x_1+x_n) + x_1 x_n & \cdots & (x^\mathrm{T}x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix} \\
R &= \frac{1}{(n-1)\,s_x^2} \begin{bmatrix} (n-1)(x^\mathrm{T}x/n) + \bar{x}(2x_1 - n\bar{x}) - x_1^2 & \cdots & -(x^\mathrm{T}x/n) + \bar{x}(x_1+x_n) - x_1 x_n \\ \vdots & \ddots & \vdots \\ -(x^\mathrm{T}x/n) + \bar{x}(x_1+x_n) - x_1 x_n & \cdots & (n-1)(x^\mathrm{T}x/n) + \bar{x}(2x_n - n\bar{x}) - x_n^2 \end{bmatrix}
\end{split}
\tag{1}
$$

where 1_n is an $n \times 1$ vector of ones, x is the $n \times 1$ single predictor variable, $\bar{x}$ is the sample mean (→ I/1.10.2) of x and s_x^2 is the sample variance (→ I/1.11.2) of x.

Proof: Simple linear regression is a special case of multiple linear regression (→ III/1.4.2) with

$$
X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \; , \tag{2}
$$

such that the simple linear regression model can also be written as

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \; . \tag{3}
$$

Moreover, we note the following equality (→ III/1.4.7):

$$
x^\mathrm{T}x - n\bar{x}^2 = (n-1)\,s_x^2 \; . \tag{4}
$$

1) The estimation matrix is given by (→ III/1.5.14)

$$
E = (X^\mathrm{T}X)^{-1}X^\mathrm{T} \tag{5}
$$

which is a $2 \times n$ matrix and can be reformulated as follows:

$$
\begin{split}
E &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} \\
&= \left(\begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \\
&= \left(\begin{bmatrix} n & n\bar{x} \\ n\bar{x} & x^\mathrm{T} x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \\
&= \frac{1}{n x^\mathrm{T} x - (n\bar{x})^2} \begin{bmatrix} x^\mathrm{T} x & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \\
&= \frac{1}{x^\mathrm{T} x - n\bar{x}^2} \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \\
&\overset{(4)}{=} \frac{1}{(n-1)\, s_x^2} \begin{bmatrix} (x^\mathrm{T} x/n)\, 1_n^\mathrm{T} - \bar{x}\, x^\mathrm{T} \\ -\bar{x}\, 1_n^\mathrm{T} + x^\mathrm{T} \end{bmatrix} \; .
\end{split} \tag{6}
$$

2) The projection matrix is given by (→ III/1.5.14)

$$
P = X (X^\mathrm{T} X)^{-1} X^\mathrm{T} = X\, E \tag{7}
$$

which is an $n \times n$ matrix and can be reformulated as follows:

$$
\begin{split}
P = X\, E &= \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \\
&= \frac{1}{(n-1)\, s_x^2} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} (x^\mathrm{T} x/n) - \bar{x} x_1 & \cdots & (x^\mathrm{T} x/n) - \bar{x} x_n \\ -\bar{x} + x_1 & \cdots & -\bar{x} + x_n \end{bmatrix} \\
&= \frac{1}{(n-1)\, s_x^2} \begin{bmatrix} (x^\mathrm{T} x/n) - 2\bar{x} x_1 + x_1^2 & \cdots & (x^\mathrm{T} x/n) - \bar{x}(x_1 + x_n) + x_1 x_n \\ \vdots & \ddots & \vdots \\ (x^\mathrm{T} x/n) - \bar{x}(x_1 + x_n) + x_1 x_n & \cdots & (x^\mathrm{T} x/n) - 2\bar{x} x_n + x_n^2 \end{bmatrix} \; .
\end{split} \tag{8}
$$

3) The residual-forming matrix is given by (→ III/1.5.14)

$$
R = I_n - X (X^\mathrm{T} X)^{-1} X^\mathrm{T} = I_n - P \tag{9}
$$

which also is an $n \times n$ matrix and can be reformulated as follows:

$$
\begin{split}
R = I_n - P &= \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \\
&\overset{(4)}{=} \frac{1}{(n-1)\, s_x^2} \begin{bmatrix} x^\mathrm{T} x - n\bar{x}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x^\mathrm{T} x - n\bar{x}^2 \end{bmatrix} \\
&- \frac{1}{(n-1)\, s_x^2} \begin{bmatrix} (x^\mathrm{T} x/n) - 2\bar{x}x_1 + x_1^2 & \cdots & (x^\mathrm{T} x/n) - \bar{x}(x_1+x_n) + x_1 x_n \\ \vdots & \ddots & \vdots \\ (x^\mathrm{T} x/n) - \bar{x}(x_1+x_n) + x_1 x_n & \cdots & (x^\mathrm{T} x/n) - 2\bar{x}x_n + x_n^2 \end{bmatrix} \\
&= \frac{1}{(n-1)\, s_x^2} \begin{bmatrix} (n-1)(x^\mathrm{T} x/n) + \bar{x}(2x_1 - n\bar{x}) - x_1^2 & \cdots & -(x^\mathrm{T} x/n) + \bar{x}(x_1+x_n) - x_1 x_n \\ \vdots & \ddots & \vdots \\ -(x^\mathrm{T} x/n) + \bar{x}(x_1+x_n) - x_1 x_n & \cdots & (n-1)(x^\mathrm{T} x/n) + \bar{x}(2x_n - n\bar{x}) - x_n^2 \end{bmatrix} .
\end{split}
\tag{10}
$$

■

1.4.16 Weighted least squares

Theorem: Given a simple linear regression model (→ III/1.4.1) with correlated observations

$$
y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{1}
$$

the parameters minimizing the weighted residual sum of squares (→ III/1.5.9) are given by

$$
\begin{split}
\hat{\beta}_0 &= \frac{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} y - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} y}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \\
\hat{\beta}_1 &= \frac{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} y - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} y}{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} x - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} x}
\end{split}
\tag{2}
$$

where 1_n is an $n \times 1$ vector of ones.

Proof: Let there be an $n \times n$ square matrix W, such that

$$
W V W^\mathrm{T} = I_n \; . \tag{3}
$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed as the matrix square root of the inverse of V:

$$
W V W = I_n \quad \Leftrightarrow \quad V = W^{-1} W^{-1} \quad \Leftrightarrow \quad V^{-1} = W W \quad \Leftrightarrow \quad W = V^{-1/2} \; . \tag{4}
$$

Because β_0 is a scalar, (1) may also be written as

$$y = \beta_0 1_n + \beta_1 x + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{5}$$

Left-multiplying (5) with W, the linear transformation theorem (→ II/4.1.13) implies that

$$Wy = \beta_0 W 1_n + \beta_1 W x + W\varepsilon, \; W\varepsilon \sim \mathcal{N}(0, \sigma^2 W V W^\mathrm{T}) \; . \tag{6}$$

Applying (3), we see that (6) is actually a linear regression model (→ III/1.5.1) with independent observations

$$\tilde{y} = \begin{bmatrix} \tilde{x}_0 & \tilde{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \tilde{\varepsilon}, \; \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \tag{7}$$

where $\tilde{y} = Wy$, $\tilde{x}_0 = W 1_n$, $\tilde{x} = Wx$ and $\tilde{\varepsilon} = W\varepsilon$, such that we can apply the ordinary least squares solution (→ III/1.5.3) giving:

$$\begin{aligned}
\hat{\beta} &= (\tilde{X}^\mathrm{T} \tilde{X})^{-1} \tilde{X}^\mathrm{T} \tilde{y} \\
&= \left(\begin{bmatrix} \tilde{x}_0^\mathrm{T} \\ \tilde{x}^\mathrm{T} \end{bmatrix} \begin{bmatrix} \tilde{x}_0 & \tilde{x} \end{bmatrix} \right)^{-1} \begin{bmatrix} \tilde{x}_0^\mathrm{T} \\ \tilde{x}^\mathrm{T} \end{bmatrix} \tilde{y} \\
&= \begin{bmatrix} \tilde{x}_0^\mathrm{T} \tilde{x}_0 & \tilde{x}_0^\mathrm{T} \tilde{x} \\ \tilde{x}^\mathrm{T} \tilde{x}_0 & \tilde{x}^\mathrm{T} \tilde{x} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x}_0^\mathrm{T} \\ \tilde{x}^\mathrm{T} \end{bmatrix} \tilde{y} \; .
\end{aligned} \tag{8}$$

Applying the inverse of a 2×2 matrix, this reformulates to:

$$\begin{aligned}
\hat{\beta} &= \frac{1}{\tilde{x}_0^\mathrm{T} \tilde{x}_0 \, \tilde{x}^\mathrm{T} \tilde{x} - \tilde{x}_0^\mathrm{T} \tilde{x} \, \tilde{x}^\mathrm{T} \tilde{x}_0} \begin{bmatrix} \tilde{x}^\mathrm{T} \tilde{x} & -\tilde{x}_0^\mathrm{T} \tilde{x} \\ -\tilde{x}^\mathrm{T} \tilde{x}_0 & \tilde{x}_0^\mathrm{T} \tilde{x}_0 \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x}_0^\mathrm{T} \\ \tilde{x}^\mathrm{T} \end{bmatrix} \tilde{y} \\
&= \frac{1}{\tilde{x}_0^\mathrm{T} \tilde{x}_0 \, \tilde{x}^\mathrm{T} \tilde{x} - \tilde{x}_0^\mathrm{T} \tilde{x} \, \tilde{x}^\mathrm{T} \tilde{x}_0} \begin{bmatrix} \tilde{x}^\mathrm{T} \tilde{x} \, \tilde{x}_0^\mathrm{T} - \tilde{x}_0^\mathrm{T} \tilde{x} \, \tilde{x}^\mathrm{T} \\ \tilde{x}_0^\mathrm{T} \tilde{x}_0 \, \tilde{x}^\mathrm{T} - \tilde{x}^\mathrm{T} \tilde{x}_0 \, \tilde{x}_0^\mathrm{T} \end{bmatrix} \tilde{y} \\
&= \frac{1}{\tilde{x}_0^\mathrm{T} \tilde{x}_0 \, \tilde{x}^\mathrm{T} \tilde{x} - \tilde{x}_0^\mathrm{T} \tilde{x} \, \tilde{x}^\mathrm{T} \tilde{x}_0} \begin{bmatrix} \tilde{x}^\mathrm{T} \tilde{x} \, \tilde{x}_0^\mathrm{T} \tilde{y} - \tilde{x}_0^\mathrm{T} \tilde{x} \, \tilde{x}^\mathrm{T} \tilde{y} \\ \tilde{x}_0^\mathrm{T} \tilde{x}_0 \, \tilde{x}^\mathrm{T} \tilde{y} - \tilde{x}^\mathrm{T} \tilde{x}_0 \, \tilde{x}_0^\mathrm{T} \tilde{y} \end{bmatrix} \; .
\end{aligned} \tag{9}$$

Applying $\tilde{x}_0 = W 1_n$, $\tilde{x} = Wx$ and $W^\mathrm{T} W = WW = V^{-1}$, we finally have

$$\begin{aligned}
\hat{\beta} &= \frac{1}{1_n^\mathrm{T} W^\mathrm{T} W 1_n \, x^\mathrm{T} W^\mathrm{T} W x - 1_n^\mathrm{T} W^\mathrm{T} W x \, x^\mathrm{T} W^\mathrm{T} W 1_n} \begin{bmatrix} x^\mathrm{T} W^\mathrm{T} W x \, 1_n^\mathrm{T} W^\mathrm{T} W y - 1_n^\mathrm{T} W^\mathrm{T} W x \, x^\mathrm{T} W^\mathrm{T} W y \\ 1_n^\mathrm{T} W^\mathrm{T} W 1_n \, x^\mathrm{T} W^\mathrm{T} W y - x^\mathrm{T} W^\mathrm{T} W 1_n \, 1_n^\mathrm{T} W^\mathrm{T} W y \end{bmatrix} \\
&= \frac{1}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \begin{bmatrix} x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} y - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} y \\ 1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} y - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} y \end{bmatrix} \\
&= \begin{bmatrix} \frac{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} y - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} y}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \\ \frac{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} y - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} y}{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} x - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} x} \end{bmatrix}
\end{aligned} \tag{10}$$

which corresponds to the weighted least squares solution (2).

■

1.4.17 Weighted least squares

Theorem: Given a simple linear regression model (→ III/1.4.1) with correlated observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{1}$$

the parameters minimizing the weighted residual sum of squares (→ III/1.5.9) are given by

$$\begin{split} \hat{\beta}_0 &= \frac{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} y - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} y}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \\ \hat{\beta}_1 &= \frac{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} y - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} y}{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} x - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} x} \end{split} \tag{2}$$

where 1_n is an $n \times 1$ vector of ones.

Proof: Simple linear regression is a special case of multiple linear regression (→ III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{3}$$

and weighted least squares estimates (→ III/1.5.21) are given by

$$\hat{\beta} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \; . \tag{4}$$

Writing out equation (4), we have

$$\begin{split} \hat{\beta} &= \left(\begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} V^{-1} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} V^{-1} y \\ &= \begin{bmatrix} 1_n^\mathrm{T} V^{-1} 1_n & 1_n^\mathrm{T} V^{-1} x \\ x^\mathrm{T} V^{-1} 1_n & x^\mathrm{T} V^{-1} x \end{bmatrix}^{-1} \begin{bmatrix} 1_n^\mathrm{T} V^{-1} y \\ x^\mathrm{T} V^{-1} y \end{bmatrix} \\ &= \frac{1}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \begin{bmatrix} x^\mathrm{T} V^{-1} x & -1_n^\mathrm{T} V^{-1} x \\ -x^\mathrm{T} V^{-1} 1_n & 1_n^\mathrm{T} V^{-1} 1_n \end{bmatrix} \begin{bmatrix} 1_n^\mathrm{T} V^{-1} y \\ x^\mathrm{T} V^{-1} y \end{bmatrix} \\ &= \frac{1}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \begin{bmatrix} x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} y - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} y \\ 1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} y - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} y \end{bmatrix} \; . \end{split} \tag{5}$$

Thus, the first entry of $\hat{\beta}$ is equal to:

$$\hat{\beta}_0 = \frac{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} y - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} y}{x^\mathrm{T} V^{-1} x \, 1_n^\mathrm{T} V^{-1} 1_n - 1_n^\mathrm{T} V^{-1} x \, x^\mathrm{T} V^{-1} 1_n} \; . \tag{6}$$

Moreover, the second entry of $\hat{\beta}$ is equal to (→ III/1.4.16):

$$\hat{\beta}_1 = \frac{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} y - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} y}{1_n^\mathrm{T} V^{-1} 1_n \, x^\mathrm{T} V^{-1} x - x^\mathrm{T} V^{-1} 1_n \, 1_n^\mathrm{T} V^{-1} x} \; . \tag{7}$$

■

1.4.18 Maximum likelihood estimation

Theorem: Given a simple linear regression model (→ III/1.5.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}$$

the maximum likelihood estimates (→ I/4.1.3) of β_0, β_1 and σ^2 are given by

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2), s_x^2 is the sample variance (→ I/1.11.2) of x and s_{xy} is the sample covariance (→ I/1.13.2) between x and y.

Proof: With the probability density function of the normal distribution (→ II/3.2.10) and probability under independence (→ I/1.3.6), the linear regression equation (1) implies the following likelihood function (→ I/5.1.2)

$$\begin{aligned} p(y|\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n p(y_i|\beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^n \mathcal{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right] \end{aligned} \tag{3}$$

and the log-likelihood function (→ I/4.1.2)

$$\begin{aligned} \mathrm{LL}(\beta_0, \beta_1, \sigma^2) &= \log p(y|\beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \; . \end{aligned} \tag{4}$$

The derivative of the log-likelihood function (4) with respect to β_0 is

$$\frac{\mathrm{dLL}(\beta_0, \beta_1, \sigma^2)}{\mathrm{d}\beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \tag{5}$$

and setting this derivative to zero gives the MLE for β_0:

$$\begin{aligned}
\frac{\mathrm{dLL}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{\mathrm{d}\beta_0} &= 0 \\
0 &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
0 &= \sum_{i=1}^n y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i \\
\hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \; .
\end{aligned} \tag{6}$$

The derivative of the log-likelihood function (4) at $\hat{\beta}_0$ with respect to β_1 is

$$\frac{\mathrm{dLL}(\hat{\beta}_0, \beta_1, \sigma^2)}{\mathrm{d}\beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \beta_1 x_i^2) \tag{7}$$

and setting this derivative to zero gives the MLE for β_1:

$$\begin{aligned}
\frac{\mathrm{dLL}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{\mathrm{d}\beta_1} &= 0 \\
0 &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \\
0 &= \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2) \\
0 &\overset{(6)}{=} \sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
0 &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
0 &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 n \bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^n x_i^2 \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \; .
\end{aligned} \tag{8}$$

The derivative of the log-likelihood function (4) at $(\hat{\beta}_0, \hat{\beta}_1)$ with respect to σ^2 is

$$\frac{\mathrm{dLL}(\hat{\beta}_0, \hat{\beta}_1, \sigma^2)}{\mathrm{d}\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{9}$$

and setting this derivative to zero gives the MLE for σ^2:

$$\begin{aligned}
\frac{\mathrm{dLL}(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)}{\mathrm{d}\sigma^2} &= 0 \\
0 &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
\frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \; .
\end{aligned} \tag{10}$$

Together, (6), (8) and (10) constitute the MLE for simple linear regression.

■

1.4.19 Maximum likelihood estimation

Theorem: Given a simple linear regression model (→ III/1.5.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}$$

the maximum likelihood estimates (→ I/4.1.3) of β_0, β_1 and σ^2 are given by

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2), s_x^2 is the sample variance (→ I/1.11.2) of x and s_{xy} is the sample covariance (→ I/1.13.2) between x and y.

Proof: Simple linear regression is a special case of multiple linear regression (→ III/1.4.2) with

$$X = \begin{bmatrix} 1_n & x \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{3}$$

and weighted least sqaures estimates (→ III/1.5.23) are given by

$$\begin{aligned}
\hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
\hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; .
\end{aligned} \tag{4}$$

Under independent observations, the covariance matrix is

$$V = I_n, \quad \text{such that} \quad V^{-1} = I_n \; . \tag{5}$$

Thus, we can write out the estimate of β

$$\begin{split}
\hat{\beta} &= \left(\begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} V^{-1} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} V^{-1} y \\
&= \left(\begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} \begin{bmatrix} 1_n & x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1_n^\mathrm{T} \\ x^\mathrm{T} \end{bmatrix} y
\end{split} \tag{6}$$

which is equal to the ordinary least squares solution for simple linear regression (→ III/1.4.4):

$$\begin{split}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \; .
\end{split} \tag{7}$$

Additionally, we can write out the estimate of σ^2:

$$\begin{split}
\hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\
&= \frac{1}{n} \left(y - \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \right)^\mathrm{T} \left(y - \begin{bmatrix} 1_n & x \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \right) \\
&= \frac{1}{n} \left(y - \hat{\beta}_0 - \hat{\beta}_1 x \right)^\mathrm{T} \left(y - \hat{\beta}_0 - \hat{\beta}_1 x \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \; .
\end{split} \tag{8}$$

■

1.4.20 t-test for intercept parameter

Theorem: Consider a simple linear regression model (→ III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}$$

and the parameter estimates (→ III/1.4.18)

$$\begin{split}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \; .
\end{split} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2) of the x_i and y_i, s_{xy} is the sample covariance (→ I/1.13.2) of the x_i and y_i and s_x^2 is the sample variance (→ I/1.11.2) of the x_i.
Then, the test statistic (→ I/4.3.5)

$$t_0 = \frac{\bar{y} - \hat{\beta}_1 \bar{x}}{\sqrt{\hat{\sigma}^2 \, \sigma_0}} \tag{3}$$

with σ_0 equal to the first diagonal element of the parameter covariance matrix (→ III/1.4.7)

$$\sigma_0 = \frac{x^\mathrm{T} x/n}{(n-1)\, s_x^2} \quad \text{where} \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4}$$

follows a t-distribution (→ II/3.3.1)

$$t_0 \sim \mathrm{t}(n-2) \tag{5}$$

under the null hypothesis (→ I/4.3.2) that the intercept parameter (→ III/1.4.1) is zero:

$$H_0: \; \beta_0 = 0 \; . \tag{6}$$

Proof: In multiple linear regression (→ III/1.5.1), the contrast-based t-test (→ III/1.5.27) is based on the t-statistic (→ I/4.3.5)

$$t = \frac{c^\mathrm{T} \hat{\beta}}{\sqrt{\hat{\sigma}^2 c^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} c}} \tag{7}$$

which follows a t-distribution (→ II/3.3.1) under the null hypothesis (→ I/4.3.2) that the scalar product of the contrast vector (→ III/1.5.25) and the regression coefficients (→ III/1.5.1) is zero:

$$t \sim \mathrm{t}(n-p), \quad \text{if} \quad c^\mathrm{T} \beta = 0 \; . \tag{8}$$

Since simple linear regression is a special case of multiple linear regression (→ III/1.4.2), in the present case we have the following quantities:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} , \; \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} , \; c_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} , \; X = \begin{bmatrix} 1_n & x \end{bmatrix} , \; V = I_n \; . \tag{9}$$

Thus, we have the null hypothesis

$$H_0: \; c_0^\mathrm{T} \beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^\mathrm{T} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_0 = 0 \tag{10}$$

and the contrast estimate

$$c_0^\mathrm{T} \hat{\beta} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^\mathrm{T} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \; . \tag{11}$$

Moreover, when deriving the distribution of ordinary least squares parameter estimates for simple linear regression with independent observations (→ III/1.4.7), we have identified the parameter covariance matrix as

$$(X^\mathrm{T} X)^{-1} = \frac{1}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \; . \tag{12}$$

Plugging (9), (11), (12) and (2) into (7), the test statistic becomes

$$\begin{split} t_0 &= \frac{c_0^\mathrm{T} \hat{\beta}}{\sqrt{\hat{\sigma}^2 \, c_0^\mathrm{T} (X^\mathrm{T} X)^{-1} c_0}} \\ &= \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^\mathrm{T}}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 1 & 0 \end{bmatrix} (X^\mathrm{T} X)^{-1} \begin{bmatrix} 1 & 0 \end{bmatrix}^\mathrm{T}}} \\ &= \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^\mathrm{T}}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 1 & 0 \end{bmatrix} \left(\frac{1}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 1 & 0 \end{bmatrix}^\mathrm{T}}} \\ &= \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2 \left(\frac{x^\mathrm{T} x/n}{(n-1)\, s_x^2} \right)}} \\ &= \frac{\bar{y} - \hat{\beta}_1 \bar{x}}{\sqrt{\hat{\sigma}^2 \, \sigma_0}} \; . \end{split} \tag{13}$$

Finally, because $X = \begin{bmatrix} 1_n & x \end{bmatrix}$ is an $n \times 2$ matrix, we have $p = 2$, such that from (8), it follows that

$$t_0 \sim \mathrm{t}(n-2), \quad \text{if} \quad \beta_0 = 0 \; . \tag{14}$$

■

1.4.21 t-test for slope parameter

Theorem: Consider a simple linear regression model (→ III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}$$

and the parameter estimates (→ III/1.4.18)

$$\begin{split} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \; . \end{split} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2) of the x_i and y_i, s_{xy} is the sample covariance (→ I/1.13.2) of the x_i and y_i and s_x^2 is the sample variance (→ I/1.11.2) of the x_i.

Then, the test statistic (→ I/4.3.5)

$$t_1 = \frac{s_{xy}/s_x^2}{\sqrt{\hat{\sigma}^2\, \sigma_1}} \tag{3}$$

with σ_1 equal to the first diagonal element of the parameter covariance matrix (→ III/1.4.7)

$$\sigma_1 = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{4}$$

follows a t-distribution (→ II/3.3.1)

$$t_1 \sim \mathrm{t}(n-2) \tag{5}$$

under the null hypothesis (→ I/4.3.2) that the slope parameter (→ III/1.4.1) is zero:

$$H_0: \; \beta_1 = 0 \; . \tag{6}$$

Proof: In multiple linear regression (→ III/1.5.1), the contrast-based t-test (→ III/1.5.27) is based on the t-statistic (→ I/4.3.5)

$$t = \frac{c^\mathrm{T}\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c}} \tag{7}$$

which follows a t-distribution (→ II/3.3.1) under the null hypothesis (→ I/4.3.2) that the scalar product of the contrast vector (→ III/1.5.25) and the regression coefficients (→ III/1.5.1) is zero:

$$t \sim \mathrm{t}(n-p), \quad \text{if} \quad c^\mathrm{T}\beta = 0 \; . \tag{8}$$

Since simple linear regression is a special case of multiple linear regression (→ III/1.4.2), in the present case we have the following quantities:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \; \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \; c_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; X = \begin{bmatrix} 1_n & x \end{bmatrix}, \; V = I_n \; . \tag{9}$$

Thus, we have the null hypothesis

$$H_0: \; c_1^\mathrm{T}\beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^\mathrm{T} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_1 = 0 \tag{10}$$

and the contrast estimate

$$c_1^\mathrm{T}\hat{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^\mathrm{T} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \; . \tag{11}$$

Moreover, when deriving the distribution of ordinary least squares parameter estimates for simple linear regression with independent observations (→ III/1.4.7), we have identified the parameter covariance matrix as

$$(X^\mathrm{T}X)^{-1} = \frac{1}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T}x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \; . \tag{12}$$

Plugging (9), (11), (12) and (2) into (7), the test statistic becomes

$$\begin{split} t_1 &= \frac{c_1^\mathrm{T}\hat{\beta}}{\sqrt{\hat{\sigma}^2 \, c_1^\mathrm{T} (X^\mathrm{T}X)^{-1} c_1}} \\ &= \frac{\begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^\mathrm{T}}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 0 & 1 \end{bmatrix} (X^\mathrm{T}X)^{-1} \begin{bmatrix} 0 & 1 \end{bmatrix}^\mathrm{T}}} \\ &= \frac{\begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^\mathrm{T}}{\sqrt{\hat{\sigma}^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \left(\frac{1}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T}x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 0 & 1 \end{bmatrix}^\mathrm{T}}} \\ &= \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{(n-1)\, s_x^2} \right)}} \\ &= \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \\ &= \frac{s_{xy}/s_x^2}{\sqrt{\hat{\sigma}^2 \, \sigma_1}} \; . \end{split} \tag{13}$$

Finally, because $X = \begin{bmatrix} 1_n & x \end{bmatrix}$ is an $n \times 2$ matrix, we have $p = 2$, such that from (8) it follows that

$$t_1 \sim \mathrm{t}(n-2), \quad \text{if} \quad \beta_1 = 0 \; . \tag{14}$$

■

1.4.22 F-test for model comparison

Theorem: Consider a simple linear regression model (→ III/1.4.1) with independent observations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \; , \tag{1}$$

and the parameter estimates (→ III/1.4.18)

$$\begin{split} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ \hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \; . \end{split} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means (→ I/1.10.2) of the x_i and y_i, s_{xy} is the sample covariance (→ I/1.13.2) of the x_i and y_i and s_x^2 is the sample variance (→ I/1.11.2) of the x_i.
Then, the test statistic (→ I/4.3.5)

$$F = \frac{s_{xy}^2/s_x^2}{\hat{\sigma}^2/(n-1)} \tag{3}$$

follows an F-distribution (→ II/3.8.1)

$$F \sim \mathrm{F}(1, n-2) \tag{4}$$

under the scenario that the data were generated using a model in which the slope parameter (→ III/1.4.1) is zero:

$$H_0: \; \beta_1 = 0 \; . \tag{5}$$

Proof: In multiple linear regression (→ III/1.5.1), the contrast-based F-test (→ III/1.5.28) is based on the F-statistic (→ I/4.3.5)

$$F = \hat{\beta}^\mathrm{T} C \left(\hat{\sigma}^2 C^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C \right)^{-1} C^\mathrm{T} \hat{\beta}/q \tag{6}$$

which follows an F-distribution (→ II/3.8.1) under the null hypothesis (→ I/4.3.2) that the product of the contrast matrix (→ III/1.5.26) $C \in \mathbb{R}^{p \times q}$ and the regression coefficients (→ III/1.5.1) is a zero vector:

$$F \sim \mathrm{F}(q, n-p), \quad \text{if} \quad C^\mathrm{T}\beta = 0_q = [0, \ldots, 0]^\mathrm{T} \; . \tag{7}$$

Since simple linear regression is a special case of multiple linear regression (→ III/1.4.2), we have the following quantities, if we want to compare the regression model against a model without the slope parameter:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \; \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \; C = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; X = \begin{bmatrix} 1_n & x \end{bmatrix}, \; V = I_n \; . \tag{8}$$

Thus, we have the null hypothesis

$$H_0: \; C^\mathrm{T}\beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^\mathrm{T} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_1 = 0 \tag{9}$$

and the contrast estimate

$$C^\mathrm{T}\hat{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^\mathrm{T} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \; . \tag{10}$$

Moreover, when deriving the distribution of ordinary least squares parameter estimates for simple linear regression with independent observations (→ III/1.4.7), we have identified the parameter covariance matrix as

$$(X^\mathrm{T} X)^{-1} = \frac{1}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \; . \tag{11}$$

Plugging (8), (10), (11) and (2) into (6), the test statistic becomes

$$\begin{split} F &= \hat{\beta}^\mathrm{T} C \left(\hat{\sigma}^2 C^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C \right)^{-1} C^\mathrm{T} \hat{\beta} / q \\ &= \left(\frac{s_{xy}}{s_x^2} \right) \left(\hat{\sigma}^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \left(\frac{1}{(n-1)\, s_x^2} \cdot \begin{bmatrix} x^\mathrm{T} x/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 0 & 1 \end{bmatrix}^\mathrm{T} \right)^{-1} \left(\frac{s_{xy}}{s_x^2} \right) / 1 \\ &= \frac{s_{xy}^2/(s_x^2)^2}{\hat{\sigma}^2/((n-1)\, s_x^2)} \\ &= \frac{s_{xy}^2/s_x^2}{\hat{\sigma}^2/(n-1)} \; . \end{split} \tag{12}$$

Finally, because $C = \begin{bmatrix} 0 & 1 \end{bmatrix}^\mathrm{T} \in \mathbb{R}^{2 \times 1}$ and $X = \begin{bmatrix} 1_n & x \end{bmatrix} \in \mathbb{R}^{n \times 2}$, we have $p = 2$ and $q = 1$, such that from (7) it follows that

$$F \sim \mathrm{F}(1, n-2), \quad \text{if} \quad \beta_1 = 0 \; . \tag{13}$$

■

1.4.23 Sum of residuals is zero

Theorem: In simple linear regression (→ III/1.4.1), the sum of the residuals (→ III/1.5.9) is zero when estimated using ordinary least squares (→ III/1.4.3).

Proof: The residuals are defined as the estimated error terms (→ III/1.4.1)

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{1}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are parameter estimates obtained using ordinary least squares (→ III/1.4.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \; . \tag{2}$$

With that, we can calculate the sum of the residuals:

$$\begin{split} \sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^n y_i - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i \\ &= n\bar{y} - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 n\bar{x} \\ &= 0 \; . \end{split} \tag{3}$$

Thus, the sum of the residuals (→ III/1.5.9) is zero under ordinary least squares (→ III/1.4.3), if the model (→ III/1.4.1) includes an intercept term β_0.

■

Sources:
- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties.

1.4.24 Correlation with covariate is zero

Theorem: In simple linear regression (→ III/1.4.1), the residuals (→ III/1.5.9) and the covariate (→ III/1.4.1) are uncorrelated (→ I/1.14.1) when estimated using ordinary least squares (→ III/1.4.3).

Proof: The residuals are defined as the estimated error terms (→ III/1.4.1)

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{1}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are parameter estimates obtained using ordinary least squares (→ III/1.4.3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \; . \tag{2}$$

With that, we can calculate the inner product of the covariate and the residuals vector:

$$
\begin{aligned}
\sum_{i=1}^n x_i \hat{\varepsilon}_i &= \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^n \left(x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2 \right) \\
&= \sum_{i=1}^n \left(x_i y_i - x_i (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i^2 \right) \\
&= \sum_{i=1}^n \left(x_i (y_i - \bar{y}) + \hat{\beta}_1 (\bar{x} x_i - x_i^2 \right) \\
&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) \\
&= \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} - n \bar{x} \bar{y} + n \bar{x} \bar{y} \right) - \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - 2n \bar{x} \bar{x} + n \bar{x}^2 \right) \\
&= \left(\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \right) - \hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n \bar{x}^2 \right) \\
&= \sum_{i=1}^n \left(x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x} \bar{y} \right) - \hat{\beta}_1 \sum_{i=1}^n \left(x_i^2 - 2 \bar{x} x_i + \bar{x}^2 \right) \\
&= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2 \\
&= (n-1) s_{xy} - \frac{s_{xy}}{s_x^2} (n-1) s_x^2 \\
&= (n-1) s_{xy} - (n-1) s_{xy} \\
&= 0 \; .
\end{aligned} \tag{3}
$$

Because an inner product of zero also implies zero correlation (→ I/1.14.1), this demonstrates that residuals (→ III/1.5.9) and covariate (→ III/1.4.1) values are uncorrelated under ordinary least squares (→ III/1.4.3).

■

Sources:
- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties.

1.4.25 Residual variance in terms of sample variance

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$
y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}
$$

and consider estimation using ordinary least squares (→ III/1.4.3). Then, residual variance (→ IV/1.1.1) and sample variance (→ I/1.11.2) are related to each other via the correlation coefficient (→ I/1.14.1):

$$\hat{\sigma}^2 = \left(1 - r_{xy}^2\right) s_y^2 \; . \tag{2}$$

Proof: The residual variance (→ IV/1.1.1) can be expressed in terms of the residual sum of squares (→ III/1.5.9):

$$\hat{\sigma}^2 = \frac{1}{n-1} \, \mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) \tag{3}$$

and the residual sum of squares for simple linear regression (→ III/1.4.13) is

$$\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) = (n-1) \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right) \; . \tag{4}$$

Combining (3) and (4), we obtain:

$$\begin{split} \hat{\sigma}^2 &= \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right) \\ &= \left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2} \right) s_y^2 \\ &= \left(1 - \left(\frac{s_{xy}}{s_x \, s_y} \right)^2 \right) s_y^2 \; . \end{split} \tag{5}$$

Using the relationship between correlation, covariance and standard deviation (→ I/1.14.1)

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{6}$$

which also holds for sample correlation, sample covariance (→ I/1.13.2) and sample standard deviation (→ I/1.16.1)

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y} \; , \tag{7}$$

we get the final result:

$$\hat{\sigma}^2 = \left(1 - r_{xy}^2\right) s_y^2 \; . \tag{8}$$

■

Sources:

- Penny, William (2006): "Relation to correlation"; in: *Mathematics for Brain Imaging*, ch. 1.2.3, p. 18, eq. 1.28; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Numerical_properties.

1.4.26 Correlation coefficient in terms of slope estimate

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}$$

and consider estimation using ordinary least squares (→ III/1.4.3). Then, correlation coefficient (→ I/1.14.3) and the estimated value of the slope parameter (→ III/1.4.1) are related to each other via the sample (→ I/1.11.2) standard deviations (→ I/1.16.1):

$$r_{xy} = \frac{s_x}{s_y} \, \hat{\beta}_1 \; . \tag{2}$$

Proof: The ordinary least squares estimate of the slope (→ III/1.4.3) is given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \; . \tag{3}$$

Using the relationship between covariance and correlation (→ I/1.13.7)

$$\mathrm{Cov}(X,Y) = \sigma_X \, \mathrm{Corr}(X,Y) \, \sigma_Y \tag{4}$$

which also holds for sample correlation (→ I/1.14.3) and sample covariance (→ I/1.13.2)

$$s_{xy} = s_x \, r_{xy} \, s_y \; , \tag{5}$$

we get the final result:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\
\hat{\beta}_1 &= \frac{s_x \, r_{xy} \, s_y}{s_x^2} \\
\hat{\beta}_1 &= \frac{s_y}{s_x} \, r_{xy} \\
\Leftrightarrow \quad r_{xy} &= \frac{s_x}{s_y} \, \hat{\beta}_1 \; .
\end{aligned} \tag{6}$$

■

Sources:

- Penny, William (2006): "Relation to correlation"; in: *Mathematics for Brain Imaging*, ch. 1.2.3, p. 18, eq. 1.27; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line.

1.4.27 Coefficient of determination in terms of correlation coefficient

Theorem: Assume a simple linear regression model (→ III/1.4.1) with independent observations

$$y = \beta_0 + \beta_1 x + \varepsilon, \; \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, \ldots, n \tag{1}$$

and consider estimation using ordinary least squares (→ III/1.4.3). Then, the coefficient of determination (→ IV/1.2.1) is equal to the squared correlation coefficient (→ I/1.14.3) between x and y:

$$R^2 = r_{xy}^2 \; . \tag{2}$$

Proof: The ordinary least squares estimates for simple linear regression (→ III/1.4.3) are

$$\begin{split} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \; . \end{split} \tag{3}$$

The coefficient of determination (→ IV/1.2.1) R^2 is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data. This can be quantified as the ratio of explained sum of squares (→ III/1.5.8) to total sum of squares (→ III/1.5.7):

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \; . \tag{4}$$

Using the explained and total sum of squares for simple linear regression (→ III/1.4.13), we have:

$$\begin{split} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \; . \end{split} \tag{5}$$

By applying (3), we can further develop the coefficient of determination:

$$\begin{split} R^2 &= \frac{\sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n \left(\hat{\beta}_1 (x_i - \bar{x}) \right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \, \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \, \frac{s_x^2}{s_y^2} \\ &= \left(\frac{s_x}{s_y} \, \hat{\beta}_1 \right)^2 \; . \end{split} \tag{6}$$

Using the relationship between correlation coefficient and slope estimate (→ III/1.4.26), we conclude:

$$R^2 = \left(\frac{s_x}{s_y} \, \hat{\beta}_1 \right)^2 = r_{xy}^2 \; . \tag{7}$$

■

Sources:

- Wikipedia (2021): "Simple linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Simple_linear_regression#Fitting_the_regression_line.
- Wikipedia (2021): "Coefficient of determination"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Coefficient_of_determination#As_squared_correlation_coefficient.
- Wikipedia (2021): "Correlation"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-10-27; URL: https://en.wikipedia.org/wiki/Correlation#Sample_correlation_coefficient.

1.5 Multiple linear regression

1.5.1 Definition

Definition: Let y be an $n \times 1$ vector and let X be an $n \times p$ matrix.
Then, a statement asserting a linear combination of X into y

$$y = X\beta + \varepsilon \; , \tag{1}$$

together with a statement asserting a normal distribution (→ II/4.1.1) for ε

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{2}$$

is called a univariate linear regression model or simply, "multiple linear regression".

- y is called "measured data", "dependent variable" or "measurements";
- X is called "design matrix", "set of independent variables" or "predictors";
- V is called "covariance matrix" or "covariance structure";
- β are called "regression coefficients" or "weights";
- ε is called "noise", "errors" or "error terms";
- σ^2 is called "noise variance" or "error variance";
- n is the number of observations;
- p is the number of predictors.

Alternatively, the linear combination may also be written as

$$y = \sum_{i=1}^{p} \beta_i x_i + \varepsilon \tag{3}$$

or, when the model includes an intercept term, as

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \tag{4}$$

which is equivalent to adding a constant regressor $x_0 = 1_n$ to the design matrix X.
When the covariance structure V is equal to the $n \times n$ identity matrix, this is called multiple linear regression with independent and identically distributed (i.i.d.) observations:

$$V = I_n \quad \Rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \Rightarrow \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{5}$$

Otherwise, it is called multiple linear regression with correlated observations.

Sources:

- Wikipedia (2020): "Linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_linear_regression.

1.5.2 Special case of general linear model

Theorem: Multiple linear regression (→ III/1.5.1) is a special case of the general linear model (→ III/1.5.1) with number of measurements $v = 1$, such that data matrix Y, regression coefficients B, noise matrix E and noise covariance Σ equate as

$$Y = y, \quad B = \beta, \quad E = \varepsilon \quad \text{and} \quad \Sigma = \sigma^2 \tag{1}$$

where y, β, ε and σ^2 are the data vector, regression coefficients, noise vector and noise variance from multiple linear regression (→ III/1.5.1).

Proof: The linear regression model with correlated errors (→ III/1.5.1) is given by:

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{2}$$

Because ε is an $n \times 1$ vector and σ^2 is scalar, we have the following identities:

$$\begin{aligned} \mathrm{vec}(\varepsilon) &= \varepsilon \\ \sigma^2 \otimes V &= \sigma^2 V \; . \end{aligned} \tag{3}$$

Thus, using the relationship between multivariate normal and matrix normal distribution (→ II/5.1.2), equation (2) can also be written as

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{MN}(0, V, \sigma^2) \; . \tag{4}$$

Comparing with the general linear model with correlated observations (→ III/2.1.1)

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; , \tag{5}$$

we finally note the equivalences given in equation (1).

■

Sources:
- Wikipedia (2022): "General linear model"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-07-21; URL: https://en.wikipedia.org/wiki/General_linear_model#Comparison_to_multiple_linear_regression.

1.5.3 Ordinary least squares

Theorem: Given a linear regression model (→ III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) are given by

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{2}$$

Proof: Let $\hat{\beta}$ be the ordinary least squares (OLS) solution and let $\hat{\varepsilon} = y - X\hat{\beta}$ be the resulting vector of residuals. Then, this vector must be orthogonal to the design matrix,

$$X^{\mathrm{T}}\hat{\varepsilon} = 0 \; , \tag{3}$$

because if it wasn't, there would be another solution $\tilde{\beta}$ giving another vector $\tilde{\varepsilon}$ with a smaller residual sum of squares. From (3), the OLS formula can be directly derived:

$$\begin{split} X^{\mathrm{T}}\hat{\varepsilon} &= 0 \\ X^{\mathrm{T}}\left(y - X\hat{\beta}\right) &= 0 \\ X^{\mathrm{T}}y - X^{\mathrm{T}}X\hat{\beta} &= 0 \\ X^{\mathrm{T}}X\hat{\beta} &= X^{\mathrm{T}}y \\ \hat{\beta} &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \; . \end{split} \tag{4}$$

■

Sources:

- Stephan, Klaas Enno (2010): "The General Linear Model (GLM)"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slides 10/11; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.

1.5.4 Ordinary least squares

Theorem: Given a linear regression model (→ III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) are given by

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \; . \tag{2}$$

Proof: The residual sum of squares (→ III/1.5.9) is defined as

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^{\mathrm{T}}\varepsilon = (y - X\beta)^{\mathrm{T}}(y - X\beta) \tag{3}$$

which can be developed into

$$\begin{split} \mathrm{RSS}(\beta) &= y^{\mathrm{T}}y - y^{\mathrm{T}}X\beta - \beta^{\mathrm{T}}X^{\mathrm{T}}y + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta \\ &= y^{\mathrm{T}}y - 2\beta^{\mathrm{T}}X^{\mathrm{T}}y + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta \; . \end{split} \tag{4}$$

The derivative of $\mathrm{RSS}(\beta)$ with respect to β is

$$\frac{\mathrm{dRSS}(\beta)}{\mathrm{d}\beta} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\beta \tag{5}$$

and setting this deriative to zero, we obtain:

$$
\begin{aligned}
\frac{\mathrm{dRSS}(\hat{\beta})}{\mathrm{d}\beta} &= 0 \\
0 &= -2X^\mathrm{T}y + 2X^\mathrm{T}X\hat{\beta} \\
X^\mathrm{T}X\hat{\beta} &= X^\mathrm{T}y \\
\hat{\beta} &= (X^\mathrm{T}X)^{-1}X^\mathrm{T}y \; .
\end{aligned} \tag{6}
$$

Since the quadratic form $y^\mathrm{T}y$ in (4) is positive, $\hat{\beta}$ minimizes $\mathrm{RSS}(\beta)$.

■

Sources:

- Wikipedia (2020): "Proofs involving ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-03; URL: https://en.wikipedia.org/wiki/Proofs_involving_ordinary_least_squares#Least_squares_estimator_for_%CE%B2.
- ad (2015): "Derivation of the Least Squares Estimator for Beta in Matrix Notation"; in: *Economic Theory Blog*, retrieved on 2021-05-27; URL: https://economictheoryblog.com/2015/02/19/ols_estimator/.

1.5.5 Ordinary least squares

Theorem: Given a linear regression model (→ III/1.5.1) with independent observations

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}
$$

the parameters minimizing the residual sum of squares (→ III/1.5.9) are given by

$$
\hat{\beta} = (X^\mathrm{T}X)^{-1}X^\mathrm{T}y \; . \tag{2}
$$

Proof: We consider the sum of squared differences between y and $X\beta$:

$$
\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^\mathrm{T}\varepsilon = (y - X\beta)^\mathrm{T}(y - X\beta) \; . \tag{3}
$$

First, we note that the residual vector $\hat{\varepsilon}$ implied by the ordinary least squares solution $\hat{\beta}$ is orthogonal to the columns of the design matrix, such that the result of their multiplication is the p-dimensional zero vector (where $X \in \mathbb{R}^{n \times p}$):

$$
\begin{aligned}
X^\mathrm{T}(y - X\hat{\beta}) &= X^\mathrm{T}y - X\hat{\beta} \\
&= X^\mathrm{T}y - X^\mathrm{T}X(X^\mathrm{T}X)^{-1}X^\mathrm{T}y \\
&= X^\mathrm{T}y - X^\mathrm{T}y \\
&= 0_p \; .
\end{aligned} \tag{4}
$$

Second, since $X^\mathrm{T}X$ is a positive semi-definite matrix (→ I/1.13.13), the following product is non-negative for each p-dimensional real vector z:

$$
z^\mathrm{T}X^\mathrm{T}Xz \geq 0 \quad \text{for each} \quad z \in \mathbb{R}^p \; . \tag{5}
$$

We continue developping the sum of squared differences from (3):

$$
\begin{split}
(y - X\beta)^\mathrm{T}(y - X\beta) &= (y - X\hat{\beta} + X\hat{\beta} - X\beta)^\mathrm{T}(y - X\hat{\beta} + X\hat{\beta} - X\beta) \\
&= \left((y - X\hat{\beta}) + X(\hat{\beta} - \beta) \right)^\mathrm{T} \left((y - X\hat{\beta}) + X(\hat{\beta} - \beta) \right) \\
&= (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) + (y - X\hat{\beta})^\mathrm{T} X(\hat{\beta} - \beta) + (\hat{\beta} - \beta)^\mathrm{T} X^\mathrm{T}(y - X\hat{\beta}) + (\hat{\beta} - \beta)^\mathrm{T} X^\mathrm{T} X(\\
&\overset{(4)}{=} (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) + 0_p^\mathrm{T}(\hat{\beta} - \beta) + (\hat{\beta} - \beta)^\mathrm{T} 0_p + (\hat{\beta} - \beta)^\mathrm{T} X^\mathrm{T} X(\hat{\beta} - \beta) \\
&= (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) + (\hat{\beta} - \beta)^\mathrm{T} X^\mathrm{T} X(\hat{\beta} - \beta) \; .
\end{split}
\tag{6}
$$

By virtue of (5), the second term on the right-hand side must be non-zero:

$$
(\hat{\beta} - \beta)^\mathrm{T} X^\mathrm{T} X(\hat{\beta} - \beta) \geq 0 \; . \tag{7}
$$

Thus, the residual sum of squares must be greater than or equal to the first term

$$
(y - X\beta)^\mathrm{T}(y - X\beta) \geq (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \tag{8}
$$

and its minimum value is reached when the the second term is zero:

$$
\begin{split}
(\hat{\beta} - \beta)^\mathrm{T} X^\mathrm{T} X(\hat{\beta} - \beta) &= 0 \\
\Leftrightarrow \quad (\hat{\beta} - \beta) &= 0 \\
\Leftrightarrow \quad \beta &= \hat{\beta} \; .
\end{split}
\tag{9}
$$

Thus, the residual sum of squares is minimized when $\beta = \hat{\beta}$:

$$
\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y = \operatorname*{arg\,min}_\beta \, (y - X\beta)^\mathrm{T}(y - X\beta) \; . \tag{10}
$$

■

Sources:

- Ostwald, Dirk (2023): “Parameterschätzung”; in: *Allgemeines Lineares Modell*, Einheit (6), Folien 10-12; URL: https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemein Lineares+Modell/6_Parametersch%C3%A4tzung.pdf.

1.5.6 Ordinary least squares for two regressors

Theorem: Consider a linear regression model (→ III/1.5.1) in which the design matrix (→ III/1.5.1) has two columns:

$$
y = X\beta + \varepsilon \quad \text{where} \quad y \in \mathbb{R}^{n \times 1} \quad \text{and} \quad X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \in \mathbb{R}^{n \times 2} \; . \tag{1}
$$

Then,
1) the ordinary least squares (→ III/1.5.3) estimates for β_1 and β_2 are given by

$$
\hat{\beta}_1 = \frac{x_2^\mathrm{T} x_2 x_1^\mathrm{T} y - x_1^\mathrm{T} x_2 x_2^\mathrm{T} y}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2 - x_1^\mathrm{T} x_2 x_2^\mathrm{T} x_1} \quad \text{and} \quad \hat{\beta}_2 = \frac{x_1^\mathrm{T} x_1 x_2^\mathrm{T} y - x_2^\mathrm{T} x_1 x_1^\mathrm{T} y}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2 - x_1^\mathrm{T} x_2 x_2^\mathrm{T} x_1} \tag{2}
$$

2) and, if the two regressors are orthogonal to each other, they simplify to

$$\hat{\beta}_1 = \frac{x_1^\mathrm{T} y}{x_1^\mathrm{T} x_1} \quad \text{and} \quad \hat{\beta}_2 = \frac{x_2^\mathrm{T} y}{x_2^\mathrm{T} x_2}, \quad \text{if} \quad x_1 \perp x_2 \; . \tag{3}$$

Proof: The model in (1) is a special case of multiple linear regression (→ III/1.5.1) and the ordinary least squares solution for multiple linear regression (→ III/1.5.3) is:

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{4}$$

1) Plugging $X = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ into this equation, we obtain:

$$\begin{split} \hat{\beta} &= \left(\begin{bmatrix} x_1^\mathrm{T} \\ x_2^\mathrm{T} \end{bmatrix} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} x_1^\mathrm{T} \\ x_2^\mathrm{T} \end{bmatrix} y \\ &= \begin{pmatrix} x_1^\mathrm{T} x_1 & x_1^\mathrm{T} x_2 \\ x_2^\mathrm{T} x_1 & x_2^\mathrm{T} x_2 \end{pmatrix}^{-1} \begin{pmatrix} x_1^\mathrm{T} y \\ x_2^\mathrm{T} y \end{pmatrix} \; . \end{split} \tag{5}$$

Using the inverse of of a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \; , \tag{6}$$

this can be further developped into

$$\begin{split} \hat{\beta} &= \frac{1}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2 - x_1^\mathrm{T} x_2 x_2^\mathrm{T} x_1} \begin{pmatrix} x_2^\mathrm{T} x_2 & -x_1^\mathrm{T} x_2 \\ -x_2^\mathrm{T} x_1 & x_1^\mathrm{T} x_1 \end{pmatrix} \begin{pmatrix} x_1^\mathrm{T} y \\ x_2^\mathrm{T} y \end{pmatrix} \\ &= \frac{1}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2 - x_1^\mathrm{T} x_2 x_2^\mathrm{T} x_1} \begin{pmatrix} x_2^\mathrm{T} x_2 x_1^\mathrm{T} y - x_1^\mathrm{T} x_2 x_2^\mathrm{T} y \\ x_1^\mathrm{T} x_1 x_2^\mathrm{T} y - x_2^\mathrm{T} x_1 x_1^\mathrm{T} y \end{pmatrix} \end{split} \tag{7}$$

which can also be written as

$$\begin{split} \hat{\beta}_1 &= \frac{x_2^\mathrm{T} x_2 x_1^\mathrm{T} y - x_1^\mathrm{T} x_2 x_2^\mathrm{T} y}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2 - x_1^\mathrm{T} x_2 x_2^\mathrm{T} x_1} \\ \hat{\beta}_2 &= \frac{x_1^\mathrm{T} x_1 x_2^\mathrm{T} y - x_2^\mathrm{T} x_1 x_1^\mathrm{T} y}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2 - x_1^\mathrm{T} x_2 x_2^\mathrm{T} x_1} \; . \end{split} \tag{8}$$

2) If two regressors are orthogonal to each other, this means that the inner product of the corresponding vectors is zero:

$$x_1 \perp x_2 \quad \Leftrightarrow \quad x_1^\mathrm{T} x_2 = x_2^\mathrm{T} x_1 = 0 \; . \tag{9}$$

Applying this to equation (8), we obtain:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{x_2^\mathrm{T} x_2 x_1^\mathrm{T} y}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2} = \frac{x_1^\mathrm{T} y}{x_1^\mathrm{T} x_1} \\
\hat{\beta}_2 &= \frac{x_1^\mathrm{T} x_1 x_2^\mathrm{T} y}{x_1^\mathrm{T} x_1 x_2^\mathrm{T} x_2} = \frac{x_2^\mathrm{T} y}{x_2^\mathrm{T} x_2} \; .
\end{aligned} \tag{10}
$$

■

1.5.7 Total sum of squares

Definition: Let there be a multiple linear regression with independent observations (→ III/1.5.1) using measured data y and design matrix X:

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{1}
$$

Then, the total sum of squares (TSS) is defined as the sum of squared deviations of the measured signal from the average signal:

$$
\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{2}
$$

Sources:
- Wikipedia (2020): "Total sum of squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Total_sum_of_squares.

1.5.8 Explained sum of squares

Definition: Let there be a multiple linear regression with independent observations (→ III/1.5.1) using measured data y and design matrix X:

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{1}
$$

Then, the explained sum of squares (ESS) is defined as the sum of squared deviations of the fitted signal from the average signal:

$$
\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2}
$$

with estimated regression coefficients $\hat{\beta}$, e.g. obtained via ordinary least squares (→ III/1.5.3).

Sources:
- Wikipedia (2020): "Explained sum of squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Explained_sum_of_squares.

1.5.9 Residual sum of squares

Definition: Let there be a multiple linear regression with independent observations (→ III/1.5.1) using measured data y and design matrix X:

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{1}$$

Then, the residual sum of squares (RSS) is defined as the sum of squared deviations of the measured signal from the fitted signal:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \tag{2}$$

with estimated regression coefficients $\hat{\beta}$, e.g. obtained via ordinary least squares (→ III/1.5.3).

Sources:
- Wikipedia (2020): "Residual sum of squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/Residual_sum_of_squares.

1.5.10 Total, explained and residual sum of squares

Theorem: Assume a linear regression model (→ III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

and let X contain a constant regressor 1_n modelling the intercept term. Then, it holds that

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS} \tag{2}$$

where TSS is the total sum of squares (→ III/1.5.7), ESS is the explained sum of squares (→ III/1.5.8) and RSS is the residual sum of squares (→ III/1.5.9).

Proof: The total sum of squares (→ III/1.5.7) is given by

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3}$$

where $\bar{y}$ is the mean across all y_i. The TSS can be rewritten as

$$
\begin{aligned}
\mathrm{TSS} &= \sum_{i=1}^{n} (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \right)^2 \\
&= \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i \right)^2 \\
&= \sum_{i=1}^{n} \left((\hat{y}_i - \bar{y})^2 + 2\,\hat{\varepsilon}_i (\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i^2 \right) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^{n} \hat{\varepsilon}_i (\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^{n} \hat{\varepsilon}_i (x_i \hat{\beta} - \bar{y}) \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \sum_{i=1}^{n} \hat{\varepsilon}_i \left(\sum_{j=1}^{p} x_{ij} \hat{\beta}_j \right) - 2 \sum_{i=1}^{n} \hat{\varepsilon}_i \, \bar{y} \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2 \sum_{j=1}^{p} \hat{\beta}_j \sum_{i=1}^{n} \hat{\varepsilon}_i x_{ij} - 2\bar{y} \sum_{i=1}^{n} \hat{\varepsilon}_i
\end{aligned} \tag{4}
$$

The fact that the design matrix includes a constant regressor ensures that

$$
\sum_{i=1}^{n} \hat{\varepsilon}_i = \hat{\varepsilon}^{\mathrm{T}} 1_n = 0 \tag{5}
$$

and because the residuals are orthogonal to the design matrix (→ III/1.5.3), we have

$$
\sum_{i=1}^{n} \hat{\varepsilon}_i x_{ij} = \hat{\varepsilon}^{\mathrm{T}} x_j = 0 \; . \tag{6}
$$

Applying (5) and (6) to (4), this becomes

$$
\mathrm{TSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{7}
$$

and, with the definitions of explained (→ III/1.5.8) and residual sum of squares (→ III/1.5.9), it is

$$
\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS} \; . \tag{8}
$$

■

Sources:

- Wikipedia (2020): “Partition of sums of squares”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-09; URL: https://en.wikipedia.org/wiki/Partition_of_sums_of_squares#Partitioning_the_sum_of_squares_in_linear_regression.

1.5.11 Estimation matrix

Definition: In multiple linear regression (→ III/1.5.1), the estimation matrix is the matrix E that results in ordinary least squares (→ III/1.5.3) or weighted least squares (→ III/1.5.21) parameter estimates when right-multiplied with the measured data:

$$Ey = \hat{\beta} \; . \tag{1}$$

1.5.12 Projection matrix

Definition: In multiple linear regression (→ III/1.5.1), the projection matrix is the matrix P that results in the fitted signal explained by estimated parameters (→ III/1.5.11) when right-multiplied with the measured data:

$$Py = \hat{y} = X\hat{\beta} \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Projection matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Overview.

1.5.13 Residual-forming matrix

Definition: In multiple linear regression (→ III/1.5.1), the residual-forming matrix is the matrix R that results in the vector of residuals left over by estimated parameters (→ III/1.5.11) when right-multiplied with the measured data:

$$Ry = \hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta} \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Projection matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Properties.

1.5.14 Estimation, projection and residual-forming matrix

Theorem: Assume a linear regression model (→ III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

and consider estimation using ordinary least squares (→ III/1.5.3). Then, the estimated parameters, fitted signal and residuals are given by

$$\begin{aligned} \hat{\beta} &= Ey \\ \hat{y} &= Py \\ \hat{\varepsilon} &= Ry \end{aligned} \tag{2}$$

where

$$
\begin{aligned}
E &= (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \\
P &= X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \\
R &= I_n - X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}}
\end{aligned} \tag{3}
$$

are the estimation matrix (→ III/1.5.11), projection matrix (→ III/1.5.12) and residual-forming matrix (→ III/1.5.13) and n is the number of observations.

Proof:
1) Ordinary least squares parameter estimates of β are defined as minimizing the residual sum of squares (→ III/1.5.9)

$$
\hat{\beta} = \operatorname*{arg\,min}_{\beta} \left[(y - X\beta)^{\mathrm{T}} (y - X\beta) \right] \tag{4}
$$

and the solution to this (→ III/1.5.3) is given by

$$
\begin{aligned}
\hat{\beta} &= (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \\
&\overset{(3)}{=} E y \; .
\end{aligned} \tag{5}
$$

2) The fitted signal is given by multiplying the design matrix with the estimated regression coefficients

$$
\hat{y} = X \hat{\beta} \tag{6}
$$

and using (5), this becomes

$$
\begin{aligned}
\hat{y} &= X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \\
&\overset{(3)}{=} P y \; .
\end{aligned} \tag{7}
$$

3) The residuals of the model are calculated by subtracting the fitted signal from the measured signal

$$
\hat{\varepsilon} = y - \hat{y} \tag{8}
$$

and using (7), this becomes

$$
\begin{aligned}
\hat{\varepsilon} &= y - X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \\
&= (I_n - X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}}) y \\
&\overset{(3)}{=} R y \; .
\end{aligned} \tag{9}
$$

■

Sources:

• Stephan, Klaas Enno (2010): "The General Linear Model (GLM)"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slide 10; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.

