The Book of Statistical Proofs

https://statproofbook.github.io/
StatProofBook@gmail.com

2020-12-02, 22:30

Contents

Chapter I

General Theorems

1 Probability theory

1.1 Random variables

1.1.1 Random experiment

Definition: A random experiment is any repeatable procedure that results in one (→ Definition I/1.1.3) out of a well-defined set of possible outcomes.

- The set of possible outcomes is called sample space.
- A set of zero or more outcomes is called a random event (→ Definition I/1.1.2).
- A function that maps from events to probabilities is called a probability function (→ Definition I/1.4.1).

Together, sample space, event space and probability function 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).

Metadata: ID: D109 | shortcut: rexp | author: JoramSoch | date: 2020-11-19, 04:10.

1.1.2 Random event

Definition: A random event E is the outcome of a random experiment (→ Definition 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 (→ Definition I/1.1.3) with two possible values (true and false, or 1 and 0). A random experiment (→ Definition I/1.1.1) with two possible outcomes is called a Bernoulli trial (→ Definition 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).

Metadata: ID: D110 | shortcut: reve | author: JoramSoch | date: 2020-11-19, 04:33.

1.1.3 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 (→ Definition I/1.1.1);
- formally, as a measurable function (→ Definition "meas-fct") X defined on a probability space (→ Definition "prob-spc") $(\Omega, \mathcal{F}, P)$ that maps from a sample space Ω to the real numbers $\mathbb{R}$ using an event space $\mathcal{F}$ and a probability function (→ Definition I/1.4.1) P;
- more broadly, as any random quantity X such as a random event (→ Definition I/1.1.2), a random scalar (→ Definition I/1.1.3), a random vector (→ Definition I/1.1.4) or a random matrix (→ Definition I/1.1.5).

Sources:

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

Metadata: ID: D65 | shortcut: rvar | author: JoramSoch | date: 2020-05-27, 22:36.

1.1.4 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 (→ Definition I/1.1.3).

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.

Metadata: ID: D66 | shortcut: rvec | author: JoramSoch | date: 2020-05-27, 22:44.

1.1.5 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 (→ Definition I/1.1.3). Equivalently, a random matrix is an $n \times p$ matrix whose columns are n-dimensional random vectors (→ Definition I/1.1.4).

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

Metadata: ID: D67 | shortcut: rmat | author: JoramSoch | date: 2020-05-27, 22:48.

1.1.6 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 (→ Definition I/1.1.3) which is equal to its expected value (→ Definition I/1.5.1)

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

or equivalently, whose variance (→ Definition I/1.6.1) 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.

Metadata: ID: D96 | shortcut: const | author: JoramSoch | date: 2020-09-09, 01:30.

1.1.7 Discrete vs. continuous

Definition: Let X be a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1);
- 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 (→ Definition I/1.4.4).

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.

Metadata: ID: D105 | shortcut: rvar-disc | author: JoramSoch | date: 2020-10-29, 04:44.

1.1.8 Univariate vs. multivariate

Definition: Let X be a random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$. Then,

- X is called a two-valued random variable or random event (→ Definition I/1.1.2), if $\mathcal{X}$ has exactly two elements, e.g. $\mathcal{X} = \left\{E, \overline{E}\right\}$ or $\mathcal{X} = \{\text{true}, \text{false}\}$ or $\mathcal{X} = \{1, 0\}$;
- X is called a univariate random variable or random scalar (→ Definition I/1.1.3), 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 (→ Definition I/1.1.4), 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 (→ Definition I/1.1.5), 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.

Metadata: ID: D106 | shortcut: rvar-uni | author: JoramSoch | date: 2020-11-06, 03:47.

1.2 Probability

1.2.1 Probability

Definition: Let E be a statement about an arbitrary event such as the outcome of a random experiment (→ Definition 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.

Metadata: ID: D48 | shortcut: prob | author: JoramSoch | date: 2020-05-10, 19:41.

1.2.2 Joint probability

Definition: Let A and B be two arbitrary statements about random variables (→ Definition I/1.1.3), 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 (→ Definition I/1.2.1) 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.

Metadata: ID: D49 | shortcut: prob-joint | author: JoramSoch | date: 2020-05-10, 19:49.

1.2.3 Marginal probability

Definition: Let A and X be two arbitrary statements about random variables (→ Definition I/1.1.3), 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 (→ Definition I/1.2.2) distribution $p(A, X)$. Then, $p(A)$ is called the marginal probability of A and,
1) if X is a discrete random variable (→ Definition I/1.1.3) 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 random variable (→ Definition I/1.1.3) 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.

Metadata: ID: D50 | shortcut: prob-marg | author: JoramSoch | date: 2020-05-10, 20:01.

1.2.4 Conditional probability

Definition: Let A and B be two arbitrary statements about random variables (→ Definition I/1.1.3), 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 (→ Definition I/1.2.2) 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 (→ Definition I/1.2.3) 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.

Metadata: ID: D51 | shortcut: prob-cond | author: JoramSoch | date: 2020-05-10, 20:06.

1.2.5 Exceedance probability

Definition: Let $X = \{X_1, \ldots, X_n\}$ be a set of n random variables (→ Definition I/1.1.3) which the joint probability distribution (→ Definition I/1.3.2) $p(X) = p(X_1, \ldots, X_n)$. Then, the exceedance probability for random variable X_i is the probability (→ Definition I/1.2.1) that X_i is larger than all other random variables $X_j, \; j \neq i$:

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

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.

Metadata: ID: D103 | shortcut: prob-exc | author: JoramSoch | date: 2020-10-22, 04:36.

1.2.6 Statistical independence

Definition: Generally speaking, random variables (→ Definition I/1.1.3) are statistically independent, if their joint probability (→ Definition I/1.2.2) can be expressed in terms of their marginal probabilities (→ Definition I/1.2.3).

1) A set of discrete random variables (→ Definition I/1.1.3) $X_1, \ldots, X_n$ 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 (→ Definition I/1.2.2) of $X_1, \ldots, X_n$ and $p(x_i)$ are the marginal probabilities (→ Definition I/1.2.3) of X_i.

2) A set of continuous random variables (→ Definition I/1.1.3) $X_1, \ldots, X_n$ 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 (→ Definition I/1.4.4) 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 (→ Definition I/1.3.2) or marginal (→ Definition I/1.3.3) cumulative distribution functions (→ Definition I/1.4.8) and f are the respective probability density functions (→ Definition I/1.4.4).

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.

Metadata: ID: D75 | shortcut: ind | author: JoramSoch | date: 2020-06-06, 07:16.

1.2.7 Conditional independence

Definition: Generally speaking, random variables (→ Definition I/1.1.3) are conditionally independent given another random variable, if they are statistically independent (→ Definition I/1.2.6) in their conditional probability distributions (→ Definition I/1.3.4) given this random variable.

1) A set of discrete random variables (→ Definition I/1.1.7) $X_1,\ldots,X_n$ 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 (→ Definition I/1.2.2) of $X_1,\ldots,X_n$ given Y and $p(x_i)$ are the marginal (conditional) probabilities (→ Definition I/1.2.3) of X_i given Y.

2) A set of continuous random variables (→ Definition I/1.1.7) $X_1,\ldots,X_n$ 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 (→ Definition I/1.4.4) 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) (→ Definition I/1.3.2) or marginal (conditional) (→ Definition I/1.3.3) cumulative distribution functions (→ Definition I/1.4.8) and f are the respective probability density functions (→ Definition I/1.4.4).

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.

Metadata: ID: D112 | shortcut: ind-cond | author: JoramSoch | date: 2020-11-19, 05:40.

1.3 Probability distributions

1.3.1 Probability distribution

Definition: Let X be a random variable (→ Definition I/1.1.3) with the set of possible outcomes $\mathcal{X}$. Then, a probability distribution of X is a mathematical function that gives the probabilities (→ Definition I/1.2.1) 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.

Metadata: ID: D55 | shortcut: dist | author: JoramSoch | date: 2020-05-17, 20:23.

1.3.2 Joint distribution

Definition: Let X and Y be random variables (→ Definition I/1.1.3) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, a joint distribution of X and Y is a probability distribution (→ Definition I/1.3.1) 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 (→ Definition I/1.1.3) is called a bivariate distribution.
- The joint distribution of a random vector (→ Definition I/1.1.4) is called a multivariate distribution.
- The joint distribution of a random matrix (→ Definition I/1.1.5) 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.

Metadata: ID: D56 | shortcut: dist-joint | author: JoramSoch | date: 2020-05-17, 20:43.

1.3.3 Marginal distribution

Definition: Let X and Y be random variables (→ Definition I/1.1.3) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, the marginal distribution of X is a probability distribution (→ Definition I/1.3.1) 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 (→ Definition I/1.3.2) of X and Y using the law of marginal probability (→ Definition I/1.2.3).

Sources:

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

Metadata: ID: D57 | shortcut: dist-marg | author: JoramSoch | date: 2020-05-17, 21:02.

1.3.4 Conditional distribution

Definition: Let X and Y be random variables (→ Definition I/1.1.3) with sets of possible outcomes $\mathcal{X}$ and $\mathcal{Y}$. Then, the conditional distribution of X given that Y is a probability distribution (→ Definition I/1.3.1) 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 (→ Definition I/1.3.2) of X and Y and the marginal distribution (→ Definition I/1.3.3) of Y using the law of conditional probability (→ Definition I/1.2.4).

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.

Metadata: ID: D58 | shortcut: dist-cond | author: JoramSoch | date: 2020-05-17, 21:25.

1.4 Probability functions

1.4.1 Probability mass function

Definition: Let X be a discrete (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$. Then, $f_X(x) : \mathbb{R} \rightarrow [0, 1]$ 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.

Metadata: ID: D9 | shortcut: pmf | author: JoramSoch | date: 2020-02-13, 19:09.