1.5.15 Symmetry of projection and residual-forming matrix

Theorem: The projection matrix (→ III/1.5.12) and the residual-forming matrix (→ III/1.5.13) are symmetric:

$$
\begin{aligned}
P^{\mathrm{T}} &= P \\
R^{\mathrm{T}} &= R \; .
\end{aligned} \tag{1}
$$

Proof: Let X be the design matrix from the linear regression model (→ III/1.5.1). Then, the matrix $X^{\mathrm{T}}X$ is symmetric, because

$$
(X^{\mathrm{T}}X)^{\mathrm{T}} = X^{\mathrm{T}}X^{\mathrm{T}^{\mathrm{T}}} = X^{\mathrm{T}}X \; . \tag{2}
$$

Thus, the inverse of $X^{\mathrm{T}}X$ is also symmetric, i.e.

$$
\left((X^{\mathrm{T}}X)^{-1}\right)^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1} \; . \tag{3}
$$

1) The projection matrix for ordinary least squares is given by (→ III/1.5.14)

$$
P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \; , \tag{4}
$$

such that

$$
\begin{aligned}
P^{\mathrm{T}} &= \left(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\right)^{\mathrm{T}} \\
&= X^{\mathrm{T}^{\mathrm{T}}} \left((X^{\mathrm{T}}X)^{-1}\right)^{\mathrm{T}} X^{\mathrm{T}} \\
&= X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \\
&\overset{(4)}{=} P \; .
\end{aligned} \tag{5}
$$

2) The residual-forming matrix for ordinary least squares is given by (→ III/1.5.14)

$$
R = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = I_n - P \; , \tag{6}
$$

such that

$$
\begin{aligned}
R^{\mathrm{T}} &= (I_n - P)^{\mathrm{T}} \\
&= I_n^{\mathrm{T}} - P^{\mathrm{T}} \\
&\overset{(5)}{=} I_n - P \\
&\overset{(6)}{=} R \; .
\end{aligned} \tag{7}
$$

■

Sources:

- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-12-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Properties.

1.5.16 Idempotence of projection and residual-forming matrix

Theorem: The projection matrix (→ III/1.5.12) and the residual-forming matrix (→ III/1.5.13) are idempotent:

$$
\begin{aligned}
P^2 &= P \\
R^2 &= R \; .
\end{aligned} \tag{1}
$$

Proof:
1) The projection matrix for ordinary least squares is given by (→ III/1.5.14)

$$
P = X(X^\mathrm{T} X)^{-1} X^\mathrm{T} \; , \tag{2}
$$

such that

$$
\begin{aligned}
P^2 &= X(X^\mathrm{T} X)^{-1} X^\mathrm{T} X (X^\mathrm{T} X)^{-1} X^\mathrm{T} \\
&= X(X^\mathrm{T} X)^{-1} X^\mathrm{T} \\
&\overset{(2)}{=} P \; .
\end{aligned} \tag{3}
$$

2) The residual-forming matrix for ordinary least squares is given by (→ III/1.5.14)

$$
R = I_n - X(X^\mathrm{T} X)^{-1} X^\mathrm{T} = I_n - P \; , \tag{4}
$$

such that

$$
\begin{aligned}
R^2 &= (I_n - P)(I_n - P) \\
&= I_n - P - P + P^2 \\
&\overset{(3)}{=} I_n - 2P + P \\
&= I_n - P \\
&\overset{(4)}{=} R \; .
\end{aligned} \tag{5}
$$

■

Sources:
- Wikipedia (2020): “Projection matrix”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Properties.

1.5.17 Independence of estimated parameters and residuals

Theorem: Assume a linear regression model (→ III/1.5.1) with correlated observations

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}
$$

and consider estimation using weighted least squares (→ III/1.5.21). Then, the estimated parameters and the vector of residuals (→ III/1.5.19) are independent from each other:

$$
\begin{aligned}
&\hat{\beta} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \quad \text{and} \\
&\hat{\varepsilon} = y - X\hat{\beta} \quad \text{ind.}
\end{aligned} \tag{2}
$$

Proof: Equation (1) implies the following distribution (→ III/1.5.19) for the random vector (→ I/1.2.3) y:

$$
\begin{aligned}
y &\sim \mathcal{N}\left(X\beta, \sigma^2 V\right) \\
&\sim \mathcal{N}\left(X\beta, \Sigma\right) \\
\text{with} \quad \Sigma &= \sigma^2 V \; .
\end{aligned} \tag{3}
$$

Note that the estimated parameters and residuals can be written as projections from the same random vector (→ III/1.5.14) y:

$$
\begin{aligned}
\hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
&= Ay \\
\text{with} \quad A &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
\hat{\varepsilon} &= y - X\hat{\beta} \\
&= (I_n - X(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1}) y \\
&= By \\
\text{with} \quad B &= (I_n - X(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1}) \; .
\end{aligned} \tag{5}
$$

Two projections AZ and BZ from the same multivariate normal (→ II/4.1.1) random vector (→ I/1.2.3) $Z \sim \mathcal{N}(\mu, \Sigma)$ are independent, if and only if the following condition holds (→ II/4.1.16):

$$
A \Sigma B^\mathrm{T} = 0 \; . \tag{6}
$$

Combining (3), (4) and (5), we check whether this is fulfilled in the present case:

$$
\begin{aligned}
A \Sigma B^\mathrm{T} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} (\sigma^2 V) (I_n - X(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1})^\mathrm{T} \\
&= \sigma^2 \left[(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} V - (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} V V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} \right] \\
&= \sigma^2 \left[(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} - (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} \right] \\
&= \sigma^2 \cdot 0_{pn} \\
&= 0 \; .
\end{aligned} \tag{7}
$$

This demonstrates that $\hat{\beta}$ and $\hat{\varepsilon}$ – and likewise, all pairs of terms separately derived (→ III/1.5.27) from $\hat{\beta}$ and $\hat{\varepsilon}$ – are statistically independent (→ I/1.3.6).

■

Sources:

- jld (2018): "Understanding t-test for linear regression"; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: https://stats.stackexchange.com/a/344008.

1.5.18 Distribution of OLS estimates, signal and residuals

Theorem: Assume a linear regression model (→ III/1.5.1) with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

and consider estimation using ordinary least squares (→ III/1.5.3). Then, the estimated parameters, fitted signal and residuals are distributed as

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}\left(\beta, \sigma^2 (X^\mathrm{T} X)^{-1}\right) \\ \hat{y} &\sim \mathcal{N}\left(X\beta, \sigma^2 P\right) \\ \hat{\varepsilon} &\sim \mathcal{N}\left(0, \sigma^2 (I_n - P)\right) \end{aligned} \tag{2}$$

where P is the projection matrix (→ III/1.5.12) for ordinary least squares (→ III/1.5.3)

$$P = X(X^\mathrm{T} X)^{-1} X^\mathrm{T} \; . \tag{3}$$

Proof: We will use the linear transformation theorem for the multivariate normal distribution (→ II/4.1.13):

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A \Sigma A^\mathrm{T}) \; . \tag{4}$$

The distributional assumption in (1) is equivalent to (→ II/4.1.16):

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \; . \tag{5}$$

Applying (4) to (5), the measured data are distributed as

$$y \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right) \; . \tag{6}$$

1) The parameter estimates from ordinary least sqaures (→ III/1.5.3) are given by

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \tag{7}$$

and thus, by applying (4) to (7), they are distributed as

$$\begin{aligned} \hat{\beta} &\sim \mathcal{N}\left(\left[(X^\mathrm{T} X)^{-1} X^\mathrm{T}\right] X\beta, \, \sigma^2 \left[(X^\mathrm{T} X)^{-1} X^\mathrm{T}\right] I_n \left[X (X^\mathrm{T} X)^{-1}\right]\right) \\ &\sim \mathcal{N}\left(\beta, \, \sigma^2 (X^\mathrm{T} X)^{-1}\right) \; . \end{aligned} \tag{8}$$

2) The fitted signal in multiple linear regression (→ III/1.5.14) is given by

$$\hat{y} = X\hat{\beta} = X(X^\mathrm{T} X)^{-1} X^\mathrm{T} y = Py \tag{9}$$

and thus, by applying (4) to (9), they are distributed as

$$\begin{aligned} \hat{y} &\sim \mathcal{N}\left(X\beta, \, \sigma^2 X (X^\mathrm{T} X)^{-1} X^\mathrm{T}\right) \\ &\sim \mathcal{N}\left(X\beta, \, \sigma^2 P\right) \; . \end{aligned} \tag{10}$$

3) The residuals of the linear regression model (→ III/1.5.14) are given by

$$\hat{\varepsilon} = y - X\hat{\beta} = \left(I_n - X(X^\mathrm{T} X)^{-1} X^\mathrm{T} \right) y = \left(I_n - P \right) y \tag{11}$$

and thus, by applying (4) to (11), they are distributed as

$$\begin{split} \hat{\varepsilon} &\sim \mathcal{N}\left(\left[I_n - X(X^\mathrm{T} X)^{-1} X^\mathrm{T} \right] X\beta, \, \sigma^2 \left[I_n - P \right] I_n \left[I_n - P \right]^\mathrm{T} \right) \\ &\sim \mathcal{N}\left(X\beta - X\beta, \, \sigma^2 \left[I_n - P \right] \left[I_n - P \right]^\mathrm{T} \right) \; . \end{split} \tag{12}$$

Because the residual-forming matrix (→ III/1.5.13) is symmetric (→ III/1.5.15) and idempotent (→ III/1.5.16), this becomes:

$$\hat{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2 (I_n - P) \right) \; . \tag{13}$$

■

Sources:
- Koch, Karl-Rudolf (2007): "Linear Model"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 4, eqs. 4.2, 4.30; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.
- Penny, William (2006): "Multiple Regression"; in: *Mathematics for Brain Imaging*, ch. 1.5, pp. 39-41, eqs. 1.106-1.110; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.

1.5.19 Distribution of WLS estimates, signal and residuals

Theorem: Assume a linear regression model (→ III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and consider estimation using weighted least squares (→ III/1.5.21). Then, the estimated parameters, fitted signal and residuals are distributed as

$$\begin{split} \hat{\beta} &\sim \mathcal{N}\left(\beta, \sigma^2 (X^\mathrm{T} V^{-1} X)^{-1} \right) \\ \hat{y} &\sim \mathcal{N}\left(X\beta, \sigma^2 (PV) \right) \\ \hat{\varepsilon} &\sim \mathcal{N}\left(0, \sigma^2 (I_n - P) V \right) \end{split} \tag{2}$$

where P is the projection matrix (→ III/1.5.12) for weighted least squares (→ III/1.5.21)

$$P = X(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} \; . \tag{3}$$

Proof: We will use the linear transformation theorem for the multivariate normal distribution (→ II/4.1.13):

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathrm{T}) \; . \tag{4}$$

Applying (4) to (1), the measured data are distributed as

$$y \sim \mathcal{N}\left(X\beta, \sigma^2 V \right) \; . \tag{5}$$

1) The parameter estimates from weighted least sqaures (→ III/1.5.21) are given by

$$\hat{\beta} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \tag{6}$$

and thus, by applying (4) to (6), they are distributed as

$$\begin{split}
\hat{\beta} &\sim \mathcal{N}\left(\left[(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} \right] X\beta, \, \sigma^2 \left[(X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} \right] V \left[V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} \right] \right) \\
&\sim \mathcal{N}\left(\beta, \, \sigma^2 (X^\mathrm{T} V^{-1} X)^{-1} \right) \; .
\end{split} \tag{7}$$

2) The fitted signal in multiple linear regression (→ III/1.5.14) is given by

$$\hat{y} = X\hat{\beta} = X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y = Py \tag{8}$$

and thus, by applying (4) to (8), they are distributed as

$$\begin{split}
\hat{y} &\sim \mathcal{N}\left(X\beta, \, \sigma^2 X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} \right) \\
&\sim \mathcal{N}\left(X\beta, \, \sigma^2 (PV) \right) \; .
\end{split} \tag{9}$$

3) The residuals of the linear regression model (→ III/1.5.14) are given by

$$\hat{\varepsilon} = y - X\hat{\beta} = \left(I_n - X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} \right) y = \left(I_n - P \right) y \tag{10}$$

and thus, by applying (4) to (10), they are distributed as

$$\begin{split}
\hat{\varepsilon} &\sim \mathcal{N}\left(\left[I_n - X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} \right] X\beta, \, \sigma^2 \left[I_n - P \right] V \left[I_n - P \right]^\mathrm{T} \right) \\
&\sim \mathcal{N}\left(X\beta - X\beta, \, \sigma^2 \left[V - V P^\mathrm{T} - PV + PVP^\mathrm{T} \right] \right) \\
&\sim \mathcal{N}\left(0, \, \sigma^2 \left[V - V V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} - X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} V + PVP^\mathrm{T} \right] \right) \\
&\sim \mathcal{N}\left(0, \, \sigma^2 \left[V - 2PV + X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} V V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} \right] \right) \\
&\sim \mathcal{N}\left(0, \, \sigma^2 \left[V - 2PV + PV \right] \right) \\
&\sim \mathcal{N}\left(0, \, \sigma^2 \left[V - PV \right] \right) \\
&\sim \mathcal{N}\left(0, \, \sigma^2 \left[I_n - P \right] V \right) \; .
\end{split} \tag{11}$$

■

Sources:

- Koch, Karl-Rudolf (2007): "Linear Model"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 4, eqs. 4.2, 4.30; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.
- Penny, William (2006): "Multiple Regression"; in: *Mathematics for Brain Imaging*, ch. 1.5, pp. 39-41, eqs. 1.106-1.110; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, eq. A.10; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.
- Soch J, Meyer AP, Allefeld C, Haynes JD (2017): "How to improve parameter estimates in GLM-based fMRI data analysis: cross-validated Bayesian model averaging"; in: *NeuroImage*, vol. 158, pp. 186-195, eq. A.2; URL: https://www.sciencedirect.com/science/article/pii/S105381191730527X; DOI: 10.1016/j.neuroimage.2017.06.056.

1.5.20 Distribution of residual sum of squares

Theorem: Assume a linear regression model (→ III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and consider estimation using weighted least squares (→ III/1.5.21). Then, the residual sum of squares (→ III/1.5.9) $\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}$, divided by the true error variance (→ III/1.5.1) σ^2, follows a chi-squared distribution (→ II/3.7.1) with $n-p$ degrees of freedom

$$\frac{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}{\sigma^2} \sim \chi^2(n-p) \tag{2}$$

where n and p are the dimensions of the $n \times p$ design matrix (→ III/1.5.1) X.

Proof: Consider an $n \times n$ square matrix W, such that

$$WVW^\mathrm{T} = I_n \; . \tag{3}$$

Then, left-multiplying the regression model in (1) with W gives

$$Wy = WX\beta + W\varepsilon, \; W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^\mathrm{T}) \tag{4}$$

which can be rewritten as

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \; \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \tag{5}$$

where $\tilde{y} = Wy$, $\tilde{X} = WX$ and $\tilde{\varepsilon} = W\varepsilon$. This implies the distribution (→ II/4.1.13)

$$\tilde{y} \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I_n) \; . \tag{6}$$

With that, we have obtained a linear regression model (→ III/1.5.1) with independent observations. Cochran's theorem for multivariate normal variables states that, for an $n \times 1$ normal random vector (→ II/4.1.1) whose covariance matrix (→ I/1.13.9) is a scalar multiple of the identity matrix, a specific squared form will follow a non-central chi-squared distribution where the degrees of freedom and the non-centrality paramter depend on the matrix in the quadratic form:

$$x \sim \mathcal{N}(\mu, \sigma^2 I_n) \quad \Rightarrow \quad y = x^\mathrm{T} A x / \sigma^2 \sim \chi^2\left(\mathrm{tr}(A), \mu^\mathrm{T} A \mu\right) \; . \tag{7}$$

First, we formulate the residuals (→ III/1.5.14) in terms of transformed measurements $\tilde{y}$:

$$\begin{aligned}
\hat{\varepsilon} &= \tilde{y} - \tilde{X}\hat{\beta} && \text{where} \quad \hat{\beta} = (\tilde{X}^\mathrm{T}\tilde{X})^{-1}\tilde{X}^\mathrm{T}\tilde{y} \\
&= (I_n - \tilde{P})\tilde{y} && \text{where} \quad \tilde{P} = \tilde{X}(\tilde{X}^\mathrm{T}\tilde{X})^{-1}\tilde{X}^\mathrm{T} \\
&= \tilde{R}\tilde{y} && \text{where} \quad \tilde{R} = I_n - \tilde{P} \; .
\end{aligned} \tag{8}$$

Next, we observe that the residual sum of squares can be represented as a quadratic form:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}{\sigma^2} = \tilde{y}^\mathrm{T}\tilde{R}^\mathrm{T}\tilde{R}\tilde{y}/\sigma^2 \tag{9}$$

Because the residual-forming matrix (→ III/1.5.13) $\tilde{R}$ is symmetric (→ III/1.5.15) and idempotent (→ III/1.5.16), we have $\tilde{R}^\mathrm{T} = \tilde{R}$ and $\tilde{R}^2 = \tilde{R}$, such that:

$$\frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{\sigma^2} = \tilde{y}^{\mathrm{T}}\tilde{R}\tilde{y}/\sigma^2 \; . \tag{10}$$

With that, we can apply Cochran's theorem given by (7) which yields

$$\begin{split} \frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{\sigma^2} &\sim \chi^2\left(\mathrm{tr}(I_n-\tilde{P}),\, \beta^{\mathrm{T}}\tilde{X}^{\mathrm{T}}\tilde{R}\tilde{X}\beta\right) \\ &\sim \chi^2\left(\mathrm{tr}(I_n)-\mathrm{tr}(\tilde{P}),\, \beta^{\mathrm{T}}\tilde{X}^{\mathrm{T}}(I_n-\tilde{P})\tilde{X}\beta\right) \\ &\sim \chi^2\left(\mathrm{tr}(I_n)-\mathrm{tr}(\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}),\, \beta^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}\tilde{X}-\tilde{X}^{\mathrm{T}}\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\tilde{X})\beta\right) \\ &\sim \chi^2\left(\mathrm{tr}(I_n)-\mathrm{tr}(\tilde{X}^{\mathrm{T}}\tilde{X}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}),\, \beta^{\mathrm{T}}(\tilde{X}^{\mathrm{T}}\tilde{X}-\tilde{X}^{\mathrm{T}}\tilde{X})\beta\right) \\ &\sim \chi^2\left(\mathrm{tr}(I_n)-\mathrm{tr}(I_p),\, \beta^{\mathrm{T}}0_{pp}\beta\right) \\ &\sim \chi^2\left(n-p,\, 0\right) \; . \end{split} \tag{11}$$

Because a non-central chi-squared distribution with non-centrality parameter of zero reduces to the central chi-squared distribution, we obtain our final result:

$$\frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{\sigma^2} \sim \chi^2(n-p) \; . \tag{12}$$

■

Sources:

- Koch, Karl-Rudolf (2007): "Estimation of the Variance Factor in Traditional Statistics"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 4.2.3, eq. 4.37; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.
- Penny, William (2006): "Estimating error variance"; in: *Mathematics for Brain Imaging*, ch. 2.2, pp. 49-51, eqs. 2.4-2.8; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.
- Wikipedia (2022): "Ordinary least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-12-13; URL: https://en.wikipedia.org/wiki/Ordinary_least_squares#Estimation.
- ocram (2022): "Why is RSS distributed chi square times n-p?"; in: *StackExchange CrossValidated*, retrieved on 2022-12-21; URL: https://stats.stackexchange.com/a/20230.

1.5.21 Weighted least squares

Theorem: Given a linear regression model (→ III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{1}$$

the parameters minimizing the weighted residual sum of squares (→ III/1.5.9) are given by

$$\hat{\beta} = (X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}y \; . \tag{2}$$

Proof: Let there be an $n \times n$ square matrix W, such that

$$WVW^{\mathrm{T}} = I_n \; . \tag{3}$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed as the matrix square root of the inverse of V:

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2} \; . \tag{4}$$

Left-multiplying the linear regression equation (1) with W, the linear transformation theorem (→ II/4.1.13) implies that

$$Wy = WX\beta + W\varepsilon, \; W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T) \; . \tag{5}$$

Applying (3), we see that (5) is actually a linear regression model (→ III/1.5.1) with independent observations

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \; \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \tag{6}$$

where $\tilde{y} = Wy$, $\tilde{X} = WX$ and $\tilde{\varepsilon} = W\varepsilon$, such that we can apply the ordinary least squares solution (→ III/1.5.3) giving

$$\begin{aligned}
\hat{\beta} &= (\tilde{X}^\mathrm{T}\tilde{X})^{-1}\tilde{X}^\mathrm{T}\tilde{y} \\
&= \left((WX)^\mathrm{T}WX\right)^{-1}(WX)^\mathrm{T}Wy \\
&= \left(X^\mathrm{T}W^\mathrm{T}WX\right)^{-1}X^\mathrm{T}W^\mathrm{T}Wy \\
&= \left(X^\mathrm{T}WWX\right)^{-1}X^\mathrm{T}WWy \\
&\overset{(4)}{=} \left(X^\mathrm{T}V^{-1}X\right)^{-1}X^\mathrm{T}V^{-1}y
\end{aligned} \tag{7}$$

which corresponds to the weighted least squares solution (2).

■

Sources:
- Stephan, Klaas Enno (2010): "The General Linear Model (GLM)"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 3, Slides 20/23; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.
- Wikipedia (2021): "Weighted least squares"; in: *Wikipedia, the free encyclopedia*, retrieved on 2021-11-17; URL: https://en.wikipedia.org/wiki/Weighted_least_squares#Motivation.

1.5.22 Weighted least squares

Theorem: Given a linear regression model (→ III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{1}$$

the parameters minimizing the weighted residual sum of squares (→ III/1.5.9) are given by

$$\hat{\beta} = (X^\mathrm{T}V^{-1}X)^{-1}X^\mathrm{T}V^{-1}y \; . \tag{2}$$

Proof: Let there be an $n \times n$ square matrix W, such that

$$WVW^\mathrm{T} = I_n \; . \tag{3}$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed the matrix square root of the inverse of V:

$$WVW = I_n \quad \Leftrightarrow \quad V = W^{-1}W^{-1} \quad \Leftrightarrow \quad V^{-1} = WW \quad \Leftrightarrow \quad W = V^{-1/2} \; . \tag{4}$$

Left-multiplying the linear regression equation (1) with W, the linear transformation theorem (→ II/4.1.13) implies that

$$Wy = WX\beta + W\varepsilon, \; W\varepsilon \sim \mathcal{N}(0, \sigma^2 WVW^T) \; . \tag{5}$$

Applying (3), we see that (5) is actually a linear regression model (→ III/1.5.1) with independent observations

$$Wy = WX\beta + W\varepsilon, \; W\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \; . \tag{6}$$

With this, we can express the weighted residual sum of squares (→ III/1.5.9) as

$$\mathrm{wRSS}(\beta) = \sum_{i=1}^{n} (W\varepsilon)_i^2 = (W\varepsilon)^\mathrm{T}(W\varepsilon) = (Wy - WX\beta)^\mathrm{T}(Wy - WX\beta) \tag{7}$$

which can be developed into

$$\begin{aligned} \mathrm{wRSS}(\beta) &= y^\mathrm{T}W^\mathrm{T}Wy - y^\mathrm{T}W^\mathrm{T}WX\beta - \beta^\mathrm{T}X^\mathrm{T}W^\mathrm{T}Wy + \beta^\mathrm{T}X^\mathrm{T}W^\mathrm{T}WX\beta \\ &= y^\mathrm{T}WWy - 2\beta^\mathrm{T}X^\mathrm{T}WWy + \beta^\mathrm{T}X^\mathrm{T}WWX\beta \\ &\overset{(4)}{=} y^\mathrm{T}V^{-1}y - 2\beta^\mathrm{T}X^\mathrm{T}V^{-1}y + \beta^\mathrm{T}X^\mathrm{T}V^{-1}X\beta \; . \end{aligned} \tag{8}$$

The derivative of wRSS(β) with respect to β is

$$\frac{\mathrm{dwRSS}(\beta)}{\mathrm{d}\beta} = -2X^\mathrm{T}V^{-1}y + 2X^\mathrm{T}V^{-1}X\beta \tag{9}$$

and setting this derivative to zero, we obtain:

$$\begin{aligned} \frac{\mathrm{dwRSS}(\hat{\beta})}{\mathrm{d}\beta} &= 0 \\ 0 &= -2X^\mathrm{T}V^{-1}y + 2X^\mathrm{T}V^{-1}X\hat{\beta} \\ X^\mathrm{T}V^{-1}X\hat{\beta} &= X^\mathrm{T}V^{-1}y \\ \hat{\beta} &= (X^\mathrm{T}V^{-1}X)^{-1}X^\mathrm{T}V^{-1}y \; . \end{aligned} \tag{10}$$

Since the quadratic form $y^\mathrm{T}V^{-1}y$ in (8) is positive, $\hat{\beta}$ minimizes wRSS(β).

■

1.5.23 Maximum likelihood estimation

Theorem: Given a linear regression model (→ III/1.5.1) with correlated observations

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{1}$$

the maximum likelihood estimates (→ I/4.1.3) of β and σ^2 are given by

$$\begin{aligned}\hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; .\end{aligned} \tag{2}$$

Proof: With the probability density function of the multivariate normal distribution (→ II/4.1.7), the linear regression equation (1) implies the following likelihood function (→ I/5.1.2)

$$\begin{aligned}p(y|\beta,\sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\ &= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp\left[-\frac{1}{2} (y - X\beta)^\mathrm{T} (\sigma^2 V)^{-1} (y - X\beta) \right]\end{aligned} \tag{3}$$

and, using $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (→ I/4.1.2)

$$\begin{aligned}\mathrm{LL}(\beta,\sigma^2) &= \log p(y|\beta,\sigma^2) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|V| \\ &\quad - \frac{1}{2\sigma^2} (y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) \; .\end{aligned} \tag{4}$$

Substituting the precision matrix $P = V^{-1}$ into (4) to ease notation, we have:

$$\begin{aligned}\mathrm{LL}(\beta,\sigma^2) = &-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|V|) \\ &- \frac{1}{2\sigma^2} \left(y^\mathrm{T} P y - 2\beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta \right) \; .\end{aligned} \tag{5}$$

The derivative of the log-likelihood function (5) with respect to β is

$$\begin{aligned}\frac{\mathrm{d}\mathrm{LL}(\beta,\sigma^2)}{\mathrm{d}\beta} &= \frac{\mathrm{d}}{\mathrm{d}\beta} \left(-\frac{1}{2\sigma^2} \left(y^\mathrm{T} P y - 2\beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta \right) \right) \\ &= \frac{1}{2\sigma^2} \frac{\mathrm{d}}{\mathrm{d}\beta} \left(2\beta^\mathrm{T} X^\mathrm{T} P y - \beta^\mathrm{T} X^\mathrm{T} P X \beta \right) \\ &= \frac{1}{2\sigma^2} \left(2X^\mathrm{T} P y - 2X^\mathrm{T} P X \beta \right) \\ &= \frac{1}{\sigma^2} \left(X^\mathrm{T} P y - X^\mathrm{T} P X \beta \right)\end{aligned} \tag{6}$$

and setting this derivative to zero gives the MLE for β:

$$\begin{aligned}\frac{\mathrm{d}\mathrm{LL}(\hat{\beta},\sigma^2)}{\mathrm{d}\beta} &= 0 \\ 0 &= \frac{1}{\sigma^2} \left(X^\mathrm{T} P y - X^\mathrm{T} P X \hat{\beta} \right) \\ 0 &= X^\mathrm{T} P y - X^\mathrm{T} P X \hat{\beta} \\ X^\mathrm{T} P X \hat{\beta} &= X^\mathrm{T} P y \\ \hat{\beta} &= \left(X^\mathrm{T} P X \right)^{-1} X^\mathrm{T} P y\end{aligned} \tag{7}$$

The derivative of the log-likelihood function (4) at $\hat{\beta}$ with respect to σ^2 is

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\beta}, \sigma^2)}{\mathrm{d}\sigma^2} &= \frac{\mathrm{d}}{\mathrm{d}\sigma^2} \left(-\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \right) \\
&= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta})
\end{aligned} \tag{8}
$$

and setting this derivative to zero gives the MLE for σ^2:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\beta}, \hat{\sigma}^2)}{\mathrm{d}\sigma^2} &= 0 \\
0 &= -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\
\frac{n}{2\hat{\sigma}^2} &= \frac{1}{2(\hat{\sigma}^2)^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\
\frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{n}{2\hat{\sigma}^2} &= \frac{2(\hat{\sigma}^2)^2}{n} \cdot \frac{1}{2(\hat{\sigma}^2)^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\
\hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta})
\end{aligned} \tag{9}
$$

Together, (7) and (9) constitute the MLE for multiple linear regression.

■

1.5.24 Maximum log-likelihood

Theorem: Consider a linear regression model (→ III/1.5.1) m with correlation structure (→ I/1.14.5) V

$$
m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}
$$

Then, the maximum log-likelihood (→ I/4.1.4) for this model is

$$
\mathrm{MLL}(m) = -\frac{n}{2} \log\left(\frac{\mathrm{RSS}}{n}\right) - \frac{n}{2} \left[1 + \log(2\pi)\right] \tag{2}
$$

under uncorrelated observations (→ III/1.5.1), i.e. if $V = I_n$, and

$$
\mathrm{MLL}(m) = -\frac{n}{2} \log\left(\frac{\mathrm{wRSS}}{n}\right) - \frac{n}{2} \left[1 + \log(2\pi)\right] - \frac{1}{2} \log|V| \; , \tag{3}
$$

in the general case, i.e. if $V \neq I_n$, where RSS is the residual sum of squares (→ III/1.5.9) and wRSS is the weighted residual sum of squares (→ III/1.5.22).

Proof: The likelihood function (→ I/5.1.2) for multiple linear regression is given by (→ III/1.5.23)

$$
\begin{aligned}
p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\
&= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp\left[-\frac{1}{2}(y - X\beta)^\mathrm{T} (\sigma^2 V)^{-1} (y - X\beta)\right] \;,
\end{aligned} \tag{4}
$$

such that, with $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (→ I/4.1.2) for this model becomes (→ III/1.5.23)

$$
\begin{aligned}
\mathrm{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|V| - \frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) \;.
\end{aligned} \tag{5}
$$

The maximum likelihood estimate for the noise variance (→ III/1.5.23) is

$$
\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \tag{6}
$$

which can also be expressed in terms of the (weighted) residual sum of squares (→ III/1.5.9) as

$$
\frac{1}{n}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) = \frac{1}{n}(Wy - WX\hat{\beta})^\mathrm{T}(Wy - WX\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(W\hat{\varepsilon})_i^2 = \frac{\mathrm{wRSS}}{n} \tag{7}
$$

where $W = V^{-1/2}$. Plugging (6) into (5), we obtain the maximum log-likelihood (→ I/4.1.4) as

$$
\begin{aligned}
\mathrm{MLL}(m) &= \mathrm{LL}(\hat{\beta}, \hat{\sigma}^2) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}\log|V| - \frac{1}{2\hat{\sigma}^2}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\mathrm{wRSS}}{n}\right) - \frac{1}{2}\log|V| - \frac{1}{2}\cdot\frac{n}{\mathrm{wRSS}}\cdot\mathrm{wRSS} \\
&= -\frac{n}{2}\log\left(\frac{\mathrm{wRSS}}{n}\right) - \frac{n}{2}\left[1 + \log(2\pi)\right] - \frac{1}{2}\log|V|
\end{aligned} \tag{8}
$$

which proves the result in (3). Assuming $V = I_n$, we have

$$
\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\mathrm{RSS}}{n} \tag{9}
$$

and

$$
\frac{1}{2}\log|V| = \frac{1}{2}\log|I_n| = \frac{1}{2}\log 1 = 0 \;, \tag{10}
$$

such that

$$
\mathrm{MLL}(m) = -\frac{n}{2}\log\left(\frac{\mathrm{RSS}}{n}\right) - \frac{n}{2}\left[1 + \log(2\pi)\right] \tag{11}
$$

which proves the result in (2). This completes the proof.

■

Sources:
- Claeskens G, Hjort NL (2008): "Akaike's information criterion"; in: *Model Selection and Model Averaging*, ex. 2.2, p. 66; URL: https://www.cambridge.org/core/books/model-selection-and-model-averaging E6F1EC77279D1223423BB64FC3A12C37; DOI: 10.1017/CBO9780511790485.

1.5.25 t-contrast

Definition: Consider a linear regression model (→ III/1.5.1) with $n \times p$ design matrix X and $p \times 1$ regression coefficients β:

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}$$

Then, a t-contrast is specified by a $p \times 1$ vector c and it entails the null hypothesis (→ I/4.3.2) that the product of this vector and the regression coefficients is zero:

$$H_0: \; c^\mathrm{T}\beta = 0 \; . \tag{2}$$

Consequently, the alternative hypothesis (→ I/4.3.3) of a two-tailed t-test (→ I/4.2.4) is

$$H_1: \; c^\mathrm{T}\beta \neq 0 \tag{3}$$

and the alternative hypothesis (→ I/4.3.3) of a one-sided t-test (→ I/4.2.4) would be

$$H_1: \; c^\mathrm{T}\beta < 0 \quad \text{or} \quad H_1: \; c^\mathrm{T}\beta > 0 \; . \tag{4}$$

Here, c is called the "contrast vector" and $c^\mathrm{T}\beta$ is called the "contrast value". With estimated regression coefficients, $c^\mathrm{T}\hat{\beta}$ is called the "estimated contrast value".

Sources:
- Stephan, Klaas Enno (2010): "Classical (frequentist) inference"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 7/9; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.

1.5.26 F-contrast

Definition: Consider a linear regression model (→ III/1.5.1) with $n \times p$ design matrix X and $p \times 1$ regression coefficients β:

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}$$

Then, an F-contrast is specified by a $p \times q$ matrix C, yielding a $q \times 1$ vector $\gamma = C^\mathrm{T}\beta$, and it entails the null hypothesis (→ I/4.3.2) that each value in this vector is zero:

$$H_0: \; \gamma_1 = 0 \wedge \ldots \wedge \gamma_q = 0 \; . \tag{2}$$

Consequently, the alternative hypothesis (→ I/4.3.3) of the statistical test (→ I/4.3.1) would be that at least one entry of this vector is non-zero:

$$H_1: \; \gamma_1 \neq 0 \vee \ldots \vee \gamma_q \neq 0 \; . \tag{3}$$

Here, C is called the "contrast matrix" and $C^\mathrm{T}\beta$ are called the "contrast values". With estimated regression coefficients, $C^\mathrm{T}\hat{\beta}$ are called the "estimated contrast values".

Sources:

- Stephan, Klaas Enno (2010): "Classical (frequentist) inference"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 23/25; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.

1.5.27 Contrast-based t-test

Theorem: Consider a linear regression model (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and a t-contrast (→ III/1.5.25) on the model parameters

$$\gamma = c^\mathrm{T}\beta \quad \text{where} \quad c \in \mathbb{R}^{p \times 1} \; . \tag{2}$$

Then, the test statistic (→ I/4.3.5)

$$t = \frac{c^\mathrm{T}\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c}} \tag{3}$$

with the parameter estimates (→ III/1.5.23)

$$\begin{aligned}
\hat{\beta} &= (X^\mathrm{T}V^{-1}X)^{-1}X^\mathrm{T}V^{-1}y \\
\hat{\sigma}^2 &= \frac{1}{n-p}(y-X\hat{\beta})^\mathrm{T}V^{-1}(y-X\hat{\beta})
\end{aligned} \tag{4}$$

follows a t-distribution (→ II/3.3.1)

$$t \sim \mathrm{t}(n-p) \tag{5}$$

under the null hypothesis (→ I/4.3.2)

$$\begin{aligned}
H_0 &: c^\mathrm{T}\beta = 0 \\
H_1 &: c^\mathrm{T}\beta > 0 \; .
\end{aligned} \tag{6}$$

Proof:
1) We know that the estimated regression coefficients in linear regression follow a multivariate normal distribution (→ III/1.5.19):

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \, \sigma^2(X^\mathrm{T}V^{-1}X)^{-1}\right) \; . \tag{7}$$

Thus, the estimated contrast value (→ III/1.5.25) $\hat{\gamma} = c^\mathrm{T}\hat{\beta}$ is distributed according to a univariate normal distribution (→ II/4.1.13):

$$\hat{\gamma} \sim \mathcal{N}\left(c^\mathrm{T}\beta, \, \sigma^2 c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c\right) \; . \tag{8}$$

Now, define the random variable z by dividing $\hat{\gamma}$ by its standard deviation:

$$z = \frac{c^{\mathrm{T}}\hat{\beta}}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}V^{-1}X)^{-1}c}} \; . \tag{9}$$

Again applying the linear transformation theorem (→ II/4.1.13), this is distributed as

$$z \sim \mathcal{N}\left(\frac{c^{\mathrm{T}}\beta}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}V^{-1}X)^{-1}c}}, 1 \right) \tag{10}$$

and thus follows a standard normal distribution (→ II/3.2.3) under the null hypothesis (→ I/4.3.2):

$$z \sim \mathcal{N}(0,1), \quad \text{if } H_0 \; . \tag{11}$$

2) We also know that the residual sum of squares (→ III/1.5.9), divided the true error variance (→ III/1.5.1)

$$v = \frac{1}{\sigma^2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{\sigma^2} = \frac{1}{\sigma^2}(y - X\hat{\beta})^{\mathrm{T}} V^{-1} (y - X\hat{\beta}) \tag{12}$$

is following a chi-squared distribution (→ III/1.5.20):

$$v \sim \chi^2(n-p) \; . \tag{13}$$

3) Because the estimated regression coefficients and the vector of residuals are independent from each other (→ III/1.5.17)

$$\hat{\beta} \quad \text{and} \quad \hat{\varepsilon} \quad \text{ind.} \tag{14}$$

and thus, the estimated contrast values are also independent from the function of the residual sum of squares

$$z = \frac{c^{\mathrm{T}}\hat{\beta}}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}V^{-1}X)^{-1}c}} \quad \text{and} \quad v = \frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{\sigma^2} \quad \text{ind.} \; , \tag{15}$$

the following quantity is, by definition, t-distributed (→ II/3.3.1)

$$t = \frac{z}{\sqrt{v/(n-p)}} \sim \mathrm{t}(n-p), \quad \text{if } H_0 \tag{16}$$

and the quantity can be evaluated as:

$$
\begin{split}
t &\overset{(16)}{=} \frac{z}{\sqrt{v/(n-p)}} \\
&\overset{(15)}{=} \frac{c^\mathrm{T}\hat{\beta}}{\sqrt{\sigma^2 c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c}} \cdot \sqrt{\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}/\sigma^2}} \\
&= \frac{c^\mathrm{T}\hat{\beta}}{\sqrt{\frac{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}{n-p} \cdot c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c}} \\
&\overset{(12)}{=} \frac{c^\mathrm{T}\hat{\beta}}{\sqrt{\frac{(y-X\hat{\beta})^\mathrm{T}V^{-1}(y-X\hat{\beta})}{n-p} \cdot c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c}} \\
&\overset{(4)}{=} \frac{c^\mathrm{T}\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}c}} \; .
\end{split}
\tag{17}
$$

This means that the null hypothesis (→ I/4.3.2) in (6) can be rejected when t from (17) is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from Student's t-distribution (→ II/3.3.1) with $n-p$ degrees of freedom using a significance level (→ I/4.3.8) α.

■

Sources:

- Stephan, Klaas Enno (2010): "Classical (frequentist) inference"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 7/9; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.
- Walter, Henrik (ed.) (2005): "Datenanalyse für funktionell bildgebende Verfahren"; in: *Funktionelle Bildgebung in Psychiatrie und Psychotherapie*, Schattauer, Stuttgart/New York, 2005, p. 40; URL: https://books.google.de/books?id=edWzKAHi7jQC&source=gbs_navlinks_s.
- jld (2018): "Understanding t-test for linear regression"; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: https://stats.stackexchange.com/a/344008.
- Soch, Joram (2020): "Distributional Transformation Improves Decoding Accuracy When Predicting Chronological Age From Structural MRI"; in: *Frontiers in Psychiatry*, vol. 11, art. 604268, eqs. 8/9; URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.604268/full; DOI: 10.3389/fpsyt.2020.604268.

1.5.28 Contrast-based F-test

Theorem: Consider a linear regression model (→ III/1.5.1)

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}
$$

and an F-contrast (→ III/1.5.26) on the model parameters

$$
\gamma = C^\mathrm{T}\beta \quad \text{where} \quad C \in \mathbb{R}^{p \times q} \; . \tag{2}
$$

Then, the test statistic (→ I/4.3.5)

$$
F = \hat{\beta}^\mathrm{T} C \left(\hat{\sigma}^2 C^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}C \right)^{-1} C^\mathrm{T}\hat{\beta}/q \tag{3}
$$

with the parameter estimates (→ III/1.5.23)

$$
\begin{aligned}
\hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
\hat{\sigma}^2 &= \frac{1}{n-p} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta})
\end{aligned} \tag{4}
$$

follows an F-distribution (→ II/3.8.1)

$$
F \sim \mathrm{F}(q, n-p) \tag{5}
$$

under the null hypothesis (→ I/4.3.2)

$$
\begin{aligned}
H_0 &: \gamma_1 = 0 \wedge \ldots \wedge \gamma_q = 0 \\
H_1 &: \gamma_1 \neq 0 \vee \ldots \vee \gamma_q \neq 0 \; .
\end{aligned} \tag{6}
$$

Proof:
1) We know that the estimated regression coefficients in linear regression follow a multivariate normal distribution (→ III/1.5.19):

$$
\hat{\beta} \sim \mathcal{N}\left(\beta, \, \sigma^2 (X^\mathrm{T} V^{-1} X)^{-1}\right) \; . \tag{7}
$$

Thus, the estimated contrast vector (→ III/1.5.26) $\hat{\gamma} = C^\mathrm{T} \hat{\beta}$ is also distributed according to a multivariate normal distribution (→ II/4.1.13):

$$
\hat{\gamma} \sim \mathcal{N}\left(C^\mathrm{T} \beta, \, \sigma^2 C^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C\right) \; . \tag{8}
$$

Substituting the noise variance σ^2 with the noise precision $\tau = 1/\sigma^2$, we can also write this down as a conditional distribution (→ I/1.5.4):

$$
\hat{\gamma} | \tau \sim \mathcal{N}\left(C^\mathrm{T} \beta, (\tau Q)^{-1}\right) \quad \text{with} \quad Q = \left(C^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C\right)^{-1} \; . \tag{9}
$$

2) We also know that the residual sum of squares (→ III/1.5.9), divided the true error variance (→ III/1.5.1)

$$
\frac{1}{\sigma^2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{\hat{\varepsilon}^\mathrm{T} \hat{\varepsilon}}{\sigma^2} = \frac{1}{\sigma^2} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \tag{10}
$$

is following a chi-squared distribution (→ III/1.5.20):

$$
\frac{\hat{\varepsilon}^\mathrm{T} \hat{\varepsilon}}{\sigma^2} = \tau \, \hat{\varepsilon}^\mathrm{T} \hat{\varepsilon} \sim \chi^2(n-p) \; . \tag{11}
$$

The chi-squared distribution is a special case of the gamma distribution (→ II/3.7.2)

$$
X \sim \chi^2(k) \quad \Rightarrow \quad X \sim \mathrm{Gam}\left(\frac{k}{2}, \frac{1}{2}\right) \tag{12}
$$

and the gamma distribution changes under multiplication (→ II/3.4.6) in the following way:

$$
X \sim \mathrm{Gam}\left(a, b\right) \quad \Rightarrow \quad cX \sim \mathrm{Gam}\left(a, \frac{b}{c}\right) \; . \tag{13}
$$

Thus, combining (12) and (13) with (11), we obtain the marginal distribution (→ I/1.5.3) of τ as:

$$\frac{1}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}\left(\tau\,\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}\right) = \tau \sim \mathrm{Gam}\left(\frac{n-p}{2}, \frac{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}{2}\right) \; . \tag{14}$$

3) Note that the joint distribution (→ I/1.5.2) of $\hat{\gamma}$ and τ is, following from (9) and (14) and by definition, a normal-gamma distribution (→ II/4.3.1):

$$\hat{\gamma}, \tau \sim \mathrm{NG}\left(C^\mathrm{T}\beta, Q, \frac{n-p}{2}, \frac{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}{2}\right) \; . \tag{15}$$

The marginal distribution of a normal-gamma distribution with respect to the normal random variable, is a multivariate t-distribution (→ II/4.3.8):

$$X, Y \sim \mathrm{NG}(\mu, \Lambda, a, b) \quad \Rightarrow \quad X \sim \mathrm{t}\left(\mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \; . \tag{16}$$

Thus, the marginal distribution (→ I/1.5.3) of $\hat{\gamma}$ is:

$$\hat{\gamma} \sim \mathrm{t}\left(C^\mathrm{T}\beta, \left(\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}Q\right)^{-1}, n-p\right) \; . \tag{17}$$

4) Because of the following relationship between the multivariate t-distribution and the F-distribution (→ II/4.2.3)

$$X \sim t(\mu, \Sigma, \nu) \quad \Rightarrow \quad (X-\mu)^\mathrm{T}\,\Sigma^{-1}(X-\mu)/n \sim F(n, \nu) \; , \tag{18}$$

the following quantity is, by definition, F-distributed (→ II/3.8.1)

$$F = \left(\hat{\gamma} - C^\mathrm{T}\beta\right)^\mathrm{T}\left(\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}Q\right)\left(\hat{\gamma} - C^\mathrm{T}\beta\right)/q \sim \mathrm{F}(q, n-p) \tag{19}$$

and under the null hypothesis (→ I/4.3.2) (6), it can be evaluated as:

$$\begin{split}
F &\overset{(19)}{=} \left(\hat{\gamma} - C^\mathrm{T}\beta\right)^\mathrm{T}\left(\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}Q\right)\left(\hat{\gamma} - C^\mathrm{T}\beta\right)/q \\
&\overset{(6)}{=} \hat{\gamma}^\mathrm{T}\left(\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}Q\right)\hat{\gamma}/q \\
&\overset{(2)}{=} \hat{\beta}^\mathrm{T}C\left(\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}Q\right)C^\mathrm{T}\hat{\beta}/q \\
&\overset{(9)}{=} \hat{\beta}^\mathrm{T}C\left(\frac{n-p}{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}\left(C^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}C\right)^{-1}\right)C^\mathrm{T}\hat{\beta}/q \\
&\overset{(10)}{=} \hat{\beta}^\mathrm{T}C\left(\frac{n-p}{(y-X\hat{\beta})^\mathrm{T}V^{-1}(y-X\hat{\beta})}\left(C^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}C\right)^{-1}\right)C^\mathrm{T}\hat{\beta}/q \\
&\overset{(4)}{=} \hat{\beta}^\mathrm{T}C\left(\frac{1}{\hat{\sigma}^2}\left(C^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}C\right)^{-1}\right)C^\mathrm{T}\hat{\beta}/q \\
&= \hat{\beta}^\mathrm{T}C\left(\hat{\sigma}^2 C^\mathrm{T}(X^\mathrm{T}V^{-1}X)^{-1}C\right)^{-1}C^\mathrm{T}\hat{\beta}/q \; .
\end{split} \tag{20}$$

This means that the null hypothesis (→ I/4.3.2) in (6) can be rejected when F from (20) is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from Fisher's F-distribution

(→ II/3.8.1) with q numerator and $n-p$ denominator degrees of freedom using a significance level (→ I/4.3.8) α.

■

Sources:

- Stephan, Klaas Enno (2010): "Classical (frequentist) inference"; in: *Methods and models for fMRI data analysis in neuroeconomics*, Lecture 4, Slides 23/25; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html.
- Koch, Karl-Rudolf (2007): "Multivariate Distributions"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, ch. 2.5, eqs. 2.202, 2.213, 2.211; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.
- jld (2018): "Understanding t-test for linear regression"; in: *StackExchange CrossValidated*, retrieved on 2022-12-13; URL: https://stats.stackexchange.com/a/344008.
- Penny, William (2006): "Comparing nested GLMs"; in: *Mathematics for Brain Imaging*, ch. 2.3, pp. 51-52, eq. 2.9; URL: https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf.

1.5.29 t-test for single regressor

Theorem: Consider a linear regression model (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

using the $n \times p$ design matrix X and the parameter estimates (→ III/1.5.23)

$$\begin{aligned} \hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n-p} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; . \end{aligned} \tag{2}$$

Then, the test statistic (→ I/4.3.5)

$$t_j = \frac{\hat{\beta}_j}{\sqrt{(\hat{\varepsilon}^\mathrm{T} V^{-1} \hat{\varepsilon}) / (n-p) \, \sigma_{jj}}} \tag{3}$$

with the $n \times 1$ vector of residuals (→ III/1.5.14)

$$\hat{\varepsilon} = y - X\hat{\beta} \tag{4}$$

and σ_{jj} equal to the j-th diagonal element of the parameter covariance matrix (→ III/1.5.19)

$$\sigma_{jj} = \left[\left(X^\mathrm{T} V^{-1} X \right)^{-1} \right]_{jj} \tag{5}$$

follows a t-distribution (→ II/3.3.1)

$$t_j \sim \mathrm{t}(n-p) \tag{6}$$

under the null hypothesis (→ I/4.3.2) that the j-th regression coefficient (→ III/1.5.1) is zero:

$$H_0: \; \beta_j = 0 \; . \tag{7}$$

Proof: This is a special case of the contrast-based t-test for multiple linear regression (→ III/1.5.27) based on the following t-statistic (→ II/3.3.1):

$$t = \frac{c^{\mathrm{T}}\hat{\beta}}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}V^{-1}X)^{-1}c}} \sim \mathrm{t}(n-p) \; . \tag{8}$$

In this special case, the contrast vector (→ III/1.5.25) is equal to the j-th elementary vector e_j (a $p \times 1$ vector of zeros, with a single 1 in the j-th entry)

$$c = e_j = \left[0, \ldots, 0, 1, 0, \ldots, 0\right]^{\mathrm{T}} \; , \tag{9}$$

such that the null hypothesis is given by

$$H_0: \; c^{\mathrm{T}}\beta = e_j^{\mathrm{T}}\beta = \beta_j = 0 \tag{10}$$

and the test statistic becomes

$$\begin{aligned} t_j &= \frac{e_j^{\mathrm{T}}\hat{\beta}}{\sqrt{\hat{\sigma}^2 e_j^{\mathrm{T}}(X^{\mathrm{T}}V^{-1}X)^{-1}e_j}} \\ &= \frac{\left[0, \ldots, 0, 1, 0, \ldots, 0\right] \left[\hat{\beta}_1, \ldots, \beta_{j-1}, \beta_j, \beta_{j+1}, \ldots, \hat{\beta}_p\right]^{\mathrm{T}}}{\sqrt{\frac{1}{n-p}(y-X\hat{\beta})^{\mathrm{T}}V^{-1}(y-X\hat{\beta}) \left[0, \ldots, 1, \ldots, 0\right] (X^{\mathrm{T}}V^{-1}X)^{-1} \left[0, \ldots, 1, \ldots, 0\right]^{\mathrm{T}}}} \\ &= \frac{\hat{\beta}_j}{\sqrt{\frac{1}{n-p}\left(\hat{\varepsilon}^{\mathrm{T}}V^{-1}\hat{\varepsilon}\right)\left[(X^{\mathrm{T}}V^{-1}X)^{-1}\right]_{jj}}} \\ &= \frac{\hat{\beta}_j}{\sqrt{\left(\hat{\varepsilon}^{\mathrm{T}}V^{-1}\hat{\varepsilon}\right)/(n-p)\,\sigma_{jj}}} \; . \end{aligned} \tag{11}$$

■

Sources:

- Ostwald, Dirk (2023): "T-Statistiken"; in: *Allgemeines Lineares Modell*, Einheit (7), Folien 20, 27; URL: https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/7_T_Statistiken-p-9968.pdf.