1.4.2 Probability mass function of strictly increasing function

Theorem: Let X be a discrete (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly increasing function on the support of X. Then, the probability mass function (→ Definition I/1.4.1) 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 (→ Definition I/1.4.1) of Y can be derived as follows:

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

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.

Metadata: ID: P184 | shortcut: pmf-sifct | author: JoramSoch | date: 2020-10-29, 05:55.

1.4.3 Probability mass function of strictly decreasing function

Theorem: Let X be a discrete (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly decreasing function on the support of X. Then, the probability mass function (→ Definition I/1.4.1) 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 (→ Definition I/1.4.1) of Y can be derived as follows:

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

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.

Metadata: ID: P187 | shortcut: pmf-sdfct | author: JoramSoch | date: 2020-11-06, 04:21.

1.4.4 Probability density function

Definition: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$. Then, $f_X(x) : \mathbb{R} \to \mathbb{R}$ 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.

Metadata: ID: D10 | shortcut: pdf | author: JoramSoch | date: 2020-02-13, 19:26.

1.4.5 Probability density function of strictly increasing function

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly increasing function on the support of X. Then, the probability density function (→ Definition I/1.4.4) 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:

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

Proof: The cumulative distribution function of a strictly increasing function (→ Proof I/1.4.9) 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 (→ Proof I/1.4.7)

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

the probability density function (→ Definition I/1.4.4) 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:

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

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.

Metadata: ID: P185 | shortcut: pdf-sifct | author: JoramSoch | date: 2020-10-29, 06:21.

1.4.6 Probability density function of strictly decreasing function

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly decreasing function on the support of X. Then, the probability density function (→ Definition I/1.4.4) of $Y = g(X)$ is given by

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

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 cumulative distribution function of a strictly decreasing function (→ Proof I/1.4.9) is

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

Note that for continuous random variables, the probability (→ Definition I/1.4.4) of point events is

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

Because the probability density function is the first derivative of the cumulative distribution function (→ Proof I/1.4.7)

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

the probability density function (→ Definition I/1.4.4) 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:

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

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.

Metadata: ID: P188 | shortcut: pdf-sdfct | author: JoramSoch | date: 2020-11-06, 05:30.

1.4.7 Probability density function in terms of cumulative distribution function

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3). Then, the probability distribution function (→ Definition I/1.4.4) of X is the first derivative of the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Proof I/1.4.12) 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.

Metadata: ID: P191 | shortcut: pdf-cdf | author: JoramSoch | date: 2020-11-12, 07:19.

1.4.8 Cumulative distribution function

Definition: The cumulative distribution function (CDF) of a random variable (→ Definition I/1.1.3) X at a given value x is defined as the probability (→ Definition I/1.2.1) that X is smaller than x:

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

1) If X is a discrete (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and the probability mass function (→ Definition I/1.4.1) $f_X(x)$, then the cumulative distribution function is the function (→ Proof I/1.4.11) $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ with

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

2) If X is a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and the probability density function (→ Definition I/1.4.4) $f_X(x)$, then the cumulative distribution function is the function (→ Proof I/1.4.12) $F_X(x) : \mathbb{R} \rightarrow [0, 1]$ 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.

Metadata: ID: D13 | shortcut: cdf | author: JoramSoch | date: 2020-02-17, 22:07.

1.4.9 Cumulative distribution function of strictly increasing function

Theorem: Let X be a random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly increasing function on the support of X. Then, the cumulative distribution function (→ Definition I/1.4.8) 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:

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

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 increasing, then $g^{-1}(y)$ is also strictly increasing. Therefore, the cumulative distribution function (→ Definition I/1.4.8) of Y can be derived as follows:
1) If y is lower than the lowest value (→ Definition I/1.11.1) 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 (→ Definition I/1.11.2) 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.

Metadata: ID: P183 | shortcut: cdf-sifct | author: JoramSoch | date: 2020-10-29, 05:35.

1.4.10 Cumulative distribution function of strictly decreasing function

Theorem: Let X be a random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $g(x)$ be a strictly decreasing function on the support of X. Then, the cumulative distribution function (→ Definition I/1.4.8) of $Y = g(X)$ is given by

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

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 (→ Definition I/1.4.8) of Y can be derived as follows:
1) If y is higher than the highest value (→ Definition I/1.11.2) Y can take, then $\mathrm{Pr}(Y \leq y) = 1$, so

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

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

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

3) If y is lower than the lowest value (→ Definition I/1.11.1) Y can take, then $\mathrm{Pr}(Y \leq y) = 0$, so

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

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.

Metadata: ID: P186 | shortcut: cdf-sdfct | author: JoramSoch | date: 2020-11-06, 04:12.

1.4.11 Cumulative distribution function of discrete random variable

Theorem: Let X be a discrete (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible values $\mathcal{X}$ and probability mass function (→ Definition I/1.4.1) $f_X(x)$. Then, the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Definition I/1.4.8) of a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) of a discrete (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) 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:

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

Sources:
- original work

Metadata: ID: P189 | shortcut: cdf-pmf | author: JoramSoch | date: 2020-11-12, 06:03.

1.4.12 Cumulative distribution function of continuous random variable

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with possible values $\mathcal{X}$ and probability density function (→ Definition I/1.4.4) $f_X(x)$. Then, the cumulative distribution function (→ Definition I/1.4.8) of X is

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

Proof: The cumulative distribution function (→ Definition I/1.4.8) of a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.4.4) of a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) 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:

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

Sources:

- original work

Metadata: ID: P190 | shortcut: cdf-pdf | author: JoramSoch | date: 2020-11-12, 06:33.

1.4.13 Quantile function

Definition: Let X be a random variable (→ Definition I/1.1.3) with the cumulative distribution function (→ Definition I/1.4.8) (CDF) $F_X(x)$. Then, the function $Q_X(p) : [0,1] \rightarrow \mathbb{R}$ 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.

Metadata: ID: D14 | shortcut: qf | author: JoramSoch | date: 2020-02-17, 22:18.

1.4.14 Quantile function in terms of cumulative distribution function

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3) with the cumulative distribution function (→ Definition I/1.4.8) $F_X(x)$. If the cumulative distribution function is strictly monotonically increasing, then the quantile function (→ Definition I/1.4.13) is identical to the inverse of $F_X(x)$:

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

Proof: The quantile function (→ Definition I/1.4.13) $Q_X(p)$ 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.

Metadata: ID: P192 | shortcut: qf-cdf | author: JoramSoch | date: 2020-11-12, 07:48.

1.4.15 Moment-generating function

Definition:
1) The moment-generating function of a random variable (→ Definition I/1.1.3) $X \in \mathbb{R}$ 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 (→ Definition I/1.1.4) $X \in \mathbb{R}^n$ 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.

Metadata: ID: D2 | shortcut: mgf | author: JoramSoch | date: 2020-01-22, 10:58.

1.4.16 Moment-generating function of linear transformation

Theorem: Let X be an $n \times 1$ random vector (→ Definition I/1.1.4) with the moment-generating function (→ Definition I/1.4.15) $M_X(t)$. 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 (→ Definition I/1.4.15) 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 (→ Definition I/1.1.4) 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.

Metadata: ID: P154 | shortcut: mgf-ltt | author: JoramSoch | date: 2020-08-19, 08:09.

1.4.17 Moment-generating function of linear combination

Theorem: Let $X_1, \ldots, X_n$ be n independent (→ Definition I/1.2.6) random variables (→ Definition I/1.1.3) with moment-generating functions (→ Definition I/1.4.15) $M_{X_i}(t)$. 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 (→ Definition I/1.4.15) X_i is

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

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 (→ Proof I/1.5.6), 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.

Metadata: ID: P155 | shortcut: mgf-lincomb | author: JoramSoch | date: 2020-08-19, 08:36.

1.4.18 Cumulant-generating function

Definition:

1) The cumulant-generating function of a random variable (→ Definition I/1.1.3) $X \in \mathbb{R}$ 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 (→ Definition I/1.1.4) $X \in \mathbb{R}^n$ 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.

Metadata: ID: D68 | shortcut: cgf | author: JoramSoch | date: 2020-05-31, 23:46.

1.4.19 Probability-generating function

Definition:
1) If X is a discrete random variable (→ Definition I/1.1.3) taking values in the non-negative integers $\{0, 1, \ldots\}$, then the probability-generating function of X is defined as

$$G_X(z) = \mathrm{E}\left[z^X\right] = \sum_{x=0}^{\infty} p(x) \, z^x \tag{1}$$

where $z \in \mathbb{C}$ and $p(x)$ is the probability mass function (→ Definition I/1.4.1) of X.
2) If X is a discrete random vector (→ Definition I/1.1.4) taking values in the n-dimensional integer lattice $x \in \{0, 1, \ldots\}^n$, then the probability-generating function of X is defined as

$$G_X(z) = \mathrm{E}\left[{z_1}^{X_1} \cdot \ldots \cdot {z_n}^{X_n}\right] = \sum_{x_1=0}^{\infty} \cdots \sum_{x_n=0}^{\infty} p(x_1, \ldots, x_n) \, {z_1}^{x_1} \cdot \ldots \cdot {z_n}^{x_n} \tag{2}$$

where $z \in \mathbb{C}^n$ and $p(x)$ is the probability mass function (→ Definition I/1.4.1) 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.

Metadata: ID: D69 | shortcut: pgf | author: JoramSoch | date: 2020-05-31, 23:59.

1.5 Expected value

1.5.1 Definition

Definition:
1) The expected value (or, mean) of a discrete random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) of X.

2) The expected value (or, mean) of a continuous random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D11 | shortcut: mean | author: JoramSoch | date: 2020-02-13, 19:38.

1.5.2 Non-negative random variable

Theorem: Let X be a non-negative random variable (→ Definition I/1.1.3). Then, the expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.4.8) of X.