1.5.30 F-test for multiple regressors

Theorem: Consider a linear regression model (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

the design matrix and regression coefficients of which are partitioned as

$$\begin{aligned} X &= \begin{bmatrix} X_0 & X_1 \end{bmatrix} \in \mathbb{R}^{n \times p} \quad \text{where} \quad X_0 \in \mathbb{R}^{n \times p_0} \quad \text{and} \quad X_1 \in \mathbb{R}^{n \times p_1} \\ \beta &= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \in \mathbb{R}^{p \times 1} \quad \text{where} \quad \beta_0 \in \mathbb{R}^{p_0 \times 1} \quad \text{and} \quad \beta_1 \in \mathbb{R}^{p_1 \times 1} \end{aligned} \tag{2}$$

with $p = p_0 + p_1$. Then, the test statistic (→ I/4.3.5)

$$F = \frac{(\hat{\varepsilon}_0^\mathrm{T} V^{-1} \hat{\varepsilon}_0 - \hat{\varepsilon}^\mathrm{T} V^{-1} \hat{\varepsilon})/p_1}{\hat{\varepsilon}^\mathrm{T} V^{-1} \hat{\varepsilon}/(n-p)} \tag{3}$$

follows an F-distribution (→ II/3.8.1)

$$F \sim \mathrm{F}(p_1, n-p) \tag{4}$$

under the null hypothesis (→ I/4.3.2) that all regression coefficients (→ III/1.5.1) β_1 are zero:

$$H_0: \; \beta_1 = 0_{p_1} \quad \Leftrightarrow \quad \beta_j = 0 \quad \text{for all} \quad j = (p_0+1), \ldots, p \; . \tag{5}$$

In (3), $\hat{\varepsilon}$ and $\hat{\varepsilon}_0$ are the residual vectors (→ III/1.5.14) when using either the full design matrix X or the reduced design matrix X_0:

$$\begin{aligned} \hat{\varepsilon} &= y - X\hat{\beta} \quad \text{with} \quad \hat{\beta} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\ \hat{\varepsilon}_0 &= y - X_0\hat{\beta}_0 \quad \text{with} \quad \hat{\beta}_0 = (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \; . \end{aligned} \tag{6}$$

Proof: This is a special case of the contrast-based F-test for multiple linear regression (→ III/1.5.28) based on the F-statistic (→ I/4.3.5)

$$F = \hat{\beta}^\mathrm{T} C \left(\hat{\sigma}^2 C^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C \right)^{-1} C^\mathrm{T} \hat{\beta}/q \tag{7}$$

which follows an F-distribution (→ II/3.8.1) under the null hypothesis (→ I/4.3.2) that the product of the contrast matrix (→ III/1.5.26) $C \in \mathbb{R}^{p \times q}$ and the regression coefficients (→ III/1.5.1) equals zero:

$$F \sim \mathrm{F}(q, n-p), \quad \text{if} \quad C^\mathrm{T}\beta = 0_q = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \; . \tag{8}$$

In (7), $\hat{\sigma}^2$ is an estimate of the noise variance (→ III/1.5.1) calculated as the weighted (→ III/1.5.21) residual sum of squares (→ III/1.5.9), divided by $n-p$:

$$\hat{\sigma}^2 = \frac{1}{n-p} (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; . \tag{9}$$

In the present case, in order to compare the full model specified by X against the reduced model specified by X_0, we have to define the contrast matrix (→ III/1.5.26) as a vertical concatenation of a zero matrix on the first p_0 components and an identity matrix on the last p_1 components of β,

$$C_1 = \begin{bmatrix} 0_{p_0,p_1} \\ I_{p_1} \end{bmatrix} \in \mathbb{R}^{p \times p_1} \; , \tag{10}$$

i.e. specify an omnibus F-contrast that tests the alternative hypothesis (→ I/4.3.3) that any of the coefficients β_1 associated with the regressors X_1 is different from zero against the null hypothesis (→ I/4.3.2) that all those coefficients are zero:

$$
\begin{aligned}
H_0 :\; C_1^{\mathrm{T}}\beta = \begin{bmatrix} 0_{p_0,p_1} \\ I_{p_1} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_1 = 0_{p_1} \quad &\Leftrightarrow \quad \beta_{1j} = 0 \quad \text{for all} \quad j = 1, \ldots, p_1 \\
\Rightarrow H_1 \Leftrightarrow \neg H_0 :\; C_1^{\mathrm{T}}\beta = \beta_1 \neq 0_{p_1} \quad &\Leftrightarrow \quad \beta_{1j} \neq 0 \quad \text{for at least one} \quad j = 1, \ldots, p_1 \;.
\end{aligned} \tag{11}
$$

Thus, plugging $C = C_1$ and $q = p_1$ into (7) and noting that $\hat{\sigma}^2$ from (9) is a scalar, we obtain:

$$
\begin{aligned}
F &= \hat{\beta}^{\mathrm{T}} C_1 \left(\hat{\sigma}^2 C_1^{\mathrm{T}} (X^{\mathrm{T}} V^{-1} X)^{-1} C_1 \right)^{-1} C_1^{\mathrm{T}} \hat{\beta} / p_1 \\
&\overset{(9)}{=} \frac{\hat{\beta}^{\mathrm{T}} C_1 \left(C_1^{\mathrm{T}} (X^{\mathrm{T}} V^{-1} X)^{-1} C_1 \right)^{-1} C_1^{\mathrm{T}} \hat{\beta} / p_1}{(y - X\hat{\beta})^{\mathrm{T}} V^{-1} (y - X\hat{\beta}) / (n-p)} \;.
\end{aligned} \tag{12}
$$

Here, we take note of the fact that the denominator in (12) is already equal to the denominator in (3):

$$
(y - X\hat{\beta})^{\mathrm{T}} V^{-1} (y - X\hat{\beta}) / (n-p) \overset{(6)}{=} \hat{\varepsilon}^{\mathrm{T}} V^{-1} \hat{\varepsilon} / (n-p) \;. \tag{13}
$$

Therefore, what remains to be shown is that the numerator in (12) is equal to the numerator in (3):

$$
\hat{\beta}^{\mathrm{T}} C_1 \left(C_1^{\mathrm{T}} (X^{\mathrm{T}} V^{-1} X)^{-1} C_1 \right)^{-1} C_1^{\mathrm{T}} \hat{\beta} / p_1 = (\hat{\varepsilon}_0^{\mathrm{T}} V^{-1} \hat{\varepsilon}_0 - \hat{\varepsilon}^{\mathrm{T}} V^{-1} \hat{\varepsilon}) / p_1 \;. \tag{14}
$$

To do this, we start with the inner-most matrix:

$$
\begin{aligned}
X^{\mathrm{T}} V^{-1} X &\overset{(2)}{=} \begin{bmatrix} X_0 & X_1 \end{bmatrix}^{\mathrm{T}} V^{-1} \begin{bmatrix} X_0 & X_1 \end{bmatrix} \\
&= \begin{bmatrix} X_0^{\mathrm{T}} \\ X_1^{\mathrm{T}} \end{bmatrix} V^{-1} \begin{bmatrix} X_0 & X_1 \end{bmatrix} \\
&= \begin{bmatrix} X_0^{\mathrm{T}} V^{-1} X_0 & X_0^{\mathrm{T}} V^{-1} X_1 \\ X_1^{\mathrm{T}} V^{-1} X_0 & X_1^{\mathrm{T}} V^{-1} X_1 \end{bmatrix} \;.
\end{aligned} \tag{15}
$$

The inverse of a block matrix is:

$$
\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1} B (D - C A^{-1} B)^{-1} C A^{-1} & -A^{-1} B (D - C A^{-1} B)^{-1} \\ -(D - C A^{-1} B)^{-1} C A^{-1} & (D - C A^{-1} B)^{-1} \end{bmatrix} \;. \tag{16}
$$

Note that, with the contrast matrix C_1, we only extract the lower-right part of the inverse block matrix, so that we have:

$$
\begin{aligned}
\left(C_1^{\mathrm{T}} (X^{\mathrm{T}} V^{-1} X)^{-1} C_1 \right)^{-1} &\overset{(10)}{=} \left(\begin{bmatrix} 0_{p_1,p_0} & I_{p_1} \end{bmatrix} (X^{\mathrm{T}} V^{-1} X)^{-1} \begin{bmatrix} 0_{p_0,p_1} \\ I_{p_1} \end{bmatrix} \right)^{-1} \\
&\overset{(16)}{=} \left(\left(X_1^{\mathrm{T}} V^{-1} X_1 - X_1^{\mathrm{T}} V^{-1} X_0 (X_0^{\mathrm{T}} V^{-1} X_0)^{-1} X_0^{\mathrm{T}} V^{-1} X_1 \right)^{-1} \right)^{-1} \\
&= X_1^{\mathrm{T}} V^{-1} X_1 - X_1^{\mathrm{T}} V^{-1} X_0 (X_0^{\mathrm{T}} V^{-1} X_0)^{-1} X_0^{\mathrm{T}} V^{-1} X_1 \;.
\end{aligned} \tag{17}
$$

We call this $p_1 \times p_1$ matrix E and note that it can be written as

$$
\begin{split}
E &\overset{(17)}{=} X_1^\mathrm{T} \left(V^{-1} - V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \right) X_1 \\
&= X_1^\mathrm{T} \left(V^{-1} - F \right) X_1
\end{split}
\tag{18}
$$

where the $n \times n$ matrix F is given as follows:

$$
F = V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \; . \tag{19}
$$

Let $\hat{\beta}_{0(X)}$ denote the first p_0 entries of $\hat{\beta}$, i.e. estimates of the coefficients belonging to X_0, but estimated with X (as opposed to $\hat{\beta}_0$ estimated with X_0 given by (6)):

$$
\hat{\beta} = \begin{bmatrix} \hat{\beta}_{0(X)} \\ \hat{\beta}_1 \end{bmatrix} \; . \tag{20}
$$

Then, it obviously holds that

$$
\begin{split}
X \hat{\beta} &\overset{(20)}{=} \begin{bmatrix} X_0 & X_1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_{0(X)} \\ \hat{\beta}_1 \end{bmatrix} \\
&= X_0 \hat{\beta}_{0(X)} + X_1 \hat{\beta}_1 \\
&\Leftrightarrow \\
X_1 \hat{\beta}_1 &= X \hat{\beta} - X_0 \hat{\beta}_{0(X)} \; .
\end{split}
\tag{21}
$$

Next, we focus on $C_1^\mathrm{T} \hat{\beta}$ which simply extracts $\hat{\beta}_1$:

$$
\begin{split}
C_1^\mathrm{T} \hat{\beta} &\overset{(20)}{=} \begin{bmatrix} 0_{p_1,p_0} & I_{p_1} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{0(X)} \\ \hat{\beta}_1 \end{bmatrix} \\
&= \hat{\beta}_1 \; .
\end{split}
\tag{22}
$$

With these identities in mind, we can get back to our main quantity of interest from (14):

$$
\begin{aligned}
&\hat{\beta}^\mathrm{T} C_1 \left(C_1^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C_1 \right)^{-1} C_1^\mathrm{T} \hat{\beta} \\
&\overset{(22)}{=} \hat{\beta}_1^\mathrm{T} \, E \, \hat{\beta}_1 \\
&\overset{(18)}{=} \hat{\beta}_1^\mathrm{T} X_1^\mathrm{T} \left(V^{-1} - V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \right) X_1 \hat{\beta}_1 \\
&\overset{(21)}{=} \left(\hat{\beta}^\mathrm{T} X^\mathrm{T} - \hat{\beta}_{0(X)}^\mathrm{T} X_0^\mathrm{T} \right) \left(V^{-1} - V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \right) \left(X \hat{\beta} - X_0 \hat{\beta}_{0(X)} \right) \\
&\overset{(19)}{=} \left(\hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} - \hat{\beta}^\mathrm{T} X^\mathrm{T} F - \hat{\beta}_{0(X)}^\mathrm{T} X_0^\mathrm{T} V^{-1} + \hat{\beta}_{0(X)}^\mathrm{T} X_0^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \right) \left(X \hat{\beta} - X_0 \hat{\beta}_{0(X)} \right) \\
&= \left(\hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} - \hat{\beta}^\mathrm{T} X^\mathrm{T} F - \hat{\beta}_{0(X)}^\mathrm{T} X_0^\mathrm{T} V^{-1} + \hat{\beta}_{0(X)}^\mathrm{T} X_0^\mathrm{T} V^{-1} \right) \left(X \hat{\beta} - X_0 \hat{\beta}_{0(X)} \right) \\
&= \left(\hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} - \hat{\beta}^\mathrm{T} X^\mathrm{T} F \right) \left(X \hat{\beta} - X_0 \hat{\beta}_{0(X)} \right) \\
&\overset{(19)}{=} \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 \hat{\beta}_{0(X)} - \hat{\beta}^\mathrm{T} X^\mathrm{T} F X \hat{\beta} + \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} X_0 \hat{\beta}_{0(X)} \\
&= \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 \hat{\beta}_{0(X)} - \hat{\beta}^\mathrm{T} X^\mathrm{T} F X \hat{\beta} + \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 \hat{\beta}_{0(X)} \\
&= \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} F X \hat{\beta} \\
&\overset{(19)}{=} \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} X \hat{\beta} \; .
\end{aligned} \tag{23}
$$

Let the residual vector of the full model be defined as given by (6)

$$
\hat{\varepsilon} = y - X \hat{\beta} \quad \Leftrightarrow \quad y = X \hat{\beta} + \hat{\varepsilon} \tag{24}
$$

and consider the term $X^\mathrm{T} V^{-1} \hat{\varepsilon}$. Using the residual-forming matrix expression of the residual vector (→ III/1.5.14), we can show that this matrix product is zero:

$$
\begin{split}
X^\mathrm{T} V^{-1} \hat{\varepsilon} &= X^\mathrm{T} V^{-1} \left(I_n - X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} \right) y \\
&= X^\mathrm{T} V^{-1} y - X^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
&= X^\mathrm{T} V^{-1} y - X^\mathrm{T} V^{-1} y \\
&= 0_p \; .
\end{split} \tag{25}
$$

From this, it follows that the product $X_0^\mathrm{T} V^{-1} \hat{\varepsilon}$ is also zero:

$$
\begin{gathered}
X^\mathrm{T} V^{-1} \hat{\varepsilon} = 0_p \\
\begin{bmatrix} X_0^\mathrm{T} \\ X_1^\mathrm{T} \end{bmatrix} V^{-1} \hat{\varepsilon} = 0_p \\
\begin{bmatrix} X_0^\mathrm{T} V^{-1} \hat{\varepsilon} \\ X_1^\mathrm{T} V^{-1} \hat{\varepsilon} \end{bmatrix} = \begin{bmatrix} 0_{p_0} \\ 0_{p_1} \end{bmatrix} \\
\Leftrightarrow \\
X_0^\mathrm{T} V^{-1} \hat{\varepsilon} = 0_{p_0} \; .
\end{gathered} \tag{26}
$$

Thus, any term containing $X_0^\mathrm{T} V^{-1} \hat{\varepsilon} = 0_{p_0}$ can be added to a sum without changing the value of this sum. Continuing from above, we therefore write:

$$
\begin{aligned}
&\hat{\beta}^\mathrm{T} C_1 \left(C_1^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C_1 \right)^{-1} C_1^\mathrm{T} \hat{\beta} \\
&\overset{(23)}{=} \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} X \hat{\beta} \\
&\overset{(26)}{=} \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} X \hat{\beta} \\
&+ 2 \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \hat{\varepsilon} + \hat{\varepsilon}^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \hat{\varepsilon} \\
&= \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - \left(X \hat{\beta} + \hat{\varepsilon} \right)^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} \left(X \hat{\beta} + \hat{\varepsilon} \right) \\
&\overset{(24)}{=} \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \; .
\end{aligned} \tag{27}
$$

In the next transformations, we will make use of the weighted least squares parameter estimates (→ III/1.5.21)

$$
\begin{aligned}
\hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
\hat{\beta}_0 &= (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y
\end{aligned} \tag{28}
$$

and the fact that matrices and their inverses cancel out:

$$
\begin{aligned}
X^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} X = I_p \\
X_0^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} &= (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} X_0 = I_{p_0} \; .
\end{aligned} \tag{29}
$$

Continuing from above, we have:

$$
\begin{aligned}
&\hat{\beta}^\mathrm{T} C_1 \left(C_1^\mathrm{T} (X^\mathrm{T} V^{-1} X)^{-1} C_1 \right)^{-1} C_1^\mathrm{T} \hat{\beta} \\
&\overset{(27)}{=} \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} - y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \\
&\overset{(28)}{=} y^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y - y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \\
&= y^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y - y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \\
&= y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y - 2 y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \\
&- y^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y + 2 y^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
&\overset{(29)}{=} y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y - 2 y^\mathrm{T} V^{-1} X_0 (X_0^\mathrm{T} V^{-1} X_0)^{-1} X_0^\mathrm{T} V^{-1} y \\
&- y^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y + 2 y^\mathrm{T} V^{-1} X (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\
&\overset{(28)}{=} \hat{\beta}_0^\mathrm{T} X_0^\mathrm{T} V^{-1} X_0 \hat{\beta}_0 - 2 y^\mathrm{T} V^{-1} X_0 \hat{\beta}_0 - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} + 2 y^\mathrm{T} V^{-1} X \hat{\beta} \\
&= y^\mathrm{T} V^{-1} y - 2 y^\mathrm{T} V^{-1} X_0 \hat{\beta}_0 + \hat{\beta}_0^\mathrm{T} X_0^\mathrm{T} V^{-1} X_0 \hat{\beta}_0 - y^\mathrm{T} V^{-1} y + 2 y^\mathrm{T} V^{-1} X \hat{\beta} - \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} \\
&= \left(y^\mathrm{T} V^{-1} y - 2 y^\mathrm{T} V^{-1} X_0 \hat{\beta}_0 + \hat{\beta}_0^\mathrm{T} X_0^\mathrm{T} V^{-1} X_0 \hat{\beta}_0 \right) - \left(y^\mathrm{T} V^{-1} y - 2 y^\mathrm{T} V^{-1} X \hat{\beta} + \hat{\beta}^\mathrm{T} X^\mathrm{T} V^{-1} X \hat{\beta} \right) \\
&= (y - X_0 \hat{\beta}_0)^\mathrm{T} V^{-1} (y - X_0 \hat{\beta}_0) - (y - X \hat{\beta})^\mathrm{T} V^{-1} (y - X \hat{\beta}) \\
&\overset{(6)}{=} \hat{\varepsilon}_0^\mathrm{T} V^{-1} \hat{\varepsilon}_0 - \hat{\varepsilon}^\mathrm{T} V^{-1} \hat{\varepsilon} \; .
\end{aligned} \tag{30}
$$

With that, it is shown that (14) is true which, together with (13), finally demonstrates that the F-value in (12) is equal to the test statistic given by (3). This completes the proof.

■

Sources:
- Ostwald, Dirk (2023): "F-Statistiken"; in: *Allgemeines Lineares Modell*, Einheit (8), Folien 20, 24; URL: https://www.ipsy.ovgu.de/ipsy_media/Methodenlehre+I/Sommersemester+2023/Allgemeines+Lineares+Modell/8_F_Statistiken-p-9972.pdf.

1.5.31 Deviance function

Theorem: Consider a linear regression model (→ III/1.5.1) m with correlation structure (→ I/1.14.5) V

$$m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}$$

Then, the deviance (→ IV/??) for this model is

$$D(\beta, \sigma^2) = \mathrm{RSS}/\sigma^2 + n \cdot \left[\log(\sigma^2) + \log(2\pi) \right] \tag{2}$$

under uncorrelated observations (→ III/1.5.1), i.e. if $V = I_n$, and

$$D(\beta, \sigma^2) = \mathrm{wRSS}/\sigma^2 + n \cdot \left[\log(\sigma^2) + \log(2\pi) \right] + \log |V| \; , \tag{3}$$

in the general case, i.e. if $V \neq I_n$, where RSS is the residual sum of squares (→ III/1.5.9) and wRSS is the weighted residual sum of squares (→ III/1.5.22).

Proof: The likelihood function (→ I/5.1.2) for multiple linear regression is given by (→ III/1.5.23)

$$\begin{split} p(y|\beta, \sigma^2) &= \mathcal{N}(y; X\beta, \sigma^2 V) \\ &= \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \cdot \exp\left[-\frac{1}{2} (y - X\beta)^\mathrm{T} (\sigma^2 V)^{-1} (y - X\beta) \right] \; , \end{split} \tag{4}$$

such that, with $|\sigma^2 V| = (\sigma^2)^n |V|$, the log-likelihood function (→ I/4.1.2) for this model becomes (→ III/1.5.23)

$$\begin{split} \mathrm{LL}(\beta, \sigma^2) &= \log p(y|\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |V| - \frac{1}{2\sigma^2} (y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) \; . \end{split} \tag{5}$$

The last term can be expressed in terms of the (weighted) residual sum of squares (→ III/1.5.9) as

$$\begin{split} -\frac{1}{2\sigma^2} (y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) &= -\frac{1}{2\sigma^2} (Wy - WX\beta)^\mathrm{T} (Wy - WX\beta) \\ &= -\frac{1}{2\sigma^2} \left(\frac{1}{n} \sum_{i=1}^{n} (W\varepsilon)_i^2 \right) = -\frac{\mathrm{wRSS}}{2\sigma^2} \end{split} \tag{6}$$

where $W = V^{-1/2}$. Plugging (6) into (5) and multiplying with -2, we obtain the deviance (→ IV/??) as

$$
\begin{aligned}
D(\beta, \sigma^2) &= -2\,\mathrm{LL}(\beta, \sigma^2) \\
&= -2 \left(-\frac{\mathrm{wRSS}}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V| \right) \\
&= \mathrm{wRSS}/\sigma^2 + n \cdot \left[\log(\sigma^2) + \log(2\pi)\right] + \log|V|
\end{aligned} \tag{7}
$$

which proves the result in (3). Assuming $V = I_n$, we have

$$
\begin{aligned}
-\frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) &= -\frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T}(y - X\beta) \\
&= -\frac{1}{2\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \right) = -\frac{\mathrm{RSS}}{2\sigma^2}
\end{aligned} \tag{8}
$$

and

$$
\frac{1}{2}\log|V| = \frac{1}{2}\log|I_n| = \frac{1}{2}\log 1 = 0 \; , \tag{9}
$$

such that

$$
D(\beta, \sigma^2) = \mathrm{RSS}/\sigma^2 + n \cdot \left[\log(\sigma^2) + \log(2\pi)\right] \tag{10}
$$

which proves the result in (2). This completes the proof.

■

1.5.32 Akaike information criterion

Theorem: Consider a linear regression model (→ III/1.5.1) m

$$
m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}
$$

Then, the Akaike information criterion (→ IV/**??**) for this model is

$$
\mathrm{AIC}(m) = n \log\left(\frac{\mathrm{wRSS}}{n} \right) + n\left[1 + \log(2\pi)\right] + \log|V| + 2(p+1) \tag{2}
$$

where wRSS is the weighted residual sum of squares (→ III/1.5.9), p is the number of regressors (→ III/1.5.1) in the design matrix X and n is the number of observations (→ III/1.5.1) in the data vector y.

Proof: The Akaike information criterion (→ IV/**??**) is defined as

$$
\mathrm{AIC}(m) = -2\,\mathrm{MLL}(m) + 2\,k \tag{3}
$$

where $\mathrm{MLL}(m)$ is the maximum log-likelihood (→ I/4.1.4) is k is the number of free parameters in m.
The maximum log-likelihood for multiple linear regression (→ III/1.5.24) is given by

$$
\mathrm{MLL}(m) = -\frac{n}{2}\log\left(\frac{\mathrm{wRSS}}{n} \right) - \frac{n}{2}\left[1 + \log(2\pi)\right] - \frac{1}{2}\log|V| \tag{4}
$$

and the number of free paramters in multiple linear regression (→ III/1.5.1) is $k = p + 1$, i.e. one for each regressor in the design matrix (→ III/1.5.1) X, plus one for the noise variance (→ III/1.5.1) σ^2.
Thus, the AIC of m follows from (3) and (4) as

$$\mathrm{AIC}(m) = n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + 2(p+1) \; . \tag{5}$$

■

Sources:
- Claeskens G, Hjort NL (2008): "Akaike's information criterion"; in: *Model Selection and Model Averaging*, ex. 2.2, p. 66; URL: https://www.cambridge.org/core/books/model-selection-and-model-averaging E6F1EC77279D1223423BB64FC3A12C37; DOI: 10.1017/CBO9780511790485.

1.5.33 Bayesian information criterion

Theorem: Consider a linear regression model (→ III/1.5.1) m

$$m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}$$

Then, the Bayesian information criterion (→ IV/??) for this model is

$$\mathrm{BIC}(m) = n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + \log(n)\,(p+1) \tag{2}$$

where wRSS is the weighted residual sum of squares (→ III/1.5.9), p is the number of regressors (→ III/1.5.1) in the design matrix X and n is the number of observations (→ III/1.5.1) in the data vector y.

Proof: The Bayesian information criterion (→ IV/??) is defined as

$$\mathrm{BIC}(m) = -2\,\mathrm{MLL}(m) + k\log(n) \tag{3}$$

where $\mathrm{MLL}(m)$ is the maximum log-likelihood (→ I/4.1.4), k is the number of free parameters in m and n is the number of observations.
The maximum log-likelihood for multiple linear regression (→ III/1.5.24) is given by

$$\mathrm{MLL}(m) = -\frac{n}{2}\log\left(\frac{\mathrm{wRSS}}{n}\right) - \frac{n}{2}\left[1 + \log(2\pi)\right] - \frac{1}{2}\log|V| \tag{4}$$

and the number of free paramters in multiple linear regression (→ III/1.5.1) is $k = p + 1$, i.e. one for each regressor in the design matrix (→ III/1.5.1) X, plus one for the noise variance (→ III/1.5.1) σ^2.
Thus, the BIC of m follows from (3) and (4) as

$$\mathrm{BIC}(m) = n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + \log(n)\,(p+1) \; . \tag{5}$$

■

1.5.34 Corrected Akaike information criterion

Theorem: Consider a linear regression model (→ III/1.5.1) m

$$m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}$$

Then, the corrected Akaike information criterion (→ IV/**??**) for this model is

$$\mathrm{AIC_c}(m) = n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + \frac{2\,n\,(p+1)}{n-p-2} \tag{2}$$

where wRSS is the weighted residual sum of squares (→ III/1.5.9), p is the number of regressors (→ III/1.5.1) in the design matrix X and n is the number of observations (→ III/1.5.1) in the data vector y.

Proof: The corrected Akaike information criterion (→ IV/**??**) is defined as

$$\mathrm{AIC_c}(m) = \mathrm{AIC}(m) + \frac{2k^2 + 2k}{n-k-1} \tag{3}$$

where $\mathrm{AIC}(m)$ is the Akaike information criterion (→ IV/**??**), k is the number of free parameters in m and n is the number of observations.
The Akaike information criterion for multiple linear regression (→ III/1.5.24) is given by

$$\mathrm{AIC}(m) = n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + 2(p+1) \tag{4}$$

and the number of free paramters in multiple linear regression (→ III/1.5.1) is $k = p+1$, i.e. one for each regressor in the design matrix (→ III/1.5.1) X, plus one for the noise variance (→ III/1.5.1) σ^2.
Thus, the corrected AIC of m follows from (3) and (4) as

$$\begin{split}
\mathrm{AIC_c}(m) &= n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + 2\,k + \frac{2k^2 + 2k}{n-k-1} \\
&= n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + \frac{2nk - 2k^2 - 2k}{n-k-1} + \frac{2k^2 + 2k}{n-k-1} \\
&= n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + \frac{2nk}{n-k-1} \\
&= n \log\left(\frac{\mathrm{wRSS}}{n}\right) + n\left[1 + \log(2\pi)\right] + \log|V| + \frac{2\,n\,(p+1)}{n-p-2}
\end{split} \tag{5}$$

.

■

Sources:

- Claeskens G, Hjort NL (2008): "Akaike's information criterion"; in: *Model Selection and Model Averaging*, ex. 2.5, p. 67; URL: https://www.cambridge.org/core/books/model-selection-and-model-averaging E6F1EC77279D1223423BB64FC3A12C37; DOI: 10.1017/CBO9780511790485.

1.6 Bayesian linear regression

1.6.1 Conjugate prior distribution

Theorem: Let

$$y = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V as well as unknown $p \times 1$ regression coefficients β and unknown noise variance σ^2.
Then, the conjugate prior (→ I/5.2.5) for this model is a normal-gamma distribution (→ II/4.3.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{2}$$

where $\tau = 1/\sigma^2$ is the inverse noise variance or noise precision.

Proof: By definition, a conjugate prior (→ I/5.2.5) is a prior distribution (→ I/5.1.3) that, when combined with the likelihood function (→ I/5.1.2), leads to a posterior distribution (→ I/5.1.7) that belongs to the same family of probability distributions (→ I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.
Equation (1) implies the following likelihood function (→ I/5.1.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T} V^{-1} (y - X\beta)\right] \tag{3}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] \tag{4}$$

using the noise precision $\tau = 1/\sigma^2$ and the $n \times n$ precision matrix $P = V^{-1}$.

Seperating constant and variable terms, we have:

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] . \tag{5}$$

Expanding the product in the exponent, we have:

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta\right)\right] . \tag{6}$$

Completing the square over β, finally gives

$$p(y|\beta, \tau) = \sqrt{\frac{|P|}{(2\pi)^n}} \cdot \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left((\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} P X (\beta - \tilde{X}y) - y^\mathrm{T} Q y + y^\mathrm{T} P y\right)\right] \tag{7}$$

where $\tilde{X} = \left(X^\mathrm{T} P X\right)^{-1} X^\mathrm{T} P$ and $Q = \tilde{X}^\mathrm{T} \left(X^\mathrm{T} P X\right) \tilde{X}$.

In other words, the likelihood function (→ I/5.1.2) is proportional to a power of τ, times an exponential of τ and an exponential of a squared form of β, weighted by τ:

$$p(y|\beta, \tau) \propto \tau^{n/2} \cdot \exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} Q y\right)\right] \cdot \exp\left[-\frac{\tau}{2}(\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} P X (\beta - \tilde{X}y)\right] \; . \tag{8}$$

The same is true for a normal-gamma distribution (→ II/4.3.1) over β and τ

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{9}$$

the probability density function of which (→ II/4.3.3)

$$p(\beta, \tau) = \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \tag{10}$$

exhibits the same proportionality

$$p(\beta, \tau) \propto \tau^{a_0 + p/2 - 1} \cdot \exp[-\tau b_0] \cdot \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \tag{11}$$

and is therefore conjugate relative to the likelihood.

■

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.12, eq. 3.112; URL: https://www.springer.com/gp/book/9780387310732.

1.6.2 Posterior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V as well as unknown $p \times 1$ regression coefficients β and unknown noise variance σ^2. Moreover, assume a normal-gamma prior distribution (→ III/1.6.1) over the model parameters β and $\tau = 1/\sigma^2$:

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a normal-gamma distribution (→ II/4.3.1)

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{4}$$

Proof: According to Bayes' theorem (→ I/5.3.1), the posterior distribution (→ I/5.1.7) is given by

$$p(\beta, \tau|y) = \frac{p(y|\beta, \tau)\, p(\beta, \tau)}{p(y)} \; . \tag{5}$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (→ I/5.1.9) to the numerator:

$$p(\beta, \tau|y) \propto p(y|\beta, \tau)\, p(\beta, \tau) = p(y, \beta, \tau) \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$p(y|\beta, \sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n |\sigma^2 V|}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T} V^{-1} (y - X\beta)\right] \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta, \tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] \tag{8}$$

using the noise precision $\tau = 1/\sigma^2$ and the $n \times n$ precision matrix (→ I/1.13.19) $P = V^{-1}$.

Combining the likelihood function (→ I/5.1.2) (8) with the prior distribution (→ I/5.1.3) (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{split} p(y, \beta, \tau) &= p(y|\beta, \tau)\, p(\beta, \tau) \\ &= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] \cdot \\ &\quad \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \cdot \\ &\quad \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \; . \end{split} \tag{9}$$

Collecting identical variables gives:

$$\begin{split} p(y, \beta, \tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left((y - X\beta)^\mathrm{T} P (y - X\beta) + (\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right)\right] \; . \end{split} \tag{10}$$

Expanding the products in the exponent gives:

$$\begin{split} p(y, \beta, \tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P| |\Lambda_0|} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0 - 1} \exp[-b_0 \tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta + \right.\right. \\ &\left.\left. \beta^\mathrm{T} \Lambda_0 \beta - \beta^\mathrm{T} \Lambda_0 \mu_0 - \mu_0^\mathrm{T} \Lambda_0 \beta + \mu_0^\mathrm{T} \Lambda_0 \mu_0\right)\right] \; . \end{split} \tag{11}$$

Completing the square over β, we finally have

$$\begin{split} p(y,\beta,\tau) = \sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}}|P||\Lambda_0|} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\ \exp\left[-\frac{\tau}{2}\left((\beta-\mu_n)^\mathrm{T}\Lambda_n(\beta-\mu_n) + (y^\mathrm{T} P y + \mu_0^\mathrm{T}\Lambda_0\mu_0 - \mu_n^\mathrm{T}\Lambda_n\mu_n)\right)\right] \end{split} \tag{12}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{split} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0\mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \; . \end{split} \tag{13}$$

Ergo, the joint likelihood is proportional to

$$p(y,\beta,\tau) \propto \tau^{p/2} \cdot \exp\left[-\frac{\tau}{2}(\beta-\mu_n)^\mathrm{T}\Lambda_n(\beta-\mu_n)\right] \cdot \tau^{a_n-1} \cdot \exp\left[-b_n\tau\right] \tag{14}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{split} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T}\Lambda_0\mu_0 - \mu_n^\mathrm{T}\Lambda_n\mu_n) \; . \end{split} \tag{15}$$

From the term in (14), we can isolate the posterior distribution over β given τ:

$$p(\beta|\tau,y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \; . \tag{16}$$

From the remaining term, we can isolate the posterior distribution over τ:

$$p(\tau|y) = \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{17}$$

Together, (16) and (17) constitute the joint (→ I/1.3.2) posterior distribution (→ I/5.1.7) of β and τ.

■

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.12, eq. 3.113; URL: https://www.springer.com/gp/book/9780387310732.

1.6.3 Log model evidence

Theorem: Let

$$m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V as well as unknown $p \times 1$ regression coefficients β and unknown noise variance σ^2. Moreover, assume a normal-gamma prior distribution (→ III/1.6.1) over the model parameters β and $\tau = 1/\sigma^2$:

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the log model evidence (→ IV/**??**) for this model is

$$\begin{split} \log p(y|m) = \frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| + \\ \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n \end{split} \tag{3}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{split} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T}Py + \Lambda_0\mu_0) \\ \Lambda_n &= X^\mathrm{T}PX + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T}Py + \mu_0^\mathrm{T}\Lambda_0\mu_0 - \mu_n^\mathrm{T}\Lambda_n\mu_n) \; . \end{split} \tag{4}$$

Proof: According to the law of marginal probability (→ I/1.3.3), the model evidence (→ I/5.1.11) for this model is:

$$p(y|m) = \iint p(y|\beta,\tau)\, p(\beta,\tau) \, \mathrm{d}\beta \, \mathrm{d}\tau \; . \tag{5}$$

According to the law of conditional probability (→ I/1.3.4), the integrand is equivalent to the joint likelihood (→ I/5.1.5):

$$p(y|m) = \iint p(y,\beta,\tau) \, \mathrm{d}\beta \, \mathrm{d}\tau \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$p(y|\beta,\sigma^2) = \mathcal{N}(y; X\beta, \sigma^2 V) = \sqrt{\frac{1}{(2\pi)^n|\sigma^2 V|}} \exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^\mathrm{T}V^{-1}(y-X\beta)\right] \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$p(y|\beta,\tau) = \mathcal{N}(y; X\beta, (\tau P)^{-1}) = \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y-X\beta)^\mathrm{T}P(y-X\beta)\right] \tag{8}$$

using the noise precision $\tau = 1/\sigma^2$ and the $n \times n$ precision matrix $P = V^{-1}$.

When deriving the posterior distribution (→ III/1.6.2) $p(\beta,\tau|y)$, the joint likelihood $p(y,\beta,\tau)$ is obtained as

$$\begin{split} p(y,\beta,\tau) = \sqrt{\frac{\tau^n|P|}{(2\pi)^n}} \sqrt{\frac{\tau^p|\Lambda_0|}{(2\pi)^p}} \frac{{b_0}^{a_0}}{\Gamma(a_0)} \tau^{a_0-1}\exp[-b_0\tau] \cdot \\ \exp\left[-\frac{\tau}{2}\left((\beta-\mu_n)^T\Lambda_n(\beta-\mu_n) + (y^TPy + \mu_0^T\Lambda_0\mu_0 - \mu_n^T\Lambda_n\mu_n)\right)\right] \; . \end{split} \tag{9}$$

Using the probability density function of the multivariate normal distribution (→ II/4.1.7), we can rewrite this as

$$
\begin{split}
p(y,\beta,\tau) = &\sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \sqrt{\frac{(2\pi)^p}{\tau^p |\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \, \exp\left[-\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] \; .
\end{split} \tag{10}
$$

Now, β can be integrated out easily:

$$
\begin{split}
\int p(y,\beta,\tau) \, \mathrm{d}\beta = &\sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\exp\left[-\frac{\tau}{2} (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n) \right] \; .
\end{split} \tag{11}
$$

Using the probability density function of the gamma distribution (→ II/3.4.7), we can rewrite this as

$$
\int p(y,\beta,\tau) \, \mathrm{d}\beta = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \frac{\Gamma(a_n)}{{b_n}^{a_n}} \, \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{12}
$$

Finally, τ can also be integrated out:

$$
\iint p(y,\beta,\tau) \, \mathrm{d}\beta \, \mathrm{d}\tau = \sqrt{\frac{|P|}{(2\pi)^n}} \sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}} \, \frac{\Gamma(a_n)}{\Gamma(a_0)} \, \frac{{b_0}^{a_0}}{{b_n}^{a_n}} = p(y|m) \; . \tag{13}
$$

Thus, the log model evidence (→ IV/**??**) of this model is given by

$$
\begin{split}
\log p(y|m) = &\frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\Lambda_0| - \frac{1}{2} \log |\Lambda_n| + \\
&\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; .
\end{split} \tag{14}
$$

■

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, ex. 3.23, eq. 3.118; URL: https://www.springer.com/gp/book/9780387310732.

1.6.4 Accuracy and complexity

Theorem: Let

$$
m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}
$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X, known $n \times n$ covariance structure V as well as unknown $p \times 1$ regression coefficients β and unknown noise variance σ^2. Moreover, assume a normal-gamma prior distribution (→ III/1.6.1) over the model parameters β and $\tau = 1/\sigma^2$:

$$p(\beta,\tau) = \mathcal{N}(\beta;\mu_0,(\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau;a_0,b_0) \; . \tag{2}$$

Then, accuracy and complexity (→ IV/**??**) of this model are

$$\begin{aligned} \mathrm{Acc}(m) = &-\frac{1}{2}\frac{a_n}{b_n}(y-X\mu_n)^\mathrm{T} P (y-X\mu_n) - \frac{1}{2}\mathrm{tr}(X^\mathrm{T} P X \Lambda_n^{-1}) \\ &+\frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{n}{2}(\psi(a_n) - \log(b_n)) \\ \mathrm{Com}(m) = &\;\frac{1}{2}\frac{a_n}{b_n}\left[(\mu_0-\mu_n)^\mathrm{T}\Lambda_0(\mu_0-\mu_n) - 2(b_n-b_0)\right] + \frac{1}{2}\mathrm{tr}(\Lambda_0\Lambda_n^{-1}) - \frac{1}{2}\log\frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\ &+ a_0\log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n-a_0)\psi(a_n) \; . \end{aligned} \tag{3}$$

where μ_n, Λ_n, a_n and b_n are the posterior hyperparameters for Bayesian linear regression (→ III/1.6.2) and P is the data precision matrix (→ I/1.13.19): $P = V^{-1}$.

Proof: Model accuracy and complexity are defined as (→ IV/**??**)

$$\begin{aligned} \mathrm{LME}(m) &= \mathrm{Acc}(m) - \mathrm{Com}(m) \\ \mathrm{Acc}(m) &= \langle \log p(y|\beta,\tau,m) \rangle_{p(\beta,\tau|y,m)} \\ \mathrm{Com}(m) &= \mathrm{KL}\left[p(\beta,\tau|y,m)\,||\,p(\beta,\tau|m)\right] \; . \end{aligned} \tag{4}$$

1) The accuracy term is the expectation (→ I/1.10.1) of the log-likelihood function (→ I/4.1.2) $\log p(y|\beta,\tau)$ with respect to the posterior distribution (→ I/5.1.7) $p(\beta,\tau|y)$. This expectation can be rewritten as:

$$\begin{aligned} \mathrm{Acc}(m) &= \iint p(\beta,\tau|y)\; \log p(y|\beta,\tau)\, \mathrm{d}\beta\, \mathrm{d}\tau \\ &= \int p(\tau|y) \int p(\beta|\tau,y)\; \log p(y|\beta,\tau)\, \mathrm{d}\beta\, \mathrm{d}\tau \\ &= \left\langle \langle \log p(y|\beta,\tau) \rangle_{p(\beta|\tau,y)} \right\rangle_{p(\tau|y)} \; . \end{aligned} \tag{5}$$

With the log-likelihood function for multiple linear regression (→ III/1.5.23), we have:

$$\begin{aligned} \mathrm{Acc}(m) &= \left\langle \left\langle \log\left(\sqrt{\frac{1}{(2\pi)^n|\sigma^2 V|}} \cdot \exp\left[-\frac{1}{2}(y-X\beta)^\mathrm{T}(\sigma^2 V)^{-1}(y-X\beta)\right] \right) \right\rangle_{p(\beta|\tau,y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \log\left(\sqrt{\frac{\tau^n|P|}{(2\pi)^n}} \cdot \exp\left[-\frac{1}{2}(y-X\beta)^\mathrm{T}(\tau P)(y-X\beta)\right] \right) \right\rangle_{p(\beta|\tau,y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \frac{1}{2}\log|P| + \frac{n}{2}\log\tau - \frac{n}{2}\log(2\pi) - \frac{1}{2}(y-X\beta)^\mathrm{T}(\tau P)(y-X\beta) \right\rangle_{p(\beta|\tau,y)} \right\rangle_{p(\tau|y)} \\ &= \left\langle \left\langle \frac{1}{2}\log|P| + \frac{n}{2}\log\tau - \frac{n}{2}\log(2\pi) - \frac{\tau}{2}\left[y^\mathrm{T} P y - 2y^\mathrm{T} P X\beta + \beta^\mathrm{T} X^\mathrm{T} P X \beta\right] \right\rangle_{p(\beta|\tau,y)} \right\rangle_{p(\tau|y)} \; . \end{aligned} \tag{6}$$

With the posterior distribution for Bayesian linear regression (→ III/1.6.2), this becomes:

$$\mathrm{Acc}(m) = \left\langle \left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} \left[y^\mathrm{T} P y - 2 y^\mathrm{T} P X \beta + \beta^\mathrm{T} X^\mathrm{T} P X \beta \right] \right\rangle_{\mathcal{N}(\beta;\mu_n,(\tau\Lambda_n)^{-1})} \right\rangle_{\mathrm{Gam}(\tau;a_n,b_n)} \tag{7}$$

If $x \sim \mathrm{N}(\mu, \Sigma)$, then its expected value is (→ II/4.1.9)

$$\langle x \rangle = \mu \tag{8}$$

and the expectation of a quadratic form is given by (→ I/1.10.9)

$$\left\langle x^\mathrm{T} A x \right\rangle = \mu^\mathrm{T} A \mu + \mathrm{tr}(A\Sigma) \; . \tag{9}$$

Thus, the model accuracy of m evaluates to:

$$\begin{split} \mathrm{Acc}(m) = &\left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \right. \\ &\left. \quad \frac{\tau}{2} \left[y^\mathrm{T} P y - 2 y^\mathrm{T} P X \mu_n + \mu_n^\mathrm{T} X^\mathrm{T} P X \mu_n + \frac{1}{\tau} \mathrm{tr}(X^\mathrm{T} P X \Lambda_n^{-1}) \right] \right\rangle_{\mathrm{Gam}(\tau;a_n,b_n)} \\ = &\left\langle \frac{1}{2} \log |P| + \frac{n}{2} \log \tau - \frac{n}{2} \log(2\pi) - \frac{\tau}{2} (y - X\mu_n)^\mathrm{T} P (y - X\mu_n) - \frac{1}{2} \mathrm{tr}(X^\mathrm{T} P X \Lambda_n^{-1}) \right\rangle_{\mathrm{Gam}(\tau;a_n,b_n)} \; . \end{split} \tag{10}$$

If $x \sim \mathrm{Gam}(a, b)$, then its expected value is (→ II/3.4.11)

$$\langle x \rangle = \frac{a}{b} \tag{11}$$

and its logarithmic expectation is given by (→ II/3.4.13)

$$\langle \log x \rangle = \psi(a) - \log(b) \; . \tag{12}$$

Thus, the model accuracy of m evaluates to

$$\begin{split} \mathrm{Acc}(m) = &- \frac{1}{2} \frac{a_n}{b_n} (y - X\mu_n)^\mathrm{T} P (y - X\mu_n) - \frac{1}{2} \mathrm{tr}(X^\mathrm{T} P X \Lambda_n^{-1}) \\ &+ \frac{1}{2} \log |P| - \frac{n}{2} \log(2\pi) + \frac{n}{2} (\psi(a_n) - \log(b_n)) \end{split} \tag{13}$$

which proofs the first part of (3).

2) The complexity penalty is the Kullback-Leibler divergence (→ I/2.5.1) of the posterior distribution (→ I/5.1.7) $p(\beta, \tau|y)$ from the prior distribution (→ I/5.1.3) $p(\beta, \tau)$. This can be rewritten as follows:

$$\begin{split} \mathrm{Com}(m) &= \iint p(\beta, \tau|y) \, \log \frac{p(\beta, \tau|y)}{p(\beta, \tau)} \, \mathrm{d}\beta \, \mathrm{d}\tau \\ &= \iint p(\beta|\tau, y) \, p(\tau|y) \, \log \left[\frac{p(\beta|\tau, y)}{p(\beta|\tau)} \, \frac{p(\tau|y)}{p(\tau)} \right] \, \mathrm{d}\beta \, \mathrm{d}\tau \\ &= \int p(\tau|y) \int p(\beta|\tau, y) \, \log \frac{p(\beta|\tau, y)}{p(\beta|\tau)} \, \mathrm{d}\beta \, \mathrm{d}\tau + \int p(\tau|y) \, \log \frac{p(\tau|y)}{p(\tau)} \int p(\beta|\tau, y) \, \mathrm{d}\beta \, \mathrm{d}\tau \\ &= \left\langle \mathrm{KL} \left[p(\beta|\tau, y) \, || \, p(\beta|\tau) \right] \right\rangle_{p(\tau|y)} + \mathrm{KL} \left[p(\tau|y) \, || \, p(\tau) \right] \; . \end{split} \tag{14}$$

With the prior distribution (→ III/1.6.1) given by (2) and the posterior distribution for Bayesian linear regression (→ III/1.6.2), this becomes:

$$\begin{aligned} \mathrm{Com}(m) &= \left\langle \mathrm{KL}\left[\mathcal{N}(\beta;\mu_n,(\tau\Lambda_n)^{-1})\,||\,\mathcal{N}(\beta;\mu_0,(\tau\Lambda_0)^{-1})\right]\right\rangle_{\mathrm{Gam}(\tau;a_n,b_n)} \\ &\quad + \mathrm{KL}\left[\mathrm{Gam}(\tau;a_n,b_n)\,||\,\mathrm{Gam}(\tau;a_0,b_0)\right] \; . \end{aligned} \tag{15}$$

With the Kullback-Leibler divergence for the multivariate normal distribution (→ II/4.1.12)

$$\mathrm{KL}[\mathcal{N}(\mu_1,\Sigma_1)\,||\,\mathcal{N}(\mu_2,\Sigma_2)] = \frac{1}{2}\left[(\mu_2-\mu_1)^\mathrm{T}\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \tag{16}$$

and the Kullback-Leibler divergence for the gamma distribution (→ II/3.4.16)

$$\mathrm{KL}[\mathrm{Gam}(a_1,b_1)\,||\,\mathrm{Gam}(a_2,b_2)] = a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \; , \tag{17}$$

the model complexity of m evaluates to:

$$\begin{aligned} \mathrm{Com}(m) &= \left\langle \frac{1}{2}\left[(\mu_0-\mu_n)^\mathrm{T}(\tau\Lambda_0)(\mu_0-\mu_n) + \mathrm{tr}((\tau\Lambda_0)(\tau\Lambda_n)^{-1}) - \log\frac{|(\tau\Lambda_n)^{-1}|}{|(\tau\Lambda_0)^{-1}|} - p\right]\right\rangle_{p(\tau|y)} \\ &\quad + a_0\,\log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n-a_0)\,\psi(a_n) - (b_n-b_0)\,\frac{a_n}{b_n} \; . \end{aligned} \tag{18}$$

Using $x \sim \mathrm{Gam}(a,b) \Rightarrow \langle x \rangle = a/b$ from (11) again, it follows that

$$\begin{aligned} \mathrm{Com}(m) &= \frac{1}{2}\frac{a_n}{b_n}\left[(\mu_0-\mu_n)^\mathrm{T}\Lambda_0(\mu_0-\mu_n)\right] + \frac{1}{2}\mathrm{tr}(\Lambda_0\Lambda_n^{-1}) - \frac{1}{2}\log\frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\ &\quad + a_0\,\log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n-a_0)\,\psi(a_n) - (b_n-b_0)\,\frac{a_n}{b_n} \; . \end{aligned} \tag{19}$$

Thus, the model complexity of m evaluates to

$$\begin{aligned} \mathrm{Com}(m) &= \frac{1}{2}\frac{a_n}{b_n}\left[(\mu_0-\mu_n)^\mathrm{T}\Lambda_0(\mu_0-\mu_n) - 2(b_n-b_0)\right] + \frac{1}{2}\mathrm{tr}(\Lambda_0\Lambda_n^{-1}) - \frac{1}{2}\log\frac{|\Lambda_0|}{|\Lambda_n|} - \frac{p}{2} \\ &\quad + a_0\log\frac{b_n}{b_0} - \log\frac{\Gamma(a_n)}{\Gamma(a_0)} + (a_n-a_0)\psi(a_n) \end{aligned} \tag{20}$$

which proofs the second part of (3).

3) A control calculation confirms that

$$\mathrm{Acc}(m) - \mathrm{Com}(m) = \mathrm{LME}(m) \tag{21}$$

where $\mathrm{LME}(m)$ is the log model evidence for Bayesian linear regression (→ III/1.6.3):

$$\begin{aligned} \log p(y|m) = \frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| + \\ \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n \end{aligned} \tag{22}$$

This requires to recognize that

$$-\frac{1}{2}\mathrm{tr}(X^\mathrm{T} P X \Lambda_n^{-1}) - \frac{1}{2}\mathrm{tr}(\Lambda_0 \Lambda_n^{-1}) + \frac{p}{2} = 0 \tag{23}$$

and

$$\frac{n}{2}(\psi(a_n) - \log(b_n)) - a_0 \log \frac{b_n}{b_0} - (a_n - a_0)\psi(a_n) = a_0 \log b_0 - a_n \log b_n \tag{24}$$

thanks to the nature of the posterior hyperparameters for Bayesian linear regression (→ III/1.6.2).

■

Sources:

- Soch J, Allefeld A (2016): "Kullback-Leibler Divergence for the Normal-Gamma Distribution"; in: *arXiv math.ST*, 1611.01437, eqs. 23/30; URL: https://arxiv.org/abs/1611.01437.
- Soch J, Allefeld C, Haynes JD (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection"; in: *NeuroImage*, vol. 141, pp. 469-489, Appendix C; URL: https://www.sciencedirect.com/science/article/pii/S1053811916303615; DOI: 10.1016/j.neuroimage
- Soch J (2018): "cvBMS and cvBMA: filling in the gaps"; in: *arXiv stat.ME*, sect. 2.2, eqs. 8-24; URL: https://arxiv.org/abs/1807.01585; DOI: 10.48550/arXiv.1807.01585.

1.6.5 Deviance information criterion

Theorem: Consider a linear regression model (→ III/1.5.1) m

$$m: \; y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \; \sigma^2 V = (\tau P)^{-1} \tag{1}$$

with a normal-gamma prior distribution (→ III/1.6.1)

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the deviance information criterion (→ IV/**??**) for this model is

$$\begin{split} \mathrm{DIC}(m) = n \cdot \log(2\pi) - n\left[2\psi(a_n) - \log(a_n) - \log(b_n)\right] - \log|P| \\ + \frac{a_n}{b_n}(y - X\mu_n)^\mathrm{T} P (y - X\mu_n) + \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1}\right) \end{split} \tag{3}$$

where μ_n and Λ_n as well as a_n and b_n are posterior parameters (→ I/5.1.7) describing the posterior distribution in Bayesian linear regression (→ III/1.6.2).

Proof: The deviance for multiple linear regression (→ III/1.5.31) is

$$D(\beta, \sigma^2) = n \cdot \log(2\pi) + n \cdot \log(\sigma^2) + \log|V| + \frac{1}{\sigma^2}(y - X\beta)^\mathrm{T} V^{-1} (y - X\beta) \tag{4}$$

which, applying the equalities $\tau = 1/\sigma^2$ and $P = V^{-1}$, becomes

$$D(\beta, \tau) = n \cdot \log(2\pi) - n \cdot \log(\tau) - \log|P| + \tau \cdot (y - X\beta)^\mathrm{T} P (y - X\beta) \; . \tag{5}$$

The deviance information criterion (→ IV/**??**) (DIC) is defined as

$$\mathrm{DIC}(m) = -2 \log p(y|\left\langle\beta\right\rangle, \left\langle\tau\right\rangle, m) + 2\, p_D \tag{6}$$

where $\log p(y|\langle\beta\rangle, \langle\tau\rangle, m)$ is the log-likelihood function (→ III/1.5.24) at the posterior expectations (→ I/1.10.1) and the "effective number of parameters" p_D is the difference between the expectation of the deviance and the deviance at the expectation (→ IV/**??**):

$$p_D = \langle D(\beta, \tau)\rangle - D(\langle\beta\rangle, \langle\tau\rangle) \; . \tag{7}$$

With that, the DIC for multiple linear regression becomes:

$$\begin{aligned} \mathrm{DIC}(m) &= -2\log p(y|\langle\beta\rangle, \langle\tau\rangle, m) + 2\, p_D \\ &= D(\langle\beta\rangle, \langle\tau\rangle) + 2\left[\langle D(\beta, \tau)\rangle - D(\langle\beta\rangle, \langle\tau\rangle)\right] \\ &= 2\,\langle D(\beta, \tau)\rangle - D(\langle\beta\rangle, \langle\tau\rangle) \; . \end{aligned} \tag{8}$$

The posterior distribution for multiple linear regression (→ III/1.6.2) is

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{9}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; . \end{aligned} \tag{10}$$

Thus, we have the following posterior expectations:

$$\langle\beta\rangle_{\beta,\tau|y} = \mu_n \tag{11}$$

$$\langle\tau\rangle_{\beta,\tau|y} = \frac{a_n}{b_n} \tag{12}$$

$$\langle\log\tau\rangle_{\beta,\tau|y} = \psi(a_n) - \log(b_n) \tag{13}$$

$$\begin{aligned} \left\langle\beta^\mathrm{T} A \beta\right\rangle_{\beta|\tau,y} &= \mu_n^\mathrm{T} A \mu_n + \mathrm{tr}\left(A(\tau\Lambda_n)^{-1}\right) \\ &= \mu_n^\mathrm{T} A \mu_n + \frac{1}{\tau}\mathrm{tr}\left(A\Lambda_n^{-1}\right) \; . \end{aligned} \tag{14}$$

In these identities, we have used the mean of the multivariate normal distribution (→ II/4.1.9), the mean of the gamma distribution (→ II/3.4.11), the logarithmic expectation of the gamma distribution (→ II/3.4.13), the expectation of a quadratic form (→ I/1.10.9) and the covariance of the multivariate normal distribution (→ II/4.1.10).
With that, the deviance at the expectation is:

$$\begin{aligned} D(\langle\beta\rangle, \langle\tau\rangle) &\overset{(5)}{=} n\cdot\log(2\pi) - n\cdot\log(\langle\tau\rangle) - \log|P| + \tau\cdot(y - X\langle\beta\rangle)^\mathrm{T} P (y - X\langle\beta\rangle) \\ &\overset{(11)}{=} n\cdot\log(2\pi) - n\cdot\log(\langle\tau\rangle) - \log|P| + \tau\cdot(y - X\mu_n)^\mathrm{T} P (y - X\mu_n) \\ &\overset{(12)}{=} n\cdot\log(2\pi) - n\cdot\log\left(\frac{a_n}{b_n}\right) - \log|P| + \frac{a_n}{b_n}\cdot(y - X\mu_n)^\mathrm{T} P (y - X\mu_n) \; . \end{aligned} \tag{15}$$

Moreover, the expectation of the deviance is:

$$
\begin{aligned}
\langle D(\beta,\tau) \rangle &\overset{(5)}{=} \left\langle n \cdot \log(2\pi) - n \cdot \log(\tau) - \log|P| + \tau \cdot (y-X\beta)^\mathrm{T} P (y-X\beta) \right\rangle \\
&= n \cdot \log(2\pi) - n \cdot \langle \log(\tau) \rangle - \log|P| + \left\langle \tau \cdot (y-X\beta)^\mathrm{T} P (y-X\beta) \right\rangle \\
&\overset{(13)}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log|P| \\
&+ \left\langle \tau \cdot \left\langle (y-X\beta)^\mathrm{T} P (y-X\beta) \right\rangle_{\beta|\tau,y} \right\rangle_{\tau|y} \\
&= n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log|P| \\
&+ \left\langle \tau \cdot \left\langle y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta \right\rangle_{\beta|\tau,y} \right\rangle_{\tau|y} \\
&\overset{(14)}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log|P| \\
&+ \left\langle \tau \cdot \left[y^\mathrm{T} P y - y^\mathrm{T} P X \mu_n - \mu_n^\mathrm{T} X^\mathrm{T} P y + \mu_n^\mathrm{T} X^\mathrm{T} P X \mu_n + \frac{1}{\tau} \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1} \right) \right] \right\rangle_{\tau|y} \\
&= n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log|P| \\
&+ \left\langle \tau \cdot (y-X\mu_n)^\mathrm{T} P (y-X\mu_n) \right\rangle_{\tau|y} + \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1} \right) \\
&\overset{(12)}{=} n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log|P| \\
&+ \frac{a_n}{b_n} \cdot (y-X\mu_n)^\mathrm{T} P (y-X\mu_n) + \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1} \right) \; .
\end{aligned} \tag{16}
$$

Finally, combining the two terms, we have:

$$
\begin{aligned}
\mathrm{DIC}(m) &\overset{(8)}{=} 2 \langle D(\beta,\tau) \rangle - D(\langle \beta \rangle, \langle \tau \rangle) \\
&\overset{(16)}{=} 2 \left[n \cdot \log(2\pi) - n \cdot [\psi(a_n) - \log(b_n)] - \log|P| \right. \\
&\left. + \frac{a_n}{b_n} \cdot (y-X\mu_n)^\mathrm{T} P (y-X\mu_n) + \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1} \right) \right] \\
&\overset{(15)}{-} \left[n \cdot \log(2\pi) - n \cdot \log\left(\frac{a_n}{b_n} \right) - \log|P| + \frac{a_n}{b_n} \cdot (y-X\mu_n)^\mathrm{T} P (y-X\mu_n) \right] \\
&= n \cdot \log(2\pi) - 2n\psi(a_n) + 2n\log(b_n) + n\log(a_n) - \log(b_n) - \log|P| \\
&+ \frac{a_n}{b_n} (y-X\mu_n)^\mathrm{T} P (y-X\mu_n) + \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1} \right) \\
&= n \cdot \log(2\pi) - n \left[2\psi(a_n) - \log(a_n) - \log(b_n) \right] - \log|P| \\
&+ \frac{a_n}{b_n} (y-X\mu_n)^\mathrm{T} P (y-X\mu_n) + \mathrm{tr}\left(X^\mathrm{T} P X \Lambda_n^{-1} \right) \; .
\end{aligned} \tag{17}
$$

This conforms to equation (3).

■

1.6.6 Maximum-a-posteriori estimation

Theorem: Let there be a linear regression model (→ III/1.5.1)

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \; \sigma^2 V = (\tau P)^{-1} \tag{1}
$$

and assume a normal-gamma prior distribution (→ III/1.6.1) over the model parameters β and $\tau = 1/\sigma^2$

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the maximum-a-posteriori estimates (→ I/5.1.8) of β and τ are

$$\begin{split} \hat{\beta}_{\mathrm{MAP}} &= (X^\mathrm{T} P X + \Lambda_0)^{-1} (X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \hat{\tau}_{\mathrm{MAP}} &= (2a_0 + n - 2) \left(2b_0 + (y - X\hat{\beta}_{\mathrm{MAP}})^\mathrm{T} P (y - X\hat{\beta}_{\mathrm{MAP}}) + (\hat{\beta}_{\mathrm{MAP}} - \mu_0)^\mathrm{T} \Lambda_0 (\hat{\beta}_{\mathrm{MAP}} - \mu_0) \right)^{-1} \end{split} \tag{3}$$

where n is the number of data points (→ III/1.5.1).

Proof: Given the prior distribution (→ I/5.1.3) in (2), the posterior distribution (→ I/5.1.7) for multiple linear regression (→ III/1.5.1) is also a normal-gamma distribution (→ III/1.6.2)

$$p(\beta, \tau | y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{4}$$

where the posterior hyperparameters (→ I/5.1.7) are equal to

$$\begin{split} \mu_n &= \Lambda_n^{-1} (X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2} (y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; . \end{split} \tag{5}$$

From this, the conditional posterior distribution over β follows as (→ II/4.3.1)

$$p(\beta | \tau, y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \tag{6}$$

and the marginal posterior distribution over τ follows as (→ II/4.3.1)

$$p(\tau | y) = \mathrm{Gam}(\tau; a_n, b_n) \; . \tag{7}$$

The mode of the multivariate normal distribution is given by

$$X \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \mathrm{mode}(X) = \mu \tag{8}$$

and the mode of the gamma distribution is given by

$$X \sim \mathrm{Gam}(a, b) \quad \Rightarrow \quad \mathrm{mode}(X) = \frac{a-1}{b} \; . \tag{9}$$

Applying (8) to (6), the maximum-a-posteriori estimate (→ I/5.1.8) of β follows as

$$\begin{split} \hat{\beta}_{\mathrm{MAP}} &= \mu_n \\ &= \Lambda_n^{-1} (X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ &= (X^\mathrm{T} P X + \Lambda_0)^{-1} (X^\mathrm{T} P y + \Lambda_0 \mu_0) \end{split} \tag{10}$$

and applying (9) to (7), the maximum-a-posteriori estimate (→ I/5.1.8) of τ follows as

$$
\begin{aligned}
\hat{\tau}_{\mathrm{MAP}} &= \frac{a_n - 1}{b_n} \\
&= \left(a_0 + \frac{n}{2} - 1\right) \left(b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n)\right)^{-1} \\
&= (2a_0 + n - 2) \left(2b_0 + y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n\right)^{-1} \\
&= (2a_0 + n - 2) \left(2b_0 + y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \hat{\beta}_{\mathrm{MAP}}^\mathrm{T} \left(X^\mathrm{T} P X + \Lambda_0\right) \hat{\beta}_{\mathrm{MAP}}\right)^{-1} \\
&= (2a_0 + n - 2) \left(2b_0 + y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \hat{\beta}_{\mathrm{MAP}}^\mathrm{T} X^\mathrm{T} P X \hat{\beta}_{\mathrm{MAP}} - \hat{\beta}_{\mathrm{MAP}}^\mathrm{T} \Lambda_0 \hat{\beta}_{\mathrm{MAP}}\right)^{-1} \\
&= (2a_0 + n - 2) \left(2b_0 + (y - X\hat{\beta}_{\mathrm{MAP}})^\mathrm{T} P (y - X\hat{\beta}_{\mathrm{MAP}}) + (\hat{\beta}_{\mathrm{MAP}} - \mu_0)^\mathrm{T} \Lambda_0 (\hat{\beta}_{\mathrm{MAP}} - \mu_0)\right)^{-1} .
\end{aligned} \tag{11}
$$

■

1.6.7 Expression of posterior parameters using error terms

Theorem: Let there be a linear regression model (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V), \; \sigma^2 V = (\tau P)^{-1} , \tag{1}$$

assume a normal-gamma prior distribution (→ III/1.6.1) over the model parameters β and $\tau = 1/\sigma^2$

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{2}$$

and consider the Bayesian posterior distribution (→ III/1.6.2) over these model parameters:

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) . \tag{3}$$

Then, the posterior hyperparameters (→ I/5.1.7) for the noise precision (→ III/1.6.1) τ can be expressed as

$$
\begin{aligned}
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}\left(\varepsilon_y^\mathrm{T} P \varepsilon_y + \varepsilon_\beta^\mathrm{T} \Lambda_0 \varepsilon_\beta\right)
\end{aligned} \tag{4}
$$

where ε_y and ε_β are the “prediction errors” and “parameter errors”

$$
\begin{aligned}
\varepsilon_y &= y - \hat{y} \\
\varepsilon_\beta &= \mu_n - \mu_0
\end{aligned} \tag{5}
$$

where $\hat{y}$ is the predicted signal (→ III/1.5.12) at the posterior mean (→ III/1.6.2) regression coefficients (→ III/1.5.1) μ_n:

$$\hat{y} = X\mu_n . \tag{6}$$

Proof: The posterior hyperparameter for Bayesian linear regression (→ III/1.6.2) are:

$$
\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{7}
$$

The shape parameter (→ II/3.4.1) a_n is given by this equation. The rate parameter (→ II/3.4.1) b_n of the posterior distribution (→ I/5.1.7) can be developped as follows:

$$
\begin{aligned}
b_n &\overset{(7)}{=} b_0 + \frac{1}{2}\left(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n\right) \\
&\overset{(7)}{=} b_0 + \frac{1}{2}\left(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} (X^\mathrm{T} P X + \Lambda_0) \mu_n\right) \\
&= b_0 + \frac{1}{2}\left(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} X^\mathrm{T} P X \mu_n - \mu_n^\mathrm{T} \Lambda_0 \mu_n\right) \\
&= b_0 + \frac{1}{2}\left((y^\mathrm{T} P y - \mu_n^\mathrm{T} X^\mathrm{T} P X \mu_n) + (\mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_0 \mu_n)\right) \\
&= b_0 + \frac{1}{2}\left((y - X\mu_n)^\mathrm{T} P (y - X\mu_n) + (\mu_0 - \mu_n)^\mathrm{T} \Lambda_0 (\mu_0 - \mu_n)\right) \\
&\overset{(6)}{=} b_0 + \frac{1}{2}\left((y - \hat{y})^\mathrm{T} P (y - \hat{y}) + (\mu_n - \mu_0)^\mathrm{T} \Lambda_0 (\mu_n - \mu_0)\right) \\
&\overset{(5)}{=} b_0 + \frac{1}{2}\left(\varepsilon_y^\mathrm{T} P \varepsilon_y + \varepsilon_\beta^\mathrm{T} \Lambda_0 \varepsilon_\beta\right) \; .
\end{aligned} \tag{8}
$$

Together with equation (**??**c), this completes the proof.

■

1.6.8 Posterior probability of alternative hypothesis

Theorem: Let there be a linear regression model (→ III/1.5.1) with normally distributed (→ II/4.1.1) errors:

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}
$$

and assume a normal-gamma (→ II/4.3.1) prior distribution (→ I/5.1.3) over the model parameters β and $\tau = 1/\sigma^2$:

$$
p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}
$$

Then, the posterior (→ I/5.1.7) probability (→ I/1.3.1) of the alternative hypothesis (→ I/4.3.3)

$$
\mathrm{H}_1 : \, c^\mathrm{T} \beta > 0 \tag{3}
$$

is given by

$$
\Pr\left(\mathrm{H}_1 | y\right) = 1 - \mathrm{T}\left(-\frac{c^\mathrm{T} \mu}{\sqrt{c^\mathrm{T} \Sigma c}}; \nu\right) \tag{4}
$$

where c is a $p \times 1$ contrast vector (→ III/1.5.25), $\mathrm{T}(x; \nu)$ is the cumulative distribution function (→ I/1.8.1) of the t-distribution (→ II/3.3.1) with ν degrees of freedom and μ, Σ and ν can be obtained from the posterior hyperparameters (→ I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (→ III/1.6.2) is given by a normal-gamma distribution (→ II/4.3.1) over β and $\tau = 1/\sigma^2$

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{5}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; . \end{aligned} \tag{6}$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (→ II/4.3.8), such that the marginal (→ I/1.5.3) posterior (→ I/5.1.7) distribution of β is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \tag{7}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{aligned} \mu &= \mu_n \\ \Sigma &= \left(\frac{a_n}{b_n} \Lambda_n \right)^{-1} \\ \nu &= 2 \, a_n \; . \end{aligned} \tag{8}$$

Define the quantity $\gamma = c^\mathrm{T} \beta$. According to the linear transformation theorem for the multivariate t-distribution, γ also follows a multivariate t-distribution (→ II/4.2.1):

$$p(\gamma|y) = t(\gamma; c^\mathrm{T} \mu, c^\mathrm{T} \Sigma \, c, \nu) \; . \tag{9}$$

Because c^T is a $1 \times p$ vector, γ is a scalar and actually has a non-standardized t-distribution (→ II/3.3.3). Therefore, the posterior probability of H_1 can be calculated using a one-dimensional integral:

$$\begin{aligned} \Pr(\mathrm{H}_1|y) &= p(\gamma > 0|y) \\ &= \int_0^{+\infty} p(\gamma|y) \, \mathrm{d}\gamma \\ &= 1 - \int_{-\infty}^0 p(\gamma|y) \, \mathrm{d}\gamma \\ &= 1 - \mathrm{T}_\mathrm{nst}(0; c^\mathrm{T} \mu, c^\mathrm{T} \Sigma \, c, \nu) \; . \end{aligned} \tag{10}$$

Using the relation between non-standardized t-distribution and standard t-distribution (→ II/3.3.4), we can finally write:

$$
\begin{aligned}
\Pr\left(\mathrm{H}_1|y\right) &= 1 - \mathrm{T}\left(\frac{(0 - c^{\mathrm{T}}\mu)}{\sqrt{c^{\mathrm{T}}\Sigma c}}; \nu\right) \\
&= 1 - \mathrm{T}\left(-\frac{c^{\mathrm{T}}\mu}{\sqrt{c^{\mathrm{T}}\Sigma c}}; \nu\right) \; .
\end{aligned} \tag{11}
$$

■

Sources:
- Koch, Karl-Rudolf (2007): "Multivariate t-distribution"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, eqs. 2.235, 2.236, 2.213, 2.210, 2.188; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

1.6.9 Posterior credibility region excluding null hypothesis

Theorem: Let there be a linear regression model (→ III/1.5.1) with normally distributed (→ II/4.1.1) errors:

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}
$$

and assume a normal-gamma (→ II/4.3.1) prior distribution (→ I/5.1.3) over the model parameters β and $\tau = 1/\sigma^2$:

$$
p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}
$$

Then, the largest posterior (→ I/5.1.7) credibility region that does not contain the omnibus null hypothesis (→ I/4.3.2)

$$
\mathrm{H}_0 : \; C^{\mathrm{T}}\beta = 0 \tag{3}
$$

is given by the credibility level

$$
(1 - \alpha) = \mathrm{F}\left(\left[\mu^{\mathrm{T}} C (C^{\mathrm{T}} \Sigma \, C)^{-1} C^{\mathrm{T}} \mu\right] / q; q, \nu\right) \tag{4}
$$

where C is a $p \times q$ contrast matrix (→ III/1.5.26), $\mathrm{F}(x; v, w)$ is the cumulative distribution function (→ I/1.8.1) of the F-distribution (→ II/3.8.1) with v numerator degrees of freedom, w denominator degrees of freedom and μ, Σ and ν can be obtained from the posterior hyperparameters (→ I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (→ III/1.6.2) is given by a normal-gamma distribution (→ II/4.3.1) over β and $\tau = 1/\sigma^2$

$$
p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{5}
$$

with the posterior hyperparameters (→ I/5.1.7)

$$
\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^{\mathrm{T}} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^{\mathrm{T}} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^{\mathrm{T}} P y + \mu_0^{\mathrm{T}} \Lambda_0 \mu_0 - \mu_n^{\mathrm{T}} \Lambda_n \mu_n) \; .
\end{aligned} \tag{6}
$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (→ II/4.3.8), such that the marginal (→ I/1.5.3) posterior (→ I/5.1.7) distribution of β is

$$p(\beta|y) = t(\beta; \mu, \Sigma, \nu) \tag{7}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{aligned} \mu &= \mu_n \\ \Sigma &= \left(\frac{a_n}{b_n} \Lambda_n \right)^{-1} \\ \nu &= 2\, a_n \; . \end{aligned} \tag{8}$$

Define the quantity $\gamma = C^\mathrm{T} \beta$. According to the linear transformation theorem for the multivariate t-distribution, γ also follows a multivariate t-distribution (→ II/4.2.1):

$$p(\gamma|y) = t(\gamma; C^\mathrm{T} \mu, C^\mathrm{T} \Sigma\, C, \nu) \; . \tag{9}$$

Because C^T is a $q \times p$ matrix, γ is a $q \times 1$ vector. The quadratic form of a multivariate t-distributed random variable has an F-distribution (→ II/4.2.3), such that we can write:

$$\mathrm{QF}(\gamma) = (\gamma - C^\mathrm{T} \mu)^\mathrm{T} (C^\mathrm{T} \Sigma\, C)^{-1} (\gamma - C^\mathrm{T} \mu)/q \; \sim \mathrm{F}(q, \nu) \; . \tag{10}$$

Therefore, the largest posterior credibility region for γ which does not contain $\gamma = 0_q$ (i.e. only touches this origin point) can be obtained by plugging QF(0) into the cumulative distribution function of the F-distribution:

$$\begin{aligned} (1-\alpha) &= \mathrm{F}\left(\mathrm{QF}(0); q, \nu \right) \\ &= \mathrm{F}\left(\left[\mu^\mathrm{T} C (C^\mathrm{T} \Sigma\, C)^{-1} C^\mathrm{T} \mu \right] / q; q, \nu \right) \; . \end{aligned} \tag{11}$$

■

Sources:
- Koch, Karl-Rudolf (2007): "Multivariate t-distribution"; in: *Introduction to Bayesian Statistics*, Springer, Berlin/Heidelberg, 2007, eqs. 2.235, 2.236, 2.213, 2.210, 2.211, 2.183; URL: https://www.springer.com/de/book/9783540727231; DOI: 10.1007/978-3-540-72726-2.

1.6.10 Combined posterior distribution from independent data sets

Theorem: Let $y = \{y_1, \ldots, y_S\}$ be a set of S conditionally independent data sets (→ I/1.3.7) assumed to follow linear regression models (→ III/1.5.1) with design matrices (→ III/1.5.1) $X_1, \ldots, X_S$, number of data points (→ III/1.5.1) $n_1, \ldots, n_S$ and precision matrices (→ III/1.6.1) $P_1, \ldots, P_n$, governed by identical regression coefficients (→ III/1.5.1) β and identical noise precision τ:

$$\begin{aligned} y_1 &= X_1 \beta + \varepsilon_1, \; \varepsilon_1 \sim \mathcal{N}(0, \sigma^2 V_1), \; \sigma^2 V_1 = (\tau P_1)^{-1} \\ &\vdots \\ y_S &= X_S \beta + \varepsilon_S, \; \varepsilon_S \sim \mathcal{N}(0, \sigma^2 V_S), \; \sigma^2 V_S = (\tau P_S)^{-1} \; . \end{aligned} \tag{1}$$

Moreover, assume a normal-gamma prior distribution (→ III/1.6.1) over the model parameters β and $\tau = 1/\sigma^2$:

$$p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau \Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \; . \tag{2}$$

Then, the combined posterior distribution (→ I/5.1.10) from observing these conditionally independent data sets (→ I/1.3.7) is also given by a normal-gamma distribution (→ II/4.3.1)

$$p(\beta, \tau | y) = \mathcal{N}(\beta; \mu_n, (\tau \Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{3}$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1} \left(\sum_{i=1}^S X_i^\mathrm{T} P_i y_i + \Lambda_0 \mu_0 \right) \\
\Lambda_n &= \sum_{i=1}^S X_i^\mathrm{T} P_i X_i + \Lambda_0 \\
a_n &= a_0 + \frac{1}{2} \sum_{i=1}^S n_i \\
b_n &= b_0 + \frac{1}{2} \left(\sum_{i=1}^S y_i^\mathrm{T} P_i y_i + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n \right) \; .
\end{aligned} \tag{4}$$

Proof: This can be seen by sequentially applying Bayes' theorem (→ I/5.3.1) for calculating the posterior distribution (→ I/5.1.9), while using the posterior after one iteration as the prior for the next iteration.

Let $\mu_0^{(i)}, \Lambda_0^{(i)}, a_0^{(i)}, b_0^{(i)}$ denote the prior hyperparameters (→ I/5.1.3) before analyzing the i-th data set, such that e.g. $\mu_0^{(1)}$ is identical to μ_0 in (2):

$$\begin{aligned}
\mu_0^{(1)} &= \mu_0 \\
\Lambda_0^{(1)} &= \Lambda_0 \\
a_0^{(1)} &= a_0 \\
b_0^{(1)} &= b_0 \; .
\end{aligned} \tag{5}$$

Moreover, let $\mu_n^{(i)}, \Lambda_n^{(i)}, a_n^{(i)}, b_n^{(i)}$ denote the posterior hyperparameters (→ I/5.1.7) after analyzing the i-th data set, such that e.g. $\mu_n^{(S)}$ is identical to μ_n in (3):

$$\begin{aligned}
\mu_n^{(S)} &= \mu_n \\
\Lambda_n^{(S)} &= \Lambda_n \\
a_n^{(S)} &= a_n \\
b_n^{(S)} &= b_n \; .
\end{aligned} \tag{6}$$

The posterior (→ I/5.1.7) after seeing the i-th data set is equal to the prior (→ I/5.1.3) before seeing the $(i+1)$-th data set, so we have the relation:

$$
\begin{aligned}
\mu_0^{(i+1)} &= \mu_n^{(i)} \\
\Lambda_0^{(i+1)} &= \Lambda_n^{(i)} \\
a_0^{(i+1)} &= a_n^{(i)} \\
b_0^{(i+1)} &= b_n^{(i)} \; .
\end{aligned} \tag{7}
$$

The posterior distribution for Bayesian linear regression when observing a single data set is given by the following hyperparameter equations (→ III/1.6.2):

$$
\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; .
\end{aligned} \tag{8}
$$

We can apply (8) to calculate the posterior hyperparameters after seeing the first data set:

$$
\begin{aligned}
\mu_n^{(1)} &= {\Lambda_n^{(1)}}^{-1} \left(X_1^\mathrm{T} P_1 y_1 + \Lambda_0^{(1)} \mu_0^{(1)} \right) \\
&= {\Lambda_n^{(1)}}^{-1} \left(X_1^\mathrm{T} P_1 y_1 + \Lambda_0 \mu_0 \right) \\
\Lambda_n^{(1)} &= X_1^\mathrm{T} P_1 X_1 + \Lambda_0^{(1)} \\
&= X_1^\mathrm{T} P_1 X_1 + \Lambda_0 \\
a_n^{(1)} &= a_0^{(1)} + \frac{1}{2} n_1 \\
&= a_0 + \frac{1}{2} n_1 \\
b_n^{(1)} &= b_0^{(1)} + \frac{1}{2} \left(y_1^\mathrm{T} P_1 y_1 + {\mu_0^{(1)}}^\mathrm{T} \Lambda_0^{(1)} \mu_0^{(1)} - {\mu_n^{(1)}}^\mathrm{T} \Lambda_n^{(1)} \mu_n^{(1)} \right) \\
&= b_0 + \frac{1}{2} \left(y_1^\mathrm{T} P_1 y_1 + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - {\mu_n^{(1)}}^\mathrm{T} \Lambda_n^{(1)} \mu_n^{(1)} \right) \; .
\end{aligned} \tag{9}
$$

These are the prior hyperparameters before seeing the second data set:

$$
\begin{aligned}
\mu_0^{(2)} &= \mu_n^{(1)} \\
\Lambda_0^{(2)} &= \Lambda_n^{(1)} \\
a_0^{(2)} &= a_n^{(1)} \\
b_0^{(2)} &= b_n^{(1)} \; .
\end{aligned} \tag{10}
$$

Thus, we can again use (8) to calculate the posterior hyperparameters after seeing the second data set:

$$
\begin{aligned}
\mu_n^{(2)} &= \Lambda_n^{(2)^{-1}} \left(X_2^\mathrm{T} P_2 y_2 + \Lambda_0^{(2)} \mu_0^{(2)} \right) \\
&= \Lambda_n^{(2)^{-1}} \left(X_2^\mathrm{T} P_2 y_2 + \Lambda_n^{(1)} \Lambda_n^{(1)^{-1}} \left(X_1^\mathrm{T} P_1 y_1 + \Lambda_0 \mu_0 \right) \right) \\
&= \Lambda_n^{(2)^{-1}} \left(X_1^\mathrm{T} P_1 y_1 + X_2^\mathrm{T} P_2 y_2 + \Lambda_0 \mu_0 \right) \\
\Lambda_n^{(2)} &= X_2^\mathrm{T} P_2 X_2 + \Lambda_0^{(2)} \\
&= X_2^\mathrm{T} P_2 X_2 + X_1^\mathrm{T} P_1 X_1 + \Lambda_0 \\
&= X_1^\mathrm{T} P_1 X_1 + X_2^\mathrm{T} P_2 X_2 + \Lambda_0 \\
a_n^{(2)} &= a_0^{(2)} + \frac{1}{2} n_2 \\
&= a_0 + \frac{1}{2} n_1 + \frac{1}{2} n_2 \\
&= a_0 + \frac{1}{2} \left(n_1 + n_2 \right) \\
b_n^{(2)} &= b_0^{(2)} + \frac{1}{2} \left(y_2^\mathrm{T} P_2 y_2 + \mu_0^{(2)^\mathrm{T}} \Lambda_0^{(2)} \mu_0^{(2)} - \mu_n^{(2)^\mathrm{T}} \Lambda_n^{(2)} \mu_n^{(2)} \right) \\
&= b_0 + \frac{1}{2} \left(y_1^\mathrm{T} P_1 y_1 + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^{(1)^\mathrm{T}} \Lambda_n^{(1)} \mu_n^{(1)} \right) + \frac{1}{2} \left(y_2^\mathrm{T} P_2 y_2 + \mu_n^{(1)^\mathrm{T}} \Lambda_n^{(1)} \mu_n^{(1)} - \mu_n^{(2)^\mathrm{T}} \Lambda_n^{(2)} \mu_n^{(2)} \right) \\
&= b_0 + \frac{1}{2} \left(y_1^\mathrm{T} P_1 y_1 + y_2^\mathrm{T} P_2 y_2 + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^{(2)^\mathrm{T}} \Lambda_n^{(2)} \mu_n^{(2)} \right) \; .
\end{aligned} \tag{11}
$$

These are the prior hyperparameters before seeing the third data set:

$$
\begin{aligned}
\mu_0^{(3)} &= \mu_n^{(2)} \\
\Lambda_0^{(3)} &= \Lambda_n^{(2)} \\
a_0^{(3)} &= a_n^{(2)} \\
b_0^{(3)} &= b_n^{(2)} \; .
\end{aligned} \tag{12}
$$

Generalizing this, we have after observing the j-th data set:

$$
\begin{aligned}
\mu_n^{(j)} &= \Lambda_n^{(j)^{-1}} \left(\sum_{i=1}^{j} X_i^\mathrm{T} P_i y_i + \Lambda_0 \mu_0 \right) \\
\Lambda_n^{(j)} &= \sum_{i=1}^{j} X_i^\mathrm{T} P_i X_i + \Lambda_0 \\
a_n^{(j)} &= a_0 + \frac{1}{2} \sum_{i=1}^{j} n_i \\
b_n^{(j)} &= b_0 + \frac{1}{2} \left(\sum_{i=1}^{j} y_i^\mathrm{T} P_i y_i + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^{(j)^\mathrm{T}} \Lambda_n^{(j)} \mu_n^{(j)} \right) \; .
\end{aligned} \tag{13}
$$

Plugging in $j = S$, we obtain the final posterior distribution:

$$
\begin{aligned}
\mu_n = \mu_n^{(S)} &= \Lambda_n^{(S)^{-1}} \left(\sum_{i=1}^{S} X_i^\mathrm{T} P_i y_i + \Lambda_0 \mu_0 \right) = \Lambda_n^{-1} \left(\sum_{i=1}^{S} X_i^\mathrm{T} P_i y_i + \Lambda_0 \mu_0 \right) \\
\Lambda_n = \Lambda_n^{(S)} &= \sum_{i=1}^{S} X_i^\mathrm{T} P_i X_i + \Lambda_0 \\
a_n = a_n^{(S)} &= a_0 + \frac{1}{2} \sum_{i=1}^{S} n_i \\
b_n = b_n^{(S)} &= b_0 + \frac{1}{2} \left(\sum_{i=1}^{S} y_i^\mathrm{T} P_i y_i + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^{(S)^\mathrm{T}} \Lambda_n^{(S)} \mu_n^{(S)} \right) \\
&= b_0 + \frac{1}{2} \left(\sum_{i=1}^{S} y_i^\mathrm{T} P_i y_i + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n \right) \; .
\end{aligned} \tag{14}
$$

This result is also compatible with the general theorem about combined posterior distributions in terms of individual posterior distributions (→ I/5.1.10) when analyzing independent data sets.

■

1.6.11 Log Bayes factor for comparison of two regression models

Theorem: Let $y = [y_1, \ldots, y_n]^\mathrm{T}$ be an $n \times 1$ vector of a measured univariate signal and consider two linear regression models (→ III/1.5.1) with design matrices (→ III/1.5.1) X_1, X_2 and precision matrices (→ III/1.6.1) P_1, P_2, entailing potentially different regression coefficients (→ III/1.5.1) β_1, β_2 and noise precisions τ_1, τ_2:

$$
\begin{aligned}
m_1 &: \; y = X_1 \beta_1 + \varepsilon_1, \; \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2 V_1), \; \sigma_1^2 V_1 = (\tau_1 P_1)^{-1} \\
m_2 &: \; y = X_2 \beta_2 + \varepsilon_2, \; \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2 V_2), \; \sigma_2^2 V_2 = (\tau_2 P_2)^{-1} \; .
\end{aligned} \tag{1}
$$

Moreover, assume normal-gamma prior distributions (→ III/1.6.1) over the model parameters β_1 and $\tau_1 = 1/\sigma_1^2$ as well as β_2 and $\tau_2 = 1/\sigma_2^2$:

$$
\begin{aligned}
p(\beta_1, \tau_1) &= \mathcal{N}\left(\beta_1; \mu_0^{(1)}, \left(\tau_1 \Lambda_0^{(1)} \right)^{-1} \right) \cdot \mathrm{Gam}\left(\tau_1; a_0^{(1)}, b_0^{(1)} \right) \\
p(\beta_2, \tau_2) &= \mathcal{N}\left(\beta_2; \mu_0^{(2)}, \left(\tau_2 \Lambda_0^{(2)} \right)^{-1} \right) \cdot \mathrm{Gam}\left(\tau_2; a_0^{(2)}, b_0^{(2)} \right) \; .
\end{aligned} \tag{2}
$$

Then, the log Bayes factor (→ IV/**??**) in favor of m_1 against m_2 is

$$
\begin{aligned}
\mathrm{LBF}_{12} = &\frac{1}{2} \log \frac{|P_1|}{|P_2|} + \frac{1}{2} \log \frac{|\Lambda_0^{(1)}|}{|\Lambda_0^{(2)}|} - \frac{1}{2} \log \frac{|\Lambda_n^{(1)}|}{|\Lambda_n^{(2)}|} \\
&+ \log \frac{\Gamma\left(a_n^{(1)} \right)}{\Gamma\left(a_0^{(1)} \right)} + a_0^{(1)} \log b_0^{(1)} - a_n^{(1)} \log b_n^{(1)} \\
&+ \log \frac{\Gamma\left(a_n^{(2)} \right)}{\Gamma\left(a_0^{(2)} \right)} - a_0^{(2)} \log b_0^{(2)} + a_n^{(2)} \log b_n^{(2)}
\end{aligned} \tag{3}
$$

where $\mu_n^{(1)}, \Lambda_n^{(1)}, a_n^{(1)}, b_n^{(1)}$ and $\mu_n^{(2)}, \Lambda_n^{(2)}, a_n^{(2)}, b_n^{(2)}$ are the posterior hyperparameters for Bayesian linear regression (→ III/1.6.2) for each of the two models which are functions of the design matrices, the precision matrices and the data vector.

Proof: For Bayesian linear regression with data vector y, design matrix X, precision matrix P and a normal-gamma prior distribution (→ III/1.6.1) over β and τ, the log model evidence is given by (→ III/1.6.3)

$$\begin{split} \log p(y|m) = \frac{1}{2}\log|P| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0| - \frac{1}{2}\log|\Lambda_n| + \\ \log\Gamma(a_n) - \log\Gamma(a_0) + a_0\log b_0 - a_n\log b_n \end{split} \tag{4}$$

where the posterior hyperparameters are equal to (→ III/1.6.2)

$$\begin{split} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T}Py + \Lambda_0\mu_0) \\ \Lambda_n &= X^\mathrm{T}PX + \Lambda_0 \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T}Py + \mu_0^\mathrm{T}\Lambda_0\mu_0 - \mu_n^\mathrm{T}\Lambda_n\mu_n) \; . \end{split} \tag{5}$$

Thus, the log model evidences (→ IV/??) for m_1 and m_2 are given by:

$$\begin{split} \mathrm{LME}(m_1) = \frac{1}{2}\log|P_1| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0^{(1)}| - \frac{1}{2}\log|\Lambda_n^{(1)}| + \\ \log\Gamma(a_n^{(1)}) - \log\Gamma(a_0^{(1)}) + a_0^{(1)}\log b_0^{(1)} - a_n^{(1)}\log b_n^{(1)} \\ \mathrm{LME}(m_2) = \frac{1}{2}\log|P_2| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Lambda_0^{(2)}| - \frac{1}{2}\log|\Lambda_n^{(2)}| + \\ \log\Gamma(a_n^{(2)}) - \log\Gamma(a_0^{(2)}) + a_0^{(2)}\log b_0^{(2)} - a_n^{(2)}\log b_n^{(2)} \; . \end{split} \tag{6}$$

The log Bayes factor is equal to the difference of two log model evidences (→ IV/??):

$$\mathrm{LBF}_{12} = \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \tag{7}$$

Plugging (6) into (7), this gives:

$$\begin{split} \mathrm{LBF}_{12} = &\frac{1}{2}\log|P_1| - \frac{1}{2}\log|P_2| \\ &+ \frac{1}{2}\log|\Lambda_0^{(1)}| - \frac{1}{2}\log|\Lambda_0^{(2)}| \\ &- \frac{1}{2}\log|\Lambda_n^{(1)}| + \frac{1}{2}\log|\Lambda_n^{(2)}| \\ &+ \log\Gamma\left(a_n^{(1)}\right) - \log\Gamma\left(a_0^{(1)}\right) + a_0^{(1)}\log b_0^{(1)} - a_n^{(1)}\log b_n^{(1)} \\ &- \log\Gamma\left(a_n^{(2)}\right) + \log\Gamma\left(a_0^{(2)}\right) - a_0^{(2)}\log b_0^{(2)} + a_n^{(2)}\log b_n^{(2)} \; . \end{split} \tag{8}$$

Applying $\log a - \log b = \log(a/b)$, we obtain:

$$
\begin{aligned}
\mathrm{LBF}_{12} = &\frac{1}{2} \log \frac{|P_1|}{|P_2|} + \frac{1}{2} \log \frac{|\Lambda_0^{(1)}|}{|\Lambda_0^{(2)}|} - \frac{1}{2} \log \frac{|\Lambda_n^{(1)}|}{|\Lambda_n^{(2)}|} \\
&+ \log \frac{\Gamma\left(a_n^{(1)}\right)}{\Gamma\left(a_0^{(1)}\right)} + a_0^{(1)} \log b_0^{(1)} - a_n^{(1)} \log b_n^{(1)} \\
&- \log \frac{\Gamma\left(a_n^{(2)}\right)}{\Gamma\left(a_0^{(2)}\right)} - a_0^{(2)} \log b_0^{(2)} + a_n^{(2)} \log b_n^{(2)} \; .
\end{aligned} \tag{9}
$$

■

1.7 Bayesian linear regression with known covariance

1.7.1 Conjugate prior distribution

Theorem: Let

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \Sigma) \tag{1}
$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X and known $n \times n$ covariance matrix Σ as well as unknown $p \times 1$ regression coefficients β. Then, the conjugate prior (→ I/5.2.5) for this model is a multivariate normal distribution (→ II/4.1.1)

$$
p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) \; . \tag{2}
$$

Proof: By definition, a conjugate prior (→ I/5.2.5) is a prior distribution (→ I/5.1.3) that, when combined with the likelihood function (→ I/5.1.2), leads to a posterior distribution (→ I/5.1.7) that belongs to the same family of probability distributions (→ I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.
Equation (1) implies the following likelihood function (→ I/5.1.2):

$$
p(y|\beta) = \mathcal{N}(y; X\beta, \Sigma) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \, \exp\left[-\frac{1}{2} (y - X\beta)^\mathrm{T} \Sigma^{-1} (y - X\beta) \right] \; . \tag{3}
$$

Expanding the product in the exponent, we have:

$$
p(y|\beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} \left(y^\mathrm{T} \Sigma^{-1} y - y^\mathrm{T} \Sigma^{-1} X\beta - \beta^\mathrm{T} X^\mathrm{T} \Sigma^{-1} y + \beta^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X\beta \right) \right] \; . \tag{4}
$$

Completing the square over β, one obtains

$$
p(y|\beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2} \left((\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X (\beta - \tilde{X}y) - y^\mathrm{T} Q y + y^\mathrm{T} \Sigma^{-1} y \right) \right] \tag{5}
$$

where $\tilde{X} = \left(X^\mathrm{T} \Sigma^{-1} X \right)^{-1} X^\mathrm{T} \Sigma^{-1}$ and $Q = \tilde{X}^\mathrm{T} \left(X^\mathrm{T} \Sigma^{-1} X \right) \tilde{X}$.

Separating constant and variable terms, we get:

$$p(y|\beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \cdot \exp\left[-\frac{1}{2}\left(y^\mathrm{T} Q y + y^\mathrm{T} \Sigma^{-1} y\right)\right] \cdot \exp\left[-\frac{1}{2}(\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X (\beta - \tilde{X}y)\right] \; . \quad (6)$$

In other words, the likelihood function (→ I/5.1.2) is proportional to an exponential of a squared form of β:

$$p(y|\beta) \propto \exp\left[-\frac{1}{2}(\beta - \tilde{X}y)^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X (\beta - \tilde{X}y)\right] \; . \quad (7)$$

The same is true for a multivariate normal distribution (→ II/4.1.1) over β

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) \quad (8)$$

the probability density function of which (→ II/4.1.7)

$$p(\beta) = \sqrt{\frac{1}{(2\pi)^p |\Sigma_0|}} \cdot \exp\left[-\frac{1}{2}(\beta - \mu_0)^\mathrm{T} \Sigma_0^{-1} (\beta - \mu_0)\right] \quad (9)$$

exhibits the same proportionality

$$p(\beta) \propto \exp\left[-\frac{1}{2}(\beta - \mu_0)^\mathrm{T} \Sigma_0^{-1} (\beta - \mu_0)\right] \quad (10)$$

and is therefore conjugate relative to the likelihood.

■

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, eq. 3.48; URL: https://www.springer.com/gp/book/9780387310732.
- Penny WD (2012): "Comparing Dynamic Causal Models using AIC, BIC and Free Energy"; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eq. 9; URL: https://www.sciencedirect.com/science/article/pii/S1053811911008160; DOI: 10.1016/j.neuroimage.2011.07.039.

1.7.2 Posterior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \Sigma) \quad (1)$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X and known $n \times n$ covariance matrix Σ as well as unknown $p \times 1$ regression coefficients β. Moreover, assume a multivariate normal distribution (→ III/1.7.1) over the model parameter β:

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) \; . \quad (2)$$

Then, the posterior distribution (→ I/5.1.7) is also a multivariate normal distribution (→ II/4.1.1)

$$p(\beta|y) = \mathcal{N}(\beta; \mu_n, \Sigma_n) \quad (3)$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$
\begin{aligned}
\mu_n &= \Sigma_n (X^\mathrm{T} \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\
\Sigma_n &= \left(X^\mathrm{T} \Sigma^{-1} X + \Sigma_0^{-1} \right)^{-1} \; .
\end{aligned} \tag{4}
$$

Proof: According to Bayes' theorem (→ I/5.3.1), the posterior distribution (→ I/5.1.7) is given by

$$
p(\beta|y) = \frac{p(y|\beta) \, p(\beta)}{p(y)} \; . \tag{5}
$$

Since $p(y)$ is just a normalization factor, the posterior is proportional (→ I/5.1.9) to the numerator:

$$
p(\beta|y) \propto p(y|\beta) \, p(\beta) = p(y,\beta) \; . \tag{6}
$$

Equation (1) implies the following likelihood function (→ I/5.1.2):

$$
p(y|\beta) = \mathcal{N}(y; X\beta, \Sigma) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \, \exp\left[-\frac{1}{2} (y-X\beta)^\mathrm{T} \Sigma^{-1} (y-X\beta) \right] \; . \tag{7}
$$

Combining the likelihood function (→ I/5.1.2) (7) with the prior distribution (→ I/5.1.3) (2) using the probability density function of the multivariate normal distribution (→ II/4.1.7), the joint likelihood (→ I/5.1.5) of the model is given by

$$
\begin{aligned}
p(y,\beta) &= p(y|\beta) \, p(\beta) \\
&= \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \, \exp\left[-\frac{1}{2} (y-X\beta)^\mathrm{T} \Sigma^{-1} (y-X\beta) \right] \cdot \\
&\quad \sqrt{\frac{1}{(2\pi)^p |\Sigma_0|}} \, \exp\left[-\frac{1}{2} (\beta-\mu_0)^\mathrm{T} \Sigma_0^{-1} (\beta-\mu_0) \right] \; .
\end{aligned} \tag{8}
$$

Collecting identical variables gives:

$$
\begin{aligned}
p(y,\beta) = &\sqrt{\frac{1}{(2\pi)^{n+p} |\Sigma| |\Sigma_0|}} \cdot \\
&\exp\left[-\frac{1}{2} \left((y-X\beta)^\mathrm{T} \Sigma^{-1} (y-X\beta) + (\beta-\mu_0)^\mathrm{T} \Sigma_0^{-1} (\beta-\mu_0) \right) \right] \; .
\end{aligned} \tag{9}
$$

Expanding the products in the exponent gives:

$$
\begin{aligned}
p(y,\beta) = &\sqrt{\frac{1}{(2\pi)^{n+p} |\Sigma| |\Sigma_0|}} \cdot \\
&\exp\left[-\frac{1}{2} \left(y^\mathrm{T} \Sigma^{-1} y - y^\mathrm{T} \Sigma^{-1} X\beta - \beta^\mathrm{T} X^\mathrm{T} \Sigma^{-1} y + \beta^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X\beta + \right. \right. \\
&\quad \left. \left. \beta^\mathrm{T} \Sigma_0^{-1} \beta - \beta^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_0^\mathrm{T} \Sigma_0^{-1} \beta + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 \right) \right] \; .
\end{aligned} \tag{10}
$$

Regrouping the terms in the exponent gives:

$$
\begin{split}
p(y,\beta) = \sqrt{\frac{1}{(2\pi)^{n+p}|\Sigma||\Sigma_0|}} \cdot \\
\exp\left[-\frac{1}{2} \left(\beta^\mathrm{T}[X^\mathrm{T}\Sigma^{-1}X + \Sigma_0^{-1}]\beta - 2\beta^\mathrm{T}[X^\mathrm{T}\Sigma^{-1}y + \Sigma_0^{-1}\mu_0] + \right. \right. \\
\left. \left. y^\mathrm{T}\Sigma^{-1}y + \mu_0^\mathrm{T}\Sigma_0^{-1}\mu_0 \right) \right] \; .
\end{split}
\tag{11}
$$

Completing the square over β, we finally have

$$
\begin{split}
p(y,\beta) = \sqrt{\frac{1}{(2\pi)^{n+p}|\Sigma||\Sigma_0|}} \cdot \\
\exp\left[-\frac{1}{2} \left((\beta-\mu_n)^\mathrm{T}\Sigma_n^{-1}(\beta-\mu_n) + (y^\mathrm{T}\Sigma^{-1}y + \mu_0^\mathrm{T}\Sigma_0^{-1}\mu_0 - \mu_n^\mathrm{T}\Sigma_n^{-1}\mu_n) \right) \right]
\end{split}
\tag{12}
$$

with the posterior hyperparameters (→ I/5.1.7)

$$
\begin{split}
\mu_n &= \Sigma_n(X^\mathrm{T}\Sigma^{-1}y + \Sigma_0^{-1}\mu_0) \\
\Sigma_n &= \left(X^\mathrm{T}\Sigma^{-1}X + \Sigma_0^{-1} \right)^{-1} \; .
\end{split}
\tag{13}
$$

Ergo, the joint likelihood is proportional to

$$
p(y,\beta) \propto \exp\left[-\frac{1}{2}(\beta-\mu_n)^\mathrm{T}\Sigma_n^{-1}(\beta-\mu_n) \right] \; ,
\tag{14}
$$

such that the posterior distribution over β is given by

$$
p(\beta|y) = \mathcal{N}(\beta; \mu_n, \Sigma_n)
\tag{15}
$$

with the posterior hyperparameters given in (13).

■

Sources:

- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161, eqs. 3.49-3.51, ex. 3.7; URL: https://www.springer.com/gp/book/9780387310732.
- Penny WD (2012): "Comparing Dynamic Causal Models using AIC, BIC and Free Energy"; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eq. 27; URL: https://www.sciencedirect.com/science/article/pii/S1053811911008160; DOI: 10.1016/j.neuroimage.2011.07.039.

1.7.3 Log model evidence

Theorem: Let

$$
m: y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \Sigma)
\tag{1}
$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X and known $n \times n$ covariance matrix Σ as well as unknown $p \times 1$ regression coefficients β. Moreover, assume a multivariate normal distribution (→ III/1.7.1) over the model parameter β:

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) \; . \tag{2}$$

Then, the log model evidence (→ IV/??) for this model is

$$\begin{split} \log p(y|m) = &-\frac{1}{2} e_y^\mathrm{T} \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) \\ &-\frac{1}{2} e_\beta^\mathrm{T} \Sigma_0^{-1} e_\beta - \frac{1}{2} \log |\Sigma_0| + \frac{1}{2} \log |\Sigma_n| \; . \end{split} \tag{3}$$

with the "prediction error" and "parameter error" terms

$$\begin{split} e_y &= y - X\mu_n \\ e_\beta &= \mu_0 - \mu_n \end{split} \tag{4}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{split} \mu_n &= \Sigma_n (X^\mathrm{T} \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\ \Sigma_n &= \left(X^\mathrm{T} \Sigma^{-1} X + \Sigma_0^{-1} \right)^{-1} \; . \end{split} \tag{5}$$

Proof: According to the law of marginal probability (→ I/1.3.3), the model evidence (→ I/5.1.11) for this model is:

$$p(y|m) = \int p(y|\beta) \, p(\beta) \, \mathrm{d}\beta \; . \tag{6}$$

According to the law of conditional probability (→ I/1.3.4), the integrand is equivalent to the joint likelihood (→ I/5.1.5):

$$p(y|m) = \int p(y,\beta) \, \mathrm{d}\beta \; . \tag{7}$$

Equation (1) implies the following likelihood function (→ I/5.1.2):

$$p(y|\beta) = \mathcal{N}(y; X\beta, \Sigma) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \, \exp\left[-\frac{1}{2} (y-X\beta)^\mathrm{T} \Sigma^{-1} (y-X\beta) \right] \; . \tag{8}$$

When deriving the posterior distribution (→ III/1.7.2) $p(\beta|y)$, the joint likelihood $p(y,\beta)$ is obtained as

$$\begin{split} p(y,\beta) = \; &\sqrt{\frac{1}{(2\pi)^{n+p} |\Sigma| |\Sigma_0|}} \cdot \\ &\exp\left[-\frac{1}{2} \left((\beta-\mu_n)^\mathrm{T} \Sigma_n^{-1} (\beta-\mu_n) + (y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} \Sigma_n^{-1} \mu_n) \right) \right] \; . \end{split} \tag{9}$$

Using the probability density function of the multivariate normal distribution (→ II/4.1.7), we can rewrite this as

$$
\begin{split}
p(y, \beta) = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \sqrt{\frac{1}{(2\pi)^p |\Sigma_0|}} \sqrt{\frac{(2\pi)^p |\Sigma_n|}{1}} \cdot \mathcal{N}(\beta; \mu_n, \Sigma_n) \cdot \\
\exp\left[-\frac{1}{2} \left(y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} \Sigma_n^{-1} \mu_n \right) \right] \; .
\end{split} \tag{10}
$$

With that, β can be integrated out easily:

$$
\int p(y, \beta) \, \mathrm{d}\beta = \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \sqrt{\frac{|\Sigma_n|}{|\Sigma_0|}} \cdot \exp\left[-\frac{1}{2} \left(y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} \Sigma_n^{-1} \mu_n \right) \right] \; . \tag{11}
$$

Now we turn to the intra-exponent term

$$
y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} \Sigma_n^{-1} \mu_n \tag{12}
$$

and plug in the posterior covariance

$$
\Sigma_n = \left(X^\mathrm{T} \Sigma^{-1} X + \Sigma_0^{-1} \right)^{-1} \; . \tag{13}
$$

This gives

$$
\begin{split}
&y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} \Sigma_n^{-1} \mu_n \\
&= y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} \left(X^\mathrm{T} \Sigma^{-1} X + \Sigma_0^{-1} \right) \mu_n \\
&= y^\mathrm{T} \Sigma^{-1} y + \mu_0^\mathrm{T} \Sigma_0^{-1} \mu_0 - \mu_n^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X \mu_n - \mu_n^\mathrm{T} \Sigma_0^{-1} \mu_n \\
&= (y - X\mu_n)^\mathrm{T} \Sigma^{-1} (y - X\mu_n) + (\mu_0 - \mu_n)^\mathrm{T} \Sigma_0^{-1} (\mu_0 - \mu_n) \\
&\overset{(4)}{=} e_y^\mathrm{T} \Sigma^{-1} e_y + e_\beta^\mathrm{T} \Sigma_0^{-1} e_\beta \; .
\end{split} \tag{14}
$$

Thus, the marginal likelihood (→ I/5.1.11) becomes

$$
p(y|m) = \int p(y, \beta) \, \mathrm{d}\beta \overset{(11)}{=} \sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \sqrt{\frac{|\Sigma_n|}{|\Sigma_0|}} \cdot \exp\left[-\frac{1}{2} \left(e_y^\mathrm{T} \Sigma^{-1} e_y + e_\beta^\mathrm{T} \Sigma_0^{-1} e_\beta \right) \right] \tag{15}
$$

and the log model evidence (→ IV/**??**) of this model is given by

$$
\begin{split}
\log p(y|m) = &- \frac{1}{2} e_y^\mathrm{T} \Sigma^{-1} e_y - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi) \\
&- \frac{1}{2} e_\beta^\mathrm{T} \Sigma_0^{-1} e_\beta - \frac{1}{2} \log |\Sigma_0| + \frac{1}{2} \log |\Sigma_n| \; .
\end{split} \tag{16}
$$

■

Sources:

- Penny WD (2012): "Comparing Dynamic Causal Models using AIC, BIC and Free Energy"; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eqs. 19-23; URL: https://www.sciencedirect.com/science/article/pii/S1053811911008160; DOI: 10.1016/j.neuroimage.2011.07.039.
- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161; URL: https://www.springer.com/gp/book/9780387310732.

1.7.4 Accuracy and complexity

Theorem: Let

$$m : y = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \Sigma) \tag{1}$$

be a linear regression model (→ III/1.5.1) with measured $n \times 1$ data vector y, known $n \times p$ design matrix X and known $n \times n$ covariance matrix Σ as well as unknown $p \times 1$ regression coefficients β. Moreover, assume a multivariate normal distribution (→ III/1.7.1) over the model parameter β:

$$p(\beta) = \mathcal{N}(\beta; \mu_0, \Sigma_0) \; . \tag{2}$$

Then, accuracy and complexity (→ IV/**??**) of this model are

$$\begin{split} \mathrm{Acc}(m) &= -\frac{1}{2} e_y^\mathrm{T} \Sigma^{-1} e_y - \frac{1}{2} \log|\Sigma| - \frac{n}{2} \log(2\pi) - \frac{1}{2} \mathrm{tr}(X^\mathrm{T} \Sigma^{-1} X \Sigma_n) \\ \mathrm{Com}(m) &= \;\; \frac{1}{2} e_\beta^\mathrm{T} \Sigma_0^{-1} e_\beta + \frac{1}{2} \log|\Sigma_0| - \frac{1}{2} \log|\Sigma_n| + \frac{1}{2} \mathrm{tr}(\Sigma_0^{-1} \Sigma_n) - \frac{p}{2} \end{split} \tag{3}$$

with the "prediction error" and "parameter error" terms

$$\begin{split} e_y &= y - X\mu_n \\ e_\beta &= \mu_0 - \mu_n \end{split} \tag{4}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{split} \mu_n &= \Sigma_n (X^\mathrm{T} \Sigma^{-1} y + \Sigma_0^{-1} \mu_0) \\ \Sigma_n &= \left(X^\mathrm{T} \Sigma^{-1} X + \Sigma_0^{-1} \right)^{-1} \; . \end{split} \tag{5}$$

Proof: Model accuracy and complexity are defined as (→ IV/**??**)

$$\begin{split} \mathrm{LME}(m) &= \mathrm{Acc}(m) - \mathrm{Com}(m) \\ \mathrm{Acc}(m) &= \left\langle \log p(y|\beta, m) \right\rangle_{p(\beta|y,m)} \\ \mathrm{Com}(m) &= \mathrm{KL}\left[p(\beta|y,m) \, || \, p(\beta|m) \right] \; . \end{split} \tag{6}$$

1) The accuracy term is the expectation (→ I/1.10.1) of the log-likelihood function (→ I/4.1.2) $\log p(y|\beta)$ with respect to the posterior distribution (→ I/5.1.7) $p(\beta|y)$:

$$\mathrm{Acc}(m) = \left\langle \log p(y|\beta) \right\rangle_{p(\beta|y)} \; . \tag{7}$$

With the likelihood function for Bayesian linear regression with known covariance (→ III/1.7.1), we have:

$$\begin{split} \mathrm{Acc}(m) &= \left\langle \log \left(\sqrt{\frac{1}{(2\pi)^n |\Sigma|}} \exp\left[-\frac{1}{2} (y - X\beta)^\mathrm{T} \Sigma^{-1} (y - X\beta) \right] \right) \right\rangle_{p(\beta|y)} \\ &= \left\langle -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (y - X\beta)^\mathrm{T} \Sigma^{-1} (y - X\beta) \right\rangle_{p(\beta|y)} \\ &= \left\langle -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \left[y^\mathrm{T} \Sigma^{-1} y - 2 y^\mathrm{T} \Sigma^{-1} X\beta + \beta^\mathrm{T} X^\mathrm{T} \Sigma^{-1} X\beta \right] \right\rangle_{p(\beta|y)} \; . \end{split} \tag{8}$$

With the posterior distribution for Bayesian linear regression with known covariance (→ III/1.7.2), this becomes:

$$\mathrm{Acc}(m) = \left\langle -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left[y^\mathrm{T}\Sigma^{-1}y - 2y^\mathrm{T}\Sigma^{-1}X\beta + \beta^\mathrm{T}X^\mathrm{T}\Sigma^{-1}X\beta\right]\right\rangle_{\mathcal{N}(\beta;\mu_n,\Sigma_n)} \; . \tag{9}$$

If $x \sim \mathrm{N}(\mu, \Sigma)$, then its expected value is (→ II/4.1.9)

$$\langle x \rangle = \mu \tag{10}$$

and the expectation of a quadratic form is given by (→ I/1.10.9)

$$\left\langle x^\mathrm{T}Ax \right\rangle = \mu^\mathrm{T}A\mu + \mathrm{tr}(A\Sigma) \; . \tag{11}$$

Thus, the model accuracy of m evaluates to

$$\begin{aligned}
\mathrm{Acc}(m) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \\
&\quad \frac{1}{2}\left[y^\mathrm{T}\Sigma^{-1}y - 2y^\mathrm{T}\Sigma^{-1}X\mu_n + \mu_n^\mathrm{T}X^\mathrm{T}\Sigma^{-1}X\mu_n + \mathrm{tr}(X^\mathrm{T}\Sigma^{-1}X\Sigma_n)\right] \\
&= -\frac{1}{2}(y-X\mu_n)^\mathrm{T}\Sigma^{-1}(y-X\mu_n) - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi) - \frac{1}{2}\mathrm{tr}(X^\mathrm{T}\Sigma^{-1}X\Sigma_n) \\
&\overset{(4)}{=} -\frac{1}{2}e_y^\mathrm{T}\Sigma^{-1}e_y - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi) - \frac{1}{2}\mathrm{tr}(X^\mathrm{T}\Sigma^{-1}X\Sigma_n)
\end{aligned} \tag{12}$$

which proofs the first part of (3).

2) The complexity penalty is the Kullback-Leibler divergence (→ I/2.5.1) of the posterior distribution (→ I/5.1.7) $p(\beta|y)$ from the prior distribution (→ I/5.1.3) $p(\beta)$:

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\beta|y)\,||\,p(\beta)\right] \; . \tag{13}$$

With the prior distribution (→ III/1.7.1) given by (2) and the posterior distribution for Bayesian linear regression with known covariance (→ III/1.7.2), this becomes:

$$\mathrm{Com}(m) = \mathrm{KL}\left[\mathcal{N}(\beta;\mu_n,\Sigma_n)\,||\,\mathcal{N}(\beta;\mu_0,\Sigma_0)\right] \; . \tag{14}$$

With the Kullback-Leibler divergence for the multivariate normal distribution (→ II/4.1.12)

$$\mathrm{KL}[\mathcal{N}(\mu_1,\Sigma_1)\,||\,\mathcal{N}(\mu_2,\Sigma_2)] = \frac{1}{2}\left[(\mu_2-\mu_1)^\mathrm{T}\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \tag{15}$$

the model complexity of m evaluates to

$$\begin{aligned}
\mathrm{Com}(m) &= \frac{1}{2}\left[(\mu_0-\mu_n)^\mathrm{T}\Sigma_0^{-1}(\mu_0-\mu_n) + \mathrm{tr}(\Sigma_0^{-1}\Sigma_n) - \log\frac{|\Sigma_n|}{|\Sigma_0|} - p\right] \\
&= \frac{1}{2}(\mu_0-\mu_n)^\mathrm{T}\Sigma_0^{-1}(\mu_0-\mu_n) + \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}\log|\Sigma_n| + \frac{1}{2}\mathrm{tr}(\Sigma_0^{-1}\Sigma_n) - \frac{p}{2} \\
&\overset{(4)}{=} \frac{1}{2}e_\beta^\mathrm{T}\Sigma_0^{-1}e_\beta + \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}\log|\Sigma_n| + \frac{1}{2}\mathrm{tr}(\Sigma_0^{-1}\Sigma_n) - \frac{p}{2}
\end{aligned} \tag{16}$$

which proofs the second part of (3).

3) A control calculation confirms that

$$\mathrm{Acc}(m) - \mathrm{Com}(m) = \mathrm{LME}(m) \tag{17}$$

where $\mathrm{LME}(m)$ is the log model evidence for Bayesian linear regression with known covariance (→ III/1.7.3):

$$\begin{split} \log p(y|m) = &-\frac{1}{2} e_y^\mathrm{T} \Sigma^{-1} e_y - \frac{1}{2} \log|\Sigma| - \frac{n}{2} \log(2\pi) \\ &- \frac{1}{2} e_\beta^\mathrm{T} \Sigma_0^{-1} e_\beta - \frac{1}{2} \log|\Sigma_0| + \frac{1}{2} \log|\Sigma_n| \; . \end{split} \tag{18}$$

This requires to recognize, based on (5), that

$$\begin{split} &- \frac{1}{2} \mathrm{tr}(X^\mathrm{T} \Sigma^{-1} X \Sigma_n) - \frac{1}{2} \mathrm{tr}(\Sigma_0^{-1} \Sigma_n) + \frac{p}{2} \\ = &- \frac{1}{2} \mathrm{tr}\left([X^\mathrm{T} \Sigma^{-1} X + \Sigma_0^{-1}] \Sigma_n \right) + \frac{p}{2} \\ = &- \frac{1}{2} \mathrm{tr}\left(\Sigma_n^{-1} \Sigma_n \right) + \frac{p}{2} \\ = &- \frac{1}{2} \mathrm{tr}\left(I_p \right) + \frac{p}{2} \\ = &- \frac{p}{2} + \frac{p}{2} \\ = &\; 0 \; . \end{split} \tag{19}$$

■

Sources:

- Penny WD (2012): "Comparing Dynamic Causal Models using AIC, BIC and Free Energy"; in: *NeuroImage*, vol. 59, iss. 2, pp. 319-330, eqs. 20-21; URL: https://www.sciencedirect.com/science/article/pii/S1053811911008160; DOI: 10.1016/j.neuroimage.2011.07.039.
- Bishop CM (2006): "Bayesian linear regression"; in: *Pattern Recognition for Machine Learning*, pp. 152-161; URL: https://www.springer.com/gp/book/9780387310732.

2 Multivariate normal data

2.1 General linear model

2.1.1 Definition

Definition: Let Y be an $n \times v$ matrix and let X be an $n \times p$ matrix. Then, a statement asserting a linear mapping from X to Y with parameters B and matrix-normally distributed (→ II/5.1.1) errors E

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

is called a multivariate linear regression model or simply, "general linear model".
- Y is called "data matrix", "set of dependent variables" or "measurements";
- X is called "design matrix", "set of independent variables" or "predictors";
- B are called "regression coefficients" or "weights";
- E is called "noise matrix" or "error terms";
- V is called "covariance across rows";
- Σ is called "covariance across columns";
- n is the number of observations;
- v is the number of measurements;
- p is the number of predictors.

When rows of Y correspond to units of time, e.g. subsequent measurements, V is called "temporal covariance". When columns of Y correspond to units of space, e.g. measurement channels, Σ is called "spatial covariance".
When the covariance matrix V is a scalar multiple of the $n \times n$ identity matrix, this is called a general linear model with independent and identically distributed (i.i.d.) observations:

$$V = \lambda I_n \quad \Rightarrow \quad E \sim \mathcal{MN}(0, \lambda I_n, \Sigma) \quad \Rightarrow \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda \Sigma) \; . \tag{2}$$

Otherwise, it is called a general linear model with correlated observations.

Sources:
- Wikipedia (2020): "General linear model"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-21; URL: https://en.wikipedia.org/wiki/General_linear_model.

2.1.2 Ordinary least squares

Theorem: Given a general linear model (→ III/2.1.1) with independent observations

$$Y = XB + E, \; E \sim \mathcal{MN}(0, \sigma^2 I_n, \Sigma) \; , \tag{1}$$

the ordinary least squares (→ III/1.5.3) parameters estimates are given by

$$\hat{B} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y \; . \tag{2}$$

Proof: Let $\hat{B}$ be the ordinary least squares (→ III/1.5.3) (OLS) solution and let $\hat{E} = Y - X\hat{B}$ be the resulting matrix of residuals. According to the exogeneity assumption of OLS, the errors have conditional mean (→ I/1.10.1) zero

$$\mathrm{E}(E|X) = 0 \; , \tag{3}$$

a direct consequence of which is that the regressors are uncorrelated with the errors

$$\mathrm{E}(X^\mathrm{T} E) = 0 \; , \tag{4}$$

which, in the finite sample, means that the residual matrix must be orthogonal to the design matrix:

$$X^\mathrm{T} \hat{E} = 0 \; . \tag{5}$$

From (5), the OLS formula can be directly derived:

$$\begin{split} X^\mathrm{T} \hat{E} &= 0 \\ X^\mathrm{T} \left(Y - X\hat{B} \right) &= 0 \\ X^\mathrm{T} Y - X^\mathrm{T} X \hat{B} &= 0 \\ X^\mathrm{T} X \hat{B} &= X^\mathrm{T} Y \\ \hat{B} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y \; . \end{split} \tag{6}$$

■

2.1.3 Weighted least squares

Theorem: Given a general linear model (→ III/2.1.1) with correlated observations

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; , \tag{1}$$

the weighted least sqaures (→ III/1.5.21) parameter estimates are given by

$$\hat{B} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y \; . \tag{2}$$

Proof: Let there be an $n \times n$ square matrix W, such that

$$W V W^\mathrm{T} = I_n \; . \tag{3}$$

Since V is a covariance matrix and thus symmetric, W is also symmetric and can be expressed as the matrix square root of the inverse of V:

$$WW = V^{-1} \quad \Leftrightarrow \quad W = V^{-1/2} \; . \tag{4}$$

Left-multiplying the linear regression equation (1) with W, the linear transformation theorem (→ II/5.1.9) implies that

$$WY = WXB + WE, \; WE \sim \mathcal{MN}(0, WVW^\mathrm{T}, \Sigma) \; . \tag{5}$$

Applying (3), we see that (5) is actually a general linear model (→ III/2.1.1) with independent observations

$$\tilde{Y} = \tilde{X}B + \tilde{E}, \; \tilde{E} \sim \mathcal{N}(0, I_n, \Sigma) \tag{6}$$

where $\tilde{Y} = WY$, $\tilde{X} = WX$ and $\tilde{E} = WE$, such that we can apply the ordinary least squares solution (→ III/2.1.2) giving

$$
\begin{split}
\hat{B} &= (\tilde{X}^\mathrm{T} \tilde{X})^{-1} \tilde{X}^\mathrm{T} \tilde{Y} \\
&= \left((WX)^\mathrm{T} WX \right)^{-1} (WX)^\mathrm{T} WY \\
&= \left(X^\mathrm{T} W^\mathrm{T} W X \right)^{-1} X^\mathrm{T} W^\mathrm{T} W Y \\
&= \left(X^\mathrm{T} W W X \right)^{-1} X^\mathrm{T} W W Y \\
&\overset{(4)}{=} \left(X^\mathrm{T} V^{-1} X \right)^{-1} X^\mathrm{T} V^{-1} Y
\end{split} \tag{7}
$$

which corresponds to the weighted least squares solution (2).

■

2.1.4 Maximum likelihood estimation

Theorem: Given a general linear model (→ III/2.1.1) with matrix-normally distributed (→ II/5.1.1) errors

$$
Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; , \tag{1}
$$

maximum likelihood estimates (→ I/4.1.3) for the unknown parameters B and Σ are given by

$$
\begin{split}
\hat{B} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y \\
\hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \; .
\end{split} \tag{2}
$$

Proof: In (1), Y is an $n \times v$ matrix of measurements (n observations, v dependent variables), X is an $n \times p$ design matrix (n observations, p independent variables) and V is an $n \times n$ covariance matrix across observations. This multivariate GLM implies the following likelihood function (→ I/5.1.2)

$$
\begin{split}
p(Y|B,\Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\
&= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y - XB)^\mathrm{T} V^{-1} (Y - XB) \right) \right]
\end{split} \tag{3}
$$

and the log-likelihood function (→ I/4.1.2)

$$
\begin{split}
\mathrm{LL}(B,\Sigma) = &\log p(Y|B,\Sigma) \\
= &-\frac{nv}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{v}{2} \log|V| \\
&- \frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} (Y - XB)^\mathrm{T} V^{-1} (Y - XB) \right] \; .
\end{split} \tag{4}
$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$
\begin{split}
\mathrm{LL}(B,\Sigma) = &- \frac{nv}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| + \frac{v}{2} \log|P| \\
&- \frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} \left(Y^\mathrm{T} P Y - Y^\mathrm{T} P X B - B^\mathrm{T} X^\mathrm{T} P Y + B^\mathrm{T} X^\mathrm{T} P X B \right) \right] \; .
\end{split} \tag{5}
$$

The derivative of the log-likelihood function (5) with respect to B is

$$
\begin{split}
\frac{\mathrm{d}\mathrm{LL}(B,\Sigma)}{\mathrm{d}B} &= \frac{\mathrm{d}}{\mathrm{d}B} \left(-\frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} \left(Y^\mathrm{T} P Y - Y^\mathrm{T} P X B - B^\mathrm{T} X^\mathrm{T} P Y + B^\mathrm{T} X^\mathrm{T} P X B \right) \right] \right) \\
&= \frac{\mathrm{d}}{\mathrm{d}B} \left(-\frac{1}{2} \mathrm{tr}\left[-2 \Sigma^{-1} Y^\mathrm{T} P X B \right] \right) + \frac{\mathrm{d}}{\mathrm{d}B} \left(-\frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} B^\mathrm{T} X^\mathrm{T} P X B \right] \right) \\
&= -\frac{1}{2} \left(-2 X^\mathrm{T} P Y \Sigma^{-1} \right) - \frac{1}{2} \left(X^\mathrm{T} P X B \Sigma^{-1} + (X^\mathrm{T} P X)^\mathrm{T} B (\Sigma^{-1})^\mathrm{T} \right) \\
&= X^\mathrm{T} P Y \Sigma^{-1} - X^\mathrm{T} P X B \Sigma^{-1}
\end{split} \tag{6}
$$

and setting this derivative to zero gives the MLE for B:

$$
\begin{split}
\frac{\mathrm{d}\mathrm{LL}(\hat{B},\Sigma)}{\mathrm{d}B} &= 0 \\
0 &= X^\mathrm{T} P Y \Sigma^{-1} - X^\mathrm{T} P X \hat{B} \Sigma^{-1} \\
0 &= X^\mathrm{T} P Y - X^\mathrm{T} P X \hat{B} \\
X^\mathrm{T} P X \hat{B} &= X^\mathrm{T} P Y \\
\hat{B} &= \left(X^\mathrm{T} P X \right)^{-1} X^\mathrm{T} P Y \; .
\end{split} \tag{7}
$$

The derivative of the log-likelihood function (4) at $\hat{B}$ with respect to Σ is

$$
\begin{split}
\frac{\mathrm{d}\mathrm{LL}(\hat{B},\Sigma)}{\mathrm{d}\Sigma} &= \frac{\mathrm{d}}{\mathrm{d}\Sigma} \left(-\frac{n}{2} \log|\Sigma| - \frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \right] \right) \\
&= -\frac{n}{2} \left(\Sigma^{-1} \right)^\mathrm{T} + \frac{1}{2} \left(\Sigma^{-1} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \, \Sigma^{-1} \right)^\mathrm{T} \\
&= -\frac{n}{2} \, \Sigma^{-1} + \frac{1}{2} \, \Sigma^{-1} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \, \Sigma^{-1}
\end{split} \tag{8}
$$

and setting this derivative to zero gives the MLE for Σ:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\hat{\Sigma})}{\mathrm{d}\Sigma} &= 0 \\
0 &= -\frac{n}{2}\,\hat{\Sigma}^{-1} + \frac{1}{2}\,\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
\frac{n}{2}\,\hat{\Sigma}^{-1} &= \frac{1}{2}\,\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n}\,\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n}\,(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\,\hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n}\,(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \; .
\end{aligned} \tag{9}
$$

Together, (7) and (9) constitute the MLE for the GLM.