Proof: Because the cumulative distribution function gives the probability of a random variable being smaller than a given value (→ Definition I/1.4.8),

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

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 (→ Definition I/1.4.4) of X, can be rewritten as

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

and by applying the definition of the expected value (→ Definition I/1.5.1), 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.

Metadata: ID: P103 | shortcut: mean-nnrvar | author: JoramSoch | date: 2020-05-18, 23:54.

1.5.3 Non-negativity

Theorem: If a random variable (→ Definition I/1.1.3) is strictly non-negative, its expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.4.1) 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 (→ Definition I/1.4.4) 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.

Metadata: ID: P52 | shortcut: mean-nonneg | author: JoramSoch | date: 2020-02-13, 20:14.

1.5.4 Linearity

Theorem: The expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.1.3) X and Y and a constant a.

Proof:
1) If X and Y are discrete random variables, the expected value (→ Definition I/1.5.1) is

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

and the law of marginal probability (→ Definition I/1.2.3) states that

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

Applying this, we have

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

as well as

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

2) If X and Y are continuous random variables, the expected value (→ Definition I/1.5.1) is

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

and the law of marginal probability (→ Definition I/1.2.3) 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, for discrete as well as for continuous 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/.

Metadata: ID: P53 | shortcut: mean-lin | author: JoramSoch | date: 2020-02-13, 21:08.

1.5.5 Monotonicity

Theorem: The expected value (→ Definition I/1.5.1) 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 (→ Proof I/1.5.4), 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 (→ Proof I/1.5.3), 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.

Metadata: ID: P54 | shortcut: mean-mono | author: JoramSoch | date: 2020-02-17, 21:00.

1.5.6 (Non-)Multiplicativity

Theorem:
1) If two random variables (→ Definition I/1.1.3) X and Y are independent (→ Definition I/1.2.6), the expected value (→ Definition I/1.5.1) is multiplicative, i.e.

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

2) If two random variables (→ Definition I/1.1.3) X and Y are dependent (→ Definition I/1.2.6), the expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.2.6), 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 (→ Definition I/1.5.1), we have

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

And applying it to the expected value for continuous random variables (→ Definition I/1.5.1), 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 (→ Definition II/1.2.1) with the following joint probability (→ Definition I/1.2.2) mass function (→ Definition I/1.4.1)

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

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.

Metadata: ID: P55 | shortcut: mean-mult | author: JoramSoch | date: 2020-02-17, 21:51.

1.5.7 Expectation of a quadratic form

Theorem: Let X be an $n \times 1$ random vector (→ Definition I/1.1.4) with mean (→ Definition I/1.5.1) μ and covariance (→ Definition I/1.7.1) Σ and let A be a symmetric $n \times n$ matrix. Then, the expectation of the quadratic form $X^\mathrm{T} AX$ is

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

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

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

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

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

Because mean and trace are linear operators (→ Proof I/1.5.4), we have

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

Note that the covariance matrix can be partitioned into expected values (→ Proof I/1.7.6)

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

such that the expected value of the quadratic form becomes

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

Finally, applying mean and covariance of X, we have

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

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

Metadata: ID: P131 | shortcut: mean-qf | author: JoramSoch | date: 2020-07-13, 21:59.

1.5.8 Law of the unconscious statistician

Theorem: Let X be a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) $f_X(x)$, the expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.4.4) $f_X(x)$, the expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.5.1) 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 (→ Definition I/1.4.8) 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 (→ Definition I/1.4.4) 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.

Metadata: ID: P138 | shortcut: mean-lotus | author: JoramSoch | date: 2020-07-22, 08:30.

1.6 Variance

1.6.1 Definition

Definition: The variance of a random variable (→ Definition I/1.1.3) X is defined as the expected value (→ Definition I/1.5.1) of the squared deviation from its expected value (→ Definition I/1.5.1):

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

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

Metadata: ID: D12 | shortcut: var | author: JoramSoch | date: 2020-02-13, 19:55.

1.6.2 Partition into expected values

Theorem: Let X be a random variable (→ Definition I/1.1.3). Then, the variance (→ Definition I/1.6.1) of X is equal to the mean (→ Definition I/1.5.1) of the square of X minus the square of the mean (→ Definition I/1.5.1) of X:

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

Proof: The variance (→ Definition I/1.6.1) 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 (→ Proof I/1.5.4), can be rewritten as

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

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

Metadata: ID: P104 | shortcut: var-mean | author: JoramSoch | date: 2020-05-19, 00:17.

1.6.3 Non-negativity

Theorem: The variance (→ Definition I/1.6.1) is always non-negative, i.e.

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

Proof: The variance (→ Definition I/1.6.1) of a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.1.3), then, because squares and probabilities are stricly 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 (→ Definition I/1.1.3), 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.

Metadata: ID: P123 | shortcut: var-nonneg | author: JoramSoch | date: 2020-06-06, 07:29.

1.6.4 Variance of a constant

Theorem: The variance (→ Definition I/1.6.1) of a constant (→ Definition I/1.1.6) is zero:

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

Proof: A constant (→ Definition I/1.1.6) is a quantity that always has the same value. Thus, if understood as a random variable (→ Definition I/1.1.3), the expected value (→ Definition I/1.5.1) of a constant is equal to itself:

$$\mathrm{E}(a) = a \; . \tag{2}$$

Plugged into the formula of the variance (→ Definition I/1.6.1), we have

$$\begin{split} \mathrm{Var}(a) &= \mathrm{E}\left[(a - \mathrm{E}(a))^2\right] \\ &= \mathrm{E}\left[(a - a)^2\right] \\ &= \mathrm{E}(0) \; . \end{split} \tag{3}$$

Applied to the formula of the expected value (→ Definition I/1.5.1), this gives

$$\mathrm{E}(0) = \sum_{x=0} x \cdot f_X(x) = 0 \cdot 1 = 0 \; . \tag{4}$$

Together, (3) and (4) imply (1).

Sources:

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

Metadata: ID: P124 | shortcut: var-const | author: JoramSoch | date: 2020-06-27, 06:44.

1.6.5 Variance equals zero

Theorem: If the variance (→ Definition I/1.6.1) of X is zero, then X is a constant (→ Definition I/1.1.6):

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

Proof: The variance (→ Definition I/1.6.1) is defined as

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

Because $(X - \mathrm{E}(X))^2$ is strictly non-negative (→ Proof I/1.5.3), the only way for the variance to become zero is, if the squared deviation is always zero:

$$(X - \mathrm{E}(X))^2 = 0 \; . \tag{3}$$

Thus, in turn, requires that X is equal to its expected value (→ Definition I/1.5.1)

$$X = \mathrm{E}(X) \tag{4}$$

which can only be the case, if X always has the same value (→ Definition I/1.1.6):

$$X = \text{const.} \tag{5}$$

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

Metadata: ID: P125 | shortcut: var-zero | author: JoramSoch | date: 2020-06-27, 07:05.

1.6.6 Invariance under addition

Theorem: The variance (→ Definition I/1.6.1) is invariant under addition of a constant (→ Definition I/1.1.6):

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

Proof: The variance (→ Definition I/1.6.1) is defined in terms of the expected value (→ Definition I/1.5.1) as

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

Using this and the linearity of the expected value (→ Proof I/1.5.4), we can derive (1) as follows:

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

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

Metadata: ID: P126 | shortcut: var-inv | author: JoramSoch | date: 2020-07-07, 05:23.

1.6.7 Scaling upon multiplication

Theorem: The variance (→ Definition I/1.6.1) scales upon multiplication with a constant (→ Definition I/1.1.6):

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

Proof: The variance (→ Definition I/1.6.1) is defined in terms of the expected value (→ Definition I/1.5.1) as

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

Using this and the linearity of the expected value (→ Proof I/1.5.4), we can derive (1) as follows:

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

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

Metadata: ID: P127 | shortcut: var-scal | author: JoramSoch | date: 2020-07-07, 05:38.

1.6.8 Variance of a sum

Theorem: The variance (→ Definition I/1.6.1) of the sum of two random variables (→ Definition I/1.1.3) equals the sum of the variances of those random variables, plus two times their covariance (→ Definition I/1.7.1):

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

Proof: The variance (→ Definition I/1.6.1) is defined in terms of the expected value (→ Definition I/1.5.1) as

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

Using this and the linearity of the expected value (→ Proof I/1.5.4), 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.

Metadata: ID: P128 | shortcut: var-sum | author: JoramSoch | date: 2020-07-07, 06:10.

1.6.9 Variance of linear combination

Theorem: The variance (→ Definition I/1.6.1) of the linear combination of two random variables (→ Definition I/1.1.3) is a function of the variances as well as the covariance (→ Definition I/1.7.1) 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 (→ Definition I/1.6.1) is defined in terms of the expected value (→ Definition I/1.5.1) as

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

Using this and the linearity of the expected value (→ Proof I/1.5.4), 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.

Metadata: ID: P129 | shortcut: var-lincomb | author: JoramSoch | date: 2020-07-07, 06:21.

1.6.10 Additivity under independence

Theorem: The variance (→ Definition I/1.6.1) is additive for independent (→ Definition I/1.2.6) random variables (→ Definition I/1.1.3):

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

Proof: The variance of the sum of two random variables (→ Proof I/1.6.8) 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 (→ Proof I/1.7.3) 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.

Metadata: ID: P130 | shortcut: var-add | author: JoramSoch | date: 2020-07-07, 06:52.

1.7 Covariance

1.7.1 Definition

Definition: The covariance of two random variables (→ Definition I/1.1.3) X and Y is defined as the expected value (→ Definition I/1.5.1) of the product of their deviations from their individual expected values (→ Definition I/1.5.1):

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

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

Metadata: ID: D70 | shortcut: cov | author: JoramSoch | date: 2020-06-02, 20:20.