■

2.1.5 Maximum log-likelihood

Theorem: Consider a general linear model (→ III/2.1.1) m with $n \times v$ data matrix Y, $n \times p$ design matrix X and $n \times n$ covariance across rows (→ III/2.1.1) V

$$
m: \; Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; . \tag{1}
$$

Then, the maximum log-likelihood (→ I/4.1.4) for this model is

$$
\mathrm{MLL}(m) = -\frac{nv}{2}\ln(2\pi) - \frac{n}{2}\ln|\hat{\Sigma}| - \frac{nv}{2} \tag{2}
$$

under uncorrelated observations (→ III/2.1.1), i.e. if $V = I_n$, and

$$
\mathrm{MLL}(m) = -\frac{nv}{2}\ln(2\pi) - \frac{n}{2}\ln|\hat{\Sigma}| - \frac{v}{2}\ln|V| - \frac{nv}{2} \; , \tag{3}
$$

in the general case, i.e. if $V \neq I_n$, where $\hat{\Sigma}$ is the maximum likelihood estimate (→ I/4.1.3) of the $v \times v$ covariance across columns (→ III/2.1.1).

Proof: The likelihood function (→ I/5.1.2) for the general linear model is given by (→ III/2.1.4)

$$
\begin{aligned}
p(Y|B,\Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\
&= \sqrt{\frac{1}{(2\pi)^{nv}|\Sigma|^n|V|^v}} \cdot \exp\left[-\frac{1}{2}\,\mathrm{tr}\left(\Sigma^{-1}(Y - XB)^\mathrm{T} V^{-1} (Y - XB)\right)\right] \; ,
\end{aligned} \tag{4}
$$

such that the log-likelihood function (→ I/4.1.2) for this model becomes (→ III/2.1.4)

$$
\mathrm{LL}(B,\Sigma) = -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{v}{2}\log|V| - \frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}(Y - XB)^\mathrm{T} V^{-1} (Y - XB)\right] \; . \tag{5}
$$

The maximum likelihood estimate for the noise covariance (→ III/2.1.4) is

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \tag{6}$$

Plugging (6) into (5), we obtain the maximum log-likelihood (→ I/4.1.4) as

$$\begin{split}
\mathrm{MLL}(m) &= \mathrm{LL}(\hat{B}, \hat{\Sigma}) \\
&= -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{v}{2}\log|V| - \frac{1}{2}\mathrm{tr}\left[\hat{\Sigma}^{-1}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\right] \\
&= -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{v}{2}\log|V| \\
&\quad - \frac{1}{2}\mathrm{tr}\left[\left(\frac{1}{n}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\right)^{-1} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B})\right] \\
&= -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{v}{2}\log|V| - \frac{n}{2}\mathrm{tr}\left[I_v\right] \\
&= -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{v}{2}\log|V| - \frac{nv}{2}
\end{split} \tag{7}$$

which proves the result in (3). Assuming $V = I_n$, we have

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^\mathrm{T} (Y - X\hat{B}) \tag{8}$$

and

$$\frac{v}{2}\log|V| = \frac{v}{2}\log|I_n| = \frac{v}{2}\log 1 = 0 \;, \tag{9}$$

such that

$$\mathrm{MLL}(m) = -\frac{nv}{2}\ln(2\pi) - \frac{n}{2}\ln|\hat{\Sigma}| - \frac{nv}{2} \tag{10}$$

which proves the result in (2). This completes the proof.

■

2.1.6 Log-likelihood ratio

Theorem: Let $Y = [y_1, \ldots, y_v]$ be an $n \times v$ data matrix and consider two general linear models (→ III/2.1.1) with design matrices (→ III/2.1.1) X_1, X_2 and row-by-row covariance matrices (→ III/2.1.1) V_1, V_2, entailing potentially different regression coefficients (→ III/2.1.1) B_1, B_2 and column-by-column covariance matrices (→ III/2.1.1) Σ_1, Σ_2:

$$\begin{split}
m_1 &: \; Y = X_1 B_1 + E_1, \; E_1 \sim \mathcal{N}(0, V_1, \Sigma_1) \\
m_2 &: \; Y = X_2 B_2 + E_2, \; E_2 \sim \mathcal{N}(0, V_2, \Sigma_2) \; .
\end{split} \tag{1}$$

Then, if the models assume the same covariance matrix across observations, i.e. if $V_1 = V_2$, the log-likelihood ratio (→ I/4.1.7) for comparing m_1 vs. m_2 is given by

$$\ln \Lambda_{12} = \frac{n}{2} \ln \frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} \tag{2}$$

where $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are the maximum likelihood estimates (→ I/4.1.3) of Σ_1 and Σ_2.

Proof: The likelihood ratio (→ I/4.1.6) between two models m_1 and m_2 with model parameters θ_1 and θ_2 and parameter spaces Θ_1 and Θ_2 is defined as the quotient of their maximized (→ I/4.1.3) likelihood functions (→ I/5.1.2):

$$\Lambda_{12} = \frac{\max_{\theta_1 \in \Theta_1} p(y|\theta_1, m_1)}{\max_{\theta_2 \in \Theta_2} p(y|\theta_2, m_2)} \; . \tag{3}$$

Thus, the log-likelihood ratio (→ I/4.1.7) is equal to the difference of the maximum log-likelihoods (→ I/4.1.4) of the two models:

$$\ln \Lambda_{12} = \ln p(y|\hat{\theta}_1, m_1) - \ln p(y|\hat{\theta}_2, m_2) \; . \tag{4}$$

The likelihood function (→ I/5.1.2) of the general linear model (→ III/2.1.1) is a matrix-normal probability density function (→ II/5.1.3):

$$\begin{split} p(Y|B,\Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y-XB)^\mathrm{T} V^{-1} (Y-XB) \right) \right] \; . \end{split} \tag{5}$$

Thus, the log-likelihood function (→ I/4.1.2) is equal to a logarithmized matrix-normal (→ II/5.1.1) density (→ I/1.7.1):

$$\begin{split} \ln p(Y|B,\Sigma) &= \ln \mathcal{MN}(Y; XB, V, \Sigma) \\ &= -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln|\Sigma| - \frac{v}{2} \ln|V| - \frac{1}{2} \mathrm{tr}\left[\Sigma^{-1} (Y-XB)^\mathrm{T} V^{-1} (Y-XB) \right] \; . \end{split} \tag{6}$$

The maximum likelihood estimates for the general linear model (→ III/2.1.4) are given by

$$\begin{split} \hat{B} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y \\ \hat{\Sigma} &= \frac{1}{n} (Y-X\hat{B})^\mathrm{T} V^{-1} (Y-X\hat{B}) \; . \end{split} \tag{7}$$

such that the last term in the maximum log-likelihood function (6) becomes

$$\begin{split} &\frac{1}{2} \mathrm{tr}\left[\hat{\Sigma}^{-1} (Y-X\hat{B})^\mathrm{T} V^{-1} (Y-X\hat{B}) \right] \\ = &\frac{1}{2} \mathrm{tr}\left[\left(\frac{1}{n} (Y-X\hat{B})^\mathrm{T} V^{-1} (Y-X\hat{B}) \right)^{-1} (Y-X\hat{B})^\mathrm{T} V^{-1} (Y-X\hat{B}) \right] \\ = &\frac{1}{2} \mathrm{tr}\left[n \left((Y-X\hat{B})^\mathrm{T} V^{-1} (Y-X\hat{B}) \right)^{-1} \left((Y-X\hat{B})^\mathrm{T} V^{-1} (Y-X\hat{B}) \right) \right] \\ = &\frac{n}{2} \mathrm{tr}\left[I_v \right] \\ = &\frac{nv}{2} \; . \end{split} \tag{8}$$

Thus, the maximum log-likelihood for the general linear model (→ III/2.1.5) is equal to

$$\ln p(Y|\hat{B},\hat{\Sigma}) = -\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln|\hat{\Sigma}| - \frac{v}{2} \ln|V| - \frac{nv}{2} \; . \tag{9}$$

Evaluating (9) for m_1 and m_2 and plugging into (4), we obtain:

$$\begin{aligned}
\ln \Lambda_{12} &= \ln p(Y|\hat{B}_1, \hat{\Sigma}_1, m_1) - \ln p(Y|\hat{B}_2, \hat{\Sigma}_2, m_2) \\
&= \left(-\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln|\hat{\Sigma}_1| - \frac{v}{2} \ln|V_1| - \frac{nv}{2} \right) \\
&- \left(-\frac{nv}{2} \ln(2\pi) - \frac{n}{2} \ln|\hat{\Sigma}_2| - \frac{v}{2} \ln|V_2| - \frac{nv}{2} \right) \\
&= -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} - \frac{v}{2} \ln \frac{|V_1|}{|V_2|} \; .
\end{aligned} \tag{10}$$

Thus, if $V_1 = V_2$, such that $\ln(|V_2|/|V_1|) = \ln(1) = 0$, the log-likelihood ratio is equal to

$$\ln \Lambda_{12} = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \; . \tag{11}$$

■

2.1.7 Mutual information

Theorem: Consider a general linear model (→ III/2.1.1) m_1 with $n \times v$ data matrix Y, $n \times p$ design matrix X and uncorrelated observations (→ III/2.1.1), i.e. $V = I_n$,

$$m_1: \; Y = XB + E_1, \; E_1 \sim \mathcal{MN}(0, I_n, \Sigma_1) \; , \tag{1}$$

as well as another model m_0 in which X has no influence on Y:

$$m_0: \; Y = E_0, \; E_0 \sim \mathcal{MN}(0, I_n, \Sigma_0) \; . \tag{2}$$

Then, the mutual information (→ I/2.4.1) of Y and X is equal to

$$I(X,Y) = -\frac{n}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} \; . \tag{3}$$

Proof: The continuous mutual information can be written in terms of marginal and conditional differential entropy (→ I/2.4.2) as follows:

$$\mathrm{I}(X,Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X) \; . \tag{4}$$

The marginal distribution of Y, unconditional on X, is given by model m_0

$$Y \sim \mathcal{MN}(0, I_n, \Sigma_0) \tag{5}$$

and the conditional distribution of Y given X is given by model m_1

$$Y \sim \mathcal{MN}(XB, I_n, \Sigma_1) \; . \tag{6}$$

Since X is constant (→ I/1.2.5) and thus only has one possible value (→ I/1.1.2), the conditional differential entropy (→ I/2.2.7) of Y given X is obtained by simply entering X into the probability distribution (→ I/1.5.1) for which the differential entropy (→ I/2.2.7) is calculated:

$$
\begin{aligned}
\mathrm{h}(Y|X) &= \int_{z \in \mathcal{X}} p(z) \cdot \mathrm{h}(Y|z) \, \mathrm{d}z \\
&= p(X) \cdot \mathrm{h}(Y|X) \\
&= \mathrm{h}\left[p(Y|X,B,\Sigma_1)\right] \; .
\end{aligned} \tag{7}
$$

The differential entropy of the matrix-normal distribution (→ II/5.1.6) is

$$
\begin{split}
& X \sim \mathcal{MN}(M,U,V) \quad \text{where} \quad X \in \mathbb{R}^{n \times p} \\
\Rightarrow \quad & \mathrm{h}(X) = \frac{np}{2} \ln(2\pi) + \frac{n}{2} \ln|V| + \frac{p}{2} \ln|U| + \frac{np}{2}
\end{split} \tag{8}
$$

such that the mutual information of Y and X becomes

$$
\begin{split}
\mathrm{I}(X,Y) &= \mathrm{h}\left[p(Y|\Sigma_0)\right] - \mathrm{h}\left[p(Y|X,B,\Sigma_1)\right] \\
&= \mathrm{h}\left[\mathcal{MN}(0,I_n,\Sigma_0)\right] - \mathrm{h}\left[\mathcal{MN}(XB,I_n,\Sigma_1)\right] \\
&= \left(\frac{nv}{2} \ln(2\pi) + \frac{n}{2} \ln|\Sigma_0| + \frac{v}{2} \ln|I_n| + \frac{nv}{2}\right) \\
&- \left(\frac{nv}{2} \ln(2\pi) + \frac{n}{2} \ln|\Sigma_1| + \frac{v}{2} \ln|I_n| + \frac{nv}{2}\right) \\
&= \frac{n}{2} \ln|\Sigma_0| - \frac{n}{2} \ln|\Sigma_1| \\
&= \frac{n}{2} \ln \frac{|\Sigma_0|}{|\Sigma_1|} \\
&= -\frac{n}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} \; .
\end{split} \tag{9}
$$

■

2.1.8 Log-likelihood ratio and estimated mutual information

Theorem: Consider a general linear model (→ III/2.1.1) m_1 with $n \times v$ data matrix Y, $n \times p$ design matrix X and uncorrelated observations (→ III/2.1.1), i.e. $V = I_n$,

$$
m_1: \; Y = XB + E_1, \; E_1 \sim \mathcal{MN}(0, I_n, \Sigma_1) \; , \tag{1}
$$

as well as another model m_0 in which X has no influence on Y:

$$
m_0: \; Y = E_0, \; E_0 \sim \mathcal{MN}(0, I_n, \Sigma_0) \; . \tag{2}
$$

Then, the log-likelihood ratio (→ I/4.1.7) of m_1 vs. m_0 is equal to the estimated mutual information (→ I/2.4.1) of X and Y:

$$
\ln \Lambda_{10} = \hat{I}(X,Y) \; . \tag{3}
$$

Proof: The maximum likelihood estimates for a general linear model (→ III/2.1.4) are

$$
\begin{aligned}
\hat{B} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y \\
\hat{\Sigma} &= \frac{1}{n} (Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \; ,
\end{aligned} \tag{4}
$$

such that, for the two models, the maximum likelihood estimates (→ I/4.1.3) are:

$$
\begin{aligned}
\hat{\Sigma}_1 &= \frac{1}{n} (Y - X\hat{B})^\mathrm{T} (Y - X\hat{B}) \quad \text{with} \quad \hat{B} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y \quad \text{and} \\
\hat{\Sigma}_0 &= \frac{1}{n} Y^\mathrm{T} Y \; .
\end{aligned} \tag{5}
$$

The log-likelihood ratio for two general linear models (→ III/2.1.6) is

$$
\ln \Lambda_{12} = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \; , \tag{6}
$$

such that in the present case, we have:

$$
\ln \Lambda_{10} = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|} \; . \tag{7}
$$

The mutual information for the general linear model (→ III/2.1.7) is

$$
I(X,Y) = -\frac{n}{2} \ln \frac{|\Sigma_1|}{|\Sigma_0|} \; , \tag{8}
$$

such that with (5), the estimated mutual information is:

$$
\hat{I}(X,Y) = -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|} \; , \tag{9}
$$

Together, (7) and (9) show that

$$
\ln \Lambda_{10} = \hat{I}(X,Y) \; . \tag{10}
$$

■

Sources:

- Friston K, Chu C, Mourão-Miranda J, Hulme O, Rees G, Penny W, Ashburner J (2008): "Bayesian decoding of brain images"; in: *NeuroImage*, vol. 39, pp. 181-205, eq. 6; URL: https://www.sciencedirect.com/science/article/abs/pii/S1053811907007203; DOI: 10.1016/j.neuroimage.2007.08.013.

2.2 Transformed general linear model

2.2.1 Definition

Definition: Let there be two general linear models (→ III/2.1.1) of measured data $Y \in \mathbb{R}^{n \times v}$ using design matrices (→ III/2.1.1) $X \in \mathbb{R}^{n \times p}$ and $X_t \in \mathbb{R}^{n \times t}$

$$
Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}
$$

$$Y = X_t \Gamma + E_t, \; E_t \sim \mathcal{MN}(0, V, \Sigma_t) \tag{2}$$

and assume that X_t can be transformed into X using a transformation matrix $T \in \mathbb{R}^{t \times p}$

$$X = X_t T \tag{3}$$

where $p < t$ and X, X_t and T have full ranks $\mathrm{rk}(X) = p$, $\mathrm{rk}(X_t) = t$ and $\mathrm{rk}(T) = p$.
Then, a linear model (→ III/2.1.1) of the parameter estimates from (2), under the assumption of (1), is called a transformed general linear model.

Sources:
- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, Appendix A; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.

2.2.2 Derivation of the distribution

Theorem: Let there be two general linear models (→ III/2.1.1) of measured data Y

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

$$Y = X_t \Gamma + E_t, \; E_t \sim \mathcal{MN}(0, V, \Sigma_t) \tag{2}$$

and a matrix T transforming X_t into X:

$$X = X_t T \; . \tag{3}$$

Then, the transformed general linear model (→ III/2.2.1) is given by

$$\hat{\Gamma} = TB + H, \; H \sim \mathcal{MN}(0, U, \Sigma) \tag{4}$$

where the covariance across rows (→ II/5.1.1) is $U = (X_t^\mathrm{T} V^{-1} X_t)^{-1}$.

Proof: The linear transformation theorem for the matrix-normal distribution (→ II/5.1.9) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^\mathrm{T}, B^\mathrm{T} V B) \; . \tag{5}$$

The weighted least squares parameter estimates (→ III/2.1.3) for (2) are given by

$$\hat{\Gamma} = (X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} Y \; . \tag{6}$$

Using (1) and (5), the distribution of Y is

$$Y \sim \mathcal{MN}(XB, V, \Sigma) \tag{7}$$

Combining (6) with (7), the distribution of $\hat{\Gamma}$ is

$$\begin{split}
\hat{\Gamma} &\sim \mathcal{MN}\left(\left[(X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} \right] XB, \left[(X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} \right] V \left[V^{-1} X_t (X_t^\mathrm{T} V^{-1} X_t)^{-1} \right], \Sigma \right) \\
&\sim \mathcal{MN}\left((X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} X_t T B, (X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} X_t (X_t^\mathrm{T} V^{-1} X_t)^{-1}, \Sigma \right) \\
&\sim \mathcal{MN}\left(TB, (X_t^\mathrm{T} V^{-1} X_t)^{-1}, \Sigma \right) \; .
\end{split} \tag{8}$$

This can also be written as

$$\hat{\Gamma} = TB + H, \; H \sim \mathcal{MN}\left(0, (X_t^\mathrm{T} V^{-1} X_t)^{-1}, \Sigma\right) \tag{9}$$

which is equivalent to (4).

■

Sources:

- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, Appendix A, Theorem 1; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.

2.2.3 Equivalence of parameter estimates

Theorem: Let there be a general linear model (→ III/2.1.1)

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

and the transformed general linear model (→ III/2.2.1)

$$\hat{\Gamma} = TB + H, \; H \sim \mathcal{MN}(0, U, \Sigma) \tag{2}$$

which are linked to each other (→ III/2.2.2) via

$$\hat{\Gamma} = (X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} Y \tag{3}$$

and

$$X = X_t T \; . \tag{4}$$

Then, the parameter estimates for B from (1) and (2) are equivalent.

Proof: The weighted least squares parameter estimates (→ III/2.1.3) for (1) are given by

$$\hat{B} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y \tag{5}$$

and the weighted least squares parameter estimates (→ III/2.1.3) for (2) are given by

$$\hat{B} = (T^\mathrm{T} U^{-1} T)^{-1} T^\mathrm{T} U^{-1} \hat{\Gamma} \; . \tag{6}$$

The covariance across rows for the transformed general linear model (→ III/2.2.2) is equal to

$$U = (X_t^\mathrm{T} V^{-1} X_t)^{-1} \; . \tag{7}$$

Applying (7), (4) and (3), the estimates in (6) can be developed into

$$\begin{aligned}
\hat{B} &\overset{(6)}{=} (T^\mathrm{T}\, U^{-1}\, T)^{-1}\, T^\mathrm{T}\, U^{-1}\, \hat{\Gamma} \\
&\overset{(7)}{=} (T^\mathrm{T} \left[X_t^\mathrm{T} V^{-1} X_t \right] T)^{-1}\, T^\mathrm{T} \left[X_t^\mathrm{T} V^{-1} X_t \right] \hat{\Gamma} \\
&\overset{(4)}{=} (X^\mathrm{T} V^{-1} X)^{-1}\, T^\mathrm{T}\, X_t^\mathrm{T} V^{-1} X_t\, \hat{\Gamma} \\
&\overset{(3)}{=} (X^\mathrm{T} V^{-1} X)^{-1}\, T^\mathrm{T}\, X_t^\mathrm{T} V^{-1} X_t \left[(X_t^\mathrm{T} V^{-1} X_t)^{-1} X_t^\mathrm{T} V^{-1} Y \right] \\
&= (X^\mathrm{T} V^{-1} X)^{-1}\, T^\mathrm{T}\, X_t^\mathrm{T} V^{-1} Y \\
&\overset{(4)}{=} (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} Y
\end{aligned} \tag{8}$$

which is equivalent to the estimates in (5).

■

Sources:
- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, Appendix A, Theorem 2; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.

2.3 Inverse general linear model

2.3.1 Definition

Definition: Let there be a general linear model (→ III/2.1.1) of measured data $Y \in \mathbb{R}^{n \times v}$ in terms of the design matrix (→ III/2.1.1) $X \in \mathbb{R}^{n \times p}$:

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; . \tag{1}$$

Then, a linear model (→ III/2.1.1) of X in terms of Y, under the assumption of (1), is called an inverse general linear model.

Sources:
- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, Appendix C; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.

2.3.2 Derivation of the distribution

Theorem: Let there be a general linear model (→ III/2.1.1) of $Y \in \mathbb{R}^{n \times v}$

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; . \tag{1}$$

Then, the inverse general linear model (→ III/2.3.1) of $X \in \mathbb{R}^{n \times p}$ is given by

$$X = YW + N, \; N \sim \mathcal{MN}(0, V, \Sigma_x) \tag{2}$$

where $W \in \mathbb{R}^{v \times p}$ is a matrix, such that $B\,W = I_p$, and the covariance across columns (→ II/5.1.1) is $\Sigma_x = W^\mathrm{T} \Sigma W$.

Proof: The linear transformation theorem for the matrix-normal distribution (→ II/5.1.9) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^{\mathrm{T}}, B^{\mathrm{T}}VB) \; . \tag{3}$$

The matrix W exists, if the rows of $B \in \mathbb{R}^{p \times v}$ are linearly independent, such that $\mathrm{rk}(B) = p$. Then, right-multiplying the model (1) with W and applying (3) yields

$$YW = XBW + EW, \; EW \sim \mathcal{MN}(0, V, W^{\mathrm{T}}\Sigma W) \; . \tag{4}$$

Employing $BW = I_p$ and rearranging, we have

$$X = YW - EW, \; EW \sim \mathcal{MN}(0, V, W^{\mathrm{T}}\Sigma W) \; . \tag{5}$$

Substituting $N = -EW$, we get

$$X = YW + N, \; N \sim \mathcal{MN}(0, V, W^{\mathrm{T}}\Sigma W) \tag{6}$$

which is equivalent to (2).

■

Sources:

- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, Appendix C, Theorem 4; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.

2.3.3 Best linear unbiased estimator

Theorem: Let there be a general linear model (→ III/2.1.1) of $Y \in \mathbb{R}^{n \times v}$

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

implying the inverse general linear model (→ III/2.3.2) of $X \in \mathbb{R}^{n \times p}$

$$X = YW + N, \; N \sim \mathcal{MN}(0, V, \Sigma_x) \; . \tag{2}$$

where

$$BW = I_p \quad \text{and} \quad \Sigma_x = W^{\mathrm{T}}\Sigma W \; . \tag{3}$$

Then, the weighted least squares solution (→ III/2.1.3) for W is the best linear unbiased estimator of W.

Proof: The linear transformation theorem for the matrix-normal distribution (→ II/5.1.9) states:

$$X \sim \mathcal{MN}(M, U, V) \quad \Rightarrow \quad Y = AXB + C \sim \mathcal{MN}(AMB + C, AUA^{\mathrm{T}}, B^{\mathrm{T}}VB) \; . \tag{4}$$

The weighted least squares parameter estimates (→ III/2.1.3) for (2) are given by

$$\hat{W} = (Y^{\mathrm{T}}V^{-1}Y)^{-1}Y^{\mathrm{T}}V^{-1}X \; . \tag{5}$$

The best linear unbiased estimator $\hat{\theta}$ of a certain quantity θ estimated from measured data y is 1) an estimator resulting from a linear operation $f(y)$, 2) whose expected value is equal to θ and 3) which has, among those satisfying 1) and 2), the minimum variance (→ I/1.11.1).

1) First, $\hat{W}$ is a linear estimator, because it is of the form $\tilde{W} = MX$ where M is an arbitrary $v \times n$ matrix.

2) Second, $\hat{W}$ is an unbiased estimator, if $\left\langle \hat{W} \right\rangle = W$. By applying (4) to (2), the distribution of $\tilde{W}$ is

$$\tilde{W} = MX \sim \mathcal{MN}(MYW, MVM^T, \Sigma_x) \tag{6}$$

which requires (→ II/5.1.4) that $MY = I_v$. This is fulfilled by any matrix

$$M = (Y^\mathrm{T} V^{-1} Y)^{-1} Y^\mathrm{T} V^{-1} + D \tag{7}$$

where D is a $v \times n$ matrix which satisfies $DY = 0$.

3) Third, the best linear unbiased estimator is the one with minimum variance (→ I/1.11.1), i.e. the one that minimizes the expected Frobenius norm

$$\mathrm{Var}\left(\tilde{W}\right) = \left\langle \mathrm{tr}\left[(\tilde{W} - W)^\mathrm{T}(\tilde{W} - W)\right] \right\rangle \; . \tag{8}$$

Using the matrix-normal distribution (→ II/5.1.1) of $\tilde{W}$ from (6)

$$\left(\tilde{W} - W\right) \sim \mathcal{MN}(0, MVM^\mathrm{T}, \Sigma_x) \tag{9}$$

and the property of the Wishart distribution (→ II/5.2.1)

$$X \sim \mathcal{MN}(0, U, V) \quad \Rightarrow \quad \left\langle XX^\mathrm{T} \right\rangle = \mathrm{tr}(V)\, U \; , \tag{10}$$

this variance (→ I/1.11.1) can be evaluated as a function of M:

$$\begin{aligned}
\mathrm{Var}\left[\tilde{W}(M)\right] &\overset{(8)}{=} \left\langle \mathrm{tr}\left[(\tilde{W} - W)^\mathrm{T}(\tilde{W} - W)\right] \right\rangle \\
&= \left\langle \mathrm{tr}\left[(\tilde{W} - W)(\tilde{W} - W)^\mathrm{T}\right] \right\rangle \\
&= \mathrm{tr}\left[\left\langle (\tilde{W} - W)(\tilde{W} - W)^\mathrm{T} \right\rangle\right] \\
&\overset{(10)}{=} \mathrm{tr}\left[\mathrm{tr}(\Sigma_x)\, MVM^\mathrm{T}\right] \\
&= \mathrm{tr}(\Sigma_x)\; \mathrm{tr}(MVM^\mathrm{T}) \; .
\end{aligned} \tag{11}$$

As a function of D and using $DY = 0$, it becomes:

$$\begin{aligned}
\mathrm{Var}\left[\tilde{W}(D)\right] &\overset{(7)}{=} \mathrm{tr}(\Sigma_x)\; \mathrm{tr}\Big[\left((Y^\mathrm{T} V^{-1} Y)^{-1} Y^\mathrm{T} V^{-1} + D\right) V \left((Y^\mathrm{T} V^{-1} Y)^{-1} Y^\mathrm{T} V^{-1} + D\right)^\mathrm{T}\Big] \\
&= \mathrm{tr}(\Sigma_x)\; \mathrm{tr}\Big[(Y^\mathrm{T} V^{-1} Y)^{-1}\, Y^\mathrm{T} V^{-1} V V^{-1} Y\; (Y^\mathrm{T} V^{-1} Y)^{-1} + \\
&\qquad (Y^\mathrm{T} V^{-1} Y)^{-1} Y^\mathrm{T} V^{-1} V D^\mathrm{T} + D V V^{-1} Y (Y^\mathrm{T} V^{-1} Y)^{-1} + D V D^\mathrm{T}\Big] \\
&= \mathrm{tr}(\Sigma_x)\; \Big[\mathrm{tr}\big((Y^\mathrm{T} V^{-1} Y)^{-1}\big) + \mathrm{tr}\big(D V D^\mathrm{T}\big)\Big] \; .
\end{aligned} \tag{12}$$

Since DVD^T is a positive-semidefinite matrix, all its eigenvalues are non-negative. Because the trace of a square matrix is the sum of its eigenvalues, the mimimum variance is achieved by $D = 0$, thus producing $\hat{W}$ as in (5).

■

Sources:

- Soch J, Allefeld C, Haynes JD (2020): "Inverse transformed encoding models – a solution to the problem of correlated trial-by-trial parameter estimates in fMRI decoding"; in: *NeuroImage*, vol. 209, art. 116449, Appendix C, Theorem 5; URL: https://www.sciencedirect.com/science/article/pii/S1053811919310407; DOI: 10.1016/j.neuroimage.2019.116449.

2.3.4 Equivalence of log-likelihood ratios

Theorem: Consider two general linear models (→ III/2.1.1)

$$
\begin{split}
m_1^{(Y)} &: Y = XB + E_1, \; E_1 \sim \mathcal{MN}(0, I_n, \Sigma_1^{(Y)}) \\
m_0^{(Y)} &: Y = E_0, \; E_0 \sim \mathcal{MN}(0, I_n, \Sigma_0^{(Y)})
\end{split} \tag{1}
$$

and two inverse general linear models (→ III/2.3.1)

$$
\begin{split}
m_1^{(X)} &: X = YW + N_1, \; N_1 \sim \mathcal{MN}(0, I_n, \Sigma_1^{(X)}) \\
m_0^{(X)} &: X = N_0, \; N_0 \sim \mathcal{MN}(0, I_n, \Sigma_0^{(X)})
\end{split} \tag{2}
$$

where $Y \in \mathbb{R}^{n \times v}$ and $X \in \mathbb{R}^{n \times p}$ are data matrices, such that $n > v$ and $n > p$. Then, the log-likelihood ratio (→ I/4.1.7) comparing the forward models (→ III/2.1.1) is equivalent to the log-likelihood ratio (→ I/4.1.7) comparing the backward models (→ III/2.3.1):

$$
\ln \Lambda_{10}^{(Y)} = \ln \Lambda_{10}^{(X)} \; . \tag{3}
$$

Proof: The maximum likelihood estimates for the general linear models (→ III/2.1.4) are

$$
\begin{split}
\hat{\Sigma}_1^{(Y)} &= \frac{1}{n}(Y - X\hat{B})^\mathrm{T}(Y - X\hat{B}) \quad \text{with} \quad \hat{B} = (X^\mathrm{T}X)^{-1}X^\mathrm{T}Y \quad \text{and} \\
\hat{\Sigma}_0^{(Y)} &= \frac{1}{n}Y^\mathrm{T}Y
\end{split} \tag{4}
$$

as well as

$$
\begin{split}
\hat{\Sigma}_1^{(X)} &= \frac{1}{n}(X - Y\hat{W})^\mathrm{T}(X - Y\hat{W}) \quad \text{with} \quad \hat{W} = (Y^\mathrm{T}Y)^{-1}Y^\mathrm{T}X \quad \text{and} \\
\hat{\Sigma}_0^{(X)} &= \frac{1}{n}X^\mathrm{T}X \; .
\end{split} \tag{5}
$$

The likelihood ratio for two general linear models (→ III/2.1.6) m_1 and m_2 is:

$$
\begin{split}
\ln \Lambda_{12} &= -\frac{n}{2} \ln \frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \\
&= -\frac{n}{2} \ln \left(|\hat{\Sigma}_2^{-1}||\hat{\Sigma}_1| \right) \\
&= -\frac{n}{2} \ln |\hat{\Sigma}_2^{-1}\hat{\Sigma}_1| \; .
\end{split} \tag{6}
$$

Thus, with (4), the log-likelihood ratio (→ I/4.1.7) of $m_1^{(Y)}$ vs. $m_0^{(Y)}$ is given as

$$
\begin{aligned}
\ln \Lambda_Y = \ln \Lambda_{10}^{(Y)} &\overset{(6)}{=} -\frac{n}{2} \ln \left| \left(\hat{\Sigma}_0^{(Y)}\right)^{-1} \hat{\Sigma}_1^{(Y)} \right| \\
&\overset{(4)}{=} -\frac{n}{2} \ln \left| \left(\frac{1}{n} Y^\mathrm{T} Y\right)^{-1} \frac{1}{n} (Y - X\hat{B})^\mathrm{T} (Y - X\hat{B}) \right| \\
&= -\frac{n}{2} \ln \left| \left(Y^\mathrm{T} Y\right)^{-1} \left(Y^\mathrm{T} Y - 2Y^\mathrm{T} X\hat{B} + \hat{B}^\mathrm{T} X^\mathrm{T} X\hat{B}\right) \right| \\
&= -\frac{n}{2} \ln \left| \left((Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} Y - 2(Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} X\hat{B} + (Y^\mathrm{T} Y)^{-1} \hat{B}^\mathrm{T} X^\mathrm{T} X\hat{B}\right) \right| \\
&\overset{(4)}{=} -\frac{n}{2} \ln \left| I_v - 2\hat{W}\hat{B} + (Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} X (X^\mathrm{T} X)^{-1} X^\mathrm{T} X (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y \right| \\
&= -\frac{n}{2} \ln \left| I_v - 2\hat{W}\hat{B} + (Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} X (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y \right| \\
&= -\frac{n}{2} \ln \left| I_v - 2\hat{W}\hat{B} + \hat{W}\hat{B} \right| \\
&= -\frac{n}{2} \ln \left| I_v - \hat{W}\hat{B} \right| \; .
\end{aligned} \tag{7}
$$

Similarly, with (5), the log-likelihood ratio (→ I/4.1.7) of $m_1^{(X)}$ vs. $m_0^{(X)}$ becomes

$$
\begin{aligned}
\ln \Lambda_X = \ln \Lambda_{10}^{(X)} &\overset{(6)}{=} -\frac{n}{2} \ln \left| \left(\hat{\Sigma}_0^{(X)}\right)^{-1} \hat{\Sigma}_1^{(X)} \right| \\
&\overset{(5)}{=} -\frac{n}{2} \ln \left| \left(\frac{1}{n} X^\mathrm{T} X\right)^{-1} \frac{1}{n} (X - Y\hat{W})^\mathrm{T} (X - Y\hat{W}) \right| \\
&= -\frac{n}{2} \ln \left| \left(X^\mathrm{T} X\right)^{-1} \left(X^\mathrm{T} X - 2X^\mathrm{T} Y\hat{W} + \hat{W}^\mathrm{T} Y^\mathrm{T} Y\hat{W}\right) \right| \\
&= -\frac{n}{2} \ln \left| \left((X^\mathrm{T} X)^{-1} X^\mathrm{T} X - 2(X^\mathrm{T} X)^{-1} X^\mathrm{T} Y\hat{W} + (X^\mathrm{T} X)^{-1} \hat{W}^\mathrm{T} Y^\mathrm{T} Y\hat{W}\right) \right| \\
&\overset{(5)}{=} -\frac{n}{2} \ln \left| I_p - 2\hat{B}\hat{W} + (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y (Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} Y (Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} X \right| \\
&= -\frac{n}{2} \ln \left| I_p - 2\hat{B}\hat{W} + (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y (Y^\mathrm{T} Y)^{-1} Y^\mathrm{T} X \right| \\
&= -\frac{n}{2} \ln \left| I_p - 2\hat{B}\hat{W} + \hat{B}\hat{W} \right| \\
&= -\frac{n}{2} \ln \left| I_p - \hat{B}\hat{W} \right| \; .
\end{aligned} \tag{8}
$$

Sylvester's determinant theorem (also known as the "Weinstein–Aronszajn identity") states that, for two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$, the following identity holds:

$$
|I_m + AB| = |I_n + BA| \; . \tag{9}
$$

Since $\hat{B} \in \mathbb{R}^{p \times v}$ and $(-\hat{W}) \in \mathbb{R}^{v \times p}$, it follows that

$$
\left| I_p - \hat{B}\hat{W} \right| = \left| I_v - \hat{W}\hat{B} \right| \tag{10}
$$

and thus, we finally have:

$$
\ln \Lambda_Y = -\frac{n}{2} \ln \left| I_v - \hat{W}\hat{B} \right| = -\frac{n}{2} \ln \left| I_p - \hat{B}\hat{W} \right| = \ln \Lambda_X \; . \tag{11}
$$

■

Sources:
- Friston K, Chu C, Mourão-Miranda J, Hulme O, Rees G, Penny W, Ashburner J (2008): "Bayesian decoding of brain images"; in: *NeuroImage*, vol. 39, pp. 181-205, p. 183; URL: https://www.sciencedirect.com/science/article/abs/pii/S1053811907007203; DOI: 10.1016/j.neuroimage.2007.08.013.
- Wikipedia (2024): "Weinstein–Aronszajn identity"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-06-28; URL: https://en.wikipedia.org/wiki/Weinstein%E2%80%93Aronszajn_identity.

2.3.5 Corresponding forward model

Definition: Let there be observations $Y \in \mathbb{R}^{n \times v}$ and $X \in \mathbb{R}^{n \times p}$ and consider a weight matrix $W = f(Y, X) \in \mathbb{R}^{v \times p}$ estimated from Y and X, such that right-multiplying Y with the weight matrix gives an estimate or prediction of X:

$$\hat{X} = YW \; . \tag{1}$$

Given that the columns of $\hat{X}$ are linearly independent, then

$$Y = \hat{X} A^{\mathrm{T}} + E \quad \text{with} \quad \hat{X}^{\mathrm{T}} E = 0 \tag{2}$$

is called the corresponding forward model relative to the weight matrix W.

Sources:
- Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014): "On the interpretation of weight vectors of linear models in multivariate neuroimaging"; in: *NeuroImage*, vol. 87, pp. 96–110, eq. 3; URL: https://www.sciencedirect.com/science/article/pii/S1053811913010914; DOI: 10.1016/j.neuroimage.2013.10.067.

2.3.6 Derivation of parameters

Theorem: Let there be observations $Y \in \mathbb{R}^{n \times v}$ and $X \in \mathbb{R}^{n \times p}$ and consider a weight matrix $W = f(Y, X) \in \mathbb{R}^{v \times p}$ predicting X from Y:

$$\hat{X} = YW \; . \tag{1}$$

Then, the parameter matrix of the corresponding forward model (→ III/2.3.5) is equal to

$$A = \Sigma_y W \Sigma_x^{-1} \tag{2}$$

with the "sample covariances (→ I/1.13.2)"

$$\begin{split} \Sigma_x &= \hat{X}^{\mathrm{T}} \hat{X} \\ \Sigma_y &= Y^{\mathrm{T}} Y \; . \end{split} \tag{3}$$

Proof: The corresponding forward model (→ III/2.3.5) is given by

$$Y = \hat{X} A^{\mathrm{T}} + E \; , \tag{4}$$

subject to the constraint that predicted X and errors E are uncorrelated:

$$\hat{X}^{\mathrm{T}} E = 0 \; . \tag{5}$$

With that, we can directly derive the parameter matrix A:

$$\begin{aligned}
Y &\overset{(4)}{=} \hat{X} A^{\mathrm{T}} + E \\
\hat{X} A^{\mathrm{T}} &= Y - E \\
\hat{X}^{\mathrm{T}} \hat{X} A^{\mathrm{T}} &= \hat{X}^{\mathrm{T}} (Y - E) \\
\hat{X}^{\mathrm{T}} \hat{X} A^{\mathrm{T}} &= \hat{X}^{\mathrm{T}} Y - \hat{X}^{\mathrm{T}} E \\
\hat{X}^{\mathrm{T}} \hat{X} A^{\mathrm{T}} &\overset{(5)}{=} \hat{X}^{\mathrm{T}} Y \\
\hat{X}^{\mathrm{T}} \hat{X} A^{\mathrm{T}} &\overset{(1)}{=} W^{\mathrm{T}} Y^{\mathrm{T}} Y \\
\Sigma_x A^{\mathrm{T}} &\overset{(3)}{=} W^{\mathrm{T}} \Sigma_y \\
A^{\mathrm{T}} &= \Sigma_x^{-1} W^{\mathrm{T}} \Sigma_y \\
A &= \Sigma_y W \Sigma_x^{-1} \; .
\end{aligned} \tag{6}$$

■

Sources:
- Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014): "On the interpretation of weight vectors of linear models in multivariate neuroimaging"; in: *NeuroImage*, vol. 87, pp. 96–110, Theorem 1; URL: https://www.sciencedirect.com/science/article/pii/S1053811913010914; DOI: 10.1016/j.neuroimage.2013.10.067.

2.3.7 Proof of existence

Theorem: Let there be observations $Y \in \mathbb{R}^{n \times v}$ and $X \in \mathbb{R}^{n \times p}$ and consider a weight matrix $W = f(Y, X) \in \mathbb{R}^{v \times p}$ predicting X from Y:

$$\hat{X} = Y W \; . \tag{1}$$

Then, there exists a corresponding forward model (→ III/2.3.5).

Proof: The corresponding forward model (→ III/2.3.5) is defined as

$$Y = \hat{X} A^{\mathrm{T}} + E \quad \text{with} \quad \hat{X}^{\mathrm{T}} E = 0 \tag{2}$$

and the parameters of the corresponding forward model (→ III/2.3.6) are equal to

$$A = \Sigma_y W \Sigma_x^{-1} \quad \text{where} \quad \Sigma_x = \hat{X}^{\mathrm{T}} \hat{X} \quad \text{and} \quad \Sigma_y = Y^{\mathrm{T}} Y \; . \tag{3}$$

1) Because the columns of $\hat{X}$ are assumed to be linearly independent by definition of the corresponding forward model (→ III/2.3.5), the matrix $\Sigma_x = \hat{X}^{\mathrm{T}} \hat{X}$ is invertible, such that A in (3) is well-defined.

2) Moreover, the solution for the matrix A satisfies the constraint of the corresponding forward model (→ III/2.3.5) for predicted X and errors E to be uncorrelated which can be shown as follows:

$$
\begin{aligned}
\hat{X}^{\mathrm{T}} E &\overset{(2)}{=} \hat{X}^{\mathrm{T}} \left(Y - \hat{X} A^{\mathrm{T}} \right) \\
&\overset{(3)}{=} \hat{X}^{\mathrm{T}} \left(Y - \hat{X}\, \Sigma_x^{-1} W^{\mathrm{T}} \Sigma_y \right) \\
&= \hat{X}^{\mathrm{T}} Y - \hat{X}^{\mathrm{T}} \hat{X}\, \Sigma_x^{-1} W^{\mathrm{T}} \Sigma_y \\
&\overset{(3)}{=} \hat{X}^{\mathrm{T}} Y - \hat{X}^{\mathrm{T}} \hat{X} \left(\hat{X}^{\mathrm{T}} \hat{X} \right)^{-1} W^{\mathrm{T}} \left(Y^{\mathrm{T}} Y \right) \\
&\overset{(1)}{=} (YW)^{\mathrm{T}} Y - W^{\mathrm{T}} \left(Y^{\mathrm{T}} Y \right) \\
&= W^{\mathrm{T}} Y^{\mathrm{T}} Y - W^{\mathrm{T}} Y^{\mathrm{T}} Y \\
&= 0 \; .
\end{aligned} \tag{4}
$$

This completes the proof.

■

Sources:

- Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014): "On the interpretation of weight vectors of linear models in multivariate neuroimaging"; in: *NeuroImage*, vol. 87, pp. 96–110, Appendix B; URL: https://www.sciencedirect.com/science/article/pii/S1053811913010914; DOI: 10.1016/j.neuroimage.2013.10.067.

2.4 Multivariate Bayesian linear regression

2.4.1 Conjugate prior distribution

Theorem: Let

$$
Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}
$$

be a general linear model (→ III/2.1.1) with measured $n \times v$ data matrix Y, known $n \times p$ design matrix X, known $n \times n$ covariance structure (→ II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (→ II/5.1.1) Σ.
Then, the conjugate prior (→ I/5.2.5) for this model is a normal-Wishart distribution (→ II/5.3.1)

$$
p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \tag{2}
$$

where $T = \Sigma^{-1}$ is the inverse noise covariance (→ I/1.13.9) or noise precision matrix (→ I/1.13.19).

Proof: By definition, a conjugate prior (→ I/5.2.5) is a prior distribution (→ I/5.1.3) that, when combined with the likelihood function (→ I/5.1.2), leads to a posterior distribution (→ I/5.1.7) that belongs to the same family of probability distributions (→ I/1.5.1). This is fulfilled when the prior density and the likelihood function are proportional to the model parameters in the same way, i.e. the model parameters appear in the same functional form in both.
Equation (1) implies the following likelihood function (→ I/5.1.2)

$$
p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \, \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y - XB)^{\mathrm{T}} V^{-1} (Y - XB) \right) \right] \tag{3}
$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B,T) = \mathcal{MN}(Y; XB, P^{-1}, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(T(Y-XB)^\mathrm{T} P (Y-XB)\right)\right] \tag{4}$$

using the $v \times v$ precision matrix (→ I/1.13.19) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (→ I/1.13.19) $P = V^{-1}$.

Seperating constant and variable terms, we have:

$$p(Y|B,T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T(Y-XB)^\mathrm{T} P (Y-XB)\right)\right] . \tag{5}$$

Expanding the product in the exponent, we have:

$$p(Y|B,T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[Y^\mathrm{T}PY - Y^\mathrm{T}PXB - B^\mathrm{T}X^\mathrm{T}PY + B^\mathrm{T}X^\mathrm{T}PXB\right]\right)\right] . \tag{6}$$

Completing the square over B, finally gives

$$p(Y|B,T) = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \cdot |T|^{n/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-\tilde{X}Y)^\mathrm{T}X^\mathrm{T}PX(B-\tilde{X}Y) - Y^\mathrm{T}QY + Y^\mathrm{T}PY\right]\right)\right] \tag{7}$$

where $\tilde{X} = \left(X^\mathrm{T}PX\right)^{-1} X^\mathrm{T}P$ and $Q = \tilde{X}^\mathrm{T}\left(X^\mathrm{T}PX\right)\tilde{X}$.

In other words, the likelihood function (→ I/5.1.2) is proportional to a power of the determinant of T, times an exponential of the trace of T and an exponential of the trace of a squared form of B, weighted by T:

$$p(Y|B,T) \propto |T|^{n/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[Y^\mathrm{T}PY - Y^\mathrm{T}QY\right]\right)\right] \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-\tilde{X}Y)^\mathrm{T}X^\mathrm{T}PX(B-\tilde{X}Y)\right]\right)\right] . \tag{8}$$

The same is true for a normal-Wishart distribution (→ II/5.3.1) over B and T

$$p(B,T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \tag{9}$$

the probability density function of which (→ II/5.3.2)

$$p(B,T) = \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(T(B-M_0)^\mathrm{T}\Lambda_0(B-M_0)\right)\right] \cdot \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \tag{10}$$

exhibits the same proportionality

$$p(B,T) \propto |T|^{(\nu_0+p-v-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\Omega_0\right)\right] \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_0)^\mathrm{T}\Lambda_0(B-M_0)\right]\right)\right] \tag{11}$$

and is therefore conjugate relative to the likelihood.

■

Sources:

- Wikipedia (2020): "Bayesian multivariate linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Bayesian_multivariate_linear_regression#Conjugate_prior_distribution.

2.4.2 Posterior distribution

Theorem: Let

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

be a general linear model (→ III/2.1.1) with measured $n \times v$ data matrix Y, known $n \times p$ design matrix X, known $n \times n$ covariance structure (→ II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (→ II/5.1.1) Σ. Moreover, assume a normal-Wishart prior distribution (→ III/2.4.1) over the model parameters B and $T = \Sigma^{-1}$:

$$p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a normal-Wishart distribution (→ II/5.3.1)

$$p(B, T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned}
M_n &= \Lambda_n^{-1}(X^\mathrm{T} P Y + \Lambda_0 M_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\
\Omega_n &= \Omega_0 + Y^\mathrm{T} P Y + M_0^\mathrm{T} \Lambda_0 M_0 - M_n^\mathrm{T} \Lambda_n M_n \\
\nu_n &= \nu_0 + n \; .
\end{aligned} \tag{4}$$

Proof: According to Bayes' theorem (→ I/5.3.1), the posterior distribution (→ I/5.1.7) is given by

$$p(B, T|Y) = \frac{p(Y|B, T) \, p(B, T)}{p(Y)} \; . \tag{5}$$

Since $p(Y)$ is just a normalization factor, the posterior is proportional (→ I/5.1.9) to the numerator:

$$p(B, T|Y) \propto p(Y|B, T) \, p(B, T) = p(Y, B, T) \; . \tag{6}$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$p(Y|B, \Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y - XB)^\mathrm{T} V^{-1} (Y - XB) \right) \right] \tag{7}$$

which, for mathematical convenience, can also be parametrized as

$$p(Y|B, T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp\left[-\frac{1}{2} \mathrm{tr}\left(T (Y - XB)^\mathrm{T} P (Y - XB) \right) \right] \tag{8}$$

using the $v \times v$ precision matrix (→ I/1.13.19) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (→ I/1.13.19) $P = V^{-1}$.

Combining the likelihood function (→ I/5.1.2) (8) with the prior distribution (→ I/5.1.3) (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$
\begin{split}
p(Y, B, T) &= p(Y|B, T)\, p(B, T) \\
&= \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \exp\left[-\frac{1}{2} \mathrm{tr}\left(T(Y - XB)^\mathrm{T} P (Y - XB)\right)\right] \cdot \\
&\quad \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \exp\left[-\frac{1}{2} \mathrm{tr}\left(T(B - M_0)^\mathrm{T} \Lambda_0 (B - M_0)\right)\right] \cdot \\
&\quad \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T\right)\right] \; .
\end{split} \tag{9}
$$

Collecting identical variables gives:

$$
\begin{split}
p(Y, B, T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T\right)\right] \cdot \\
&\exp\left[-\frac{1}{2} \mathrm{tr}\left(T\left[(Y - XB)^\mathrm{T} P (Y - XB) + (B - M_0)^\mathrm{T} \Lambda_0 (B - M_0)\right]\right)\right] \; .
\end{split} \tag{10}
$$

Expanding the products in the exponent gives:

$$
\begin{split}
p(Y, B, T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T\right)\right] \cdot \\
&\exp\left[-\frac{1}{2} \mathrm{tr}\left(T\left[Y^\mathrm{T} P Y - Y^\mathrm{T} P X B - B^\mathrm{T} X^\mathrm{T} P Y + B^\mathrm{T} X^\mathrm{T} P X B + \right.\right.\right. \\
&\qquad\qquad \left.\left.\left. B^\mathrm{T} \Lambda_0 B - B^\mathrm{T} \Lambda_0 M_0 - M_0^\mathrm{T} \Lambda_0 B + M_0^\mathrm{T} \Lambda_0 \mu_0\right]\right)\right] \; .
\end{split} \tag{11}
$$

Completing the square over B, we finally have

$$
\begin{split}
p(Y, B, T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0 - v - 1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T\right)\right] \cdot \\
&\exp\left[-\frac{1}{2} \mathrm{tr}\left(T\left[(B - M_n)^\mathrm{T} \Lambda_n (B - M_n) + (Y^\mathrm{T} P Y + M_0^\mathrm{T} \Lambda_0 M_0 - M_n^\mathrm{T} \Lambda_n M_n)\right]\right)\right] \; .
\end{split} \tag{12}
$$

with the posterior hyperparameters (→ I/5.1.7)

$$
\begin{split}
M_n &= \Lambda_n^{-1} (X^\mathrm{T} P Y + \Lambda_0 M_0) \\
\Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \; .
\end{split} \tag{13}
$$

Ergo, the joint likelihood is proportional to

$$p(Y,B,T) \propto |T|^{p/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_n)^\mathrm{T}\Lambda_n(B-M_n)\right]\right)\right] \cdot |T|^{(\nu_n-v-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_n T\right)\right] \quad (14)$$

with the posterior hyperparameters (→ I/5.1.7)

$$\begin{split} \Omega_n &= \Omega_0 + Y^\mathrm{T} P Y + M_0^\mathrm{T} \Lambda_0 M_0 - M_n^\mathrm{T} \Lambda_n M_n \\ \nu_n &= \nu_0 + n \; . \end{split} \quad (15)$$

From the term in (14), we can isolate the posterior distribution over B given T:

$$p(B|T,Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \; . \quad (16)$$

From the remaining term, we can isolate the posterior distribution over T:

$$p(T|Y) = \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \; . \quad (17)$$

Together, (16) and (17) constitute the joint (→ I/1.3.2) posterior distribution (→ I/5.1.7) of B and T.

■

Sources:
- Wikipedia (2020): "Bayesian multivariate linear regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-09-03; URL: https://en.wikipedia.org/wiki/Bayesian_multivariate_linear_regression#Posterior_distribution.

2.4.3 Log model evidence

Theorem: Let

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \quad (1)$$

be a general linear model (→ III/2.1.1) with measured $n \times v$ data matrix Y, known $n \times p$ design matrix X, known $n \times n$ covariance structure (→ II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (→ II/5.1.1) Σ. Moreover, assume a normal-Wishart prior distribution (→ III/2.4.1) over the model parameters B and $T = \Sigma^{-1}$:

$$p(B,T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \; . \quad (2)$$

Then, the log model evidence (→ IV/**??**) for this model is

$$\begin{split} \log p(Y|m) = &\frac{v}{2}\log|P| - \frac{nv}{2}\log(2\pi) + \frac{v}{2}\log|\Lambda_0| - \frac{v}{2}\log|\Lambda_n| + \\ &\frac{\nu_0}{2}\log\left|\frac{1}{2}\Omega_0\right| - \frac{\nu_n}{2}\log\left|\frac{1}{2}\Omega_n\right| + \log\Gamma_v\left(\frac{\nu_n}{2}\right) - \log\Gamma_v\left(\frac{\nu_0}{2}\right) \end{split} \quad (3)$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$
\begin{aligned}
M_n &= \Lambda_n^{-1}(X^{\mathrm{T}} P Y + \Lambda_0 M_0) \\
\Lambda_n &= X^{\mathrm{T}} P X + \Lambda_0 \\
\Omega_n &= \Omega_0 + Y^{\mathrm{T}} P Y + M_0^{\mathrm{T}} \Lambda_0 M_0 - M_n^{\mathrm{T}} \Lambda_n M_n \\
\nu_n &= \nu_0 + n \; .
\end{aligned} \tag{4}
$$

Proof: According to the law of marginal probability (→ I/1.3.3), the model evidence (→ I/5.1.11) for this model is:

$$
p(Y|m) = \iint p(Y|B,T) \, p(B,T) \, \mathrm{d}B \, \mathrm{d}T \; . \tag{5}
$$

According to the law of conditional probability (→ I/1.3.4), the integrand is equivalent to the joint likelihood (→ I/5.1.5):

$$
p(Y|m) = \iint p(Y,B,T) \, \mathrm{d}B \, \mathrm{d}T \; . \tag{6}
$$

Equation (1) implies the following likelihood function (→ I/5.1.2)

$$
p(Y|B,\Sigma) = \mathcal{MN}(Y; XB, V, \Sigma) = \sqrt{\frac{1}{(2\pi)^{nv} |\Sigma|^n |V|^v}} \, \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} (Y-XB)^{\mathrm{T}} V^{-1} (Y-XB) \right) \right] \tag{7}
$$

which, for mathematical convenience, can also be parametrized as

$$
p(Y|B,T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \, \exp\left[-\frac{1}{2} \mathrm{tr}\left(T (Y-XB)^{\mathrm{T}} P (Y-XB) \right) \right] \tag{8}
$$

using the $v \times v$ precision matrix (→ I/1.13.19) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (→ I/1.13.19) $P = V^{-1}$.

When deriving the posterior distribution (→ III/2.4.2) $p(B,T|Y)$, the joint likelihood $p(Y,B,T)$ is obtained as

$$
\begin{split}
p(Y,B,T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T \right) \right] \cdot \\
&\exp\left[-\frac{1}{2} \mathrm{tr}\left(T \left[(B-M_n)^{\mathrm{T}} \Lambda_n (B-M_n) + (Y^{\mathrm{T}} P Y + M_0^{\mathrm{T}} \Lambda_0 M_0 - M_n^{\mathrm{T}} \Lambda_n M_n) \right] \right) \right] \; .
\end{split} \tag{9}
$$

Using the probability density function of the matrix-normal distribution (→ II/5.1.3), we can rewrite this as

$$
\begin{split}
p(Y,B,T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{(2\pi)^{pv}}{|T|^p |\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T \right) \right] \cdot \\
&\mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left(T \left[Y^{\mathrm{T}} P Y + M_0^{\mathrm{T}} \Lambda_0 M_0 - M_n^{\mathrm{T}} \Lambda_n M_n \right] \right) \right] \; .
\end{split} \tag{10}
$$

Now, B can be integrated out easily:

$$\begin{split} \int p(Y,B,T) \, \mathrm{d}B = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \cdot \\ &\exp\left[-\frac{1}{2} \mathrm{tr}\left(T\left[\Omega_0 + Y^\mathrm{T} P Y + M_0^\mathrm{T} \Lambda_0 M_0 - M_n^\mathrm{T} \Lambda_n M_n \right] \right) \right] \; . \end{split} \tag{11}$$

Using the probability density function of the Wishart distribution, we can rewrite this as

$$\int p(Y,B,T) \, \mathrm{d}B = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \sqrt{\frac{2^{\nu_n v}}{|\Omega_n|^{\nu_n}}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \; . \tag{12}$$

Finally, T can also be integrated out:

$$\iint p(Y,B,T) \, \mathrm{d}B \, \mathrm{d}T = \sqrt{\frac{|P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{\left|\frac{1}{2}\Omega_0\right|^{\nu_0}}{\left|\frac{1}{2}\Omega_n\right|^{\nu_n}}} \frac{\Gamma_v\left(\frac{\nu_n}{2}\right)}{\Gamma_v\left(\frac{\nu_0}{2}\right)} = p(y|m) \; . \tag{13}$$

Thus, the log model evidence (→ IV/**??**) of this model is given by

$$\begin{split} \log p(Y|m) = &\frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ &\frac{\nu_0}{2} \log \left| \frac{1}{2}\Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2}\Omega_n \right| + \log \Gamma_v\left(\frac{\nu_n}{2}\right) - \log \Gamma_v\left(\frac{\nu_0}{2}\right) \; . \end{split} \tag{14}$$

■

3 Count data

3.1 Binomial observations

3.1.1 Definition

Definition: An ordered pair (n, y) with $n \in \mathbb{N}$ and $y \in \mathbb{N}_0$, where y is the number of successes in n trials, constitutes a set of binomial observations.