1.7.2 Partition into expected values

Theorem: Let X and Y be random variables (→ Definition I/1.1.3). Then, the covariance (→ Definition I/1.7.1) of X and Y is equal to the mean (→ Definition I/1.5.1) of the product of X and Y minus the product of the means (→ Definition I/1.5.1) of X and Y:

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

Proof: The covariance (→ Definition I/1.7.1) 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 (→ Proof I/1.5.4), 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.

Metadata: ID: P118 | shortcut: cov-mean | author: JoramSoch | date: 2020-06-02, 20:50.

1.7.3 Covariance under independence

Theorem: Let X and Y be independent (→ Definition I/1.2.6) random variables (→ Definition I/1.1.3). Then, the covariance (→ Definition I/1.7.1) 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 (→ Proof I/1.7.2) 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 (→ Proof I/1.5.6):

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

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.

Metadata: ID: P158 | shortcut: cov-ind | author: JoramSoch | date: 2020-09-03, 06:05.

1.7.4 Relationship to correlation

Theorem: Let X and Y be random variables (→ Definition I/1.1.3). Then, the covariance (→ Definition I/1.7.1) of X and Y is equal to the product of their correlation (→ Definition I/1.8.1) and the standard deviations (→ Definition I/1.10.1) of X and Y:

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

Proof: The correlation (→ Definition I/1.8.1) 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 (→ Definition I/1.7.1) to give

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

Sources:
- original work

Metadata: ID: P119 | shortcut: cov-corr | author: JoramSoch | date: 2020-06-02, 21:00.

1.7.5 Covariance matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ be a random vector (→ Definition I/1.1.4). Then, the covariance matrix of X is defined as the $n \times n$ matrix in which the entry (i,j) is the covariance (→ Definition I/1.7.1) 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.

Metadata: ID: D72 | shortcut: covmat | author: JoramSoch | date: 2020-06-06, 04:24.

1.7.6 Covariance matrix and expected values

Theorem: Let X be a random vector (→ Definition I/1.1.4). Then, the covariance matrix (→ Definition I/1.7.5) of X is equal to the mean (→ Definition I/1.5.1) of the outer product of X with itself minus the outer product of the mean (→ Definition I/1.5.1) 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 (→ Definition I/1.7.5) of X is defined as

$$\Sigma_{XX} = \begin{bmatrix} \mathrm{E}\,[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] & \ldots & \mathrm{E}\,[(X_1 - \mathrm{E}[X_1])(X_n - \mathrm{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \mathrm{E}\,[(X_n - \mathrm{E}[X_n])(X_1 - \mathrm{E}[X_1])] & \ldots & \mathrm{E}\,[(X_n - \mathrm{E}[X_n])(X_n - \mathrm{E}[X_n])] \end{bmatrix} \tag{2}$$

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 (→ Proof I/1.5.4), 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.

Metadata: ID: P120 | shortcut: covmat-mean | author: JoramSoch | date: 2020-06-06, 05:31.

1.7.7 Covariance matrix and correlation matrix

Theorem: Let X be a random vector (→ Definition I/1.1.4). Then, the covariance matrix (→ Definition I/1.7.5) of X can be expressed in terms of its correlation matrix (→ Definition I/1.8.2) 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 (→ Definition I/1.10.1) 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 (→ Definition I/1.8.2), 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 (→ Definition I/1.7.5).

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.

Metadata: ID: P121 | shortcut: covmat-corrmat | author: JoramSoch | date: 2020-06-06, 06:02.

1.7.8 Precision matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ be a random vector (→ Definition I/1.1.4). Then, the precision matrix of X is defined as the inverse of the covariance matrix (→ Definition I/1.7.5) 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).

Metadata: ID: D74 | shortcut: precmat | author: JoramSoch | date: 2020-06-06, 05:08.

1.7.9 Precision matrix and correlation matrix

Theorem: Let X be a random vector (→ Definition I/1.1.4). Then, the precision matrix (→ Definition I/1.7.8) of X can be expressed in terms of its correlation matrix (→ Definition I/1.8.2) 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 (→ Definition I/1.10.1) 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 (→ Definition I/1.7.8) is defined as the inverse of the covariance matrix (→ Definition I/1.7.5)

$$\Lambda_{XX} = \Sigma_{XX}^{-1} \tag{3}$$

and the relation between covariance matrix and correlation matrix (→ Proof I/1.7.7) 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

$$\begin{split} \Lambda_{XX} &\overset{(3)}{=} \Sigma_{XX}^{-1} \\ &\overset{(4)}{=} \left(\mathrm{D}_X \cdot \mathrm{C}_{XX} \cdot \mathrm{D}_X \right)^{-1} \\ &\overset{(6)}{=} \mathrm{D}_X^{-1} \cdot \mathrm{C}_{XX}^{-1} \cdot \mathrm{D}_X^{-1} \\ &\overset{(7)}{=} \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \cdot \mathrm{C}_{XX}^{-1} \cdot \begin{bmatrix} \frac{1}{\sigma_{X_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sigma_{X_n}} \end{bmatrix} \end{split} \tag{8}$$

which conforms to equation (1).

Sources:
- original work

Metadata: ID: P122 | shortcut: precmat-corrmat | author: JoramSoch | date: 2020-06-06, 06:28.

1.8 Correlation

1.8.1 Definition

Definition: The correlation of two random variables (→ Definition I/1.1.3) X and Y, also called Pearson product-moment correlation coefficient (PPMCC), is defined as the ratio of the covariance (→ Definition I/1.7.1) of X and Y relative to the product of their standard deviations (→ Definition I/1.10.1):

$$\mathrm{Corr}(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)} \sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y]) \right]}{\sqrt{\mathrm{E}\left[(X-\mathrm{E}[X])^2 \right]} \sqrt{\mathrm{E}\left[(Y-\mathrm{E}[Y])^2 \right]}} \; . \tag{1}$$

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.

Metadata: ID: D71 | shortcut: corr | author: JoramSoch | date: 2020-06-02, 20:34.

1.8.2 Correlation matrix

Definition: Let $X = [X_1, \ldots, X_n]^\mathrm{T}$ be a random vector (→ Definition I/1.1.4). Then, the correlation matrix of X is defined as the $n \times n$ matrix in which the entry (i,j) is the correlation (→ Definition I/1.8.1) 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.

Metadata: ID: D73 | shortcut: corrmat | author: JoramSoch | date: 2020-06-06, 04:56.

1.9 Measures of central tendency

1.9.1 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 (→ Definition "samp") from a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.1.3) with cumulative distribution function (→ Definition I/1.4.8) $F_X(x)$. 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.

Metadata: ID: D101 | shortcut: med | author: JoramSoch | date: 2020-10-15, 10:53.

1.9.2 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 (→ Definition "samp") from a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.1.3) with probability mass function (→ Definition I/1.4.1) or probability density function (→ Definition I/1.4.4) $f_X(x)$. 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).

Metadata: ID: D102 | shortcut: mode | author: JoramSoch | date: 2020-10-15, 11:10.

1.10 Measures of statistical dispersion

1.10.1 Standard deviation

Definition: The standard deviation σ of a random variable (→ Definition I/1.1.3) X with expected value (→ Definition I/1.5.1) μ is defined as the square root of the variance (→ Definition I/1.6.1), 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.

Metadata: ID: D94 | shortcut: std | author: JoramSoch | date: 2020-09-03, 05:43.

1.10.2 Full width at half maximum

Definition: Let X be a continuous random variable (→ Definition I/1.1.3) with a unimodal probability density function (→ Definition I/1.4.4) $f_X(x)$ and mode (→ Definition I/1.9.2) 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.

Metadata: ID: D91 | shortcut: fwhm | author: JoramSoch | date: 2020-08-19, 05:40.

1.11 Further summary statistics

1.11.1 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 (→ Definition "samp") from a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.1.3) 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.

Metadata: ID: D107 | shortcut: min | author: JoramSoch | date: 2020-11-12, 05:25.

1.11.2 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 (→ Definition "samp") from a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.1.3) 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.

Metadata: ID: D108 | shortcut: max | author: JoramSoch | date: 2020-11-12, 05:33.

1.12 Further moments

1.12.1 Moment

Definition: Let X be a random variable (→ Definition I/1.1.3), let c be a constant (→ Definition I/1.1.6) and let n be a positive integer. Then, the n-th moment of X about c is defined as the expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.12.3) $\mu_n' = \mu_n(0)$;
- the n-th central moment (→ Definition I/1.12.6) $\mu_n = \mu_n(\mu)$;
- the n-th standardized moment (→ Definition I/1.12.9) $\mu_n^* = \mu_n/\sigma^n$.

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.

Metadata: ID: D90 | shortcut: mom | author: JoramSoch | date: 2020-08-19, 05:24.

1.12.2 Moment in terms of moment-generating function

Theorem: Let X be a scalar random variable (→ Definition I/1.1.3) with the moment-generating function (→ Definition I/1.4.15) $M_X(t)$. Then, the n-th raw moment (→ Definition I/1.12.3) 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 (→ Definition I/1.4.15), 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 (→ Proof I/1.5.4), we have:

$$
\begin{aligned}
M_X^{(n)}(t) &= \frac{\mathrm{d}^n}{\mathrm{d}t^n} \sum_{m=0}^{\infty} \mathrm{E}\left(\frac{t^m X^m}{m!}\right) \\
&= \sum_{m=0}^{\infty} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \frac{t^m}{m!} \mathrm{E}\left(X^m\right) \; .
\end{aligned} \tag{5}
$$

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.

Metadata: ID: P153 | shortcut: mom-mgf | author: JoramSoch | date: 2020-08-19, 07:51.

1.12.3 Raw moment

Definition: Let X be a random variable (→ Definition I/1.1.3) 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 (→ Definition I/1.12.1) 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.