3.1.2 Binomial test

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Then, the null hypothesis (→ I/4.3.2)

$$H_0 : \; p = p_0 \tag{2}$$

is rejected (→ I/4.3.1) at significance level (→ I/4.3.8) α, if

$$y \leq c_1 \quad \text{or} \quad y \geq c_2 \tag{3}$$

where c_1 is the largest integer value, such that

$$\sum_{x=0}^{c_1} \mathrm{Bin}(x; n, p_0) \leq \frac{\alpha}{2} \; , \tag{4}$$

and c_2 is the smallest integer value, such that

$$\sum_{x=c_2}^{n} \mathrm{Bin}(x; n, p_0) \leq \frac{\alpha}{2} \; , \tag{5}$$

where $\mathrm{Bin}(x; n, p)$ is the probability mass function of the binomial distribution (→ II/1.3.2):

$$\mathrm{Bin}(x; n, p) = \binom{n}{x} p^x \, (1-p)^{n-x} \; . \tag{6}$$

Proof: The alternative hypothesis (→ I/4.3.3) relative to H_0 for a two-sided test (→ I/4.3.4) is

$$H_1 : \; p \neq p_0 \; . \tag{7}$$

We can use y as a test statistic (→ I/4.3.5). Its sampling distribution (→ I/1.5.5) is given by (1). The cumulative distribution function (→ I/1.8.1) (CDF) of the test statistic under the null hypothesis is thus equal to the cumulative distribution function of a binomial distribution with success probability (→ II/1.3.1) p_0:

$$\Pr(y \leq z | H_0) = \sum_{x=0}^{z} \mathrm{Bin}(x; n, p_0) = \sum_{x=0}^{z} \binom{n}{x} p_0^x \, (1-p_0)^{n-x} \; . \tag{8}$$

The critical value (→ I/4.3.9) is the value of y, such that the probability of observing this or more extreme values of the test statistic is equal to or smaller than α. Since H_0 and H_1 define a two-tailed test, we need two critical values y_1 and y_2 that satisfy

$$\begin{aligned}\alpha &\geq \Pr(y \in \{0, \ldots, y_1\} \cup \{y_2, \ldots, n\} \,|H_0) \\ &= \Pr(y \leq y_1|H_0) + \Pr(y \geq y_2|H_0) \\ &= \Pr(y \leq y_1|H_0) + (1 - \Pr(y \leq (y_2-1)|H_0) \;.\end{aligned} \tag{9}$$

Given the test statistic's CDF in (8), this is fulfilled by the values c_1 and c_2 defined in (4) and (5). Thus, the null hypothesis H_0 can be rejected (→ I/4.3.9), if the observed test statistic is inside the rejection region (→ I/4.3.1):

$$y \in \{0, \ldots, c_1\} \cup \{c_2, \ldots, n\} \;. \tag{10}$$

This is equivalent to (3) and thus completes the proof.

■

Sources:
- Wikipedia (2023): "Binomial test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-12-16; URL: https://en.wikipedia.org/wiki/Binomial_test#Usage.
- Wikipedia (2023): "Binomialtest"; in: *Wikipedia – Die freie Enzyklopädie*, retrieved on 2023-12-16; URL: https://de.wikipedia.org/wiki/Binomialtest#Signifikanzniveau_und_kritische_Werte.

3.1.3 Maximum likelihood estimation

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n,p) \;. \tag{1}$$

Then, the maximum likelihood estimator (→ I/4.1.3) of p is

$$\hat{p} = \frac{y}{n} \;. \tag{2}$$

Proof: With the probability mass function of the binomial distribution (→ II/1.3.2), equation (1) implies the following likelihood function (→ I/5.1.2):

$$\begin{aligned}\mathrm{p}(y|p) &= \mathrm{Bin}(y; n, p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \;.\end{aligned} \tag{3}$$

Thus, the log-likelihood function (→ I/4.1.2) is given by

$$\begin{aligned}\mathrm{LL}(p) &= \log \mathrm{p}(y|p) \\ &= \log \binom{n}{y} + y \log p + (n-y) \log(1-p) \;.\end{aligned} \tag{4}$$

The derivative of the log-likelihood function (4) with respect to p is

$$\frac{\mathrm{dLL}(p)}{\mathrm{d}p} = \frac{y}{p} - \frac{n-y}{1-p} \tag{5}$$

and setting this derivative to zero gives the MLE for p:

$$\begin{aligned}
\frac{\mathrm{dLL}(p)}{\mathrm{d}\hat{p}} &= 0 \\
0 &= \frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} \\
\frac{n-y}{1-\hat{p}} &= \frac{y}{\hat{p}} \\
(n-y)\,\hat{p} &= y\,(1-\hat{p}) \\
n\,\hat{p} - y\,\hat{p} &= y - y\,\hat{p} \\
n\,\hat{p} &= y \\
\hat{p} &= \frac{y}{n} \; .
\end{aligned} \tag{6}$$

■

Sources:
- Wikipedia (2022): “Binomial distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-11-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Statistical_inference.

3.1.4 Maximum log-likelihood

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the maximum log-likelihood (→ I/4.1.4) for this model is

$$\begin{split}
\mathrm{MLL} = \log\Gamma(n+1) - \log\Gamma(y+1) - \log\Gamma(n-y+1) \\
- n\log(n) + y\log(y) + (n-y)\log(n-y) \; .
\end{split} \tag{2}$$

Proof: The log-likelihood function for binomial data (→ III/3.1.3) is given by

$$\mathrm{LL}(p) = \log\binom{n}{y} + y\log p + (n-y)\log(1-p) \tag{3}$$

and the maximum likelihood estimate of the success probability (→ III/3.1.3) p is

$$\hat{p} = \frac{y}{n} \; . \tag{4}$$

Plugging (4) into (3), we obtain the maximum log-likelihood (→ I/4.1.4) of the binomial observation model in (1) as

$$
\begin{aligned}
\mathrm{MLL} &= \mathrm{LL}(\hat{p}) \\
&= \log \binom{n}{y} + y \log \left(\frac{y}{n}\right) + (n-y) \log \left(1 - \frac{y}{n}\right) \\
&= \log \binom{n}{y} + y \log \left(\frac{y}{n}\right) + (n-y) \log \left(\frac{n-y}{n}\right) \\
&= \log \binom{n}{y} + y \log(y) + (n-y) \log(n-y) - n \log(n) \; .
\end{aligned} \tag{5}
$$

With the definition of the binomial coefficient

$$
\binom{n}{k} = \frac{n!}{k! \, (n-k)!} \tag{6}
$$

and the definition of the gamma function

$$
\Gamma(n) = (n-1)! \; , \tag{7}
$$

the MLL finally becomes

$$
\begin{split}
\mathrm{MLL} = \log \Gamma(n+1) - \log \Gamma(y+1) - \log \Gamma(n-y+1) \\
- \, n \log(n) + y \log(y) + (n-y) \log(n-y) \; .
\end{split} \tag{8}
$$

■

3.1.5 Maximum-a-posteriori estimation

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$
y \sim \mathrm{Bin}(n, p) \; . \tag{1}
$$

Moreover, assume a beta prior distribution (→ III/3.1.6) over the model parameter p:

$$
\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}
$$

Then, the maximum-a-posteriori estimate (→ I/5.1.8) of p is

$$
\hat{p}_{\mathrm{MAP}} = \frac{\alpha_0 + y - 1}{\alpha_0 + \beta_0 + n - 2} \; . \tag{3}
$$

Proof: Given the prior distribution (→ I/5.1.3) in (2), the posterior distribution (→ I/5.1.7) for binomial observations (→ III/3.1.1) is also a beta distribution (→ III/3.1.7)

$$
\mathrm{p}(p|y) = \mathrm{Bet}(p; \alpha_n, \beta_n) \tag{4}
$$

where the posterior hyperparameters (→ I/5.1.7) are equal to

$$
\begin{aligned}
\alpha_n &= \alpha_0 + y \\
\beta_n &= \beta_0 + (n-y) \; .
\end{aligned} \tag{5}
$$

The mode of the beta distribution is given by:

$$X \sim \mathrm{Bet}(\alpha, \beta) \quad \Rightarrow \quad \mathrm{mode}(X) = \frac{\alpha - 1}{\alpha + \beta - 2} \; . \tag{6}$$

Applying (6) to (4) with (5), the maximum-a-posteriori estimate (→ I/5.1.8) of p follows as:

$$\begin{aligned} \hat{p}_{\mathrm{MAP}} &= \frac{\alpha_n - 1}{\alpha_n + \beta_n - 2} \\ &\overset{(5)}{=} \frac{\alpha_0 + y - 1}{\alpha_0 + y + \beta_0 + (n - y) - 2} \\ &= \frac{\alpha_0 + y - 1}{\alpha_0 + \beta_0 + n - 2} \; . \end{aligned} \tag{7}$$

■

3.1.6 Conjugate prior distribution

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Then, the conjugate prior (→ I/5.2.5) for the model parameter p is a beta distribution (→ II/3.9.1):

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Proof: With the probability mass function of the binomial distribution (→ II/1.3.2), the likelihood function (→ I/5.1.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{3}$$

In other words, the likelihood function is proportional to a power of p times a power of $(1 - p)$:

$$\mathrm{p}(y|p) \propto p^y \, (1-p)^{n-y} \; . \tag{4}$$

The same is true for a beta distribution over p

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \tag{5}$$

the probability density function of which (→ II/3.9.3)

$$\mathrm{p}(p) = \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \tag{6}$$

exhibits the same proportionality

$$\mathrm{p}(p) \propto p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \tag{7}$$

and is therefore conjugate relative to the likelihood.

■

Sources:
- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters.

3.1.7 Posterior distribution

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Moreover, assume a beta prior distribution (→ III/3.1.6) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a beta distribution (→ II/3.9.1)

$$\mathrm{p}(p|y) = \mathrm{Bet}(p; \alpha_n, \beta_n) \; . \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{split} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n-y) \; . \end{split} \tag{4}$$

Proof: With the probability mass function of the binomial distribution (→ II/1.3.2), the likelihood function (→ I/5.1.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{5}$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{split} \mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \cdot \frac{1}{B(\alpha_0,\beta_0)} \, p^{\alpha_0-1} \, (1-p)^{\beta_0-1} \\ &= \frac{1}{B(\alpha_0,\beta_0)} \binom{n}{y} \, p^{\alpha_0+y-1} \, (1-p)^{\beta_0+(n-y)-1} \; . \end{split} \tag{6}$$

Note that the posterior distribution is proportional to the joint likelihood (→ I/5.1.9):

$$\mathrm{p}(p|y) \propto \mathrm{p}(y,p) \; . \tag{7}$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n-y)$, the posterior distribution is therefore proportional to

$$\mathrm{p}(p|y) \propto p^{\alpha_n-1} \, (1-p)^{\beta_n-1} \tag{8}$$

which, when normalized to one, results in the probability density function of the beta distribution (→ II/3.9.3):

$$\mathrm{p}(p|y) = \frac{1}{B(\alpha_n,\beta_n)} \, p^{\alpha_n-1} \, (1-p)^{\beta_n-1} = \mathrm{Bet}(p; \alpha_n, \beta_n) \; . \tag{9}$$

■

Sources:
- Wikipedia (2020): "Binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-23; URL: https://en.wikipedia.org/wiki/Binomial_distribution#Estimation_of_parameters.

3.1.8 Log model evidence

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Moreover, assume a beta prior distribution (→ III/3.1.6) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Then, the log model evidence (→ IV/**??**) for this model is

$$\begin{split} \log \mathrm{p}(y|m) = \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \\ + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) \; . \end{split} \tag{3}$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{split} \alpha_n &= \alpha_0 + y \\ \beta_n &= \beta_0 + (n-y) \; . \end{split} \tag{4}$$

Proof: With the probability mass function of the binomial distribution (→ II/1.3.2), the likelihood function (→ I/5.1.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y} \, p^y \, (1-p)^{n-y} \; . \tag{5}$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{split} \mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \cdot \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0-1} \, (1-p)^{\beta_0-1} \\ &= \binom{n}{y} \, \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0+y-1} \, (1-p)^{\beta_0+(n-y)-1} \; . \end{split} \tag{6}$$

Note that the model evidence is the marginal density of the joint likelihood (→ I/5.1.11):

$$\mathrm{p}(y) = \int \mathrm{p}(y,p) \, \mathrm{d}p \; . \tag{7}$$

Setting $\alpha_n = \alpha_0 + y$ and $\beta_n = \beta_0 + (n-y)$, the joint likelihood can also be written as

$$
\mathrm{p}(y,p) = \binom{n}{y} \frac{1}{B(\alpha_0,\beta_0)} \frac{B(\alpha_n,\beta_n)}{1} \frac{1}{B(\alpha_n,\beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} \; . \tag{8}
$$

Using the probability density function of the beta distribution (→ II/3.9.3), p can now be integrated out easily

$$
\begin{split}
\mathrm{p}(y) &= \int \binom{n}{y} \frac{1}{B(\alpha_0,\beta_0)} \frac{B(\alpha_n,\beta_n)}{1} \frac{1}{B(\alpha_n,\beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} \, \mathrm{d}p \\
&= \binom{n}{y} \frac{B(\alpha_n,\beta_n)}{B(\alpha_0,\beta_0)} \int \mathrm{Bet}(p; \alpha_n, \beta_n) \, \mathrm{d}p \\
&= \binom{n}{y} \frac{B(\alpha_n,\beta_n)}{B(\alpha_0,\beta_0)} \; ,
\end{split} \tag{9}
$$

such that the log model evidence (→ IV/??) (LME) is shown to be

$$
\log \mathrm{p}(y|m) = \log \binom{n}{y} + \log B(\alpha_n,\beta_n) - \log B(\alpha_0,\beta_0) \; . \tag{10}
$$

With the definition of the binomial coefficient

$$
\binom{n}{k} = \frac{n!}{k! \, (n-k)!} \tag{11}
$$

and the definition of the gamma function

$$
\Gamma(n) = (n-1)! \; , \tag{12}
$$

the LME finally becomes

$$
\begin{split}
\log \mathrm{p}(y|m) = \log \Gamma(n+1) - \log \Gamma(y+1) - \log \Gamma(n-y+1) \\
+ \log B(\alpha_n,\beta_n) - \log B(\alpha_0,\beta_0) \; .
\end{split} \tag{13}
$$

■

Sources:

- Wikipedia (2020): "Beta-binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-24; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#Motivation_and_derivation.

3.1.9 Log Bayes factor

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$
y \sim \mathrm{Bin}(n,p) \; . \tag{1}
$$

Moreover, assume two statistical models (→ I/5.1.4), one assuming that p is 0.5 (null model (→ I/4.3.2)), the other imposing a beta distribution (→ III/3.1.6) as the prior distribution (→ I/5.1.3) on the model parameter p (alternative (→ I/4.3.3)):

$$
\begin{aligned}
m_0 &: \; y \sim \mathrm{Bin}(n,p), \; p = 0.5 \\
m_1 &: \; y \sim \mathrm{Bin}(n,p), \; p \sim \mathrm{Bet}(\alpha_0, \beta_0) \; .
\end{aligned} \tag{2}
$$

Then, the log Bayes factor (→ IV/**??**) in favor of m_1 against m_0 is

$$
\mathrm{LBF}_{10} = \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) - n \log \left(\frac{1}{2} \right) \tag{3}
$$

where $B(x,y)$ is the beta function and α_n and β_n are the posterior hyperparameters for binomial observations (→ III/3.1.7) which are functions of the number of trials (→ II/1.3.1) n and the number of successes (→ II/1.3.1) y.

Proof: The log Bayes factor is equal to the difference of two log model evidences (→ IV/**??**):

$$
\mathrm{LBF}_{12} = \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \tag{4}
$$

The LME of the alternative m_1 is equal to the log model evidence for binomial observations (→ III/3.1.8):

$$
\mathrm{LME}(m_1) = \log p(y|m_1) = \log \binom{n}{y} + \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) \; . \tag{5}
$$

Because the null model m_0 has no free parameter, its log model evidence (→ IV/**??**) (logarithmized marginal likelihood (→ I/5.1.11)) is equal to the log-likelihood function for binomial observations (→ III/3.1.3) at the value $p = 0.5$:

$$
\begin{aligned}
\mathrm{LME}(m_0) = \log p(y|p=0.5) &= \log \binom{n}{y} + y \log(0.5) + (n-y) \log(1-0.5) \\
&= \log \binom{n}{y} + n \log \left(\frac{1}{2} \right) \; .
\end{aligned} \tag{6}
$$

Subtracting the two LMEs from each other, the LBF emerges as

$$
\mathrm{LBF}_{10} = \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) - n \log \left(\frac{1}{2} \right) \tag{7}
$$

where the posterior hyperparameters (→ I/5.1.7) are given by (→ III/3.1.7)

$$
\begin{aligned}
\alpha_n &= \alpha_0 + y \\
\beta_n &= \beta_0 + (n-y)
\end{aligned} \tag{8}
$$

with the number of trials (→ II/1.3.1) n and the number of successes (→ II/1.3.1) y.

■

3.1.10 Posterior probability

Theorem: Let y be the number of successes resulting from n independent trials with unknown success probability p, such that y follows a binomial distribution (→ II/1.3.1):

$$y \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Moreover, assume two statistical models (→ I/5.1.4), one assuming that p is 0.5 (null model (→ I/4.3.2)), the other imposing a beta distribution (→ III/3.1.6) as the prior distribution (→ I/5.1.3) on the model parameter p (alternative (→ I/4.3.3)):

$$\begin{split} m_0 &: \; y \sim \mathrm{Bin}(n,p), \; p = 0.5 \\ m_1 &: \; y \sim \mathrm{Bin}(n,p), \; p \sim \mathrm{Bet}(\alpha_0, \beta_0) \; . \end{split} \tag{2}$$

Then, the posterior probability (→ IV/**??**) of the alternative model (→ I/4.3.3) is given by

$$p(m_1|y) = \frac{1}{1 + 2^{-n} \left[B(\alpha_0, \beta_0) / B(\alpha_n, \beta_n) \right]} \tag{3}$$

where $B(x,y)$ is the beta function and α_n and β_n are the posterior hyperparameters for binomial observations (→ III/3.1.7) which are functions of the number of trials (→ II/1.3.1) n and the number of successes (→ II/1.3.1) y.

Proof: The posterior probability for one of two models is a function of the log Bayes factor in favor of this model (→ IV/**??**):

$$p(m_1|y) = \frac{\exp(\mathrm{LBF}_{12})}{\exp(\mathrm{LBF}_{12}) + 1} \; . \tag{4}$$

The log Bayes factor in favor of the alternative model for binomial observations (→ III/3.1.9) is given by

$$\mathrm{LBF}_{10} = \log B(\alpha_n, \beta_n) - \log B(\alpha_0, \beta_0) - n \log \left(\frac{1}{2} \right) \; . \tag{5}$$

and the corresponding Bayes factor (→ IV/**??**), i.e. exponentiated log Bayes factor (→ IV/**??**), is equal to

$$\mathrm{BF}_{10} = \exp(\mathrm{LBF}_{10}) = 2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \; . \tag{6}$$

Thus, the posterior probability of the alternative, assuming a prior distribution over the probability p, compared to the null model, assuming a fixed probability $p = 0.5$, follows as

$$\begin{split} p(m_1|y) &\overset{(4)}{=} \frac{\exp(\mathrm{LBF}_{10})}{\exp(\mathrm{LBF}_{10}) + 1} \\ &\overset{(6)}{=} \frac{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)}}{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} + 1} \\ &= \frac{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)}}{2^n \cdot \frac{B(\alpha_n, \beta_n)}{B(\alpha_0, \beta_0)} \left(1 + 2^{-n} \frac{B(\alpha_0, \beta_0)}{B(\alpha_n, \beta_n)} \right)} \\ &= \frac{1}{1 + 2^{-n} \left[B(\alpha_0, \beta_0) / B(\alpha_n, \beta_n) \right]} \end{split} \tag{7}$$

where the posterior hyperparameters (→ I/5.1.7) are given by (→ III/3.1.7)

$$
\begin{aligned}
\alpha_n &= \alpha_0 + y \\
\beta_n &= \beta_0 + (n - y)
\end{aligned} \tag{8}
$$

with the number of trials (→ II/1.3.1) n and the number of successes (→ II/1.3.1) y.

■

3.2 Multinomial observations

3.2.1 Definition

Definition: An ordered pair (n, y) with $n \in \mathbb{N}$ and $y = [y_1, \ldots, y_k] \in \mathbb{N}_0^{1 \times k}$, where y_i is the number of observations for the i-th out of k categories obtained in n trials, $i = 1, \ldots, k$, constitutes a set of multinomial observations.

3.2.2 Multinomial test

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$
y \sim \mathrm{Mult}(n, p) \; . \tag{1}
$$

Then, the null hypothesis (→ I/4.3.2)

$$
H_0: \; p = p_0 = [p_{01}, \ldots, p_{0k}] \tag{2}
$$

is rejected (→ I/4.3.1) at significance level (→ I/4.3.8) α, if

$$
\mathrm{Pr}_\mathrm{sig} = \sum_{x: \, \mathrm{Pr}_0(x) \leq \mathrm{Pr}_0(y)} \mathrm{Pr}_0(x) < \alpha \tag{3}
$$

where $\mathrm{Pr}_0(x)$ is the probability of observing the numbers of occurences $x = [x_1, \ldots, x_k]$ under the null hypothesis:

$$
\mathrm{Pr}_0(x) = n! \prod_{j=1}^{k} \frac{p_{0j}^{x_j}}{x_j!} \; . \tag{4}
$$

Proof: The alternative hypothesis (→ I/4.3.3) relative to H_0 is

$$
H_1: \; p_j \neq p_{0j} \quad \text{for at least one} \quad j = 1, \ldots, k \; . \tag{5}
$$

We can use y as a test statistic (→ I/4.3.5). Its sampling distribution (→ I/1.5.5) is given by (1). The probability mass function (→ I/1.6.1) (PMF) of the test statistic under the null hypothesis is thus equal to the probability mass function of the multionomial distribution (→ II/2.2.2) with category probabilities (→ II/2.2.1) p_0:

$$
\Pr(y = x | H_0) = \mathrm{Mult}(x; n, p_0) = \binom{n}{x_1, \ldots, x_k} \prod_{j=1}^{k} {p_j}^{x_j} \; . \tag{6}
$$

The multinomial coefficient in this equation is equal to

$$\binom{n}{k_1, \ldots, k_m} = \frac{n!}{k_1! \cdot \ldots \cdot k_m!} \; , \tag{7}$$

such that the probability of observing the counts y, given H_0, is

$$\mathrm{Pr}(y|H_0) = n! \prod_{j=1}^{k} \frac{p_{0i}^{y_j}}{y_j!} \; . \tag{8}$$

The probability of observing any other set of counts x, given H_0, is

$$\mathrm{Pr}(x|H_0) = n! \prod_{j=1}^{k} \frac{p_{0i}^{x_j}}{x_j!} \; . \tag{9}$$

The p-value (→ I/4.3.10) is the probability of observing a value of the test statistic (→ I/4.3.5) that is as extreme or more extreme then the actually observed test statistic. Any set of counts x might be considered as extreme or more extreme than the actually observed counts y, if the former is equally probable or less probably than the latter:

$$\mathrm{Pr}_0(x) \leq \mathrm{Pr}_0(y) \; . \tag{10}$$

Thus, the p-value (→ I/4.3.10) for the data in (1) is equal to

$$p = \sum_{x\colon \mathrm{Pr}_0(x) \leq \mathrm{Pr}_0(y)} \mathrm{Pr}_0(x) \tag{11}$$

and the null hypothesis in (2) is rejected (→ I/4.3.1), if

$$p < \alpha \; . \tag{12}$$

■

Sources:
- Wikipedia (2023): "Multinomial test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2023-12-23; URL: https://en.wikipedia.org/wiki/Multinomial_test.

3.2.3 Maximum likelihood estimation

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Then, the maximum likelihood estimator (→ I/4.1.3) of p is

$$\hat{p} = \frac{1}{n} y, \quad \text{i.e.} \quad \hat{p}_j = \frac{y_j}{n} \quad \text{for all} \quad j = 1, \ldots, k \; . \tag{2}$$

Proof: Note that the marginal distribution of each element in a multinomial random vector is a binomial distribution

$$X \sim \mathrm{Mult}(n, p) \quad \Rightarrow \quad X_j \sim \mathrm{Bin}(n, p_j) \quad \text{for all} \quad j = 1, \ldots, k \; . \tag{3}$$

Thus, combining (1) with (3), we have

$$y_j \sim \mathrm{Bin}(n, p_j) \tag{4}$$

which implies the likelihood function (→ III/3.1.3)

$$\mathrm{p}(y|p_j) = \mathrm{Bin}(y_j; n, p_j) = \binom{n}{y_j} p_j^{y_j} \, (1-p_j)^{n-y_j} \; . \tag{5}$$

To this, we can apply maximum likelihood estimation for binomial observations (→ III/3.1.3), such that the MLE for each p_j is

$$\hat{p}_j = \frac{y_j}{n} \; . \tag{6}$$

■

3.2.4 Maximum log-likelihood

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Then, the maximum log-likelihood (→ I/4.1.4) for this model is

$$\mathrm{MLL} = \log \Gamma(n+1) - \sum_{j=1}^{k} \log \Gamma(y_j + 1) - n \log(n) + \sum_{j=1}^{k} y_j \log(y_j) \; . \tag{2}$$

Proof: With the probability mass function of the multinomial distribution (→ II/2.2.2), equation (1) implies the following likelihood function (→ I/5.1.2):

$$\begin{split} \mathrm{p}(y|p) &= \mathrm{Mult}(y; n, p) \\ &= \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \; . \end{split} \tag{3}$$

Thus, the log-likelihood function (→ I/4.1.2) is given by

$$\begin{split} \mathrm{LL}(p) &= \log \mathrm{p}(y|p) \\ &= \log \binom{n}{y_1, \ldots, y_k} + \sum_{j=1}^{k} y_j \log(p_j) \; . \end{split} \tag{4}$$

The maximum likelihood estimates of the category probabilities (→ III/3.2.3) p are

$$\hat{p} = [\hat{p}_1, \ldots, \hat{p}_k] \quad \text{with} \quad \hat{p}_j = \frac{y_j}{n} \quad \text{for all} \quad j = 1, \ldots, k \; . \tag{5}$$

Plugging (5) into (4), we obtain the maximum log-likelihood (→ I/4.1.4) of the multinomial observation model in (1) as

$$
\begin{aligned}
\mathrm{MLL} &= \mathrm{LL}(\hat{p}) \\
&= \log \binom{n}{y_1, \ldots, y_k} + \sum_{j=1}^{k} y_j \log\left(\frac{y_j}{n}\right) \\
&= \log \binom{n}{y_1, \ldots, y_k} + \sum_{j=1}^{k} \left[y_j \log(y_j) - y_j \log(n) \right] \\
&= \log \binom{n}{y_1, \ldots, y_k} + \sum_{j=1}^{k} y_j \log(y_j) - \sum_{j=1}^{k} y_j \log(n) \\
&= \log \binom{n}{y_1, \ldots, y_k} + \sum_{j=1}^{k} y_j \log(y_j) - n \log(n) \; .
\end{aligned} \tag{6}
$$

With the definition of the multinomial coefficient

$$
\binom{n}{k_1, \ldots, k_m} = \frac{n!}{k_1! \cdot \ldots \cdot k_m!} \tag{7}
$$

and the definition of the gamma function

$$
\Gamma(n) = (n-1)! \; , \tag{8}
$$

the MLL finally becomes

$$
\mathrm{MLL} = \log \Gamma(n+1) - \sum_{j=1}^{k} \log \Gamma(y_j + 1) - n \log(n) + \sum_{j=1}^{k} y_j \log(y_j) \; . \tag{9}
$$

■

3.2.5 Maximum-a-posteriori estimation

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$
y \sim \mathrm{Mult}(n, p) \; . \tag{1}
$$

Moreover, assume a Dirichlet prior distribution (→ III/3.2.6) over the model parameter p:

$$
\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}
$$

Then, the maximum-a-posteriori estimates (→ I/5.1.8) of p are

$$
\hat{p}_{\mathrm{MAP}} = \frac{\alpha_0 + y - 1}{\sum_{j=1}^{k} \alpha_{0j} + n - k} \; . \tag{3}
$$

Proof: Given the prior distribution (→ I/5.1.3) in (2), the posterior distribution (→ I/5.1.7) for multinomial observations (→ III/3.2.1) is also a Dirichlet distribution (→ III/3.2.7)

$$\mathrm{p}(p|y) = \mathrm{Dir}(p; \alpha_n) \tag{4}$$

where the posterior hyperparameters (→ I/5.1.7) are equal to

$$\alpha_{nj} = \alpha_{0j} + y_j, \; j = 1, \ldots, k \; . \tag{5}$$

The mode of the Dirichlet distribution is given by:

$$X \sim \mathrm{Dir}(\alpha) \quad \Rightarrow \quad \mathrm{mode}(X_i) = \frac{\alpha_i - 1}{\sum_j \alpha_j - k} \; . \tag{6}$$

Applying (6) to (4) with (5), the maximum-a-posteriori estimates (→ I/5.1.8) of p follow as

$$\begin{aligned}
\hat{p}_{i,\mathrm{MAP}} &= \frac{\alpha_{ni} - 1}{\sum_j \alpha_{nj} - k} \\
&\overset{(5)}{=} \frac{\alpha_{0i} + y_i - 1}{\sum_j (\alpha_{0j} + y_j) - k} \\
&= \frac{\alpha_{0i} + y_i - 1}{\sum_j \alpha_{0j} + \sum_j y_j - k} \; .
\end{aligned} \tag{7}$$

Since $y_1 + \ldots + y_k = n$ by definition (→ III/3.2.1), this becomes

$$\hat{p}_{i,\mathrm{MAP}} = \frac{\alpha_{0i} + y_i - 1}{\sum_j \alpha_{0j} + n - k} \tag{8}$$

which, using the $1 \times k$ vectors (→ III/3.2.1) y, p and α_0, can be written as:

$$\hat{p}_{\mathrm{MAP}} = \frac{\alpha_0 + y - 1}{\sum_{j=1}^{k} \alpha_{0j} + n - k} \; . \tag{9}$$

■

3.2.6 Conjugate prior distribution

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Then, the conjugate prior (→ I/5.2.5) for the model parameter p is a Dirichlet distribution (→ II/4.4.1):

$$\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}$$

Proof: With the probability mass function of the multinomial distribution (→ II/2.2.2), the likelihood function (→ I/5.1.2) implied by (1) is given by

$$\mathrm{p}(y|p) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \; . \tag{3}$$

In other words, the likelihood function is proportional to a product of powers of the entries of the vector p:

$$\mathrm{p}(y|p) \propto \prod_{j=1}^{k} {p_j}^{y_j} \; . \tag{4}$$

The same is true for a Dirichlet distribution over p

$$\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \tag{5}$$

the probability density function of which (→ II/4.4.2)

$$\mathrm{p}(p) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \tag{6}$$

exhibits the same proportionality

$$\mathrm{p}(p) \propto \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \tag{7}$$

and is therefore conjugate relative to the likelihood.

■

Sources:

- Wikipedia (2020): "Dirichlet distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-11; URL: https://en.wikipedia.org/wiki/Dirichlet_distribution#Conjugate_to_categorical/multinomi

3.2.7 Posterior distribution

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Moreover, assume a Dirichlet prior distribution (→ III/3.2.6) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a Dirichlet distribution (→ II/4.4.1)

$$\mathrm{p}(p|y) = \mathrm{Dir}(p; \alpha_n) \; . \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\alpha_{nj} = \alpha_{0j} + y_j, \; j = 1, \ldots, k \; . \tag{4}$$

Proof: With the probability mass function of the multinomial distribution (→ II/2.2.2), the likelihood function (→ I/5.1.2) implied by (1) is given by

$$
\mathrm{p}(y|p) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \; . \tag{5}
$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$
\begin{split}
\mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\
&= \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \\
&= \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}+y_j-1} \; .
\end{split} \tag{6}
$$

Note that the posterior distribution is proportional to the joint likelihood (→ I/5.1.9):

$$
\mathrm{p}(p|y) \propto \mathrm{p}(y,p) \; . \tag{7}
$$

Setting $\alpha_{nj} = \alpha_{0j} + y_j$, the posterior distribution is therefore proportional to

$$
\mathrm{p}(p|y) \propto \prod_{j=1}^{k} {p_j}^{\alpha_{nj}-1} \tag{8}
$$

which, when normalized to one, results in the probability density function of the Dirichlet distribution (→ II/4.4.2):

$$
\mathrm{p}(p|y) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{nj})} \prod_{j=1}^{k} {p_j}^{\alpha_{nj}-1} = \mathrm{Dir}(p; \alpha_n) \; . \tag{9}
$$

■

Sources:
- Wikipedia (2020): “Dirichlet distribution”; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-11; URL: https://en.wikipedia.org/wiki/Dirichlet_distribution#Conjugate_to_categorical/multinomi

3.2.8 Log model evidence

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$
y \sim \mathrm{Mult}(n,p) \; . \tag{1}
$$

Moreover, assume a Dirichlet prior distribution (→ III/3.2.6) over the model parameter p:

$$
\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}
$$

Then, the log model evidence (→ IV/**??**) for this model is

$$
\begin{aligned}
\log \mathrm{p}(y|m) = \log \Gamma(n+1) &- \sum_{j=1}^{k} \log \Gamma(y_j+1) \\
&+ \log \Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right) \\
&+ \sum_{j=1}^{k} \log \Gamma(\alpha_{nj}) - \sum_{j=1}^{k} \log \Gamma(\alpha_{0j}) \; .
\end{aligned} \tag{3}
$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$
\alpha_{nj} = \alpha_{0j} + y_j, \; j = 1, \ldots, k \; . \tag{4}
$$

Proof: With the probability mass function of the multinomial distribution (→ II/2.2.2), the likelihood function (→ I/5.1.2) implied by (1) is given by

$$
\mathrm{p}(y|p) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \; . \tag{5}
$$

Combining the likelihood function (5) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$
\begin{aligned}
\mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\
&= \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \\
&= \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}+y_j-1} \; .
\end{aligned} \tag{6}
$$

Note that the model evidence is the marginal density of the joint likelihood:

$$
\mathrm{p}(y) = \int \mathrm{p}(y,p) \, \mathrm{d}p \; . \tag{7}
$$

Setting $\alpha_{nj} = \alpha_{0j} + y_j$, the joint likelihood can also be written as

$$
\mathrm{p}(y,p) = \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{nj})} \prod_{j=1}^{k} {p_j}^{\alpha_{nj}-1} \; . \tag{8}
$$

Using the probability density function of the Dirichlet distribution (→ II/4.4.2), p can now be integrated out easily

$$
\begin{aligned}
\mathrm{p}(y) &= \int \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)}{\prod_{j=1}^k \Gamma(\alpha_{nj})} \prod_{j=1}^k {p_j}^{\alpha_{nj}-1} \, \mathrm{d}p \\
&= \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \int \mathrm{Dir}(p; \alpha_n) \, \mathrm{d}p \\
&= \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^k \alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^k \alpha_{nj}\right)} \frac{\prod_{j=1}^k \Gamma(\alpha_{nj})}{\prod_{j=1}^k \Gamma(\alpha_{0j})} \; ,
\end{aligned} \tag{9}
$$

such that the log model evidence (→ IV/**??**) (LME) is shown to be

$$
\begin{split}
\log \mathrm{p}(y|m) = \log \binom{n}{y_1, \ldots, y_k} + \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\
+ \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) \; .
\end{split} \tag{10}
$$

With the definition of the multinomial coefficient

$$
\binom{n}{k_1, \ldots, k_m} = \frac{n!}{k_1! \cdot \ldots \cdot k_m!} \tag{11}
$$

and the definition of the gamma function

$$
\Gamma(n) = (n-1)! \; , \tag{12}
$$

the LME finally becomes

$$
\begin{split}
\log \mathrm{p}(y|m) = \log \Gamma(n+1) - \sum_{j=1}^k \log \Gamma(y_j+1) \\
+ \log \Gamma\left(\sum_{j=1}^k \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^k \alpha_{nj}\right) \\
+ \sum_{j=1}^k \log \Gamma(\alpha_{nj}) - \sum_{j=1}^k \log \Gamma(\alpha_{0j}) \; .
\end{split} \tag{13}
$$

■

3.2.9 Log Bayes factor

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$y \sim \mathrm{Mult}(n,p) \; . \tag{1}$$

Moreover, assume two statistical models (→ I/5.1.4), one assuming that each p_j is $1/k$ (null model (→ I/4.3.2)), the other imposing a Dirichlet distribution (→ III/3.2.6) as the prior distribution (→ I/5.1.3) on the model parameters $p_1, \ldots, p_k$ (alternative (→ I/4.3.3)):

$$\begin{split} m_0 &: \; y \sim \mathrm{Mult}(n,p), \; p = [1/k, \ldots, 1/k] \\ m_1 &: \; y \sim \mathrm{Mult}(n,p), \; p \sim \mathrm{Dir}(\alpha_0) \; . \end{split} \tag{2}$$

Then, the log Bayes factor (→ IV/??) in favor of m_1 against m_0 is

$$\begin{split} \mathrm{LBF}_{10} = \log \Gamma\left(\sum_{j=1}^{k} \alpha_{0j} \right) - \log \Gamma\left(\sum_{j=1}^{k} \alpha_{nj} \right) \\ + \sum_{j=1}^{k} \log \Gamma(\alpha_{nj}) - \sum_{j=1}^{k} \log \Gamma(\alpha_{0j}) - n \log \left(\frac{1}{k} \right) \end{split} \tag{3}$$

where $\Gamma(x)$ is the gamma function and α_n are the posterior hyperparameters for multinomial observations (→ III/3.2.7) which are functions of the numbers of observations (→ II/2.2.1) $y_1, \ldots, y_k$.

Proof: The log Bayes factor is equal to the difference of two log model evidences (→ IV/??):

$$\mathrm{LBF}_{12} = \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \tag{4}$$

The LME of the alternative m_1 is equal to the log model evidence for multinomial observations (→ III/3.2.8):

$$\begin{split} \mathrm{LME}(m_1) = \log p(y|m_1) &= \log \Gamma(n+1) - \sum_{j=1}^{k} \log \Gamma(y_j+1) \\ &+ \log \Gamma\left(\sum_{j=1}^{k} \alpha_{0j} \right) - \log \Gamma\left(\sum_{j=1}^{k} \alpha_{nj} \right) \\ &+ \sum_{j=1}^{k} \log \Gamma(\alpha_{nj}) - \sum_{j=1}^{k} \log \Gamma(\alpha_{0j}) \; . \end{split} \tag{5}$$

Because the null model m_0 has no free parameter, its log model evidence (→ IV/??) (logarithmized marginal likelihood (→ I/5.1.11)) is equal to the log-likelihood function for multinomial observations (→ III/3.2.3) at the value $p_0 = [1/k, \ldots, 1/k]$:

$$\begin{split} \mathrm{LME}(m_0) = \log p(y|p=p_0) &= \log \binom{n}{y_1, \ldots, y_k} + \sum_{j=1}^{k} y_j \log \left(\frac{1}{k} \right) \\ &= \log \binom{n}{y_1, \ldots, y_k} + n \log \left(\frac{1}{k} \right) \; . \end{split} \tag{6}$$

Subtracting the two LMEs from each other, the LBF emerges as

$$
\begin{aligned}
\mathrm{LBF}_{10} = \log\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right) - \log\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right) \\
+ \sum_{j=1}^{k}\log\Gamma(\alpha_{nj}) - \sum_{j=1}^{k}\log\Gamma(\alpha_{0j}) - n\log\left(\frac{1}{k}\right)
\end{aligned} \tag{7}
$$

where the posterior hyperparameters (→ I/5.1.7) are given by (→ III/3.2.7)

$$
\begin{aligned}
\alpha_n &= \alpha_0 + y \\
&= [\alpha_{01}, \ldots, \alpha_{0k}] + [y_1, \ldots, y_k] \\
&= [\alpha_{01} + y_1, \ldots, \alpha_{0k} + y_k] \\
\text{i.e.} \quad \alpha_{nj} &= \alpha_{0j} + y_j \quad \text{for all} \quad j = 1, \ldots, k
\end{aligned} \tag{8}
$$

with the numbers of observations (→ II/2.2.1) $y_1, \ldots, y_k$.

■

3.2.10 Posterior probability

Theorem: Let $y = [y_1, \ldots, y_k]$ be the number of observations in k categories resulting from n independent trials with unknown category probabilities $p = [p_1, \ldots, p_k]$, such that y follows a multinomial distribution (→ II/2.2.1):

$$
y \sim \mathrm{Mult}(n, p) \; . \tag{1}
$$

Moreover, assume two statistical models (→ I/5.1.4), one assuming that each p_j is $1/k$ (null model (→ I/4.3.2)), the other imposing a Dirichlet distribution (→ III/3.2.6) as the prior distribution (→ I/5.1.3) on the model parameters $p_1, \ldots, p_k$ (alternative (→ I/4.3.3)):

$$
\begin{aligned}
m_0 &: \; y \sim \mathrm{Mult}(n, p), \; p = [1/k, \ldots, 1/k] \\
m_1 &: \; y \sim \mathrm{Mult}(n, p), \; p \sim \mathrm{Dir}(\alpha_0) \; .
\end{aligned} \tag{2}
$$

Then, the posterior probability (→ IV/??) of the alternative model (→ I/4.3.3) is given by

$$
p(m_1|y) = \frac{1}{1 + k^{-n} \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{0j})}{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}} \tag{3}
$$

where $\Gamma(x)$ is the gamma function and α_n are the posterior hyperparameters for multinomial observations (→ III/3.2.7) which are functions of the numbers of observations (→ II/2.2.1) $y_1, \ldots, y_k$.

Proof: The posterior probability for one of two models is a function of the log Bayes factor in favor of this model (→ IV/??):

$$
p(m_1|y) = \frac{\exp(\mathrm{LBF}_{12})}{\exp(\mathrm{LBF}_{12}) + 1} \; . \tag{4}
$$

The log Bayes factor in favor of the alternative model for multinomial observations (→ III/3.2.9) is given by

$$
\begin{split}
\mathrm{LBF}_{10} = \log\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right) - \log\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right) \\
+ \sum_{j=1}^{k}\log\Gamma(\alpha_{nj}) - \sum_{j=1}^{k}\log\Gamma(\alpha_{0j}) - n\log\left(\frac{1}{k}\right)
\end{split}
\tag{5}
$$

and the corresponding Bayes factor (→ IV/**??**), i.e. exponentiated log Bayes factor (→ IV/**??**), is equal to

$$
\mathrm{BF}_{10} = \exp(\mathrm{LBF}_{10}) = k^n \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}{\prod_{j=1}^{k}\Gamma(\alpha_{0j})} \; . \tag{6}
$$

Thus, the posterior probability of the alternative, assuming a prior distribution over the probabilities $p_1, \ldots, p_k$, compared to the null model, assuming fixed probabilities $p = [1/k, \ldots, 1/k]$, follows as

$$
\begin{split}
p(m_1|y) &\overset{(4)}{=} \frac{\exp(\mathrm{LBF}_{10})}{\exp(\mathrm{LBF}_{10})+1} \\
&\overset{(6)}{=} \frac{k^n \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}{\prod_{j=1}^{k}\Gamma(\alpha_{0j})}}{k^n \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}{\prod_{j=1}^{k}\Gamma(\alpha_{0j})} + 1} \\
&= \frac{k^n \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}{\prod_{j=1}^{k}\Gamma(\alpha_{0j})}}{k^n \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}{\prod_{j=1}^{k}\Gamma(\alpha_{0j})}\left(1 + k^{-n} \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{0j})}{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}\right)} \\
&= \frac{1}{1 + k^{-n} \cdot \frac{\Gamma\left(\sum_{j=1}^{k}\alpha_{nj}\right)}{\Gamma\left(\sum_{j=1}^{k}\alpha_{0j}\right)} \cdot \frac{\prod_{j=1}^{k}\Gamma(\alpha_{0j})}{\prod_{j=1}^{k}\Gamma(\alpha_{nj})}}
\end{split}
\tag{7}
$$

where the posterior hyperparameters (→ I/5.1.7) are given by (→ III/3.2.7)

$$
\alpha_n = \alpha_0 + y, \quad \text{i.e.} \quad \alpha_{nj} = \alpha_{0j} + y_j \tag{8}
$$

with the numbers of observations (→ II/2.2.1) $y_1, \ldots, y_k$.

■

3.3 Poisson-distributed data

3.3.1 Definition

Definition: Poisson-distributed data are defined as a set of observed counts $y = \{y_1, \ldots, y_n\}$, independent and identically distributed according to a Poisson distribution (→ II/1.5.1) with rate λ:

$$y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{1}$$

Sources:
- Wikipedia (2020): "Poisson distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Poisson_distribution#Parameter_estimation.

3.3.2 Maximum likelihood estimation

Theorem: Let there be a Poisson-distributed data (→ III/3.3.1) set $y = \{y_1, \ldots, y_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the maximum likelihood estimate (→ I/4.1.3) for the rate parameter λ is given by

$$\hat{\lambda} = \bar{y} \tag{2}$$

where $\bar{y}$ is the sample mean (→ I/1.10.2)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{3}$$

Proof: The likelihood function (→ I/5.1.2) for each observation is given by the probability mass function of the Poisson distribution (→ II/1.5.2)

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \tag{4}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \; . \tag{5}$$

Thus, the log-likelihood function (→ I/4.1.2) is

$$\mathrm{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \tag{6}$$

which can be developed into

$$
\begin{split}
\mathrm{LL}(\lambda) &= \sum_{i=1}^n \log \left[\frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \\
&= \sum_{i=1}^n \left[y_i \cdot \log(\lambda) - \lambda - \log(y_i!) \right] \\
&= -\sum_{i=1}^n \lambda + \sum_{i=1}^n y_i \cdot \log(\lambda) - \sum_{i=1}^n \log(y_i!) \\
&= -n\lambda + \log(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^n \log(y_i!)
\end{split} \tag{7}
$$

The derivatives of the log-likelihood with respect to λ are

$$
\begin{split}
\frac{\mathrm{dLL}(\lambda)}{\mathrm{d}\lambda} &= \frac{1}{\lambda} \sum_{i=1}^n y_i - n \\
\frac{\mathrm{d}^2\mathrm{LL}(\lambda)}{\mathrm{d}\lambda^2} &= -\frac{1}{\lambda^2} \sum_{i=1}^n y_i \; .
\end{split} \tag{8}
$$

Setting the first derivative to zero, we obtain:

$$
\begin{split}
\frac{\mathrm{dLL}(\hat{\lambda})}{\mathrm{d}\lambda} &= 0 \\
0 &= \frac{1}{\hat{\lambda}} \sum_{i=1}^n y_i - n \\
\hat{\lambda} &= \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} \; .
\end{split} \tag{9}
$$

Plugging this value into the second derivative, we confirm:

$$
\begin{split}
\frac{\mathrm{d}^2\mathrm{LL}(\hat{\lambda})}{\mathrm{d}\lambda^2} &= -\frac{1}{\bar{y}^2} \sum_{i=1}^n y_i \\
&= -\frac{n \cdot \bar{y}}{\bar{y}^2} \\
&= -\frac{n}{\bar{y}} < 0 \; .
\end{split} \tag{10}
$$

This demonstrates that the estimate $\hat{\lambda} = \bar{y}$ maximizes the likelihood $p(y|\lambda)$.

■

3.3.3 Conjugate prior distribution

Theorem: Let there be a Poisson-distributed data (→ III/3.3.1) set $y = \{y_1, \ldots, y_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the conjugate prior (→ I/5.2.5) for the model parameter λ is a gamma distribution (→ II/3.4.1):

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \tag{3}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \; . \tag{4}$$

Resolving the product in the likelihood function, we have

$$\begin{aligned} p(y|\lambda) &= \prod_{i=1}^{n} \frac{1}{y_i!} \cdot \prod_{i=1}^{n} \lambda^{y_i} \cdot \prod_{i=1}^{n} \exp\left[-\lambda\right] \\ &= \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \cdot \lambda^{n\bar{y}} \cdot \exp\left[-n\lambda\right] \end{aligned} \tag{5}$$

where $\bar{y}$ is the mean (→ I/1.10.2) of y:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{6}$$

In other words, the likelihood function is proportional to a power of λ times an exponential of λ:

$$p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp\left[-n\lambda\right] \; . \tag{7}$$

The same is true for a gamma distribution over λ

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \tag{8}$$

the probability density function of which (→ II/3.4.7)

$$p(\lambda) = \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \tag{9}$$

exhibits the same proportionality

$$p(\lambda) \propto \lambda^{a_0 - 1} \cdot \exp[-b_0 \lambda] \tag{10}$$

and is therefore conjugate relative to the likelihood.

■

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14ff.; URL: http://www.stat.columbia.edu/~gelman/book/.

3.3.4 Posterior distribution

Theorem: Let there be a Poisson-distributed data (→ III/3.3.1) set $y = \{y_1, \ldots, y_n\}$:

$$y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{1}$$

Moreover, assume a gamma prior distribution (→ III/3.3.3) over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a gamma distribution (→ II/3.4.1)

$$p(\lambda|y) = \mathrm{Gam}(\lambda; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n \; . \end{aligned} \tag{4}$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \tag{5}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \; . \tag{6}$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda)\, p(\lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \; . \end{aligned} \tag{7}$$

Resolving the product in the joint likelihood, we have

$$
\begin{aligned}
p(y,\lambda) &= \prod_{i=1}^{n} \frac{1}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \lambda^{n\bar{y}} \exp\left[-n\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0+n\lambda)\right]
\end{aligned} \tag{8}
$$

where $\bar{y}$ is the mean (→ I/1.10.2) of y:

$$
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{9}
$$

Note that the posterior distribution is proportional to the joint likelihood (→ I/5.1.9):

$$
p(\lambda|y) \propto p(y,\lambda) \; . \tag{10}
$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n$, the posterior distribution is therefore proportional to

$$
p(\lambda|y) \propto \lambda^{a_n-1} \cdot \exp\left[-b_n\lambda\right] \tag{11}
$$

which, when normalized to one, results in the probability density function of the gamma distribution (→ II/3.4.7):

$$
p(\lambda|y) = \frac{{b_n}^{a_n}}{\Gamma(a_0)} \lambda^{a_n-1} \exp\left[-b_n\lambda\right] = \mathrm{Gam}(\lambda; a_n, b_n) \; . \tag{12}
$$

■

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.15; URL: http://www.stat.columbia.edu/~gelman/book/.

3.3.5 Log model evidence

Theorem: Let there be a Poisson-distributed data (→ III/3.3.1) set $y = \{y_1, \ldots, y_n\}$:

$$
y_i \sim \mathrm{Poiss}(\lambda), \quad i = 1, \ldots, n \; . \tag{1}
$$

Moreover, assume a gamma prior distribution (→ III/3.3.3) over the model parameter λ:

$$
p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}
$$

Then, the log model evidence (→ IV/**??**) for this model is

$$
\log p(y|m) = -\sum_{i=1}^{n} \log y_i! + \log\Gamma(a_n) - \log\Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; . \tag{3}
$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$
\begin{aligned}
a_n &= a_0 + n\bar{y} \\
b_n &= b_0 + n \; .
\end{aligned} \tag{4}
$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$
p(y_i|\lambda) = \mathrm{Poiss}(y_i;\lambda) = \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \tag{5}
$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$
p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \; . \tag{6}
$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$
\begin{aligned}
p(y,\lambda) &= p(y|\lambda)\, p(\lambda) \\
&= \prod_{i=1}^{n} \frac{\lambda^{y_i} \cdot \exp\left[-\lambda\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \; .
\end{aligned} \tag{7}
$$

Resolving the product in the joint likelihood, we have

$$
\begin{aligned}
p(y,\lambda) &= \prod_{i=1}^{n} \frac{1}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \lambda^{n\bar{y}} \exp\left[-n\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \\
&= \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0+n\lambda)\right]
\end{aligned} \tag{8}
$$

where $\bar{y}$ is the mean (→ I/1.10.2) of y:

$$
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{9}
$$

Note that the model evidence is the marginal density of the joint likelihood (→ I/5.1.11):

$$
p(y) = \int p(y,\lambda) \, \mathrm{d}\lambda \; . \tag{10}
$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n$, the joint likelihood can also be written as

$$
p(y,\lambda) = \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp\left[-b_n\lambda\right] \; . \tag{11}
$$

Using the probability density function of the gamma distribution (→ II/3.4.7), λ can now be integrated out easily

$$
\begin{aligned}
\mathrm{p}(y) &= \int \prod_{i=1}^{n} \left(\frac{1}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp\left[-b_n \lambda\right] \, \mathrm{d}\lambda \\
&= \prod_{i=1}^{n} \left(\frac{1}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \int \mathrm{Gam}(\lambda; a_n, b_n) \, \mathrm{d}\lambda \\
&= \prod_{i=1}^{n} \left(\frac{1}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \; ,
\end{aligned} \tag{12}
$$

such that the log model evidence (→ IV/**??**) is shown to be

$$
\log p(y|m) = -\sum_{i=1}^{n} \log y_i! + \log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; . \tag{13}
$$

■

3.4 Poisson distribution with exposure values

3.4.1 Definition

Definition: A Poisson distribution with exposure values is defined as a set of observed counts $y = \{y_1, \ldots, y_n\}$, independently distributed according to a Poisson distribution (→ II/1.5.1) with common rate λ and a set of concurrent exposures $x = \{x_1, \ldots, x_n\}$:

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Sources:
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14; URL: http://www.stat.columbia.edu/~gelman/book/.

3.4.2 Maximum likelihood estimation

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ III/3.4.1):

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Then, the maximum likelihood estimate (→ I/4.1.3) for the rate parameter λ is given by

$$
\hat{\lambda} = \frac{\bar{y}}{\bar{x}} \tag{2}
$$

where $\bar{y}$ and $\bar{x}$ are the sample means (→ I/1.10.2)

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; .
\end{aligned} \tag{3}
$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$
p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \tag{4}
$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$
p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \; . \tag{5}
$$

Thus, the log-likelihood function (→ I/4.1.2) is

$$
\mathrm{LL}(\lambda) = \log p(y|\lambda) = \log \left[\prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \right] \tag{6}
$$

which can be developed into

$$
\begin{aligned}
\mathrm{LL}(\lambda) &= \sum_{i=1}^{n} \log \left[\frac{(\lambda x_i)^{y_i} \cdot \exp[-\lambda x_i]}{y_i!} \right] \\
&= \sum_{i=1}^{n} \left[y_i \cdot \log(\lambda x_i) - \lambda x_i - \log(y_i!) \right] \\
&= -\sum_{i=1}^{n} \lambda x_i + \sum_{i=1}^{n} y_i \cdot \left[\log(\lambda) + \log(x_i) \right] - \sum_{i=1}^{n} \log(y_i!) \\
&= -\lambda \sum_{i=1}^{n} x_i + \log(\lambda) \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log(y_i!) \\
&= -n\bar{x}\lambda + n\bar{y} \log(\lambda) + \sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log(y_i!)
\end{aligned} \tag{7}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means from equation (3).

The derivatives of the log-likelihood with respect to λ are

$$
\begin{aligned}
\frac{\mathrm{dLL}(\lambda)}{\mathrm{d}\lambda} &= -n\bar{x} + \frac{n\bar{y}}{\lambda} \\
\frac{\mathrm{d}^2\mathrm{LL}(\lambda)}{\mathrm{d}\lambda^2} &= -\frac{n\bar{y}}{\lambda^2} \; .
\end{aligned} \tag{8}
$$

Setting the first derivative to zero, we obtain:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\lambda})}{\mathrm{d}\lambda} &= 0 \\
0 &= -n\bar{x} + \frac{n\bar{y}}{\hat{\lambda}} \\
\hat{\lambda} &= \frac{n\bar{y}}{n\bar{x}} = \frac{\bar{y}}{\bar{x}} \; .
\end{aligned} \tag{9}
$$

Plugging this value into the second derivative, we confirm:

$$
\begin{aligned}
\frac{\mathrm{d}^2\mathrm{LL}(\hat{\lambda})}{\mathrm{d}\lambda^2} &= -\frac{n\bar{y}}{\hat{\lambda}^2} \\
&= -\frac{n \cdot \bar{y}}{(\bar{y}/\bar{x})^2} \\
&= -\frac{n \cdot \bar{x}^2}{\bar{y}} < 0 \; .
\end{aligned} \tag{10}
$$

This demonstrates that the estimate $\hat{\lambda} = \bar{y}/\bar{x}$ maximizes the likelihood $p(y|\lambda)$.

■

3.4.3 Conjugate prior distribution

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ III/3.4.1):

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Then, the conjugate prior (→ I/5.2.5) for the model parameter λ is a gamma distribution (→ II/3.4.1):

$$
p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}
$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$
p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{3}
$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$
p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{4}
$$

Resolving the product in the likelihood function, we have

$$
\begin{aligned}
p(y|\lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \cdot \prod_{i=1}^{n} \lambda^{y_i} \cdot \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!}\right) \cdot \lambda^{\sum_{i=1}^{n} y_i} \cdot \exp\left[-\lambda \sum_{i=1}^{n} x_i\right] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!}\right) \cdot \lambda^{n\bar{y}} \cdot \exp\left[-n\bar{x}\lambda\right]
\end{aligned} \tag{5}
$$

where $\bar{y}$ and $\bar{x}$ are the means (→ I/1.10.2) of y and x respectively:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; .
\end{aligned} \tag{6}
$$

In other words, the likelihood function is proportional to a power of λ times an exponential of λ:

$$
p(y|\lambda) \propto \lambda^{n\bar{y}} \cdot \exp\left[-n\bar{x}\lambda\right] \; . \tag{7}
$$

The same is true for a gamma distribution over λ

$$
p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \tag{8}
$$

the probability density function of which (→ II/3.4.7)

$$
p(\lambda) = \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \tag{9}
$$

exhibits the same proportionality

$$
p(\lambda) \propto \lambda^{a_0-1} \cdot \exp[-b_0\lambda] \tag{10}
$$

and is therefore conjugate relative to the likelihood.

■

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.14ff.; URL: http://www.stat.columbia.edu/~gelman/book/.

3.4.4 Posterior distribution

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ III/3.4.1):

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Moreover, assume a gamma prior distribution (→ III/3.4.3) over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ I/5.1.7) is also a gamma distribution (→ II/3.4.1)

$$p(\lambda|y) = \mathrm{Gam}(\lambda; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ I/5.1.7) are given by

$$\begin{aligned} a_n &= a_0 + n\bar{y} \\ b_n &= b_0 + n\bar{x} \; . \end{aligned} \tag{4}$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{5}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{6}$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{aligned} p(y, \lambda) &= p(y|\lambda) \, p(\lambda) \\ &= \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \; . \end{aligned} \tag{7}$$

Resolving the product in the joint likelihood, we have

$$\begin{aligned} p(y, \lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^{n} y_i} \exp\left[-\lambda \sum_{i=1}^{n} x_i \right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0 + n\bar{x})\lambda\right] \end{aligned} \tag{8}$$

where $\bar{y}$ and $\bar{x}$ are the means (→ I/1.10.2) of y and x respectively:

$$
\begin{aligned}
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; .
\end{aligned} \tag{9}
$$

Note that the posterior distribution is proportional to the joint likelihood (→ I/5.1.9):

$$
p(\lambda|y) \propto p(y,\lambda) \; . \tag{10}
$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the posterior distribution is therefore proportional to

$$
p(\lambda|y) \propto \lambda^{a_n-1} \cdot \exp\left[-b_n \lambda\right] \tag{11}
$$

which, when normalized to one, results in the probability density function of the gamma distribution (→ II/3.4.7):

$$
p(\lambda|y) = \frac{{b_n}^{a_n}}{\Gamma(a_0)} \lambda^{a_n-1} \exp\left[-b_n \lambda\right] = \mathrm{Gam}(\lambda; a_n, b_n) \; . \tag{12}
$$

■

Sources:

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014): "Other standard single-parameter models"; in: *Bayesian Data Analysis*, 3rd edition, ch. 2.6, p. 45, eq. 2.15; URL: http://www.stat.columbia.edu/~gelman/book/.