Metadata: ID: D97 | shortcut: mom-raw | author: JoramSoch | date: 2020-10-08, 03:31.

1.12.4 First raw moment is mean

Theorem: The first raw moment (→ Definition I/1.12.3) equals the mean (→ Definition I/1.5.1), i.e.

$$\mu_1' = \mu \; . \tag{1}$$

Proof: The first raw moment (→ Definition I/1.12.3) of a random variable (→ Definition I/1.1.3) X is defined as

$$\mu_1' = \mathrm{E}\left[(X-0)^1\right] \tag{2}$$

which is equal to the expected value (→ Definition I/1.5.1) of X:

$$\mu_1' = \mathrm{E}\left[X\right] = \mu \; . \tag{3}$$

Sources:
- original work

Metadata: ID: P171 | shortcut: momraw-1st | author: JoramSoch | date: 2020-10-08, 04:19.

1.12.5 Second raw moment and variance

Theorem: The second raw moment (→ Definition I/1.12.3) can be expressed as

$$\mu_2' = \mathrm{Var}(X) + \mathrm{E}(X)^2 \tag{1}$$

where $\mathrm{Var}(X)$ is the variance (→ Definition I/1.6.1) of X and $\mathrm{E}(X)$ is the expected value (→ Definition I/1.5.1) of X.

Proof: The second raw moment (→ Definition I/1.12.3) of a random variable (→ Definition I/1.1.3) X is defined as

$$\mu_2' = \mathrm{E}\left[(X-0)^2\right] \; . \tag{2}$$

Using the partition of variance into expected values (→ Proof I/1.6.2)

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \; , \tag{3}$$

the second raw moment can be rearranged into:

$$\mu_2' \overset{(2)}{=} \mathrm{E}(X^2) \overset{(3)}{=} \mathrm{Var}(X) + \mathrm{E}(X)^2 \; . \tag{4}$$

Sources:
- original work

Metadata: ID: P172 | shortcut: momraw-2nd | author: JoramSoch | date: 2020-10-08, 05:05.

1.12.6 Central moment

Definition: Let X be a random variable (→ Definition I/1.1.3) with expected value (→ Definition I/1.5.1) μ and let n be a positive integer. Then, the n-th central moment of X is defined as the n-th moment (→ Definition I/1.12.1) 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.

Metadata: ID: D98 | shortcut: mom-cent | author: JoramSoch | date: 2020-10-08, 03:37.

1.12.7 First central moment is zero

Theorem: The first central moment (→ Definition I/1.12.6) is zero, i.e.

$$\mu_1 = 0 \; . \tag{1}$$

Proof: The first central moment (→ Definition I/1.12.6) of a random variable (→ Definition I/1.1.3) X with mean (→ Definition I/1.5.1) μ is defined as

$$\mu_1 = \mathrm{E}\left[(X-\mu)^1\right] \; . \tag{2}$$

Due to the linearity of the expected value (→ Proof I/1.5.4) 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.

Metadata: ID: P167 | shortcut: momcent-1st | author: JoramSoch | date: 2020-09-09, 07:51.

1.12.8 Second central moment is variance

Theorem: The second central moment (→ Definition I/1.12.6) equals the variance (→ Definition I/1.6.1), i.e.

$$\mu_2 = \mathrm{Var}(X) \; . \tag{1}$$

Proof: The second central moment (→ Definition I/1.12.6) of a random variable (→ Definition I/1.1.3) X with mean (→ Definition I/1.5.1) μ is defined as

$$\mu_2 = \mathrm{E}\left[(X-\mu)^2\right] \tag{2}$$

which is equivalent to the definition of the variance (→ Definition I/1.6.1):

$$\mu_2 = \mathrm{E}\left[(X-\mathrm{E}(X))^2\right] = \mathrm{Var}(X) \; . \tag{3}$$

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.

Metadata: ID: P173 | shortcut: momcent-2nd | author: JoramSoch | date: 2020-10-08, 05:13.

1.12.9 Standardized moment

Definition: Let X be a random variable (→ Definition I/1.1.3) with expected value (→ Definition I/1.5.1) μ and standard deviation (→ Definition I/1.10.1) σ and let n be a positive integer. Then, the n-th standardized moment of X is defined as the n-th moment (→ Definition I/1.12.1) 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.

Metadata: ID: D99 | shortcut: mom-stand | author: JoramSoch | date: 2020-10-08, 03:47.

2 Information theory

2.1 Shannon entropy

2.1.1 Definition

Definition: Let X be a discrete random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and the (observed or assumed) probability mass function (→ Definition I/1.4.1) $p(x) = f_X(x)$. 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.

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.

Metadata: ID: D15 | shortcut: ent | author: JoramSoch | date: 2020-02-19, 17:36.

2.1.2 Non-negativity

Theorem: The entropy of a discrete random variable (→ Definition I/1.1.3) is a non-negative number:

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

Proof: The entropy of a discrete random variable (→ Definition 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 (→ Definition I/1.4.1) 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.

Metadata: ID: P57 | shortcut: ent-nonneg | author: JoramSoch | date: 2020-02-19, 19:10.

2.1.3 Concavity

Theorem: The entropy (→ Definition I/2.1.1) is concave in the probability mass function (→ Definition I/1.4.1) 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 (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let $u(x)$ be the probability mass function (→ Definition I/1.4.1) of a discrete uniform distribution (→ Definition II/1.1.1) on $X \in \mathcal{X}$. Then, the entropy (→ Definition I/2.1.1) of an arbitrary probability mass function (→ Definition I/1.4.1) $p(x)$ 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 (→ Definition I/2.5.1), the probability mass function of the discrete uniform distribution (→ Proof II/1.1.2) and the total sum over the probability mass function (→ Definition I/1.4.1).
Note that the KL divergence is convex (→ Proof I/2.5.5) in the pair of probability distributions (→ Definition I/1.3.1) (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.

Metadata: ID: P149 | shortcut: ent-conc | author: JoramSoch | date: 2020-08-11, 08:29.

2.1.4 Conditional entropy

Definition: Let X and Y be discrete random variables (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and probability mass functions (→ Definition I/1.4.1) $p(x)$ 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 (→ Definition 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.

Metadata: ID: D17 | shortcut: ent-cond | author: JoramSoch | date: 2020-02-19, 18:08.

2.1.5 Joint entropy

Definition: Let X and Y be discrete random variables (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and joint probability (→ Definition I/1.2.2) mass function (→ Definition I/1.4.1) $p(x, y)$. 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.

Metadata: ID: D18 | shortcut: ent-joint | author: JoramSoch | date: 2020-02-19, 18:18.

2.1.6 Cross-entropy

Definition: Let X be a discrete random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ Definition I/1.3.1) on X with the probability mass functions (→ Definition I/1.4.1) $p(x)$ 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.

Metadata: ID: D85 | shortcut: ent-cross | author: JoramSoch | date: 2020-07-28, 02:51.

2.1.7 Convexity of cross-entropy

Theorem: The cross-entropy (→ Definition I/2.1.6) is convex in the probability distribution (→ Definition I/1.3.1) 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 (→ Proof 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 (→ Proof I/2.5.5) in the pair of probability distributions (→ Definition I/1.3.1) (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

$$\begin{aligned} \mathrm{KL}[\lambda p + (1-\lambda) p || \lambda q_1 + (1-\lambda) q_2] &\leq \lambda \mathrm{KL}[p||q_1] + (1-\lambda) \mathrm{KL}[p||q_2] \\ \mathrm{KL}[p || \lambda q_1 + (1-\lambda) q_2] &\leq \lambda \mathrm{KL}[p||q_1] + (1-\lambda) \mathrm{KL}[p||q_2] \end{aligned} \tag{4}$$

and applying equation (2), we have

$$\begin{aligned} \mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] - \mathrm{H}[p] &\leq \lambda \left(\mathrm{H}[p, q_1] - \mathrm{H}[p] \right) + (1-\lambda) \left(\mathrm{H}[p, q_2] - \mathrm{H}[p] \right) \\ \mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] - \mathrm{H}[p] &\leq \lambda \mathrm{H}[p, q_1] + (1-\lambda) \mathrm{H}[p, q_2] - \mathrm{H}[p] \\ \mathrm{H}[p, \lambda q_1 + (1-\lambda) q_2] &\leq \lambda \mathrm{H}[p, q_1] + (1-\lambda) \mathrm{H}[p, q_2] \end{aligned} \tag{5}$$

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.

Metadata: ID: P150 | shortcut: entcross-conv | author: JoramSoch | date: 2020-08-11, 09:16.

2.1.8 Gibbs' inequality

Theorem: Let X be a discrete random variable (→ Definition I/1.1.3) and consider two probability distributions (→ Definition I/1.3.1) with probability mass functions (→ Definition I/1.4.1) $p(x)$ and $q(x)$. Then, Gibbs' inequality states that the entropy (→ Definition I/2.1.1) of X according to P is smaller than or equal to the cross-entropy (→ Definition 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 of all x for which $p(x)$ is non-zero. Then, proving (1) requires to show that

$$\sum_{x \in I} p(x) \, \ln \frac{p(x)}{q(x)} \geq 0 \; . \tag{3}$$

Because $\ln x \leq x - 1$, i.e. $-\ln x \geq 1 - x$, for all $x > 0$, with equality only if $x = 1$, we can say about the left-hand side that

$$\begin{aligned} \sum_{x \in I} p(x) \, \ln \frac{p(x)}{q(x)} &\geq \sum_{x \in I} p(x) \left(1 - \frac{p(x)}{q(x)} \right) \\ &= \sum_{x \in I} p(x) - \sum_{x \in I} q(x) \; . \end{aligned} \tag{4}$$

Finally, since $p(x)$ and $q(x)$ are probability mass functions (→ Definition I/1.4.1), we have

$$
\begin{aligned}
0 \leq p(x) \leq 1, \quad & \sum_{x \in I} p(x) = 1 \quad \text{and} \\
0 \leq q(x) \leq 1, \quad & \sum_{x \in I} q(x) \leq 1 \; ,
\end{aligned} \tag{5}
$$

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.

Metadata: ID: P164 | shortcut: gibbs-ineq | author: JoramSoch | date: 2020-09-09, 02:18.

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

$$
\begin{split}
\sum_{i=1}^n a_i \, \ln \frac{a_i}{b_i} &= \sum_{i=1}^n b_i \, f\left(\frac{a_i}{b_i} \right) \\
&= b \sum_{i=1}^n \frac{b_i}{b} \, f\left(\frac{a_i}{b_i} \right) \; .
\end{split} \tag{3}
$$

Because $f(x)$ is a convex function and

$$
\begin{split}
\frac{b_i}{b} &\geq 0 \\
\sum_{i=1}^n \frac{b_i}{b} &= 1 \; ,
\end{split} \tag{4}
$$

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.

Metadata: ID: P165 | shortcut: logsum-ineq | author: JoramSoch | date: 2020-09-09, 02:46.

2.2 Differential entropy

2.2.1 Definition

Definition: Let X be a continuous random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and the (estimated or assumed) probability density function (→ Definition I/1.4.4) $p(x) = f_X(x)$. 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.

Metadata: ID: D16 | shortcut: dent | author: JoramSoch | date: 2020-02-19, 17:53.

2.2.2 Negativity

Theorem: Unlike its discrete analogue (→ Proof I/2.1.2), the differential entropy (→ Definition I/2.2.1) can become negative.

Proof: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition 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 (→ Proof II/3.1.2) is:

$$f_X(x) = 2 \quad \text{for} \quad 0 \leq x \leq \frac{1}{2} \; . \tag{2}$$

Thus, the differential entropy (→ Definition I/2.2.1) follows as

$$\begin{aligned}
\mathrm{h}(X) &= -\int_{\mathcal{X}} f_X(x) \log_b f_X(x) \, \mathrm{d}x \\
&= -\int_0^{\frac{1}{2}} 2 \, \log_b(2) \, \mathrm{d}x \\
&= -\log_b(2) \int_0^{\frac{1}{2}} 2 \, \mathrm{d}x \\
&= -\log_b(2) \, [2x]_0^{\frac{1}{2}} \\
&= -\log_b(2)
\end{aligned} \tag{3}$$

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.

Metadata: ID: P68 | shortcut: dent-neg | author: JoramSoch | date: 2020-03-02, 20:32.

2.2.3 Invariance under addition

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3). Then, the differential entropy (→ Definition 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 (→ Definition 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 (→ Definition I/1.4.4) 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 (→ Proof I/1.4.5) 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:

$$\begin{aligned} \mathrm{h}(Y) &= -\int_{\mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y \\ &\overset{(4)}{=} -\int_{\mathcal{Y}} f_X(y-c) \log f_X(y-c) \, \mathrm{d}y \end{aligned} \tag{5}$$

Substituting $x = y - c$, such that $y = x + c$, this yields:

$$\begin{aligned} \mathrm{h}(Y) &= -\int_{\{y-c \,|\, y \in \mathcal{Y}\}} f_X(x+c-c) \log f_X(x+c-c) \, \mathrm{d}(x+c) \\ &= -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x \\ &\overset{(2)}{=} \mathrm{h}(X) \; . \end{aligned} \tag{6}$$

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.

Metadata: ID: P199 | shortcut: dent-inv | author: JoramSoch | date: 2020-12-02, 16:11.

2.2.4 Addition upon multiplication

Theorem: Let X be a continuous (→ Definition I/1.1.7) random variable (→ Definition I/1.1.3). Then, the differential entropy (→ Definition 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 (→ Definition 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 (→ Definition I/1.4.4) 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 (→ Proof I/1.4.5) 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 (→ Proof I/1.4.6) 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:

$$\begin{aligned} \mathrm{h}(Y) &= -\int_{\mathcal{Y}} f_Y(y) \log f_Y(y) \, \mathrm{d}y \\ &\overset{(6)}{=} -\int_{\mathcal{Y}} \frac{1}{|a|} \, f_X\left(\frac{y}{a}\right) \log\left[\frac{1}{|a|} \, f_X\left(\frac{y}{a}\right)\right] \, \mathrm{d}y \end{aligned} \tag{7}$$

Substituting $x = y/a$, such that $y = ax$, this yields:

$$\begin{aligned} \mathrm{h}(Y) &= -\int_{\{y/a \,|\, y \in \mathcal{Y}\}} \frac{1}{|a|} \, f_X\left(\frac{ax}{a}\right) \log\left[\frac{1}{|a|} \, f_X\left(\frac{ax}{a}\right)\right] \, \mathrm{d}(ax) \\ &= -\int_{\mathcal{X}} f_X(x) \log\left[\frac{1}{|a|} \, f_X(x)\right] \, \mathrm{d}x \\ &= -\int_{\mathcal{X}} f_X(x) \left[\log f_X(x) - \log|a|\right] \, \mathrm{d}x \\ &= -\int_{\mathcal{X}} f_X(x) \log f_X(x) \, \mathrm{d}x + \log|a| \int_{\mathcal{X}} f_X(x) \, \mathrm{d}x \\ &\overset{(2)}{=} \mathrm{h}(X) + \log|a| \; . \end{aligned} \tag{8}$$

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.

Metadata: ID: P200 | shortcut: dent-add | author: JoramSoch | date: 2020-12-02, 16:39.

2.2.5 Conditional differential entropy

Definition: Let X and Y be continuous random variables (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and probability density functions (→ Definition I/1.4.4) $p(x)$ and $p(y)$. Then, the conditional differential entropy of Y given X or, differential entropy of Y conditioned on X, is defined as

$$\mathrm{h}(Y|X) = \int_{x \in \mathcal{X}} p(x) \cdot \mathrm{h}(Y|X = x) \tag{1}$$

where $\mathrm{h}(Y|X = x)$ is the (marginal) differential entropy (→ Definition I/2.2.1) of Y, evaluated at x.

Sources:
- original work

Metadata: ID: D34 | shortcut: dent-cond | author: JoramSoch | date: 2020-03-21, 12:27.

2.2.6 Joint differential entropy

Definition: Let X and Y be continuous random variables (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and $\mathcal{Y}$ and joint probability (→ Definition I/1.2.2) density function (→ Definition I/1.4.4) $p(x,y)$. 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.

Sources:
- original work

Metadata: ID: D35 | shortcut: dent-joint | author: JoramSoch | date: 2020-03-21, 12:37.

2.2.7 Differential cross-entropy

Definition: Let X be a continuous random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ Definition I/1.3.1) on X with the probability density functions (→ Definition I/1.4.4) $p(x)$ 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.

Metadata: ID: D86 | shortcut: dent-cross | author: JoramSoch | date: 2020-07-28, 03:03.

2.3 Discrete mutual information

2.3.1 Definition

Definition:
1) The mutual information of two discrete random variables (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition I/1.2.2) mass function of X and Y.

2) The mutual information of two continuous random variables (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition I/1.2.2) 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.

Metadata: ID: D19 | shortcut: mi | author: JoramSoch | date: 2020-02-19, 18:35.

2.3.2 Relation to marginal and conditional entropy

Theorem: Let X and Y be discrete random variables (→ Definition I/1.1.3) with the joint probability (→ Definition I/1.2.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition 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 entropies (→ Definition I/2.1.1) of X and Y and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies (→ Definition I/2.1.4).

Proof: The mutual information (→ Definition 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 (→ Definition I/1.2.4), 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 (→ Definition I/1.2.3), 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 (→ Definition I/2.1.1) and conditional (→ Definition 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.

Metadata: ID: P19 | shortcut: dmi-mce | author: JoramSoch | date: 2020-01-13, 18:20.

2.3.3 Relation to marginal and joint entropy

Theorem: Let X and Y be discrete random variables (→ Definition I/1.1.3) with the joint probability (→ Definition I/1.2.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition 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 (→ Definition I/2.1.1) of X and Y and $\mathrm{H}(X,Y)$ is the joint entropy (→ Definition I/2.1.5).

Proof: The mutual information (→ Definition 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:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \left(\sum_y p(x,y) \right) \log p(x) - \sum_y \left(\sum_x p(x,y) \right) \log p(y) \; . \quad (4)$$

Applying the law of marginal probability (→ Definition I/1.2.3), i.e. $p(x) = \sum_y p(x,y)$, we get:

$$\mathrm{I}(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x p(x) \log p(x) - \sum_y p(y) \log p(y) \; . \quad (5)$$

Now considering the definitions of marginal (→ Definition I/2.1.1) and joint (→ Definition I/2.1.5) entropy

$$\begin{aligned} \mathrm{H}(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ \mathrm{H}(X,Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y) \; , \end{aligned} \quad (6)$$

we can finally show:

$$\begin{aligned} \mathrm{I}(X,Y) &= -\mathrm{H}(X,Y) + \mathrm{H}(X) + \mathrm{H}(Y) \\ &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \; . \end{aligned} \quad (7)$$

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.

Metadata: ID: P20 | shortcut: dmi-mje | author: JoramSoch | date: 2020-01-13, 21:53.

2.3.4 Relation to joint and conditional entropy

Theorem: Let X and Y be discrete random variables (→ Definition I/1.1.3) with the joint probability (→ Definition I/1.2.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition I/2.4.1) of X and Y can be expressed as

$$\mathrm{I}(X,Y) = \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \quad (1)$$

where $\mathrm{H}(X,Y)$ is the joint entropy (→ Definition I/2.1.5) of X and Y and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are the conditional entropies (→ Definition I/2.1.4).