3.4.5 Log model evidence

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ III/3.4.1):

$$
y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}
$$

Moreover, assume a gamma prior distribution (→ III/3.4.3) over the model parameter λ:

$$
p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}
$$

Then, the log model evidence (→ IV/**??**) for this model is

$$
\begin{split}
\log p(y|m) = &\sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i ! + \\
&\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; .
\end{split} \tag{3}
$$

where the posterior hyperparameters (→ I/5.1.7) are given by

$$
\begin{aligned}
a_n &= a_0 + n\bar{y} \\
a_n &= a_0 + n\bar{x} \; .
\end{aligned} \tag{4}
$$

Proof: With the probability mass function of the Poisson distribution (→ II/1.5.2), the likelihood function (→ I/5.1.2) for each observation implied by (1) is given by

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda x_i) = \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \tag{5}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \; . \tag{6}$$

Combining the likelihood function (6) with the prior distribution (2), the joint likelihood (→ I/5.1.5) of the model is given by

$$\begin{split} p(y,\lambda) &= p(y|\lambda)\, p(\lambda) \\ &= \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \cdot \exp\left[-\lambda x_i\right]}{y_i!} \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \; . \end{split} \tag{7}$$

Resolving the product in the joint likelihood, we have

$$\begin{split} p(y,\lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^{n} y_i} \exp\left[-\lambda \sum_{i=1}^{n} x_i \right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0 \lambda] \\ &= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0+n\bar{y}-1} \cdot \exp\left[-(b_0+n\bar{x})\lambda\right] \end{split} \tag{8}$$

where $\bar{y}$ and $\bar{x}$ are the means (→ I/1.10.2) of y and x respectively:

$$\begin{split} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \; . \end{split} \tag{9}$$

Note that the model evidence is the marginal density of the joint likelihood (→ I/5.1.11):

$$p(y) = \int p(y,\lambda) \, \mathrm{d}\lambda \; . \tag{10}$$

Setting $a_n = a_0 + n\bar{y}$ and $b_n = b_0 + n\bar{x}$, the joint likelihood can also be written as

$$p(y,\lambda) = \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n-1} \exp\left[-b_n \lambda\right] \; . \tag{11}$$

Using the probability density function of the gamma distribution (→ II/3.4.7), λ can now be integrated out easily

$$
\begin{aligned}
\mathrm{p}(y) &= \int \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_n)}{{b_n}^{a_n}} \cdot \frac{{b_n}^{a_n}}{\Gamma(a_n)} \lambda^{a_n - 1} \exp\left[-b_n \lambda\right] \, \mathrm{d}\lambda \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \int \mathrm{Gam}(\lambda; a_n, b_n) \, \mathrm{d}\lambda \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{\Gamma(a_n)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_n}^{a_n}} \; ,
\end{aligned} \tag{12}
$$

such that the log model evidence (→ IV/**??**) is shown to be

$$
\begin{split}
\log p(y|m) = &\sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i! + \\
&\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; .
\end{split} \tag{13}
$$

■

4 Frequency data

4.1 Beta-distributed data

4.1.1 Definition

Definition: Beta-distributed data are defined as a set of proportions $y = \{y_1, \ldots, y_n\}$ with $y_i \in [0, 1]$, $i = 1, \ldots, n$, independent and identically distributed according to a beta distribution (→ II/3.9.1) with shapes α and β:

$$y_i \sim \mathrm{Bet}(\alpha, \beta), \quad i = 1, \ldots, n \; . \tag{1}$$

4.1.2 Method of moments

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed counts independent and identically distributed according to a beta distribution (→ II/3.9.1) with shapes α and β:

$$y_i \sim \mathrm{Bet}(\alpha, \beta), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the method-of-moments estimates (→ I/4.1.8) for the shape parameters α and β are given by

$$\begin{aligned} \hat{\alpha} &= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \\ \hat{\beta} &= (1-\bar{y}) \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \end{aligned} \tag{2}$$

where $\bar{y}$ is the sample mean (→ I/1.10.2) and $\bar{v}$ is the unbiased sample variance (→ I/1.11.2):

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \end{aligned} \tag{3}$$

Proof: Mean (→ II/3.9.6) and variance (→ II/3.9.7) of the beta distribution (→ II/3.9.1) in terms of the parameters α and β are given by

$$\begin{aligned} \mathrm{E}(X) &= \frac{\alpha}{\alpha+\beta} \\ \mathrm{Var}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \end{aligned} \tag{4}$$

Thus, matching the moments (→ I/4.1.8) requires us to solve the following equation system for α and β:

$$\begin{aligned} \bar{y} &= \frac{\alpha}{\alpha+\beta} \\ \bar{v} &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \end{aligned} \tag{5}$$

From the first equation, we can deduce:

$$
\begin{split}
\bar{y}(\alpha+\beta) &= \alpha \\
\alpha\bar{y} + \beta\bar{y} &= \alpha \\
\beta\bar{y} &= \alpha - \alpha\bar{y} \\
\beta &= \frac{\alpha}{\bar{y}} - \alpha \\
\beta &= \alpha \left(\frac{1}{\bar{y}} - 1 \right) \; .
\end{split} \tag{6}
$$

If we define $q = 1/\bar{y} - 1$ and plug (6) into the second equation, we have:

$$
\begin{split}
\bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha + \alpha q)^2 (\alpha + \alpha q + 1)} \\
&= \frac{\alpha^2 q}{(\alpha(1+q))^2 (\alpha(1+q) + 1)} \\
&= \frac{q}{(1+q)^2 (\alpha(1+q) + 1)} \\
&= \frac{q}{\alpha(1+q)^3 + (1+q)^2} \\
q &= \bar{v} \left[\alpha(1+q)^3 + (1+q)^2 \right] \; .
\end{split} \tag{7}
$$

Noting that $1 + q = 1/\bar{y}$ and $q = (1 - \bar{y})/\bar{y}$, one obtains for α:

$$
\begin{split}
\frac{1-\bar{y}}{\bar{y}} &= \bar{v} \left[\frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \right] \\
\frac{1-\bar{y}}{\bar{y}\,\bar{v}} &= \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \\
\frac{\bar{y}^3(1-\bar{y})}{\bar{y}\,\bar{v}} &= \alpha + \bar{y} \\
\alpha &= \frac{\bar{y}^2(1-\bar{y})}{\bar{v}} - \bar{y} \\
&= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \; .
\end{split} \tag{8}
$$

Plugging this into equation (6), one obtains for β:

$$
\begin{split}
\beta &= \bar{y} \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \cdot \left(\frac{1-\bar{y}}{\bar{y}} \right) \\
&= (1-\bar{y}) \left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1 \right) \; .
\end{split} \tag{9}
$$

Together, (8) and (9) constitute the method-of-moment estimates of α and β.

■

Sources:

- Wikipedia (2020): "Beta distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-20; URL: https://en.wikipedia.org/wiki/Beta_distribution#Method_of_moments.

4.2 Dirichlet-distributed data

4.2.1 Definition

Definition: Dirichlet-distributed data are defined as a set of vectors of proportions $y = \{y_1, \ldots, y_n\}$ where

$$\begin{aligned} y_i &= [y_{i1}, \ldots, y_{ik}], \\ y_{ij} &\in [0,1] \quad \text{and} \\ \sum_{j=1}^{k} y_{ij} &= 1 \end{aligned} \tag{1}$$

for all $i = 1, \ldots, n$ (and $j = 1, \ldots, k$) and each y_i is independent and identically distributed according to a Dirichlet distribution (→ II/4.4.1) with concentration parameters $\alpha = [\alpha_1, \ldots, \alpha_k]$:

$$y_i \sim \mathrm{Dir}(\alpha), \quad i = 1, \ldots, n \; . \tag{2}$$

4.2.2 Maximum likelihood estimation

Theorem: Let there be a Dirichlet-distributed data (→ III/4.2.1) set $y = \{y_1, \ldots, y_n\}$:

$$y_i \sim \mathrm{Dir}(\alpha), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the maximum likelihood estimate (→ I/4.1.3) for the concentration parameters α can be obtained by iteratively computing

$$\alpha_j^{(\mathrm{new})} = \psi^{-1}\left[\psi\left(\sum_{j=1}^{k} \alpha_j^{(\mathrm{old})}\right) + \log \bar{y}_j\right] \tag{2}$$

where $\psi(x)$ is the digamma function and $\log \bar{y}_j$ is given by:

$$\log \bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} \log y_{ij} \; . \tag{3}$$

Proof: The likelihood function (→ I/5.1.2) for each observation is given by the probability density function of the Dirichlet distribution (→ II/4.4.2)

$$p(y_i|\alpha) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} {y_{ij}}^{\alpha_j - 1} \tag{4}$$

and because observations are independent (→ I/1.3.6), the likelihood function for all observations is the product of the individual ones:

$$p(y|\alpha) = \prod_{i=1}^{n} p(y_i|\alpha) = \prod_{i=1}^{n} \left[\frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} {y_{ij}}^{\alpha_j - 1}\right] \; . \tag{5}$$

Thus, the log-likelihood function (→ I/4.1.2) is

$$\mathrm{LL}(\alpha) = \log p(y|\alpha) = \log \prod_{i=1}^{n} \left[\frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} {y_{ij}}^{\alpha_j - 1} \right] \tag{6}$$

which can be developed into

$$\begin{aligned}
\mathrm{LL}(\alpha) &= \sum_{i=1}^{n} \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \sum_{i=1}^{n} \sum_{j=1}^{k} \log \Gamma(\alpha_j) + \sum_{i=1}^{n} \sum_{j=1}^{k} (\alpha_j - 1) \log y_{ij} \\
&= n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - n \sum_{j=1}^{k} \log \Gamma(\alpha_j) + n \sum_{j=1}^{k} (\alpha_j - 1) \, \frac{1}{n} \sum_{i=1}^{n} \log y_{ij} \\
&= n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - n \sum_{j=1}^{k} \log \Gamma(\alpha_j) + n \sum_{j=1}^{k} (\alpha_j - 1) \, \log \bar{y}_j
\end{aligned} \tag{7}$$

where we have specified

$$\log \bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} \log y_{ij} \; . \tag{8}$$

The derivative of the log-likelihood with respect to a particular parameter α_j is

$$\begin{aligned}
\frac{\mathrm{dLL}(\alpha)}{\mathrm{d}\alpha_j} &= \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - n \sum_{j=1}^{k} \log \Gamma(\alpha_j) + n \sum_{j=1}^{k} (\alpha_j - 1) \, \log \bar{y}_j \right] \\
&= \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) \right] - \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n \log \Gamma(\alpha_j) \right] + \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n (\alpha_j - 1) \, \log \bar{y}_j \right] \\
&= n \psi\left(\sum_{j=1}^{k} \alpha_j\right) - n \psi(\alpha_j) + n \log \bar{y}_j
\end{aligned} \tag{9}$$

where we have used the digamma function

$$\psi(x) = \frac{\mathrm{d} \log \Gamma(x)}{\mathrm{d}x} \; . \tag{10}$$

Setting this derivative to zero, we obtain:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\alpha)}{\mathrm{d}\alpha_j} &= 0 \\
0 &= n\psi\left(\sum_{j=1}^{k} \alpha_j\right) - n\psi(\alpha_j) + n\log \bar{y}_j \\
0 &= \psi\left(\sum_{j=1}^{k} \alpha_j\right) - \psi(\alpha_j) + \log \bar{y}_j \\
\psi(\alpha_j) &= \psi\left(\sum_{j=1}^{k} \alpha_j\right) + \log \bar{y}_j \\
\alpha_j &= \psi^{-1}\left[\psi\left(\sum_{j=1}^{k} \alpha_j\right) + \log \bar{y}_j\right] \; .
\end{aligned} \tag{11}
$$

In the following, we will use a fixed-point iteration to maximize $\mathrm{LL}(\alpha)$. Given an initial guess for α, we construct a lower bound on the likelihood function (7) which is tight at α. The maximum of this bound is computed and it becomes the new guess. Because the Dirichlet distribution (→ II/4.4.1) belongs to the exponential family, the log-likelihood function is convex in α ánd the maximum is the only stationary point, such that the procedure is guaranteed to converge to the maximum.
In our case, we use a bound on the gamma function

$$
\begin{aligned}
\Gamma(x) &\geq \Gamma(\hat{x}) \cdot \exp\left[(x-\hat{x})\, \psi(\hat{x})\right] \\
\log\Gamma(x) &\geq \log\Gamma(\hat{x}) + (x-\hat{x})\, \psi(\hat{x})
\end{aligned} \tag{12}
$$

and apply it to $\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)$ in (7) to yield

$$
\begin{aligned}
\frac{1}{n}\mathrm{LL}(\alpha) &= \log\Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \sum_{j=1}^{k} \log\Gamma(\alpha_j) + \sum_{j=1}^{k} (\alpha_j - 1)\, \log \bar{y}_j \\
\frac{1}{n}\mathrm{LL}(\alpha) &\geq \log\Gamma\left(\sum_{j=1}^{k} \hat{\alpha}_j\right) + \left(\sum_{j=1}^{k} \alpha_j - \sum_{j=1}^{k} \hat{\alpha}_j\right) \psi\left(\sum_{j=1}^{k} \hat{\alpha}_j\right) - \sum_{j=1}^{k} \log\Gamma(\alpha_j) + \sum_{j=1}^{k} (\alpha_j - 1)\, \log \bar{y}_j \\
\frac{1}{n}\mathrm{LL}(\alpha) &\geq \left(\sum_{j=1}^{k} \alpha_j\right) \psi\left(\sum_{j=1}^{k} \hat{\alpha}_j\right) - \sum_{j=1}^{k} \log\Gamma(\alpha_j) + \sum_{j=1}^{k} (\alpha_j - 1)\, \log \bar{y}_j + \text{const.}
\end{aligned} \tag{13}
$$

which leads to the following fixed-point iteration using (11):

$$
\alpha_j^{(\mathrm{new})} = \psi^{-1}\left[\psi\left(\sum_{j=1}^{k} \alpha_j^{(\mathrm{old})}\right) + \log \bar{y}_j\right] \; . \tag{14}
$$

■

Sources:

- Minka TP (2012): "Estimating a Dirichlet distribution"; in: *Papers by Tom Minka*, retrieved on 2020-10-22; URL: https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf.

4.3 Beta-binomial data

4.3.1 Definition

Definition: Beta-binomial data are defined as a set of counts $y = \{y_1, \ldots, y_N\}$ with $y_i \in \mathbb{N}$, $i = 1, \ldots, N$, independent and identically distributed according to a beta-binomial distribution (→ II/1.4.1) with number of trials n as well as shapes α and β:

$$y_i \sim \mathrm{BetBin}(n, \alpha, \beta), \quad i = 1, \ldots, N \; . \tag{1}$$

4.3.2 Method of moments

Theorem: Let $y = \{y_1, \ldots, y_N\}$ be a set of observed counts independent and identically distributed according to a beta-binomial distribution (→ II/1.4.1) with number of trials n as well as parameters α and β:

$$y_i \sim \mathrm{BetBin}(n, \alpha, \beta), \quad i = 1, \ldots, N \; . \tag{1}$$

Then, the method-of-moments estimates (→ I/4.1.8) for the parameters α and β are given by

$$\begin{aligned} \hat{\alpha} &= \frac{n m_1 - m_2}{n\left(\frac{m_2}{m_1} - m_1 - 1\right) + m_1} \\ \hat{\beta} &= \frac{(n - m_1)\left(n - \frac{m_2}{m_1}\right)}{n\left(\frac{m_2}{m_1} - m_1 - 1\right) + m_1} \end{aligned} \tag{2}$$

where m_1 and m_2 are the first two raw sample moments (→ I/1.18.3):

$$\begin{aligned} m_1 &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ m_2 &= \frac{1}{N} \sum_{i=1}^{N} y_i^2 \; . \end{aligned} \tag{3}$$

Proof: The first two raw moments of the beta-binomial distribution in terms of the parameters α and β are given by

$$\begin{aligned} \mu_1 &= \frac{n\alpha}{\alpha + \beta} \\ \mu_2 &= \frac{n\alpha(n\alpha + \beta + n)}{(\alpha + \beta)(n\alpha + \beta + 1)} \end{aligned} \tag{4}$$

Thus, matching the moments (→ I/4.1.8) requires us to solve the following equation system for α and β:

$$\begin{aligned} m_1 &= \frac{n\alpha}{\alpha + \beta} \\ m_2 &= \frac{n\alpha(n\alpha + \beta + n)}{(\alpha + \beta)(n\alpha + \beta + 1)} \; . \end{aligned} \tag{5}$$

From the first equation, we can deduce:

$$
\begin{aligned}
m_1(\alpha+\beta) &= n\alpha \\
m_1\alpha + m_1\beta &= n\alpha \\
m_1\beta &= n\alpha - m_1\alpha \\
\beta &= \frac{n\alpha}{m_1} - \alpha \\
\beta &= \alpha\left(\frac{n}{m_1} - 1\right) \; .
\end{aligned}
\tag{6}
$$

If we define $q = n/m_1 - 1$ and plug (6) into the second equation, we have:

$$
\begin{aligned}
m_2 &= \frac{n\alpha(n\alpha + \alpha q + n)}{(\alpha + \alpha q)(\alpha + \alpha q + 1)} \\
&= \frac{n\alpha(\alpha(n+q) + n)}{\alpha(1+q)(\alpha(1+q)+1)} \\
&= \frac{n(\alpha(n+q)+n)}{(1+q)(\alpha(1+q)+1)} \\
&= \frac{n(\alpha(n+q)+n)}{\alpha(1+q)^2 + (1+q)} \; .
\end{aligned}
\tag{7}
$$

Noting that $1 + q = n/m_1$ and expanding the fraction with m_1, one obtains:

$$
\begin{aligned}
m_2 &= \frac{n\left(\alpha\left(\frac{n}{m_1} + n - 1\right) + n\right)}{n\left(\alpha\frac{n}{m_1^2} + \frac{1}{m_1}\right)} \\
m_2 &= \frac{\alpha\left(n + nm_1 - m_1\right) + nm_1}{\alpha\frac{n}{m_1} + 1} \\
m_2\left(\frac{\alpha n}{m_1} + 1\right) &= \alpha\left(n + nm_1 - m_1\right) + nm_1 \\
\alpha\left(n\frac{m_2}{m_1} - (n + nm_1 - m_1)\right) &= nm_1 - m_2 \\
\alpha\left(n\left(\frac{m_2}{m_1} - m_1 - 1\right) + m_1\right) &= nm_1 - m_2 \\
\alpha &= \frac{nm_1 - m_2}{n\left(\frac{m_2}{m_1} - m_1 - 1\right) + m_1} \; .
\end{aligned}
\tag{8}
$$

Plugging this into equation (6), one obtains for β:

$$
\begin{aligned}
\beta &= \alpha \left(\frac{n}{m_1} - 1 \right) \\
\beta &= \left(\frac{n m_1 - m_2}{n \left(\frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \right) \left(\frac{n}{m_1} - 1 \right) \\
\beta &= \frac{n^2 - n m_1 - n \frac{m_2}{m_1} + m_2}{n \left(\frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \\
\hat{\beta} &= \frac{(n - m_1) \left(n - \frac{m_2}{m_1} \right)}{n \left(\frac{m_2}{m_1} - m_1 - 1 \right) + m_1} \; .
\end{aligned} \tag{9}
$$

Together, (8) and (9) constitute the method-of-moment estimates of α and β.

■

Sources:

- statisticsmatt (2022): "Method of Moments Estimation Beta Binomial Distribution"; in: *YouTube*, retrieved on 2022-10-07; URL: https://www.youtube.com/watch?v=18PWnWJsPnA.
- Wikipedia (2022): "Beta-binomial distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2022-10-07; URL: https://en.wikipedia.org/wiki/Beta-binomial_distribution#Method_of_moments.

5 Categorical data

5.1 Logistic regression

5.1.1 Definition

Definition: A logistic regression model is given by a set of binary observations $y_i \in \{0, 1\}, i = 1, \ldots, n$, a set of predictors $x_j \in \mathbb{R}^n, j = 1, \ldots, p$, a base b and the assumption that the log-odds are a linear combination of the predictors:

$$l_i = x_i \beta + \varepsilon_i, \; i = 1, \ldots, n \tag{1}$$

where l_i are the log-odds that $y_i = 1$

$$l_i = \log_b \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} \tag{2}$$

and x_i is the i-th row of the $n \times p$ matrix

$$X = [x_1, \ldots, x_p] \; . \tag{3}$$

Within this model,
- y are called "categorical observations" or "dependent variable";
- X is called "design matrix" or "set of independent variables";
- β are called "regression coefficients" or "weights";
- ε_i is called "noise" or "error term";
- n is the number of observations;
- p is the number of predictors.

Sources:
- Wikipedia (2020): "Logistic regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-06-28; URL: https://en.wikipedia.org/wiki/Logistic_regression#Logistic_model.

5.1.2 Probability and log-odds

Theorem: Assume a logistic regression model (→ III/5.1.1)

$$l_i = x_i \beta + \varepsilon_i, \; i = 1, \ldots, n \tag{1}$$

where x_i are the predictors corresponding to the i-th observation y_i and l_i are the log-odds that $y_i = 1$.
Then, the log-odds in favor of $y_i = 1$ against $y_i = 0$ can also be expressed as

$$l_i = \log_b \frac{p(x_i|y_i = 1)\, p(y_i = 1)}{p(x_i|y_i = 0)\, p(y_i = 0)} \tag{2}$$

where $p(x_i|y_i)$ is a likelihood function (→ I/5.1.2) consistent with (1), $p(y_i)$ are prior probabilities (→ I/5.1.3) for $y_i = 1$ and $y_i = 0$ and where b is the base used to form the log-odds l_i.

Proof: Using Bayes' theorem (→ I/5.3.1) and the law of marginal probability (→ I/1.3.3), the posterior probabilities (→ I/5.1.7) for $y_i = 1$ and $y_i = 0$ are given by

$$
\begin{aligned}
p(y_i = 1|x_i) &= \frac{p(x_i|y_i = 1)\, p(y_i = 1)}{p(x_i|y_i = 1)\, p(y_i = 1) + p(x_i|y_i = 0)\, p(y_i = 0)} \\
p(y_i = 0|x_i) &= \frac{p(x_i|y_i = 0)\, p(y_i = 0)}{p(x_i|y_i = 1)\, p(y_i = 1) + p(x_i|y_i = 0)\, p(y_i = 0)} \; .
\end{aligned} \tag{3}
$$

Calculating the log-odds from the posterior probabilties, we have

$$
\begin{aligned}
l_i &= \log_b \frac{p(y_i = 1|x_i)}{p(y_i = 0|x_i)} \\
&= \log_b \frac{p(x_i|y_i = 1)\, p(y_i = 1)}{p(x_i|y_i = 0)\, p(y_i = 0)} \; .
\end{aligned} \tag{4}
$$

■

Sources:

- Bishop, Christopher M. (2006): "Linear Models for Classification"; in: *Pattern Recognition for Machine Learning*, ch. 4, p. 197, eq. 4.58; URL: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf.

5.1.3 Log-odds and probability

Theorem: Assume a logistic regression model (→ III/5.1.1)

$$
l_i = x_i \beta + \varepsilon_i, \; i = 1, \ldots, n \tag{1}
$$

where x_i are the predictors corresponding to the i-th observation y_i and l_i are the log-odds that $y_i = 1$.
Then, the probability that $y_i = 1$ is given by

$$
\mathrm{Pr}(y_i = 1) = \frac{1}{1 + b^{-(x_i \beta + \varepsilon_i)}} \tag{2}
$$

where b is the base used to form the log-odds l_i.

Proof: Let us denote $\mathrm{Pr}(y_i = 1)$ as p_i. Then, the log-odds are

$$
l_i = \log_b \frac{p_i}{1 - p_i} \tag{3}
$$

and using (1), we have

$$
\begin{aligned}
\log_b \frac{p_i}{1-p_i} &= x_i\beta + \varepsilon_i \\
\frac{p_i}{1-p_i} &= b^{x_i\beta+\varepsilon_i} \\
p_i &= \left(b^{x_i\beta+\varepsilon_i}\right)(1-p_i) \\
p_i\left(1+b^{x_i\beta+\varepsilon_i}\right) &= b^{x_i\beta+\varepsilon_i} \\
p_i &= \frac{b^{x_i\beta+\varepsilon_i}}{1+b^{x_i\beta+\varepsilon_i}} \\
p_i &= \frac{b^{x_i\beta+\varepsilon_i}}{b^{x_i\beta+\varepsilon_i}\left(1+b^{-(x_i\beta+\varepsilon_i)}\right)} \\
p_i &= \frac{1}{1+b^{-(x_i\beta+\varepsilon_i)}}
\end{aligned} \tag{4}
$$

which proves the identity given by (2).

■

Sources:

- Wikipedia (2020): "Logistic regression"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-03; URL: https://en.wikipedia.org/wiki/Logistic_regression#Logistic_model.

Chapter IV

Model Selection

1 Goodness-of-fit measures

1.1 Residual variance

1.1.1 Definition

Definition: Let there be a linear regression model (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

with measured data y, known design matrix X and covariance structure V as well as unknown regression coefficients β and noise variance σ^2.
Then, an estimate of the noise variance σ^2 is called the "residual variance" $\hat{\sigma}^2$, e.g. obtained via maximum likelihood estimation (→ I/4.1.3).

1.1.2 Maximum likelihood estimator is biased (p = 1)

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of independent normally distributed (→ II/3.2.1) observations with unknown mean (→ I/1.10.1) μ and variance (→ I/1.11.1) σ^2:

$$y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Then,
1) the maximum likelihood estimator (→ I/4.1.3) of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2}$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3}$$

2) and $\hat{\sigma}^2$ is a biased estimator of σ^2

$$\mathrm{E}\left[\hat{\sigma}^2\right] \neq \sigma^2 \; , \tag{4}$$

more precisely:

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \frac{n-1}{n} \sigma^2 \; . \tag{5}$$

Proof:
1) This is equivalent to the maximum likelihood estimator for the univariate Gaussian with unknown variance (→ III/1.1.2) and a special case of the maximum likelihood estimator for multiple linear regression (→ III/1.5.23) in which $X = 1_n$ and $\hat{\beta} = \bar{y}$:

$$\begin{split}
\hat{\sigma}^2 &= \frac{1}{n}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \\
&= \frac{1}{n}(y - 1_n\bar{y})^\mathrm{T}(y - 1_n\bar{y}) \\
&= \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \; .
\end{split} \tag{6}$$

2) The expectation (→ I/1.10.1) of the maximum likelihood estimator (→ I/4.1.3) can be developed as follows:

$$
\begin{aligned}
\mathrm{E}\left[\hat{\sigma}^2\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\left(y_i^2-2y_i\bar{y}+\bar{y}^2\right)\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}y_i^2-2\sum_{i=1}^{n}y_i\bar{y}+\sum_{i=1}^{n}\bar{y}^2\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}y_i^2-2n\bar{y}^2+n\bar{y}^2\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathrm{E}\left[y_i^2\right]-n\mathrm{E}\left[\bar{y}^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[y_i^2\right]-\mathrm{E}\left[\bar{y}^2\right]
\end{aligned} \tag{7}
$$

Due to the partition of variance into expected values (→ I/1.11.3)

$$
\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \;, \tag{8}
$$

we have

$$
\begin{aligned}
\mathrm{Var}(y_i) &= \mathrm{E}(y_i^2) - \mathrm{E}(y_i)^2 \\
\mathrm{Var}(\bar{y}) &= \mathrm{E}(\bar{y}^2) - \mathrm{E}(\bar{y})^2 \;,
\end{aligned} \tag{9}
$$

such that (7) becomes

$$
\mathrm{E}\left[\hat{\sigma}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}(y_i)+\mathrm{E}(y_i)^2\right) - \left(\mathrm{Var}(\bar{y})+\mathrm{E}(\bar{y})^2\right) \;. \tag{10}
$$

From (1), it follows that

$$
\mathrm{E}(y_i) = \mu \quad \text{and} \quad \mathrm{Var}(y_i) = \sigma^2 \;. \tag{11}
$$

The expectation (→ I/1.10.1) of $\bar{y}$ given by (3) is

$$
\begin{aligned}
\mathrm{E}\left[\bar{y}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[y_i\right] \\
&\overset{(11)}{=} \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n \cdot \mu \\
&= \mu \; .
\end{aligned} \tag{12}
$$

The variance of $\bar{y}$ given by (3) is

$$
\begin{aligned}
\mathrm{Var}\left[\bar{y}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[y_i\right] \\
&\overset{(11)}{=} \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\
&= \frac{1}{n}\sigma^2 \; .
\end{aligned} \tag{13}
$$

Plugging (11), (12) and (13) into (10), we have

$$
\begin{aligned}
\mathrm{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\sum_{i=1}^{n} \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\
\mathrm{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n} \cdot n \cdot \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\
\mathrm{E}\left[\hat{\sigma}^2\right] &= \sigma^2 + \mu^2 - \frac{1}{n}\sigma^2 - \mu^2 \\
\mathrm{E}\left[\hat{\sigma}^2\right] &= \frac{n-1}{n}\sigma^2
\end{aligned} \tag{14}
$$

which proves the bias given by (5).

■

Sources:

- Liang, Dawen (????): "Maximum Likelihood Estimator for Variance is Biased: Proof", retrieved on 2020-02-24; URL: https://dawenl.github.io/files/mle_biased.pdf.

1.1.3 Maximum likelihood estimator is biased (p > 1)

Theorem: Consider a linear regression model (→ III/1.5.1) with known design matrix X, known covariance structure V, unknown regression parameters β and unknown noise variance σ^2:

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}
$$

Then,
1) the maximum likelihood estimator (→ I/4.1.3) of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \tag{2}$$

where

$$\hat{\beta} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \tag{3}$$

2) and $\hat{\sigma}^2$ is a biased estimator of σ^2

$$\mathrm{E}\left[\hat{\sigma}^2\right] \neq \sigma^2 \; , \tag{4}$$

more precisely:

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \frac{n-p}{n} \sigma^2 \; . \tag{5}$$

Proof:
1) This follows from maximum likelihood estimation for multiple linear regression (→ III/1.5.23) and is a special case (→ III/1.5.2) of maximum likelihood estimation for the general linear model (→ III/2.1.4) in which $Y = y$, $B = \beta$ and $\Sigma = \sigma^2$:

$$\begin{split} \hat{\sigma}^2 &= \frac{1}{n}(Y - X\hat{B})^\mathrm{T} V^{-1} (Y - X\hat{B}) \\ &= \frac{1}{n}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; . \end{split} \tag{6}$$

2) We know that the residual sum of squares, divided by the true noise variance, is following a chi-squared distribution (→ III/1.5.20):

$$\begin{split} \frac{\hat{\varepsilon}^\mathrm{T}\hat{\varepsilon}}{\sigma^2} &\sim \chi^2(n-p) \\ \text{where} \quad \hat{\varepsilon}^\mathrm{T}\hat{\varepsilon} &= (y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; . \end{split} \tag{7}$$

Thus, combining (7) and (6), we have:

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p) \; . \tag{8}$$

Using the relationship between chi-squared distribution and gamma distribution (→ II/3.7.2)

$$X \sim \chi^2(k) \quad \Rightarrow \quad cX \sim \mathrm{Gam}\left(\frac{k}{2}, \frac{1}{2c}\right) \; , \tag{9}$$

we can deduce from (8) that

$$\hat{\sigma}^2 = \frac{\sigma^2}{n} \cdot \frac{n\hat{\sigma}^2}{\sigma^2} \sim \mathrm{Gam}\left(\frac{n-p}{2}, \frac{n}{2\sigma^2}\right) \; . \tag{10}$$

Using the expected value of the gamma distribution (→ II/3.4.11)

$$X \sim \mathrm{Gam}(a,b) \quad \Rightarrow \quad \mathrm{E}(X) = \frac{a}{b} \; , \tag{11}$$

we can deduce from (10) that

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \frac{\frac{n-p}{2}}{\frac{n}{2\sigma^2}} = \frac{n-p}{n}\sigma^2 \tag{12}$$

which proves the relationship given by (5).

■

Sources:

- ocram (2022): "Why is RSS distributed chi square times n-p?"; in: *StackExchange CrossValidated*, retrieved on 2022-12-21; URL: https://stats.stackexchange.com/a/20230.

1.1.4 Construction of unbiased estimator (p = 1)

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of independent normally distributed (→ II/3.2.1) observations with unknown mean (→ I/1.10.1) μ and variance (→ I/1.11.1) σ^2:

$$y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

An unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2_{\text{unb}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \; . \tag{2}$$

Proof: It can be shown that (→ IV/1.1.2) the maximum likelihood estimator (→ I/4.1.3) of σ^2

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \tag{3}$$

is a biased estimator in the sense that

$$\mathbb{E}\left[\hat{\sigma}^2_{\text{MLE}}\right] = \frac{n-1}{n}\sigma^2 \; . \tag{4}$$

From (4), it follows that

$$\begin{split} \mathbb{E}\left[\frac{n}{n-1}\hat{\sigma}^2_{\text{MLE}}\right] &= \frac{n}{n-1}\mathbb{E}\left[\hat{\sigma}^2_{\text{MLE}}\right] \\ &\overset{(4)}{=} \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 \\ &= \sigma^2 \; , \end{split} \tag{5}$$

such that an unbiased estimator can be constructed as

$$\begin{split} \hat{\sigma}^2_{\text{unb}} &= \frac{n}{n-1}\hat{\sigma}^2_{\text{MLE}} \\ &\overset{(3)}{=} \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \; . \end{split} \tag{6}$$

■

Sources:
- Liang, Dawen (????): "Maximum Likelihood Estimator for Variance is Biased: Proof", retrieved on 2020-02-25; URL: https://dawenl.github.io/files/mle_biased.pdf.

1.1.5 Construction of unbiased estimator (p > 1)

Theorem: Consider a linear regression model (→ III/1.5.1) with known design matrix X, known covariance structure V, unknown regression parameters β and unknown noise variance σ^2:

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}$$

An unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \tag{2}$$

where

$$\hat{\beta} = (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \; . \tag{3}$$

Proof: It can be shown that (→ IV/1.1.3) the maximum likelihood estimator (→ I/4.1.3) of σ^2

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \tag{4}$$

is a biased estimator in the sense that

$$\mathbb{E}\left[\hat{\sigma}^2_{\mathrm{MLE}}\right] = \frac{n-p}{n}\sigma^2 \; . \tag{5}$$

From (5), it follows that

$$\begin{split} \mathbb{E}\left[\frac{n}{n-p}\hat{\sigma}^2_{\mathrm{MLE}}\right] &= \frac{n}{n-p}\mathbb{E}\left[\hat{\sigma}^2_{\mathrm{MLE}}\right] \\ &\overset{(5)}{=} \frac{n}{n-p} \cdot \frac{n-p}{n}\sigma^2 \\ &= \sigma^2 \; , \end{split} \tag{6}$$

such that an unbiased estimator can be constructed as

$$\begin{split} \hat{\sigma}^2_{\mathrm{unb}} &= \frac{n}{n-p}\hat{\sigma}^2_{\mathrm{MLE}} \\ &\overset{(4)}{=} \frac{n}{n-p} \cdot \frac{1}{n}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \\ &= \frac{1}{n-p}(y - X\hat{\beta})^\mathrm{T} V^{-1} (y - X\hat{\beta}) \; . \end{split} \tag{7}$$

■

1.2 R-squared

1.2.1 Definition

Definition: Let there be a linear regression model (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with measured data y, known design matrix X as well as unknown regression coefficients β and noise variance σ^2.
Then, the proportion of the variance of the dependent variable y ("total variance (→ III/1.5.7)") that can be predicted from the independent variables X ("explained variance (→ III/1.5.8)") is called "coefficient of determination", "R-squared" or R^2.

Sources:

- Wikipedia (2020): "Coefficient of determination"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-25; URL: https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship.

1.2.2 Derivation of R² and adjusted R²

Theorem: Given a linear regression model (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with n independent observations and p independent variables,
1) the coefficient of determination (→ IV/1.2.1) is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{2}$$

2) the adjusted coefficient of determination is

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \tag{3}$$

where the residual (→ III/1.5.9) and total sum of squares (→ III/1.5.7) are

$$\begin{split} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \text{TSS} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \;, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \end{split} \tag{4}$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares (→ III/1.5.3) estimates.

Proof: The coefficient of determination (→ IV/1.2.1) R^2 is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data.

1) If we define the explained sum of squares (→ III/1.5.8) as

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \; , \tag{5}$$

then R^2 is given by

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \; . \tag{6}$$

which is equal to

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \; , \tag{7}$$

because (→ III/1.5.10) TSS = ESS + RSS.

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2} \; . \tag{8}$$

If we replace the variance estimates by their unbiased estimators (→ IV/1.1.5), we obtain

$$R^2_{\mathrm{adj}} = 1 - \frac{\frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}/\mathrm{df}_r}{\mathrm{TSS}/\mathrm{df}_t} \tag{9}$$

where $\mathrm{df}_r = n - p$ and $\mathrm{df}_t = n - 1$ are the residual and total degrees of freedom.

This gives the adjusted R^2 which adjusts R^2 for the number of explanatory variables.

■

Sources:

- Wikipedia (2019): "Coefficient of determination"; in: *Wikipedia, the free encyclopedia*, retrieved on 2019-12-06; URL: https://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2.

1.2.3 Relationship to residual variance

Theorem: Given a linear regression model with independent observations (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the coefficient of determination (→ IV/1.2.1) can be expressed in terms of residual variances (→ IV/1.1.1) as

$$R^2 = 1 - \frac{(n-p) \cdot \hat{\sigma}^2}{(n-1) \cdot s^2} \tag{2}$$

where n is the number of observations, p is the number of predictors, $\hat{\sigma}^2$ is an unbiased estimate of the noise variance (→ IV/1.1.5) σ^2 and s^2 is the unbiased sample variance of y.

Proof: The coefficient of determination (→ IV/1.2.2) is given by

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{3}$$

where RSS is the residual sum of squares (→ III/1.5.9)

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{where} \quad \hat{y} = X\hat{\beta} \tag{4}$$

and TSS is the total sum of squares (→ III/1.5.7)

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \; . \tag{5}$$

Note that the residual sum of squares can be written as:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (X\hat{\beta})_i)^2 = (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \; . \tag{6}$$

The unbiased estimate of the noise variance (→ IV/1.1.5) is

$$\hat{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \tag{7}$$

and the unbiased sample variance of the dependent variable is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 \; , \tag{8}$$

Combining (3), (4) and (5), the coefficient of determination can be rewritten as follows:

$$\begin{split}
R^2 &\overset{(3)}{=} 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \\
&\overset{(5)}{=} 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\
&\overset{(6)}{=} 1 - \frac{(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\
&= 1 - \frac{(n-p) \cdot \frac{1}{n-p}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta})}{(n-1) \cdot \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2} \\
&\overset{(7)}{=} 1 - \frac{(n-p) \cdot \hat{\sigma}^2}{(n-1) \cdot \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2} \\
&\overset{(8)}{=} 1 - \frac{(n-p) \cdot \hat{\sigma}^2}{(n-1) \cdot s^2} \; .
\end{split} \tag{9}$$

This completes the proof.

■

1.2.4 Relationship to maximum log-likelihood

Theorem: Given a linear regression model with independent observations (→ III/1.5.1)

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the coefficient of determination (→ IV/1.2.1) can be expressed in terms of the maximum log-likelihood (→ I/4.1.4) as

$$R^2 = 1 - (\exp[\Delta\mathrm{MLL}])^{-2/n} \tag{2}$$

where n is the number of observations and ΔMLL is the difference in maximum log-likelihood between the model given by (1) and a linear regression model with only a constant regressor.

Proof: First, we express the maximum log-likelihood (→ I/4.1.4) (MLL) of a linear regression model in terms of its residual sum of squares (→ III/1.5.9) (RSS). The model in (1) implies the following log-likelihood function (→ I/4.1.2)

$$\mathrm{LL}(\beta, \sigma^2) = \log p(y|\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\mathrm{T}(y - X\beta) \; , \tag{3}$$

such that maximum likelihood estimates are (→ III/1.5.23)

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \tag{5}$$

and the residual sum of squares (→ III/1.5.9) is

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{\varepsilon}_i = \hat{\varepsilon}^\mathrm{T}\hat{\varepsilon} = (y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) = n \cdot \hat{\sigma}^2 \; . \tag{6}$$

Since $\hat{\beta}$ and $\hat{\sigma}^2$ are maximum likelihood estimates (→ I/4.1.3), plugging them into the log-likelihood function gives the maximum log-likelihood:

$$\mathrm{MLL} = \mathrm{LL}(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \; . \tag{7}$$

With (6) for the first $\hat{\sigma}^2$ and (5) for the second $\hat{\sigma}^2$, the MLL becomes

$$\mathrm{MLL} = -\frac{n}{2}\log(\mathrm{RSS}) - \frac{n}{2}\log\left(\frac{2\pi}{n}\right) - \frac{n}{2} \; . \tag{8}$$

Second, we establish the relationship between maximum log-likelihood (MLL) and coefficient of determination (R^2). Consider the two models

$$\begin{aligned} m_0 &: \; X_0 = 1_n \\ m_1 &: \; X_1 = X \end{aligned} \tag{9}$$

For m_1, the residual sum of squares is given by (6); and for m_0, the residual sum of squares is equal to the total sum of squares (→ III/1.5.7):

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \tag{10}$$

Using (8), we can therefore write

$$\Delta\mathrm{MLL} = \mathrm{MLL}(m_1) - \mathrm{MLL}(m_0) = -\frac{n}{2}\log(\mathrm{RSS}) + \frac{n}{2}\log(\mathrm{TSS}) \; . \tag{11}$$

Exponentiating both sides of the equation, we have:

$$
\begin{aligned}
\exp[\Delta\mathrm{MLL}] &= \exp\left[-\frac{n}{2}\log(\mathrm{RSS}) + \frac{n}{2}\log(\mathrm{TSS})\right] \\
&= \left(\exp\left[\log(\mathrm{RSS}) - \log(\mathrm{TSS})\right]\right)^{-n/2} \\
&= \left(\frac{\exp[\log(\mathrm{RSS})]}{\exp[\log(\mathrm{TSS})]}\right)^{-n/2} \\
&= \left(\frac{\mathrm{RSS}}{\mathrm{TSS}}\right)^{-n/2} \; .
\end{aligned} \tag{12}
$$

Taking both sides to the power of $-2/n$ and subtracting from 1, we have

$$
\begin{aligned}
(\exp[\Delta\mathrm{MLL}])^{-2/n} &= \frac{\mathrm{RSS}}{\mathrm{TSS}} \\
1 - (\exp[\Delta\mathrm{MLL}])^{-2/n} &= 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = R^2
\end{aligned} \tag{13}
$$

which proves the identity given above.

■

1.2.5 Statistical significance test for R²

Theorem: Consider a linear regression model (→ III/1.5.1) with known design matrix X, known covariance structure V, unknown regression parameters β and unknown noise variance σ^2:

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; . \tag{1}
$$

Further assume that X contains a constant regressor (→ III/1.5.1). Then, the coefficient of determination (→ IV/1.2.1) can be used to calculate a test statistic (→ I/4.3.5)

$$
F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \tag{2}
$$

where n and p are the dimensions of the design matrix (→ III/1.5.1) X, and this test statistic follows an F-distribution (→ II/3.8.1)

$$
F \sim \mathrm{F}(p-1, n-p) \tag{3}
$$

under the null hypothesis (→ I/4.3.2) that the true coefficient of determination (→ IV/1.2.1) is zero

$$
H_0: \; R^2 = 0 \; . \tag{4}
$$

Proof: Consider two linear regression models (→ III/1.5.1) for the same measured data y, with design matrices $X = X_0 \in \mathbb{R}^{n \times p_0}$ and $X = [X_0, X_1] \in \mathbb{R}^{n \times p}$ as well as regression coefficients $\beta = \beta_0 \in \mathbb{R}^{p_0 \times 1}$ and $\beta = \left[\beta_0^\mathrm{T}, \beta_1^\mathrm{T}\right]^\mathrm{T} \in \mathbb{R}^{p \times 1}$.
Then, under the null hypothesis that all regression coefficients (→ III/1.5.1) β_1 associated with X_1 are zero

$$
H_0: \; \beta_1 = 0_{p-p_0} \quad \Leftrightarrow \quad \beta_i = 0 \quad \text{for all} \quad j = p_0+1, \ldots, p \; , \tag{5}
$$

the omnibus F-statistic follows an F-distribution (→ III/1.5.30)

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n-p)} \sim \mathrm{F}(p - p_0, n - p) \tag{6}$$

where RSS_0 and RSS are the residual sums of squares (→ III/1.5.9) of the null model with X_0 and the full model with X_0 nested in X, after regression coefficients have been estimated with weighted least squares (→ III/1.5.21) or maximum likelihood (→ III/1.5.23).

Since by the requirements of our theorem, X contains a constant regressor, we can assume the following design matrices without loss of generality:

$$X_0 = 1_n \in \mathbb{R}^{n \times 1} \quad \text{and} \quad X = [1_n, X_1] \in \mathbb{R}^{n \times p} \; . \tag{7}$$

Thus, since a single constant regressor estimates the mean and considering the definition of the total sum of squares (→ III/1.5.7) TSS, we in our case have:

$$\mathrm{RSS}_0 = \mathrm{TSS} \quad \text{and} \quad p_0 = 1 \; . \tag{8}$$

The coefficient of determination is given by (→ IV/1.2.2)

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{9}$$

which can also be written as (→ III/1.5.10)

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \; . \tag{10}$$

If all regression coefficients β_1 associated with X_1 are zero, then the true R^2 is zero, because there is no variance explained beyond the constant regressor, the explained sum of squares (→ III/1.5.8) ESS is zero and the residual sum of squares (→ III/1.5.9) RSS is equal to the total sum of squares (→ III/1.5.7) TSS.

Then, by virtue of (6), we get the following F-statistic:

$$\begin{split}
F &\overset{(6)}{=} \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n-p)} \\
&\overset{(8)}{=} \frac{(\mathrm{TSS} - \mathrm{RSS})/(p - 1)}{\mathrm{RSS}/(n-p)} \\
&= \frac{\frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}}/(p-1)}{\frac{\mathrm{RSS}}{\mathrm{TSS}}/(n-p)} \\
&= \frac{\left(1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\right)/(p-1)}{\left(1 - \left(1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\right)\right)/(n-p)} \\
&\overset{(9)}{=} \frac{\left(R^2\right)/(p-1)}{\left(1 - R^2\right)/(n-p)} \; .
\end{split} \tag{11}$$

This means that the null hypothesis (→ I/4.3.2) can be rejected when F as a function of R^2 is as extreme or more extreme than the critical value (→ I/4.3.9) obtained from the F-distribution (→ II/3.8.1) with $p - 1$ denominator and $n - p$ numerator degrees of freedom using a significance level (→ I/4.3.8) α.

■

Sources:
- Alecos Papadopoulos (2014): "What is the distribution of R² in linear regression under the null hypothesis?"; in: *StackExchange CrossValidated*, retrieved on 2024-03-15; URL: https://stats.stackexchange.com/a/130082.

1.3 F-statistic

1.3.1 Definition

Definition: Consider two linear regression models (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$
\begin{split}
m_1 &: \; y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \\
m_0 &: \; y = X_0\beta_0 + \varepsilon_0, \; \varepsilon_{0i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2)
\end{split} \tag{1}
$$

operating on identical measured data y, but with different design matrices $X \in \mathbb{R}^{n \times p}$ and $X_0 \in \mathbb{R}^{n \times p_0}$ and thus different regression coefficients $\beta \in \mathbb{R}^{p \times 1}$ and $\beta_0 \in \mathbb{R}^{p_0 \times 1}$. Furthermore, let the design matrix of the null model be fully contained in the design matrix of the full model:

$$
X = \left[\begin{array}{cc} X_0 & X_1 \end{array} \right] \; . \tag{2}
$$

Then, the F-statistic for model comparison is defined as the ratio of the difference in residual sum of squares (→ III/1.5.9) between the two models, divided by the difference in number of parameters (→ III/1.5.1), to the residual sum of squares (→ III/1.5.9) of the full model, divided by the number of degrees of freedom:

$$
F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n-p)} \; . \tag{3}
$$

Sources:
- Wikipedia (2024): "F-test"; in: *Wikipedia, the free encyclopedia*, retrieved on 2024-03-15; URL: https://en.wikipedia.org/wiki/F-test#Regression_problems.

1.3.2 Relationship to coefficient of determination

Theorem: Let there be a linear regression model (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; . \tag{1}
$$

Then, the F-statistic (→ IV/1.3.1) for comparing this model against a null model containing only a constant regressor (→ III/1.5.1) $x_0 = 1_n$ can be expressed in terms of the coefficient of determination (→ IV/1.2.1)

$$
F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \tag{2}
$$

and vice versa

$$R^2 = 1 - \frac{1}{F \cdot \frac{n-p}{p-1} + 1} \tag{3}$$

where n and p are the dimensions of the design matrix $X \in \mathbb{R}^{n \times p}$.

Proof: Consider two linear regression models (→ III/1.5.1) for the same measured data y, one using design matrix X from (1) and the other with design matrix $X_0 = 1_n \in \mathbb{R}^{n \times 1}$. Then, RSS is the residual sum of squares (→ III/1.5.9) of the model in (1) and the residual sum of squares for the model using X_0 is equal to the total sum of squares (→ III/1.5.7).

1) Thus, the F-statistic (→ IV/1.3.1)

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n-p)} \tag{4}$$

becomes

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/(p-1)}{\mathrm{RSS}/(n-p)} \; . \tag{5}$$

From this, we can derive F in terms of R^2:

$$\begin{split} F &= \frac{(\mathrm{TSS} - \mathrm{RSS})/(p-1)}{\mathrm{RSS}/(n-p)} \\ &= \frac{\frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}}/(p-1)}{\frac{\mathrm{RSS}}{\mathrm{TSS}}/(n-p)} \\ &= \frac{\left(1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\right)/(p-1)}{\left(1 - \left(1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\right)\right)/(n-p)} \\ &= \frac{\left(R^2\right)/(p-1)}{\left(1 - R^2\right)/(n-p)} \; . \end{split} \tag{6}$$

2) Rearranging this equation, we can derive R^2 in terms of F:

$$
\begin{aligned}
F &= \frac{\left(R^2\right)/(p-1)}{\left(1-R^2\right)/(n-p)} \\
F \cdot \frac{n-p}{p-1} &= \frac{R^2}{(1-R^2)} \\
F \cdot \frac{n-p}{p-1} \cdot (1-R^2) &= R^2 \\
F \cdot \frac{n-p}{p-1} - F \cdot \frac{n-p}{p-1} \cdot R^2 &= R^2 \\
F \cdot \frac{n-p}{p-1} \cdot R^2 + R^2 &= F \cdot \frac{n-p}{p-1} \\
R^2 \left(F \cdot \frac{n-p}{p-1} \cdot +1 \right) &= F \cdot \frac{n-p}{p-1} \\
R^2 &= \frac{F \cdot \frac{n-p}{p-1}}{F \cdot \frac{n-p}{p-1} \cdot +1} \\
R^2 &= \frac{F \cdot \frac{n-p}{p-1} + 1 - 1}{F \cdot \frac{n-p}{p-1} \cdot +1} \\
R^2 &= \frac{F \cdot \frac{n-p}{p-1} + 1}{F \cdot \frac{n-p}{p-1} \cdot +1} - \frac{1}{F \cdot \frac{n-p}{p-1} \cdot +1} \\
R^2 &= 1 - \frac{1}{F \cdot \frac{n-p}{p-1} \cdot +1}
\end{aligned} \tag{7}
$$

This completes the proof.

■

Sources:

- Alecos Papadopoulos (2014): "What is the distribution of R² in linear regression under the null hypothesis?"; in: *StackExchange CrossValidated*, retrieved on 2024-03-15; URL: https://stats.stackexchange.com/a/130082.

1.3.3 Relationship to maximum log-likelihood

Theorem: Consider two linear regression models (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$
\begin{aligned}
m_1 &: \; y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \\
m_0 &: \; y = X_0\beta_0 + \varepsilon_0, \; \varepsilon_{0i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2) \; .
\end{aligned} \tag{1}
$$

Then, the F-statistic (→ IV/1.3.1) can be expressed in terms of the maximum log-likelihood (→ I/4.1.4) as

$$
F = \left[\left(\exp[\Delta \text{MLL}] \right)^{2/n} - 1 \right] \cdot \frac{n-p}{p-p_0} \tag{2}
$$

where n, p and p_0 are the dimensions of the design matrices $X = [X_0, X_1] \in \mathbb{R}^{n \times p}$ and $X_0 \in \mathbb{R}^{n \times p_0}$ and ΔMLL is the difference in maximum log-likelihood between the two models given by (1)

Proof: Under the conditions mentioned in the theorem, the F-statistic is defined in terms of the residual sum of squares (→ IV/1.3.1) as

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p-p_0)}{\mathrm{RSS}/(n-p)} \; . \tag{3}$$

We also know that the maximum log-likelihood can be expressed in terms of residual sum of squares (→ III/1.5.24):

$$\mathrm{MLL}(m) = -\frac{n}{2}\log\left(\frac{\mathrm{RSS}}{n}\right) - \frac{n}{2}\left[1+\log(2\pi)\right] \; . \tag{4}$$

Based on this, we see that the difference of the maximum log-likelihoods develops into

$$\begin{split} \Delta\mathrm{MLL} &= \mathrm{MLL}(m_1) - \mathrm{MLL}(m_0) \\ &= \left(-\frac{n}{2}\log\left(\frac{\mathrm{RSS}}{n}\right) - \frac{n}{2}\left[1+\log(2\pi)\right]\right) \\ &- \left(-\frac{n}{2}\log\left(\frac{\mathrm{RSS}_0}{n}\right) - \frac{n}{2}\left[1+\log(2\pi)\right]\right) \\ &= -\frac{n}{2}\log\left(\frac{\mathrm{RSS}}{n}\right) + \frac{n}{2}\log\left(\frac{\mathrm{RSS}_0}{n}\right) \; . \end{split} \tag{5}$$

Finally, we simply perform algebraic operations on both sides (→ IV/1.2.4) until we reach the F-statistic on the right side. We start by exponentiating the MLL difference:

$$
\begin{aligned}
\exp[\Delta\mathrm{MLL}] &= \exp\left[-\frac{n}{2}\log(\mathrm{RSS}/n) + \frac{n}{2}\log(\mathrm{RSS}_0/n)\right] \\
\exp[\Delta\mathrm{MLL}] &= \left(\exp\left[\log(\mathrm{RSS}/n) - \log(\mathrm{RSS}_0/n)\right]\right)^{-n/2} \\
\exp[\Delta\mathrm{MLL}] &= \left(\frac{\exp[\log(\mathrm{RSS}/n)]}{\exp[\log(\mathrm{RSS}_0/n)]}\right)^{-n/2} \\
\exp[\Delta\mathrm{MLL}] &= \left(\frac{\mathrm{RSS}/n}{\mathrm{RSS}_0/n}\right)^{-n/2} \\
\exp[\Delta\mathrm{MLL}] &= \left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}}\right)^{n/2} \\
(\exp[\Delta\mathrm{MLL}])^{2/n} &= \frac{\mathrm{RSS}_0}{\mathrm{RSS}} \\
(\exp[\Delta\mathrm{MLL}])^{2/n} - 1 &= \frac{\mathrm{RSS}_0}{\mathrm{RSS}} - 1 \\
(\exp[\Delta\mathrm{MLL}])^{2/n} - 1 &= \frac{\mathrm{RSS}_0}{\mathrm{RSS}} - \frac{\mathrm{RSS}}{\mathrm{RSS}} \\
\left[(\exp[\Delta\mathrm{MLL}])^{2/n} - 1\right]\cdot\frac{n-p}{p-p_0} &= \left[\frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}}\right]\cdot\frac{n-p}{p-p_0} \\
\left[(\exp[\Delta\mathrm{MLL}])^{2/n} - 1\right]\cdot\frac{n-p}{p-p_0} &= \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p-p_0)}{\mathrm{RSS}/(n-p)} \\
\left[(\exp[\Delta\mathrm{MLL}])^{2/n} - 1\right]\cdot\frac{n-p}{p-p_0} &= F \; .
\end{aligned} \tag{6}
$$

This completes the proof.

■

1.4 Signal-to-noise ratio

1.4.1 Definition

Definition: Let there be a linear regression model (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}
$$

with measured data y, known design matrix X as well as unknown regression coefficients β and noise variance σ^2.

Given estimated regression coefficients (→ III/1.5.23) $\hat{\beta}$ and residual variance (→ IV/1.1.1) $\hat{\sigma}^2$, the signal-to-noise ratio (SNR) is defined as the ratio of estimated signal variance to estimated noise variance:

$$
\mathrm{SNR} = \frac{\mathrm{Var}(X\hat{\beta})}{\hat{\sigma}^2} \; . \tag{2}
$$

Sources:

- Soch J, Allefeld C (2018): "MACS – a new SPM toolbox for model assessment, comparison and selection"; in: *Journal of Neuroscience Methods*, vol. 306, pp. 19-31, eq. 6; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

1.4.2 Relationship to coefficient of determination

Theorem: Let there be a linear regression model (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

and parameter estimates obtained with ordinary least squares (→ III/1.5.3)

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{2}$$

Then, the signal-to noise ratio (→ IV/1.4.1) can be expressed in terms of the coefficient of determination (→ IV/1.2.1)

$$\mathrm{SNR} = \frac{R^2}{1-\mathrm{R}^2} \tag{3}$$

and vice versa

$$R^2 = \frac{\mathrm{SNR}}{1+\mathrm{SNR}} \; , \tag{4}$$

if the predicted signal mean is equal to the actual signal mean.

Proof: The signal-to-noise ratio (→ IV/1.4.1) (SNR) is defined as

$$\mathrm{SNR} = \frac{\mathrm{Var}(X\hat{\beta})}{\hat{\sigma}^2} = \frac{\mathrm{Var}(\hat{y})}{\hat{\sigma}^2} \; . \tag{5}$$

Writing out the sample variances (→ I/1.11.2), we have

$$\mathrm{SNR} = \frac{\frac{1}{n}\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \; . \tag{6}$$

Note that it is irrelevant whether we use the biased estimator of the variance (→ IV/1.1.2) (dividing by n) or the unbiased estimator fo the variance (→ IV/1.1.4) (dividing by $n-1$), because the relevant terms cancel out.

If the predicted signal mean is equal to the actual signal mean – which is the case when variable regressors in X have mean zero, such that they are orthogonal to a constant regressor in X –, this means that $\bar{\hat{y}} = \bar{y}$, such that

$$\mathrm{SNR} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \; . \tag{7}$$

Then, the SNR can be written in terms of the explained (→ III/1.5.8), residual (→ III/1.5.9) and total sum of squares (→ III/1.5.7):

$$\mathrm{SNR} = \frac{\mathrm{ESS}}{\mathrm{RSS}} = \frac{\mathrm{ESS}/\mathrm{TSS}}{\mathrm{RSS}/\mathrm{TSS}} \; . \tag{8}$$

With the derivation of the coefficient of determination (→ IV/1.2.2), this becomes

$$\mathrm{SNR} = \frac{R^2}{1-R^2} \; . \tag{9}$$

Rearranging this equation for the coefficient of determination (→ IV/1.2.1), we have

$$R^2 = \frac{\mathrm{SNR}}{1+\mathrm{SNR}} \; , \tag{10}$$

■

1.4.3 Relationship to maximum log-likelihood

Theorem: Given a linear regression model (→ III/1.5.1) with independent (→ I/1.3.6) observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; ; \tag{1}$$

the signal-to-noise ratio (→ IV/1.4.1) can be expressed in terms of the maximum log-likelihood (→ I/4.1.4) as

$$\mathrm{SNR} = \left(\exp[\Delta\mathrm{MLL}]\right)^{2/n} - 1 \; , \tag{2}$$

where n is the number of observations and ΔMLL is the difference in maximum log-likelihood between the model given by (1) and a linear regression model with only a constant regressor.
This holds, if the predicted signal mean is equal to the actual signal mean

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}(X\hat{\beta})_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{3}$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares (→ III/1.5.3) estimates.