Proof: The existence of the joint probability mass function (→ Definition I/1.4.1) ensures that the mutual information (→ Definition I/2.4.1) is defined:

$$\mathrm{I}(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \; . \quad (2)$$

The relation of mutual information to conditional entropy (→ Proof 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 (→ Proof 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

$$\begin{split} \mathrm{I}(X,Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) + \mathrm{H}(Y) - \mathrm{H}(Y|X) - \mathrm{H}(X) - \mathrm{H}(Y) + \mathrm{H}(X,Y) \\ &= \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X) \end{split} \tag{7}$$

which proves the identity given above.

Sources:
- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-01-13; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P21 | shortcut: dmi-jce | author: JoramSoch | date: 2020-01-13, 22:17.

2.4 Continuous mutual information

2.4.1 Definition

Definition:
1) The mutual information of two discrete random variables (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition I/1.2.2) mass function of X and Y.
2) The mutual information of two continuous random variables (→ Definition I/1.1.3) 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 (→ Definition I/1.4.1) of X and Y and $p(x,y)$ is the joint probability (→ Definition I/1.2.2) 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.

Metadata: ID: D19 | shortcut: mi | author: JoramSoch | date: 2020-02-19, 18:35.

2.4.2 Relation to marginal and conditional differential entropy

Theorem: Let X and Y be continuous random variables (→ Definition I/1.1.3) with the joint probability (→ Definition I/1.2.2) $p(x, y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition 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 (→ Definition I/2.2.1) of X and Y and $\mathrm{h}(X|Y)$ and $\mathrm{h}(Y|X)$ are the conditional differential entropies (→ Definition I/2.2.5).

Proof: The mutual information (→ Definition 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 (→ Definition I/1.2.4), 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 (→ Definition I/1.2.3), 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 (→ Definition I/2.2.1) and conditional (→ Definition I/2.2.5) differential entropy

$$\begin{aligned} \mathrm{h}(X) &= -\int_{\mathcal{X}} p(x) \log p(x) \, \mathrm{d}x \\ \mathrm{h}(X|Y) &= \int_{\mathcal{Y}} p(y) \, \mathrm{h}(X|Y = y) \, \mathrm{d}y \; , \end{aligned} \tag{7}$$

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.

Metadata: ID: P58 | shortcut: cmi-mcde | author: JoramSoch | date: 2020-02-21, 16:53.

2.4.3 Relation to marginal and joint differential entropy

Theorem: Let X and Y be continuous random variables (→ Definition I/1.1.3) with the joint probability (→ Definition I/1.2.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition 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 (→ Definition I/2.2.1) of X and Y and $\mathrm{h}(X,Y)$ is the joint differential entropy (→ Definition I/2.2.6).

Proof: The mutual information (→ Definition 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 (→ Definition I/1.2.3), 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 (→ Definition I/2.2.1) and joint (→ Definition I/2.2.6) 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.

Metadata: ID: P59 | shortcut: cmi-mjde | author: JoramSoch | date: 2020-02-21, 17:13.

2.4.4 Relation to joint and conditional differential entropy

Theorem: Let X and Y be continuous random variables (→ Definition I/1.1.3) with the joint probability (→ Definition I/1.2.2) $p(x,y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, the mutual information (→ Definition 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 (→ Definition I/2.2.6) of X and Y and $\mathrm{h}(X|Y)$ and $\mathrm{h}(Y|X)$ are the conditional differential entropies (→ Definition I/2.2.5).

Proof: The existence of the joint probability density function (→ Definition I/1.4.4) ensures that the mutual information (→ Definition 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 (→ Proof 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 (→ Proof 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

$$\begin{aligned}
\mathrm{I}(X,Y) &= \mathrm{h}(X) - \mathrm{h}(X|Y) + \mathrm{h}(Y) - \mathrm{h}(Y|X) - \mathrm{h}(X) - \mathrm{h}(Y) + \mathrm{h}(X,Y) \\
&= \mathrm{h}(X,Y) - \mathrm{h}(X|Y) - \mathrm{h}(Y|X)
\end{aligned} \tag{7}$$

which proves the identity given above.

Sources:
- Wikipedia (2020): "Mutual information"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-02-21; URL: https://en.wikipedia.org/wiki/Mutual_information#Relation_to_conditional_and_joint_entropy.

Metadata: ID: P60 | shortcut: cmi-jcde | author: JoramSoch | date: 2020-02-21, 17:23.

2.5 Kullback-Leibler divergence

2.5.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ Definition I/1.3.1) 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 (→ Definition I/1.4.1) 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 (→ Definition I/1.4.4) of 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.

Metadata: ID: D52 | shortcut: kl | author: JoramSoch | date: 2020-05-10, 20:20.

2.5.2 Non-negativity

Theorem: The Kullback-Leibler divergence (→ Definition 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 (→ Definition 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 (→ Proof I/2.1.8) states that the entropy (→ Definition I/2.1.1) of a probability distribution is always less than or equal to the cross-entropy (→ Definition 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.

Metadata: ID: P117 | shortcut: kl-nonneg | author: JoramSoch | date: 2020-05-31, 23:43.

2.5.3 Non-negativity

Theorem: The Kullback-Leibler divergence (→ Definition 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 (→ Definition 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 (→ Proof 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 (→ Definition I/1.4.1), such that

$$
\begin{aligned}
p(x) \geq 0, \quad \sum_{x \in \mathcal{X}} p(x) = 1 \quad \text{and} \\
q(x) \geq 0, \quad \sum_{x \in \mathcal{X}} q(x) = 1 \; ,
\end{aligned} \tag{4}
$$

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.

Metadata: ID: P166 | shortcut: kl-nonneg2 | author: JoramSoch | date: 2020-09-09, 07:02.

2.5.4 Non-symmetry

Theorem: The Kullback-Leibler divergence (→ Definition I/2.5.1) is non-symmetric, i.e.

$$
\mathrm{KL}[P||Q] \neq \mathrm{KL}[Q||P] \tag{1}
$$

for some [probability distributions](dist) P and Q.

Proof: Let $X \in \mathcal{X} = \{0, 1, 2\}$ be a discrete random variable (→ Definition I/1.1.3) and consider the two [probability distributions](dist)

$$
\begin{aligned}
P &: \; X \sim \mathrm{Bin}(2, 0.5) \\
Q &: \; X \sim \mathcal{U}(0, 2)
\end{aligned} \tag{2}
$$

where $\mathrm{Bin}(n, p)$ indicates a binomial distribution (→ Definition II/1.3.1) and $\mathcal{U}(a, b)$ indicates a discrete uniform distribution (→ Definition II/1.1.1).
Then, the probability mass function of the binomial distribution (→ Proof II/1.3.2) entails that

$$
p(x) = \left\{
\begin{array}{rl}
1/4 \; , & \text{if} \; x = 0 \\
1/2 \; , & \text{if} \; x = 1 \\
1/4 \; , & \text{if} \; x = 2
\end{array}
\right. \tag{3}
$$

and the probability mass function of the discrete uniform distribution (→ Proof II/1.1.2) entails that

$$
q(x) = \frac{1}{3} \; , \tag{4}
$$

such that the Kullback-Leibler divergence (→ Definition I/2.5.1) of P from Q is

$$
\begin{split}
\mathrm{KL}[P||Q] &= \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} \\
&= \frac{1}{4} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} + \frac{1}{4} \log \frac{3}{4} \\
&= \frac{1}{2} \log \frac{3}{4} + \frac{1}{2} \log \frac{3}{2} \\
&= \frac{1}{2} \left(\log \frac{3}{4} + \log \frac{3}{2} \right) \\
&= \frac{1}{2} \log \left(\frac{3}{4} \cdot \frac{3}{2} \right) \\
&= \frac{1}{2} \log \frac{9}{8} = 0.0589
\end{split} \tag{5}
$$

and the Kullback-Leibler divergence (→ Definition I/2.5.1) of Q from P is

$$
\begin{split}
\mathrm{KL}[Q||P] &= \sum_{x \in \mathcal{X}} q(x) \cdot \log \frac{q(x)}{p(x)} \\
&= \frac{1}{3} \log \frac{4}{3} + \frac{1}{3} \log \frac{2}{3} + \frac{1}{3} \log \frac{4}{3} \\
&= \frac{1}{3} \left(\log \frac{4}{3} + \log \frac{2}{3} + \log \frac{4}{3} \right) \\
&= \frac{1}{3} \log \left(\frac{4}{3} \cdot \frac{2}{3} \cdot \frac{4}{3} \right) \\
&= \frac{1}{3} \log \frac{32}{27} = 0.0566
\end{split} \tag{6}
$$

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.

Metadata: ID: P147 | shortcut: kl-nonsymm | author: JoramSoch | date: 2020-08-11, 06:57.

2.5.5 Convexity

Theorem: The Kullback-Leibler divergence (→ Definition I/2.5.1) is convex in the pair of probability distributions (→ Definition I/1.3.1) (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 (→ Definition 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 (→ Proof 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.

Metadata: ID: P148 | shortcut: kl-conv | author: JoramSoch | date: 2020-08-11, 07:30.

2.5.6 Additivity for independent distributions

Theorem: The Kullback-Leibler divergence (→ Definition 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 (→ Definition I/1.2.6) distributions (→ Definition I/1.3.1) with the joint distribution (→ Definition I/1.3.2) 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 (→ Definition 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.

Metadata: ID: P116 | shortcut: kl-add | author: JoramSoch | date: 2020-05-31, 23:26.

2.5.7 Invariance under parameter transformation

Theorem: The Kullback-Leibler divergence (→ Definition 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 (→ Definition I/1.4.4) of the probability distributions (→ Definition I/1.3.1) P and Q on the continuous random variable (→ Definition I/1.1.3) X.