Proof: Under the conditions mentioned in the theorem, the signal-to-noise ratio can be expressed in terms of the coefficient of determination (→ IV/1.4.2) as

$$\mathrm{SNR} = \frac{R^2}{1-\mathrm{R}^2} \tag{4}$$

and R-squared can be expressed in terms of maximum likelihood (→ IV/1.2.4) as

$$R^2 = 1 - \left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n} \; . \tag{5}$$

Plugging (5) into (4), we obtain:

$$\begin{split} \mathrm{SNR} &= \frac{1 - \left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n}}{\left(\exp[\Delta\mathrm{MLL}]\right)^{-2/\mathrm{n}}} \\ &= \frac{1}{\left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n}} - \frac{\left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n}}{\left(\exp[\Delta\mathrm{MLL}]\right)^{-2/\mathrm{n}}} \\ &= \left(\exp[\Delta\mathrm{MLL}]\right)^{2/n} - 1 \; . \end{split} \tag{6}$$

■

2 Bayesian model selection

Chapter V

Appendix

1 Proof by Number

P22	bern-mean	Mean of the Bernoulli distribution	JoramSoch	2020-01-16	150
P23	bin-mean	Mean of the binomial distribution	JoramSoch	2020-01-16	156
P24	cat-mean	Mean of the categorical distribution	JoramSoch	2020-01-16	169
P25	mult-mean	Mean of the multinomial distribution	JoramSoch	2020-01-16	172
P26	matn-mvn	Equivalence of matrix-normal distribution and multivariate normal distribution	JoramSoch	2020-01-20	326
P27	poiss-mle	Maximum likelihood estimation for Poisson-distributed data	JoramSoch	2020-01-20	572
P28	beta-mome	Method of moments for beta-distributed data	JoramSoch	2020-01-22	585
P29	bin-prior	Conjugate prior distribution for binomial observations	JoramSoch	2020-01-23	554
P30	bin-post	Posterior distribution for binomial observations	JoramSoch	2020-01-24	555
P31	bin-lme	Log model evidence for binomial observations	JoramSoch	2020-01-24	556
P33	norm-pdf	Probability density function of the normal distribution	JoramSoch	2020-01-27	194
P34	mvn-pdf	Probability density function of the multivariate normal distribution	JoramSoch	2020-01-27	290
P35	mvn-marg	Marginal distributions of the multivariate normal distribution	JoramSoch	2020-01-29	298
P36	ng-marg	Marginal distributions of the normal-gamma distribution	JoramSoch	2020-01-29	316
P37	cuni-pdf	Probability density function of the continuous uniform distribution	JoramSoch	2020-01-31	176
P38	cuni-cdf	Cumulative distribution function of the continuous uniform distribution	JoramSoch	2020-01-02	177
P39	cuni-qf	Quantile function of the continuous uniform distribution	JoramSoch	2020-01-02	178
P40	mlr-ols2	Ordinary least squares for multiple linear regression	JoramSoch	2020-02-03	453
P41	poissexp-prior	Conjugate prior distribution for the Poisson distribution with exposure values	JoramSoch	2020-02-04	580

P42	poissexp-post	Posterior distribution for the Poisson distribution with exposure values	JoramSoch	2020-02-04	581
P43	poissexp-lme	Log model evidence for the Poisson distribution with exposure values	JoramSoch	2020-02-04	583
P44	ng-pdf	Probability density function of the normal-gamma distribution	JoramSoch	2020-02-07	309
P45	gam-pdf	Probability density function of the gamma distribution	JoramSoch	2020-02-08	223
P46	exp-pdf	Probability density function of the exponential distribution	JoramSoch	2020-02-08	234
P47	exp-mean	Mean of the exponential distribution	JoramSoch	2020-02-10	237
P48	exp-cdf	Cumulative distribution function of the exponential distribution	JoramSoch	2020-02-11	235
P49	exp-med	Median of the exponential distribution	JoramSoch	2020-02-11	238
P50	exp-qf	Quantile function of the exponential distribution	JoramSoch	2020-02-12	236
P51	exp-mode	Mode of the exponential distribution	kantundpeterpan	2020-02-12	238
P52	mean-nonneg	Non-negativity of the expected value	JoramSoch	2020-02-13	42
P53	mean-lin	Linearity of the expected value	JoramSoch	2020-02-13	43
P54	mean-mono	Monotonicity of the expected value	JoramSoch	2020-02-17	45
P55	mean-mult	(Non-)Multiplicativity of the expected value	JoramSoch	2020-02-17	45
P56	ci-wilks	Construction of confidence intervals using Wilks' theorem	JoramSoch	2020-02-19	114
P57	ent-nonneg	Non-negativity of the Shannon entropy	JoramSoch	2020-02-19	85
P58	cmi-mcde	Relation of continuous mutual information to marginal and conditional differential entropy	JoramSoch	2020-02-21	101
P59	cmi-mjde	Relation of continuous mutual information to marginal and joint differential entropy	JoramSoch	2020-02-21	103

P60	cmi-jcde	Relation of continuous mutual information to joint and conditional differential entropy	JoramSoch	2020-02-21	104
P61	resvar-bias	Maximum likelihood estimator of variance is biased	JoramSoch	2020-02-24	598
P62	resvar-unb	Construction of unbiased estimator for variance	JoramSoch	2020-02-25	602
P63	snr-rsq	Relationship between signal-to-noise ratio and R^2	JoramSoch	2020-02-26	615
P68	dent-neg	Differential entropy can be negative	JoramSoch	2020-03-02	91
P69	exp-gam	Exponential distribution is a special case of gamma distribution	JoramSoch	2020-03-02	233
P70	matn-pdf	Probability density function of the matrix-normal distribution	JoramSoch	2020-03-02	327
P71	norm-mgf	Moment-generating function of the normal distribution	JoramSoch	2020-03-03	195
P72	logreg-lonp	Log-odds and probability in logistic regression	JoramSoch	2020-03-03	594
P75	mlr-mat	Transformation matrices for ordinary least squares	JoramSoch	2020-03-09	460
P76	mlr-pss	Partition of sums of squares for multiple linear regression	JoramSoch	2020-03-09	458
P77	mlr-wls	Weighted least squares for multiple linear regression	JoramSoch	2020-03-11	469
P78	mlr-mle	Maximum likelihood estimation for multiple linear regression	JoramSoch	2020-03-11	471
P79	mult-prior	Conjugate prior distribution for multinomial observations	JoramSoch	2020-03-11	564
P80	mult-post	Posterior distribution for multinomial observations	JoramSoch	2020-03-11	565
P81	mult-lme	Log model evidence for multinomial observations	JoramSoch	2020-03-11	566
P82	cuni-mean	Mean of the continuous uniform distribution	JoramSoch	2020-03-16	179
P83	cuni-med	Median of the continuous uniform distribution	JoramSoch	2020-03-16	180

P84	cuni-med	Mode of the continuous uniform distribution	JoramSoch	2020-03-16	180
P85	norm-cdf	Cumulative distribution function of the normal distribution	JoramSoch	2020-03-20	196
P86	norm-cdfwerf	Expression of the cumulative distribution function of the normal distribution without the error function	JoramSoch	2020-03-20	198
P87	norm-qf	Quantile function of the normal distribution	JoramSoch	2020-03-20	201
P88	mvn-cond	Conditional distributions of the multivariate normal distribution	JoramSoch	2020-03-20	299
P89	jl-lfnprior	Joint likelihood is the product of likelihood function and prior density	JoramSoch	2020-05-05	126
P90	post-jl	Posterior density is proportional to joint likelihood	JoramSoch	2020-05-05	127
P91	ml-jl	Marginal likelihood is a definite integral of joint likelihood	JoramSoch	2020-05-05	128
P92	mvn-kl	Kullback-Leibler divergence for the multivariate normal distribution	JoramSoch	2020-05-05	295
P93	gam-kl	Kullback-Leibler divergence for the gamma distribution	JoramSoch	2020-05-05	232
P94	beta-pdf	Probability density function of the beta distribution	JoramSoch	2020-05-05	260
P95	dir-pdf	Probability density function of the Dirichlet distribution	JoramSoch	2020-05-05	321
P96	bern-pmf	Probability mass function of the Bernoulli distribution	JoramSoch	2020-05-11	150
P97	bin-pmf	Probability mass function of the binomial distribution	JoramSoch	2020-05-11	154
P98	cat-pmf	Probability mass function of the categorical distribution	JoramSoch	2020-05-11	169
P99	mult-pmf	Probability mass function of the multinomial distribution	JoramSoch	2020-05-11	171
P100	mvn-dent	Differential entropy of the multivariate normal distribution	JoramSoch	2020-05-14	294
P101	norm-dent	Differential entropy of the normal distribution	JoramSoch	2020-05-14	209

P102	poiss-pmf	Probability mass function of the Poisson distribution	JoramSoch	2020-05-14	166
P103	mean-nnrvar	Expected value of a non-negative random variable	JoramSoch	2020-05-18	41
P104	var-mean	Partition of variance into expected values	JoramSoch	2020-05-19	55
P105	logreg-pnlo	Probability and log-odds in logistic regression	JoramSoch	2020-05-19	593
P106	glm-ols	Ordinary least squares for the general linear model	JoramSoch	2020-05-19	524
P107	glm-wls	Weighted least squares for the general linear model	JoramSoch	2020-05-19	525
P108	gam-mean	Mean of the gamma distribution	JoramSoch	2020-05-19	226
P109	gam-var	Variance of the gamma distribution	JoramSoch	2020-05-19	227
P110	gam-logmean	Logarithmic expectation of the gamma distribution	JoramSoch	2020-05-25	228
P111	norm-snorm	Relationship between normal distribution and standard normal distribution	JoramSoch	2020-05-26	186
P112	gam-sgam	Relationship between gamma distribution and standard gamma distribution	JoramSoch	2020-05-26	220
P113	kl-ent	Relation of discrete Kullback-Leibler divergence to Shannon entropy	JoramSoch	2020-05-27	111
P114	kl-dent	Relation of continuous Kullback-Leibler divergence to differential entropy	JoramSoch	2020-05-27	111
P115	kl-inv	Invariance of the Kullback-Leibler divergence under parameter transformation	JoramSoch	2020-05-28	110
P116	kl-add	Additivity of the Kullback-Leibler divergence for independent distributions	JoramSoch	2020-05-31	109
P117	kl-nonneg	Non-negativity of the Kullback-Leibler divergence	JoramSoch	2020-05-31	105
P118	cov-mean	Partition of covariance into expected values	JoramSoch	2020-06-02	63

P119	cov-corr	Relationship between covariance and correlation	JoramSoch	2020-06-02	65
P120	covmat-mean	Partition of a covariance matrix into expected values	JoramSoch	2020-06-06	67
P121	covmat-corrmat	Relationship between covariance matrix and correlation matrix	JoramSoch	2020-06-06	71
P122	precmat-corrmat	Relationship between precision matrix and correlation matrix	JoramSoch	2020-06-06	73
P123	var-nonneg	Non-negativity of the variance	JoramSoch	2020-06-06	55
P124	var-const	Variance of constant is zero	JoramSoch	2020-06-27	56
P126	var-inv	Invariance of the variance under addition of a constant	JoramSoch	2020-07-07	57
P127	var-scal	Scaling of the variance upon multiplication with a constant	JoramSoch	2020-07-07	57
P128	var-sum	Variance of the sum of two random variables	JoramSoch	2020-07-07	58
P129	var-lincomb	Variance of the linear combination of two random variables	JoramSoch	2020-07-07	58
P130	var-add	Additivity of the variance for independent random variables	JoramSoch	2020-07-07	59
P131	mean-qf	Expectation of a quadratic form	JoramSoch	2020-07-13	48
P133	blr-pp	Posterior probability of the alternative hypothesis for Bayesian linear regression	JoramSoch	2020-07-17	506
P134	blr-pcr	Posterior credibility region against the omnibus null hypothesis for Bayesian linear regression	JoramSoch	2020-07-17	508
P135	mlr-idem	Projection matrix and residual-forming matrix are idempotent	JoramSoch	2020-07-22	463
P136	mlr-wls2	Weighted least squares for multiple linear regression	JoramSoch	2020-07-22	470
P138	mean-lotus	Law of the unconscious statistician	JoramSoch	2020-07-22	51
P140	duni-pmf	Probability mass function of the discrete uniform distribution	JoramSoch	2020-07-28	144
P141	duni-cdf	Cumulative distribution function of the discrete uniform distribution	JoramSoch	2020-07-28	144

P142	duni-qf	Quantile function of the discrete uniform distribution	JoramSoch	2020-07-28	146
P144	matn-trans	Transposition of a matrix-normal random variable	JoramSoch	2020-08-03	331
P145	matn-ltt	Linear transformation theorem for the matrix-normal distribution	JoramSoch	2020-08-03	332
P146	ng-cond	Conditional distributions of the normal-gamma distribution	JoramSoch	2020-08-05	318
P147	kl-nonsymm	Non-symmetry of the Kullback-Leibler divergence	JoramSoch	2020-08-11	106
P148	kl-conv	Convexity of the Kullback-Leibler divergence	JoramSoch	2020-08-11	108
P149	ent-conc	Concavity of the Shannon entropy	JoramSoch	2020-08-11	86
P150	entcross-conv	Convexity of the cross-entropy	JoramSoch	2020-08-11	88
P151	poiss-mean	Mean of the Poisson distribution	JoramSoch	2020-08-19	166
P152	norm-fwhm	Full width at half maximum for the normal distribution	JoramSoch	2020-08-19	206
P153	mom-mgf	Moment in terms of moment-generating function	JoramSoch	2020-08-19	80
P154	mgf-ltt	Linear transformation theorem for the moment-generating function	JoramSoch	2020-08-19	36
P155	mgf-lincomb	Moment-generating function of linear combination of independent random variables	JoramSoch	2020-08-19	37
P158	cov-ind	Covariance of independent random variables	JoramSoch	2020-09-03	64
P159	mblr-prior	Conjugate prior distribution for multivariate Bayesian linear regression	JoramSoch	2020-09-03	543
P160	mblr-post	Posterior distribution for multivariate Bayesian linear regression	JoramSoch	2020-09-03	545
P161	mblr-lme	Log model evidence for multivariate Bayesian linear regression	JoramSoch	2020-09-03	547
P162	wald-pdf	Probability density function of the Wald distribution	tomfaulkenberry	2020-09-04	266
P164	gibbs-ineq	Gibbs' inequality	JoramSoch	2020-09-09	89

P165	logsum-ineq	Log sum inequality	JoramSoch	2020-09-09	90
P166	kl-nonneg2	Non-negativity of the Kullback-Leibler divergence	JoramSoch	2020-09-09	106
P167	momcent-1st	First central moment is zero	JoramSoch	2020-09-09	83
P168	wald-mgf	Moment-generating function of the Wald distribution	tomfaulkenberry	2020-09-13	266
P169	wald-mean	Mean of the Wald distribution	tomfaulkenberry	2020-09-13	268
P170	wald-var	Variance of the Wald distribution	tomfaulkenberry	2020-09-13	269
P171	momraw-1st	First raw moment is mean	JoramSoch	2020-10-08	82
P172	momraw-2nd	Relationship between second raw moment, variance and mean	JoramSoch	2020-10-08	82
P173	momcent-2nd	Second central moment is variance	JoramSoch	2020-10-08	83
P174	chi2-gam	Chi-squared distribution is a special case of gamma distribution	kjpetrykowski	2020-10-12	252
P175	chi2-mom	Moments of the chi-squared distribution	kjpetrykowski	2020-10-13	254
P176	norm-snorm2	Relationship between normal distribution and standard normal distribution	JoramSoch	2020-10-15	188
P177	gam-sgam2	Relationship between gamma distribution and standard gamma distribution	JoramSoch	2020-10-15	221
P178	gam-cdf	Cumulative distribution function of the gamma distribution	JoramSoch	2020-10-15	224
P179	gam-xlogx	Expected value of x times ln(x) for a gamma distribution	JoramSoch	2020-10-15	230
P180	norm-snorm3	Relationship between normal distribution and standard normal distribution	JoramSoch	2020-10-22	188
P181	dir-ep	Exceedance probabilities for the Dirichlet distribution	JoramSoch	2020-10-22	323
P182	dir-mle	Maximum likelihood estimation for Dirichlet-distributed data	JoramSoch	2020-10-22	587

P183	cdf-sifct	Cumulative distribution function of a strictly increasing function of a random variable	JoramSoch	2020-10-29	29
P184	pmf-sifct	Probability mass function of a strictly increasing function of a discrete random variable	JoramSoch	2020-10-29	19
P185	pdf-sifct	Probability density function of a strictly increasing function of a continuous random variable	JoramSoch	2020-10-29	22
P186	cdf-sdfct	Cumulative distribution function of a strictly decreasing function of a random variable	JoramSoch	2020-11-06	29
P187	pmf-sdfct	Probability mass function of a strictly decreasing function of a discrete random variable	JoramSoch	2020-11-06	19
P188	pdf-sdfct	Probability density function of a strictly decreasing function of a continuous random variable	JoramSoch	2020-11-06	23
P189	cdf-pmf	Cumulative distribution function in terms of probability mass function of a discrete random variable	JoramSoch	2020-11-12	30
P190	cdf-pdf	Cumulative distribution function in terms of probability density function of a continuous random variable	JoramSoch	2020-11-12	31
P191	pdf-cdf	Probability density function is first derivative of cumulative distribution function	JoramSoch	2020-11-12	27
P192	qf-cdf	Quantile function is inverse of strictly monotonically increasing cumulative distribution function	JoramSoch	2020-11-12	34
P193	norm-kl	Kullback-Leibler divergence for the normal distribution	JoramSoch	2020-11-19	210
P194	gam-qf	Quantile function of the gamma distribution	JoramSoch	2020-11-19	225
P195	beta-cdf	Cumulative distribution function of the beta distribution	JoramSoch	2020-11-19	262
P196	norm-gi	Gaussian integral	JoramSoch	2020-11-25	193
P197	chi2-pdf	Probability density function of the chi-squared distribution	JoramSoch	2020-11-25	253

P198	beta-mgf	Moment-generating function of the beta distribution	JoramSoch	2020-11-25	261
P199	dent-inv	Invariance of the differential entropy under addition of a constant	JoramSoch	2020-12-02	92
P200	dent-add	Addition of the differential entropy upon multiplication with a constant	JoramSoch	2020-12-02	92
P201	ug-prior	Conjugate prior distribution for the univariate Gaussian	JoramSoch	2021-03-03	351
P202	ug-post	Posterior distribution for the univariate Gaussian	JoramSoch	2021-03-03	353
P203	ug-lme	Log model evidence for the univariate Gaussian	JoramSoch	2021-03-03	356
P204	ug-ttest1	One-sample t-test for independent observations	JoramSoch	2021-03-12	346
P205	ug-ttest2	Two-sample t-test for independent observations	JoramSoch	2021-03-12	347
P206	ug-ttestp	Paired t-test for dependent observations	JoramSoch	2021-03-12	349
P207	ugkv-mle	Maximum likelihood estimation for the univariate Gaussian with known variance	JoramSoch	2021-03-24	360
P208	ugkv-ztest1	One-sample z-test for independent observations	JoramSoch	2021-03-24	361
P209	ugkv-ztest2	Two-sample z-test for independent observations	JoramSoch	2021-03-24	362
P210	ugkv-ztestp	Paired z-test for dependent observations	JoramSoch	2021-03-24	364
P211	ugkv-prior	Conjugate prior distribution for the univariate Gaussian with known variance	JoramSoch	2021-03-24	364
P212	ugkv-post	Posterior distribution for the univariate Gaussian with known variance	JoramSoch	2021-03-24	366
P213	ugkv-lme	Log model evidence for the univariate Gaussian with known variance	JoramSoch	2021-03-24	369
P214	ugkv-anc	Accuracy and complexity for the univariate Gaussian with known variance	JoramSoch	2021-03-24	370

P215	ugkv-lbf	Log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	371
P216	ugkv-lbfmean	Expectation of the log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	373
P217	ugkv-cvlme	Cross-validated log model evidence for the univariate Gaussian with known variance	JoramSoch	2021-03-24	374
P218	ugkv-cvlbf	Cross-validated log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	377
P219	ugkv-cvlbfmean	Expectation of the cross-validated log Bayes factor for the univariate Gaussian with known variance	JoramSoch	2021-03-24	378
P220	cdf-pit	Probability integral transform using cumulative distribution function	JoramSoch	2021-04-07	32
P221	cdf-itm	Inverse transformation method using cumulative distribution function	JoramSoch	2021-04-07	32
P222	cdf-dt	Distributional transformation using cumulative distribution function	JoramSoch	2021-04-07	33
P223	ug-mle	Maximum likelihood estimation for the univariate Gaussian	JoramSoch	2021-04-16	344
P224	poissexp-mle	Maximum likelihood estimation for the Poisson distribution with exposure values	JoramSoch	2021-04-16	578
P225	poiss-prior	Conjugate prior distribution for Poisson-distributed data	JoramSoch	2020-04-21	573
P226	poiss-post	Posterior distribution for Poisson-distributed data	JoramSoch	2020-04-21	574
P227	poiss-lme	Log model evidence for Poisson-distributed data	JoramSoch	2020-04-21	576
P228	beta-mean	Mean of the beta distribution	JoramSoch	2021-04-29	263
P229	beta-var	Variance of the beta distribution	JoramSoch	2021-04-29	264
P230	poiss-var	Variance of the Poisson distribution	JoramSoch	2021-04-29	167
P231	mvt-f	Relationship between multivariate t-distribution and F-distribution	JoramSoch	2021-05-04	306

P232	nst-t	Relationship between non-standardized t-distribution and t-distribution	JoramSoch	2021-05-11	216
P233	norm-chi2	Relationship between normal distribution and chi-squared distribution	JoramSoch	2021-05-20	189
P234	norm-t	Relationship between normal distribution and t-distribution	JoramSoch	2021-05-27	191
P235	norm-lincomb	Linear combination of independent normal random variables	JoramSoch	2021-06-02	213
P236	mvn-ind	Necessary and sufficient condition for independence of multivariate normal random variables	JoramSoch	2021-06-02	302
P237	ng-mean	Mean of the normal-gamma distribution	JoramSoch	2021-07-08	310
P238	ng-dent	Differential entropy of the normal-gamma distribution	JoramSoch	2021-07-08	313
P239	gam-dent	Differential entropy of the gamma distribution	JoramSoch	2021-07-14	231
P240	ug-anc	Accuracy and complexity for the univariate Gaussian	JoramSoch	2021-07-14	358
P241	prob-ind	Probability under statistical independence	JoramSoch	2021-07-23	8
P242	prob-exc	Probability under mutual exclusivity	JoramSoch	2021-07-23	9
P243	prob-mon	Monotonicity of probability	JoramSoch	2021-07-30	10
P244	prob-emp	Probability of the empty set	JoramSoch	2021-07-30	11
P245	prob-comp	Probability of the complement	JoramSoch	2021-07-30	12
P246	prob-range	Range of probability	JoramSoch	2021-07-30	12
P247	prob-add	Addition law of probability	JoramSoch	2021-07-30	13
P248	prob-tot	Law of total probability	JoramSoch	2021-08-08	14
P249	prob-exh	Probability of exhaustive events	JoramSoch	2021-08-08	14
P250	norm-maxent	Normal distribution maximizes differential entropy for fixed variance	JoramSoch	2020-08-25	212
P251	norm-extr	Extreme points of the probability density function of the normal distribution	JoramSoch	2020-08-25	207

P252	norm-infl	Inflection points of the probability density function of the normal distribution	JoramSoch	2020-08-26	208
P253	pmf-invfct	Probability mass function of an invertible function of a random vector	JoramSoch	2021-08-30	20
P254	pdf-invfct	Probability density function of an invertible function of a continuous random vector	JoramSoch	2021-08-30	24
P255	pdf-linfct	Probability density function of a linear function of a continuous random vector	JoramSoch	2021-08-30	26
P256	cdf-sumind	Cumulative distribution function of a sum of independent random variables	JoramSoch	2021-08-30	28
P257	pmf-sumind	Probability mass function of a sum of independent discrete random variables	JoramSoch	2021-08-30	18
P258	pdf-sumind	Probability density function of a sum of independent discrete random variables	JoramSoch	2021-08-30	21
P259	cf-fct	Characteristic function of a function of a random variable	JoramSoch	2021-09-22	35
P260	mgf-fct	Moment-generating function of a function of a random variable	JoramSoch	2021-09-22	36
P261	dent-addvec	Addition of the differential entropy upon multiplication with invertible matrix	JoramSoch	2021-10-07	94
P262	dent-noninv	Non-invariance of the differential entropy under change of variables	JoramSoch	2021-10-07	95
P263	t-pdf	Probability density function of the t-distribution	JoramSoch	2021-10-12	217
P264	f-pdf	Probability density function of the F-distribution	JoramSoch	2021-10-12	256
P265	tglm-dist	Distribution of the transformed general linear model	JoramSoch	2021-10-21	534
P266	tglm-para	Equivalence of parameter estimates from the transformed general linear model	JoramSoch	2021-10-21	535

P267	iglm-dist	Distribution of the inverse general linear model	JoramSoch	2021-10-21	536
P268	iglm-blue	Best linear unbiased estimator for the inverse general linear model	JoramSoch	2021-10-21	537
P269	cfm-para	Parameters of the corresponding forward model	JoramSoch	2021-10-21	541
P270	cfm-exist	Existence of a corresponding forward model	JoramSoch	2021-10-21	542
P271	slr-ols	Ordinary least squares for simple linear regression	JoramSoch	2021-10-27	412
P272	slr-olsmean	Expectation of parameter estimates for simple linear regression	JoramSoch	2021-10-27	416
P273	slr-olsvar	Variance of parameter estimates for simple linear regression	JoramSoch	2021-10-27	418
P274	slr-meancent	Effects of mean-centering on parameter estimates for simple linear regression	JoramSoch	2021-10-27	423
P275	slr-comp	The regression line goes through the center of mass point	JoramSoch	2021-10-27	425
P276	slr-ressum	The sum of residuals is zero in simple linear regression	JoramSoch	2021-10-27	446
P277	slr-rescorr	The residuals and the covariate are uncorrelated in simple linear regression	JoramSoch	2021-10-27	447
P278	slr-resvar	Relationship between residual variance and sample variance in simple linear regression	JoramSoch	2021-10-27	448
P279	slr-corr	Relationship between correlation coefficient and slope estimate in simple linear regression	JoramSoch	2021-10-27	449
P280	slr-rsq	Relationship between coefficient of determination and correlation coefficient in simple linear regression	JoramSoch	2021-10-27	450
P281	slr-mlr	Simple linear regression is a special case of multiple linear regression	JoramSoch	2021-11-09	411
P282	slr-olsdist	Distribution of parameter estimates for simple linear regression	JoramSoch	2021-11-09	420

P283	slr-proj	Projection of a data point to the regression line	JoramSoch	2021-11-09	425
P284	slr-sss	Sums of squares for simple linear regression	JoramSoch	2021-11-09	427
P285	slr-mat	Transformation matrices for simple linear regression	JoramSoch	2021-11-09	432
P286	slr-wls	Weighted least squares for simple linear regression	JoramSoch	2021-11-16	434
P287	slr-mle	Maximum likelihood estimation for simple linear regression	JoramSoch	2021-11-16	437
P288	slr-ols2	Ordinary least squares for simple linear regression	JoramSoch	2021-11-16	414
P289	slr-wls2	Weighted least squares for simple linear regression	JoramSoch	2021-11-16	436
P290	slr-mle2	Maximum likelihood estimation for simple linear regression	JoramSoch	2021-11-16	439
P291	mean-tot	Law of total expectation	JoramSoch	2021-11-26	51
P292	var-tot	Law of total variance	JoramSoch	2021-11-26	59
P293	cov-tot	Law of total covariance	JoramSoch	2021-11-26	65
P294	dir-kl	Kullback-Leibler divergence for the Dirichlet distribution	JoramSoch	2021-12-02	321
P295	wish-kl	Kullback-Leibler divergence for the Wishart distribution	JoramSoch	2021-12-02	336
P296	matn-kl	Kullback-Leibler divergence for the matrix-normal distribution	JoramSoch	2021-12-02	330
P297	matn-samp	Sampling from the matrix-normal distribution	JoramSoch	2021-12-07	335
P298	mean-tr	Expected value of the trace of a matrix	JoramSoch	2021-12-07	47
P299	corr-z	Correlation coefficient in terms of standard scores	JoramSoch	2021-12-14	75
P300	corr-range	Correlation always falls between -1 and +1	JoramSoch	2021-12-14	74
P301	bern-var	Variance of the Bernoulli distribution	JoramSoch	2022-01-20	151

P302	bin-var	Variance of the binomial distribution	JoramSoch	2022-01-20	156
P303	bern-varrange	Range of the variance of the Bernoulli distribution	JoramSoch	2022-01-27	151
P304	bin-varrange	Range of the variance of the binomial distribution	JoramSoch	2022-01-27	157
P305	mlr-mll	Maximum log-likelihood for multiple linear regression	JoramSoch	2022-02-04	473
P306	lognorm-med	Median of the log-normal distribution	majapavlo	2022-02-07	248
P307	mlr-aic	Akaike information criterion for multiple linear regression	JoramSoch	2022-02-11	489
P308	mlr-bic	Bayesian information criterion for multiple linear regression	JoramSoch	2022-02-11	490
P309	mlr-aicc	Corrected Akaike information criterion for multiple linear regression	JoramSoch	2022-02-11	491
P310	lognorm-pdf	Probability density function of the log-normal distribution	majapavlo	2022-02-13	243
P311	lognorm-mode	Mode of the log-normal distribution	majapavlo	2022-02-13	248
P312	mlr-dev	Deviance for multiple linear regression	JoramSoch	2022-03-01	488
P313	blr-dic	Deviance information criterion for multiple linear regression	JoramSoch	2022-03-01	501
P317	mle-bias	Maximum likelihood estimation can result in biased estimates	JoramSoch	2022-03-18	116
P318	pval-h0	The p-value follows a uniform distribution under the null hypothesis	JoramSoch	2022-03-18	123
P319	prob-exh2	Probability of exhaustive events	JoramSoch	2022-03-27	15
P320	slr-olscorr	Parameter estimates for simple linear regression are uncorrelated after mean-centering	JoramSoch	2022-04-14	422
P321	norm-probstd	Probability of normal random variable being within standard deviations from its mean	JoramSoch	2022-05-08	200
P322	mult-cov	Covariance matrix of the multinomial distribution	adkipnis	2022-05-11	173

P323	nw-pdf	Probability density function of the normal-Wishart distribution	JoramSoch	2022-05-14	338
P324	ng-nw	Normal-gamma distribution is a special case of normal-Wishart distribution	JoramSoch	2022-05-20	308
P325	lognorm-cdf	Cumulative distribution function of the log-normal distribution	majapavlo	2022-06-29	244
P326	lognorm-qf	Quantile function of the log-normal distribution	majapavlo	2022-07-09	245
P327	nw-mean	Mean of the normal-Wishart distribution	JoramSoch	2022-07-14	340
P328	gam-wish	Gamma distribution is a special case of Wishart distribution	JoramSoch	2022-07-14	219
P329	mlr-glm	Multiple linear regression is a special case of the general linear model	JoramSoch	2022-07-21	452
P330	mvn-matn	Multivariate normal distribution is a special case of matrix-normal distribution	JoramSoch	2022-07-31	285
P331	norm-mvn	Normal distribution is a special case of multivariate normal distribution	JoramSoch	2022-08-19	186
P332	t-mvt	t-distribution is a special case of multivariate t-distribution	JoramSoch	2022-08-25	215
P333	mvt-pdf	Probability density function of the multivariate t-distribution	JoramSoch	2022-09-02	305
P334	bern-ent	Entropy of the Bernoulli distribution	JoramSoch	2022-09-02	152
P335	bin-ent	Entropy of the binomial distribution	JoramSoch	2022-09-02	158
P336	cat-ent	Entropy of the categorical distribution	JoramSoch	2022-09-09	170
P337	mult-ent	Entropy of the multinomial distribution	JoramSoch	2022-09-09	174
P338	cat-cov	Covariance matrix of the categorical distribution	JoramSoch	2022-09-09	170
P339	mvn-mean	Mean of the multivariate normal distribution	JoramSoch	2022-09-15	292
P340	mvn-cov	Covariance matrix of the multivariate normal distribution	JoramSoch	2022-09-15	293

P341	matn-mean	Mean of the matrix-normal distribution	JoramSoch	2022-09-15	328
P342	matn-cov	Covariance matrices of the matrix-normal distribution	JoramSoch	2022-09-15	328
P343	matn-marg	Marginal distributions for the matrix-normal distribution	JoramSoch	2022-09-15	333
P344	matn-dent	Differential entropy for the matrix-normal distribution	JoramSoch	2022-09-22	329
P345	ng-cov	Covariance and variance of the normal-gamma distribution	JoramSoch	2022-09-22	311
P346	ng-samp	Sampling from the normal-gamma distribution	JoramSoch	2022-09-22	319
P347	covmat-inv	Invariance of the covariance matrix under addition of constant vector	JoramSoch	2022-09-22	69
P348	covmat-scal	Scaling of the covariance matrix upon multiplication with constant matrix	JoramSoch	2022-09-22	70
P349	covmat-sum	Covariance matrix of the sum of two random vectors	JoramSoch	2022-09-26	71
P350	covmat-symm	Symmetry of the covariance matrix	JoramSoch	2022-09-26	67
P351	covmat-psd	Positive semi-definiteness of the covariance matrix	JoramSoch	2022-09-26	68
P352	cov-var	Self-covariance equals variance	JoramSoch	2022-09-26	64
P353	cov-symm	Symmetry of the covariance	JoramSoch	2022-09-26	63
P354	lognorm-mean	Mean of the log-normal distribution	majapavlo	2022-10-02	246
P355	lognorm-var	Variance of the log-normal distribution	majapavlo	2022-10-02	250
P356	beta-chi2	Relationship between chi-squared distribution and beta distribution	JoramSoch	2022-10-07	258
P357	betabin-mome	Method of moments for beta-binomial data	JoramSoch	2022-10-07	590
P358	bin-margcond	Marginal distribution of a conditional binomial distribution	JoramSoch	2022-10-07	160

P359	mean-prodsqr	Square of expectation of product is less than or equal to product of expectation of squares	JoramSoch	2022-10-11	49
P360	pgf-mean	Probability-generating function is expectation of function of random variable	JoramSoch	2022-10-11	38
P361	pgf-zero	Value of the probability-generating function for argument zero	JoramSoch	2022-10-11	39
P362	pgf-one	Value of the probability-generating function for argument one	JoramSoch	2022-10-11	39
P363	bin-pgf	Probability-generating function of the binomial distribution	JoramSoch	2022-10-11	155
P364	betabin-pmf	Probability mass function of the beta-binomial distribution	JoramSoch	2022-10-20	162
P365	betabin-pmfitogf	Expression of the probability mass function of the beta-binomial distribution using only the gamma function	JoramSoch	2022-10-20	164
P366	betabin-cdf	Cumulative distribution function of the beta-binomial distribution	JoramSoch	2022-10-22	164
P369	anova1-ols	Ordinary least squares for one-way analysis of variance	JoramSoch	2022-11-06	380
P370	anova1-f	F-test for main effect in one-way analysis of variance	JoramSoch	2022-11-06	382
P371	anova2-ols	Ordinary least squares for two-way analysis of variance	JoramSoch	2022-11-06	391
P372	anova2-fme	F-test for main effect in two-way analysis of variance	JoramSoch	2022-11-10	403
P373	anova2-fia	F-test for interaction in two-way analysis of variance	JoramSoch	2022-11-11	406
P374	anova2-fgm	F-test for grand mean in two-way analysis of variance	JoramSoch	2022-11-11	407
P375	anova1-repara	Reparametrization for one-way analysis of variance	JoramSoch	2022-11-15	387
P376	anova1-pss	Partition of sums of squares in one-way analysis of variance	JoramSoch	2022-11-15	381

P377	anova1-fols	F-statistic in terms of ordinary least squares estimates in one-way analysis of variance	JoramSoch	2022-11-15	386
P378	anova2-cochran	Application of Cochran's theorem to two-way analysis of variance	JoramSoch	2022-11-16	397
P379	anova2-pss	Partition of sums of squares in two-way analysis of variance	JoramSoch	2022-11-16	395
P380	anova2-fols	F-statistics in terms of ordinary least squares estimates in two-way analysis of variance	JoramSoch	2022-11-16	409
P381	bin-mle	Maximum likelihood estimation for binomial observations	JoramSoch	2022-11-23	551
P382	bin-mll	Maximum log-likelihood for binomial observations	JoramSoch	2022-11-24	552
P383	bin-lbf	Log Bayes factor for binomial observations	JoramSoch	2022-11-25	557
P384	bin-pp	Posterior probability of the alternative model for binomial observations	JoramSoch	2022-11-26	559
P385	mult-mle	Maximum likelihood estimation for multinomial observations	JoramSoch	2022-12-02	561
P386	mult-mll	Maximum log-likelihood for multinomial observations	JoramSoch	2022-12-02	562
P387	mult-lbf	Log Bayes factor for multinomial observations	JoramSoch	2022-12-02	568
P388	mult-pp	Posterior probability of the alternative model for multinomial observations	JoramSoch	2022-12-02	570
P389	mlr-wlsdist	Distributions of estimated parameters, fitted signal and residuals in multiple linear regression upon weighted least squares	JoramSoch	2022-12-13	466
P390	mlr-rssdist	Distribution of residual sum of squares in multiple linear regression with weighted least squares	JoramSoch	2022-12-13	468
P391	mlr-t	t-test for multiple linear regression using contrast-based inference	JoramSoch	2022-12-13	476
P392	mlr-f	F-test for multiple linear regression using contrast-based inference	JoramSoch	2022-12-13	478

P393	mlr-ind	Independence of estimated parameters and residuals in multiple linear regression	JoramSoch	2022-12-13	463
P394	mvn-indprod	Independence of products of multivariate normal random vector	JoramSoch	2022-12-13	304
P395	mvn-chi2	Relationship between multivariate normal distribution and chi-squared distribution	JoramSoch	2022-12-20	286
P396	cuni-var	Variance of the continuous uniform distribution	JoramSoch	2022-12-20	181
P397	cuni-dent	Differential entropy of the continuous uniform distribution	JoramSoch	2022-12-20	182
P398	resvar-biasp	Maximum likelihood estimator of variance in multiple linear regression is biased	JoramSoch	2022-12-21	600
P399	mlr-symm	Projection matrix and residual-forming matrix are symmetric	JoramSoch	2022-12-22	462
P400	mlr-olsdist	Distributions of estimated parameters, fitted signal and residuals in multiple linear regression upon ordinary least squares	JoramSoch	2022-12-23	465
P401	exp-var	Variance of the exponential distribution	majapavlo	2023-01-23	239
P402	exg-pdf	Probability density function of the ex-Gaussian distribution	tomfaulkenberry	2023-04-18	275
P403	exp-mgf	Moment-generating function of the exponential distribution	tomfaulkenberry	2023-04-19	234
P404	exg-mgf	Moment-generating function of the ex-Gaussian distribution	tomfaulkenberry	2023-04-19	277
P405	exg-mean	Mean of the ex-Gaussian distribution	tomfaulkenberry	2023-04-19	277
P406	exg-var	Variance of the ex-Gaussian distribution	tomfaulkenberry	2023-04-19	278
P407	skew-mean	Partition of skewness into expected values	tomfaulkenberry	2023-04-20	61
P408	exg-skew	Skewness of the ex-Gaussian distribution	tomfaulkenberry	2023-04-21	280

P409	exp-skew	Skewness of the exponential distribution	tomfaulkenberry	2023-04-24	240
P410	duni-ent	Entropy of the discrete uniform distribution	JoramSoch	2023-08-11	146
P411	duni-maxent	Discrete uniform distribution maximizes entropy for finite support	JoramSoch	2023-08-18	148
P412	cuni-maxent	Continuous uniform distribution maximizes differential entropy for fixed range	JoramSoch	2023-08-25	184
P413	post-ind	Combined posterior distributions in terms of individual posterior distributions obtained from conditionally independent data	JoramSoch	2023-09-01	127
P416	bvn-pdf	Probability density function of the bivariate normal distribution	JoramSoch	2023-09-22	287
P417	bvn-pdfcorr	Probability density function of the bivariate normal distribution in terms of correlation coefficient	JoramSoch	2023-09-29	288
P418	mlr-olstr	Ordinary least squares for multiple linear regression with two regressors	JoramSoch	2023-10-06	456
P419	bern-kl	Kullback-Leibler divergence for the Bernoulli distribution	JoramSoch	2023-10-13	153
P420	bin-kl	Kullback-Leibler divergence for the binomial distribution	JoramSoch	2023-10-20	159
P421	wald-skew	Skewness of the Wald distribution	tomfaulkenberry	2023-10-24	270
P422	cuni-kl	Kullback-Leibler divergence for the continuous uniform distribution	JoramSoch	2023-10-27	183
P423	wald-mome	Method of moments for Wald-distributed data	tomfaulkenberry	2023-10-30	273
P424	exg-mome	Method of moments for ex-Gaussian-distributed data	tomfaulkenberry	2023-10-30	282
P425	duni-kl	Kullback-Leibler divergence for the discrete uniform distribution	JoramSoch	2023-11-17	147
P426	gam-scal	Scaling of a random variable following the gamma distribution	JoramSoch	2023-11-24	222
P427	bin-map	Maximum-a-posteriori estimation for binomial observations	JoramSoch	2023-12-01	553

P428	mult-map	Maximum-a-posteriori estimation for multinomial observations	JoramSoch	2023-12-08	563
P429	bin-test	Binomial test	JoramSoch	2023-12-16	550
P430	mult-test	Multinomial test	JoramSoch	2023-12-23	560
P431	blr-anc	Accuracy and complexity for Bayesian linear regression	JoramSoch	2024-01-12	497
P432	blrkc-prior	Conjugate prior distribution for Bayesian linear regression with known covariance	JoramSoch	2024-01-19	515
P433	blrkc-post	Posterior distribution for Bayesian linear regression with known covariance	JoramSoch	2024-01-19	516
P434	blrkc-lme	Log model evidence for Bayesian linear regression with known covariance	JoramSoch	2024-01-19	518
P435	blrkc-anc	Accuracy and complexity for Bayesian linear regression with known covariance	JoramSoch	2024-01-19	521
P436	mvn-mgf	Moment-generating function of the multivariate normal distribution	JoramSoch	2024-02-16	291
P437	gam-mgf	Moment-generating function of the gamma distribution	JoramSoch	2024-02-23	223
P438	lpsr-spsr	The log probability scoring rule is a strictly proper scoring rule	KarahanS	2024-02-28	136
P439	resvar-unbp	Construction of unbiased estimator for variance in multiple linear regression	JoramSoch	2024-03-08	603
P440	rsq-resvar	Expression of R^2 in terms of residual variances	JoramSoch	2024-03-08	605
P441	rsq-test	Statistical significance test for the coefficient of determinantion based on an omnibus F-test	JoramSoch	2024-03-08	608
P442	fstat-rsq	Relationship between F-statistic and R^2	JoramSoch	2024-03-15	610
P443	fstat-mll	Relationship between F-statistic and maximum log-likelihood	JoramSoch	2024-03-28	612
P444	snr-mll	Relationship between signal-to-noise ratio and maximum log-likelihood	JoramSoch	2024-03-28	616

P445	bsr-spsr	Brier scoring rule is strictly proper scoring rule	KarahanS	2024-03-28	140
P446	blr-posterr	Expression of the noise precision posterior for Bayesian linear regression using prediction and parameter errors	JoramSoch	2024-04-05	505
P447	blr-postind	Combined posterior distribution for Bayesian linear regression when analyzing conditionally independent data sets	JoramSoch	2024-04-12	509
P448	blr-map	Maximum-a-posteriori estimation for Bayesian linear regression	JoramSoch	2024-04-19	503
P449	blr-lbf	Log Bayes factor for Bayesian linear regression	JoramSoch	2024-04-26	513
P450	mlr-tsingle	Specific t-test for single regressor in multiple linear regression	JoramSoch	2024-05-03	481
P451	slr-tint	Statistical test for intercept parameter in simple linear regression model	JoramSoch	2024-05-10	440
P452	slr-tslo	Statistical test for slope parameter in simple linear regression model	JoramSoch	2024-05-17	442
P453	slr-fcomp	Statistical test for comparing simple linear regression models with and without slope parameter	JoramSoch	2024-05-24	444
P454	mlr-fomnibus	Omnibus F-test for multiple regressors in multiple linear regression	JoramSoch	2024-05-31	482
P455	glm-llr	Log-likelihood ratio for the general linear model	JoramSoch	2024-06-07	529
P456	glm-mll	Maximum log-likelihood for the general linear model	JoramSoch	2024-06-14	528
P457	glm-mi	Mutual information of dependent and independent variables in the general linear model	JoramSoch	2024-06-21	531
P458	glm-llrmi	Equivalence of log-likelihood ratio and mutual information for the general linear model	JoramSoch	2024-06-21	532
P459	iglm-llrs	Equivalence of log-likelihood ratios for regular and inverse general linear model	JoramSoch	2024-06-28	539

P460	ug-fev	F-test for equality of variances in two independent samples	JoramSoch	2024-07-05	350
P461	slr-pss	Partition of sums of squares for simple linear regression	JoramSoch	2024-07-12	429
P462	mlr-ols3	Ordinary least squares for multiple linear regression	JoramSoch	2024-07-18	454

2 Definition by Number

ID	Shortcut	Definition	Author	Date	Page
D1	mvn	Multivariate normal distribution	JoramSoch	2020-01-22	285
D2	mgf	Moment-generating function	JoramSoch	2020-01-22	35
D3	cuni	Continuous uniform distribution	JoramSoch	2020-01-27	176
D4	norm	Normal distribution	JoramSoch	2020-01-27	185
D5	ng	Normal-gamma distribution	JoramSoch	2020-01-27	307
D6	matn	Matrix-normal distribution	JoramSoch	2020-01-27	326
D7	gam	Gamma distribution	JoramSoch	2020-02-08	219
D8	exp	Exponential distribution	JoramSoch	2020-02-08	233
D9	pmf	Probability mass function	JoramSoch	2020-02-13	17
D10	pdf	Probability density function	JoramSoch	2020-02-13	20
D11	mean	Expected value	JoramSoch	2020-02-13	41
D12	var	Variance	JoramSoch	2020-02-13	54
D13	cdf	Cumulative distribution function	JoramSoch	2020-02-17	27
D14	qf	Quantile function	JoramSoch	2020-02-17	34
D15	ent	Shannon entropy	JoramSoch	2020-02-19	85
D16	dent	Differential entropy	JoramSoch	2020-02-19	91
D17	ent-cond	Conditional entropy	JoramSoch	2020-02-19	87
D18	ent-joint	Joint entropy	JoramSoch	2020-02-19	87
D19	mi	Mutual information	JoramSoch	2020-02-19	101
D19	mi	Mutual information	JoramSoch	2020-02-19	101
D20	resvar	Residual variance	JoramSoch	2020-02-25	598
D21	rsq	Coefficient of determination	JoramSoch	2020-02-25	604
D22	snr	Signal-to-noise ratio	JoramSoch	2020-02-25	614
D27	gm	Generative model	JoramSoch	2020-03-03	125
D28	lf	Likelihood function	JoramSoch	2020-03-03	125
D28	lf	Likelihood function	JoramSoch	2020-03-03	125
D29	prior	Prior distribution	JoramSoch	2020-03-03	125
D30	fpm	Full probability model	JoramSoch	2020-03-03	125

D31	jl	Joint likelihood	JoramSoch	2020-03-03	126
D32	post	Posterior distribution	JoramSoch	2020-03-03	126
D33	ml	Marginal likelihood	JoramSoch	2020-03-03	128
D34	dent-cond	Conditional differential entropy	JoramSoch	2020-03-21	97
D35	dent-joint	Joint differential entropy	JoramSoch	2020-03-21	97
D36	mlr	Multiple linear regression	JoramSoch	2020-03-21	451
D37	tss	Total sum of squares	JoramSoch	2020-03-21	457
D38	ess	Explained sum of squares	JoramSoch	2020-03-21	457
D39	rss	Residual sum of squares	JoramSoch	2020-03-21	458
D40	glm	General linear model	JoramSoch	2020-03-21	524
D41	poiss-data	Poisson-distributed data	JoramSoch	2020-03-22	571
D42	poissexp	Poisson distribution with exposure values	JoramSoch	2020-03-22	578
D43	wish	Wishart distribution	JoramSoch	2020-03-22	336
D44	bern	Bernoulli distribution	JoramSoch	2020-03-22	149
D45	bin	Binomial distribution	JoramSoch	2020-03-22	154
D46	cat	Categorical distribution	JoramSoch	2020-03-22	169
D47	mult	Multinomial distribution	JoramSoch	2020-03-22	171
D48	prob	Probability	JoramSoch	2020-05-10	5
D49	prob-joint	Joint probability	JoramSoch	2020-05-10	5
D50	prob-marg	Law of marginal probability	JoramSoch	2020-05-10	5
D51	prob-cond	Law of conditional probability	JoramSoch	2020-05-10	6
D52	kl	Kullback-Leibler divergence	JoramSoch	2020-05-10	104
D53	beta	Beta distribution	JoramSoch	2020-05-10	258
D54	dir	Dirichlet distribution	JoramSoch	2020-05-10	321
D55	dist	Probability distribution	JoramSoch	2020-05-17	16
D56	dist-joint	Joint probability distribution	JoramSoch	2020-05-17	16
D57	dist-marg	Marginal probability distribution	JoramSoch	2020-05-17	17
D58	dist-cond	Conditional probability distribution	JoramSoch	2020-05-17	17
D59	llf	Log-likelihood function	JoramSoch	2020-05-17	116
D60	mle	Maximum likelihood estimation	JoramSoch	2020-05-15	116

D61	mll	Maximum log-likelihood	JoramSoch	2020-05-15	116
D62	poiss	Poisson distribution	JoramSoch	2020-05-25	166
D63	snorm	Standard normal distribution	JoramSoch	2020-05-26	186
D64	sgam	Standard gamma distribution	JoramSoch	2020-05-26	220
D65	rvar	Random variable	JoramSoch	2020-05-27	3
D66	rvec	Random vector	JoramSoch	2020-05-27	3
D67	rmat	Random matrix	JoramSoch	2020-05-27	4
D68	cgf	Cumulant-generating function	JoramSoch	2020-05-31	40
D69	pgf	Probability-generating function	JoramSoch	2020-05-31	38
D70	cov	Covariance	JoramSoch	2020-06-02	62
D71	corr	Correlation	JoramSoch	2020-06-02	74
D72	covmat	Covariance matrix	JoramSoch	2020-06-06	66
D73	corrmat	Correlation matrix	JoramSoch	2020-06-06	76
D74	precmat	Precision matrix	JoramSoch	2020-06-06	72
D75	ind	Statistical independence	JoramSoch	2020-06-06	7
D76	logreg	Logistic regression	JoramSoch	2020-06-28	593
D77	beta-data	Beta-distributed data	JoramSoch	2020-06-28	585
D78	bin-data	Binomial observations	JoramSoch	2020-07-07	550
D79	mult-data	Multinomial observations	JoramSoch	2020-07-07	560
D81	emat	Estimation matrix	JoramSoch	2020-07-22	460
D82	pmat	Projection matrix	JoramSoch	2020-07-22	460
D83	rfmat	Residual-forming matrix	JoramSoch	2020-07-22	460
D85	ent-cross	Cross-entropy	JoramSoch	2020-07-28	88
D86	dent-cross	Differential cross-entropy	JoramSoch	2020-07-28	97
D88	duni	Discrete uniform distribution	JoramSoch	2020-07-28	144
D90	mom	Moment	JoramSoch	2020-08-19	79
D91	fwhm	Full width at half maximum	JoramSoch	2020-08-19	78
D94	std	Standard deviation	JoramSoch	2020-09-03	78
D95	wald	Wald distribution	tomfaulkenberry	2020-09-04	265
D96	const	Constant	JoramSoch	2020-09-09	4

D97	mom-raw	Raw moment	JoramSoch	2020-10-08	81
D98	mom-cent	Central moment	JoramSoch	2020-10-08	82
D99	mom-stand	Standardized moment	JoramSoch	2020-10-08	84
D100	chi2	Chi-squared distribution	kjpetrykowski	2020-10-13	252
D101	med	Median	JoramSoch	2020-10-15	77
D102	mode	Mode	JoramSoch	2020-10-15	78
D103	prob-exc	Exceedance probability	JoramSoch	2020-10-22	9
D104	dir-data	Dirichlet-distributed data	JoramSoch	2020-10-22	587
D105	rvar-disc	Discrete and continuous random variable	JoramSoch	2020-10-29	4
D106	rvar-uni	Univariate and multivariate random variable	JoramSoch	2020-11-06	4
D107	min	Minimum	JoramSoch	2020-11-12	79
D108	max	Maximum	JoramSoch	2020-11-12	79
D109	rexp	Random experiment	JoramSoch	2020-11-19	2
D110	reve	Random event	JoramSoch	2020-11-19	3
D112	ind-cond	Conditional independence	JoramSoch	2020-11-19	7
D116	prior-flat	Flat, hard and soft prior distribution	JoramSoch	2020-12-02	129
D117	prior-uni	Uniform and non-uniform prior distribution	JoramSoch	2020-12-02	129
D118	prior-inf	Informative and non-informative prior distribution	JoramSoch	2020-12-02	130
D119	prior-emp	Empirical and theoretical prior distribution	JoramSoch	2020-12-02	130
D120	prior-conj	Conjugate and non-conjugate prior distribution	JoramSoch	2020-12-02	130
D121	prior-maxent	Maximum entropy prior distribution	JoramSoch	2020-12-02	131
D122	prior-eb	Empirical Bayes prior distribution	JoramSoch	2020-12-02	131
D123	prior-ref	Reference prior distribution	JoramSoch	2020-12-02	131
D124	ug	Univariate Gaussian	JoramSoch	2021-03-03	344
D125	h0	Null hypothesis	JoramSoch	2021-03-12	120

D126	h1	Alternative hypothesis	JoramSoch	2021-03-12	121
D127	hyp	Statistical hypothesis	JoramSoch	2021-03-19	118
D128	hyp-simp	Simple and composite hypothesis	JoramSoch	2021-03-19	118
D129	hyp-point	Point and set hypothesis	JoramSoch	2021-03-19	119
D130	test	Statistical hypothesis test	JoramSoch	2021-03-19	120
D131	tstat	Test statistic	JoramSoch	2021-03-19	121
D132	size	Size of a statistical test	JoramSoch	2021-03-19	122
D133	alpha	Significance level	JoramSoch	2021-03-19	122
D134	cval	Critical value	JoramSoch	2021-03-19	123
D135	pval	p-value	JoramSoch	2021-03-19	123
D136	ugkv	Univariate Gaussian with known variance	JoramSoch	2021-03-23	359
D137	power	Power of a statistical test	JoramSoch	2021-03-31	122
D138	hyp-tail	One-tailed and two-tailed hypothesis	JoramSoch	2021-03-31	119
D139	test-tail	One-tailed and two-tailed test	JoramSoch	2021-03-31	121
D140	dist-samp	Sampling distribution	JoramSoch	2021-03-31	17
D141	cdf-joint	Joint cumulative distribution function	JoramSoch	2020-04-07	33
D142	mean-samp	Sample mean	JoramSoch	2021-04-16	41
D143	var-samp	Sample variance	JoramSoch	2021-04-16	54
D144	cov-samp	Sample covariance	ciaranmci	2021-04-21	62
D145	prec	Precision	JoramSoch	2020-04-21	60
D146	f	F-distribution	JoramSoch	2020-04-21	255
D147	t	t-distribution	JoramSoch	2021-04-21	214
D148	mvt	Multivariate t-distribution	JoramSoch	2020-04-21	305
D149	eb	Empirical Bayes	JoramSoch	2021-04-29	133
D150	vb	Variational Bayes	JoramSoch	2021-04-29	134
D151	mome	Method-of-moments estimation	JoramSoch	2021-04-29	117
D152	nst	Non-standardized t-distribution	JoramSoch	2021-05-20	215

D153	covmat-samp	Sample covariance matrix	JoramSoch	2021-05-20	66
D154	mean-rvec	Expected value of a random vector	JoramSoch	2021-07-08	53
D155	mean-rmat	Expected value of a random matrix	JoramSoch	2021-07-08	54
D156	exc	Mutual exclusivity	JoramSoch	2021-07-23	9
D157	suni	Standard uniform distribution	JoramSoch	2021-07-23	176
D158	prob-ax	Kolmogorov axioms of probability	JoramSoch	2021-07-30	10
D159	cf	Characteristic function	JoramSoch	2021-09-22	34
D160	tglm	Transformed general linear model	JoramSoch	2021-10-21	533
D161	iglm	Inverse general linear model	JoramSoch	2021-10-21	536
D162	cfm	Corresponding forward model	JoramSoch	2021-10-21	541
D163	slr	Simple linear regression	JoramSoch	2021-10-27	410
D164	regline	Regression line	JoramSoch	2021-10-27	424
D165	samp-spc	Sample space	JoramSoch	2021-11-26	2
D166	eve-spc	Event space	JoramSoch	2021-11-26	2
D167	prob-spc	Probability space	JoramSoch	2021-11-26	2
D168	corr-samp	Sample correlation coefficient	JoramSoch	2021-12-14	75
D169	corrmat-samp	Sample correlation matrix	JoramSoch	2021-12-14	76
D170	lognorm	Log-normal distribution	majapavlo	2022-02-07	242
D173	mse	Mean squared error	JoramSoch	2022-03-27	113
D174	ci	Confidence interval	JoramSoch	2022-03-27	114
D175	nw	Normal-Wishart distribution	JoramSoch	2022-05-14	338
D176	covmat-cross	Cross-covariance matrix	JoramSoch	2022-09-26	70
D177	betabin	Beta-binomial distribution	JoramSoch	2022-10-20	162
D178	betabin-data	Beta-binomial data	JoramSoch	2022-10-20	590
D181	anova1	One-way analysis of variance	JoramSoch	2022-11-06	379
D182	anova2	Two-way analysis of variance	JoramSoch	2022-11-06	390
D183	trss	Treatment sum of squares	JoramSoch	2022-12-14	380
D184	iass	Interaction sum of squares	JoramSoch	2022-12-14	391

D185	tcon	t-contrast for contrast-based inference in multiple linear regression	JoramSoch	2022-12-16	475
D186	fcon	F-contrast for contrast-based inference in multiple linear regression	JoramSoch	2022-12-16	475
D187	exg	ex-Gaussian distribution	tomfaulkenberry	2023-04-18	274
D188	skew	Skewness	tomfaulkenberry	2023-04-20	61
D189	bvn	Bivariate normal distribution	JoramSoch	2023-09-22	287
D190	skew-samp	Sample skewness	tomfaulkenberry	2023-10-30	61
D191	map	Maximum-a-posteriori estimation	JoramSoch	2023-12-01	126
D192	sr	Scoring rule	KarahanS	2024-02-28	135
D193	psr	Proper scoring rule	KarahanS	2024-02-28	135
D194	spsr	Strictly proper scoring rule	KarahanS	2024-02-28	135
D195	lpsr	Log probability scoring rule	KarahanS	2024-02-28	136
D196	fstat	F-statistic	JoramSoch	2024-03-15	610
D197	bsr	Brier scoring rule	KarahanS	2024-03-23	139
D198	lr	Likelihood ratio	JoramSoch	2024-06-14	117
D199	llr	Log-likelihood ratio	JoramSoch	2024-06-14	117

3 Proof by Topic

E

F

G

N

W

4 Definition by Topic