Proof: The continuous Kullback-Leibler divergence (→ Definition 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:

$$
\begin{split}
\mathrm{KL}[p(x)||q(x)] &= \int_a^b p(x) \cdot \log \frac{p(x)}{q(x)}\,\mathrm{d}x \\
&= \int_{y(a)}^{y(b)} p(y)\,\frac{\mathrm{d}y}{\mathrm{d}x} \cdot \log \left(\frac{p(y)\,\frac{\mathrm{d}y}{\mathrm{d}x}}{q(y)\,\frac{\mathrm{d}y}{\mathrm{d}x}} \right) \mathrm{d}x \\
&= \int_{y(a)}^{y(b)} p(y) \cdot \log \frac{p(y)}{q(y)}\,\mathrm{d}y \\
&= \mathrm{KL}[p(y)||q(y)] \; .
\end{split} \tag{5}
$$

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

Metadata: ID: P115 | shortcut: kl-inv | author: JoramSoch | date: 2020-05-28, 00:18.

2.5.8 Relation to discrete entropy

Theorem: Let X be a discrete random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ Definition I/1.3.1) on X. Then, the Kullback-Leibler divergence (→ Definition 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 (→ Definition I/2.1.6) of P and Q and $\mathrm{H}(P)$ is the marginal entropy (→ Definition I/2.1.1) of P.

Proof: The discrete Kullback-Leibler divergence (→ Definition 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 (→ Definition I/1.4.1) 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 (→ Definition I/2.1.1) and cross-entropy (→ Definition 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.

Metadata: ID: P113 | shortcut: kl-ent | author: JoramSoch | date: 2020-05-27, 23:20.

2.5.9 Relation to differential entropy

Theorem: Let X be a continuous random variable (→ Definition I/1.1.3) with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions (→ Definition I/1.3.1) on X. Then, the Kullback-Leibler divergence (→ Definition 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 (→ Definition I/2.2.7) of P and Q and $\mathrm{h}(P)$ is the marginal differential entropy (→ Definition I/2.2.1) of P.

Proof: The continuous Kullback-Leibler divergence (→ Definition 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 (→ Definition I/1.4.4) 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 (→ Definition I/2.2.1) and differential cross-entropy (→ Definition I/2.2.7)

$$\begin{aligned} \mathrm{h}(P) &= -\int_{\mathcal{X}} p(x)\, \log p(x)\, \mathrm{d}x \\ \mathrm{h}(P,Q) &= -\int_{\mathcal{X}} p(x)\, \log q(x)\, \mathrm{d}x \;, \end{aligned} \tag{4}$$

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.

Metadata: ID: P114 | shortcut: kl-dent | author: JoramSoch | date: 2020-05-27, 23:32.

3 Estimation theory

3.1 Point estimates

3.1.1 Partition of the mean squared error into bias and variance

Theorem: The mean squared error (→ Definition "mse") 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 (→ Definition I/1.6.1) is given by

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] \tag{2}$$

and the bias (→ Definition "bias") is given by

$$\mathrm{Bias}(\hat{\theta}, \theta) = \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) \; . \tag{3}$$

Proof: The mean squared error (MSE) is defined as (→ Definition "mse") the expected value (→ Definition I/1.5.1) of the squared deviation of the estimated value $\hat{\theta}$ from the true value θ of a parameter, over all values $\hat{\theta}$:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \; . \tag{4}$$

This formula can be evaluated in the following way:

$$\begin{split}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta}) + \mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2 + 2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \mathbb{E}_{\hat{\theta}}\left[2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\right] + \mathbb{E}_{\hat{\theta}}\left[\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \; .
\end{split} \tag{5}$$

Because $\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta$ is constant as a function of $\hat{\theta}$, we have:

$$\begin{split}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\mathbb{E}_{\hat{\theta}}\left[\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \; .
\end{split} \tag{6}$$

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.

Metadata: ID: P5 | shortcut: mse-bnv | author: JoramSoch | date: 2019-11-27, 14:26.

3.2 Interval estimates

3.2.1 Construction of confidence intervals using Wilks' theorem

Theorem: Let m be a generative model (→ Definition I/5.1.1) for measured data y with model parameters θ, consisting of a parameter of interest ϕ and nuisance parameters λ:

$$m : p(y|\theta) = \mathcal{D}(y;\theta), \quad \theta = \{\phi, \lambda\} \; . \tag{1}$$

Further, let $\hat{\theta}$ be an estimate of θ, obtained using maximum-likelihood-estimation (→ Definition I/4.1.3):

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta), \quad \hat{\theta} = \left\{ \hat{\phi}, \hat{\lambda} \right\} \; . \tag{2}$$

Then, an asymptotic confidence interval (→ Definition "ci") 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 (→ Definition II/3.5.1) with 1 degree of freedom (→ Definition "dof").

Proof: The confidence interval (→ Definition "ci") is defined as the interval that, under infinitely repeated random experiments (→ Definition I/1.1.1), contains the true parameter value with a certain probability.
Let us define the likelihood ratio (→ Definition "lr")

$$\Lambda(\phi) = \frac{p(y|\phi,\hat{\lambda})}{p(y|\hat{\phi},\hat{\lambda})} \tag{4}$$

and compute the log-likelihood ratio (→ Definition "llr")

$$\log \Lambda(\phi) = \log p(y|\phi,\hat{\lambda}) - \log p(y|\hat{\phi},\hat{\lambda}) \; . \tag{5}$$

[Wilks' theorem](llr-wilks) 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 (→ Definition II/3.5.1), if the null hypothesis is true:

$$H_0 : \theta \in \Theta_0 \quad \Rightarrow \quad -2 \log \frac{\max_{\theta \in \Theta_0} p(y|\theta)}{\max_{\theta \in \Theta_1} p(y|\theta)} \sim \chi^2_{\Delta k} \tag{6}$$

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 (→ Definition I/1.4.13) $\chi^2_{k,p}$ of the chi-squared distribution (→ Definition II/3.5.1), 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.

Metadata: ID: P56 | shortcut: ci-wilks | author: JoramSoch | date: 2020-02-19, 17:15.

4 Frequentist statistics

4.1 Likelihood theory

4.1.1 Likelihood function

Definition: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ. Then, the probability density function (→ Definition I/1.4.4) of the distribution of y given θ is called the likelihood function of m:

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D28 | shortcut: lf | author: JoramSoch | date: 2020-03-03, 15:50.

4.1.2 Log-likelihood function

Definition: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ. Then, the logarithm of the probability density function (→ Definition I/1.4.4) of the distribution of y given θ is called the log-likelihood function (→ Definition I/5.1.2) of m:

$$\mathrm{LL}_m(\theta) = \log p(y|\theta, m) = \log \mathcal{D}(y; \theta) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D59 | shortcut: llf | author: JoramSoch | date: 2020-05-17, 22:52.

4.1.3 Maximum likelihood estimation

Definition: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ. Then, the parameter values maximizing the likelihood function (→ Definition I/5.1.2) or log-likelihood function (→ Definition 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".

Sources:
- original work

Metadata: ID: D60 | shortcut: mle | author: JoramSoch | date: 2020-05-15, 23:05.

4.1.4 Maximum log-likelihood

Definition: Let there be a generative model (→ Definition 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 (→ Definition 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 (→ Definition I/4.1.3) into the log-likelihood function (→ Definition I/4.1.2).

Sources:
- original work

Metadata: ID: D61 | shortcut: mll | author: JoramSoch | date: 2020-05-15, 23:13.

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 of y given θ is called a generative model m:

$$m : \; y \sim \mathcal{D}(\theta) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D27 | shortcut: gm | author: JoramSoch | date: 2020-03-03, 15:50.

5.1.2 Likelihood function

Definition: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ. Then, the probability density function (→ Definition I/1.4.4) of the distribution of y given θ is called the likelihood function of m:

$$\mathcal{L}_m(\theta) = p(y|\theta, m) = \mathcal{D}(y; \theta) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D28 | shortcut: lf | author: JoramSoch | date: 2020-03-03, 15:50.

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 (→ Definition I/1.4.4) is called the prior density:

$$p(\theta|m) = \mathcal{D}(\theta; \lambda) \; . \tag{2}$$

Sources:
- original work

Metadata: ID: D29 | shortcut: prior | author: JoramSoch | date: 2020-03-03, 16:09.

5.1.4 Full probability model

Definition: Consider measured data y and some unknown latent parameters θ. The combination of a generative model (→ Definition I/5.1.1) for y and a prior distribution (→ Definition 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:
- original work

Metadata: ID: D30 | shortcut: fpm | author: JoramSoch | date: 2020-03-03, 16:16.

5.1.5 Joint likelihood

Definition: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (→ Definition I/5.1.3) on θ. Then, the joint probability (→ Definition I/1.2.2) density function (→ Definition I/1.4.4) of y and θ is called the joint likelihood:

$$p(y,\theta|m) = p(y|\theta,m) \, p(\theta|m) \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D31 | shortcut: jl | author: JoramSoch | date: 2020-03-03, 16:36.

5.1.6 Joint likelihood is product of likelihood and prior

Theorem: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (→ Definition I/5.1.3) on θ. Then, the joint likelihood (→ Definition I/5.1.5) is equal to the product of likelihood function (→ Definition I/5.1.2) and prior density (→ Definition I/5.1.3):

$$p(y,\theta|m) = p(y|\theta,m) \, p(\theta|m) \; . \tag{1}$$

Proof: The joint likelihood (→ Definition I/5.1.5) is defined as the joint probability (→ Definition I/1.2.2) density function (→ Definition I/1.4.4) of data y and parameters θ:

$$p(y,\theta|m) \; . \tag{2}$$

Applying the law of conditional probability (→ Definition I/1.2.4), we have:

$$\begin{gathered} p(y|\theta,m) = \frac{p(y,\theta|m)}{p(\theta|m)} \\ \Leftrightarrow \\ p(y,\theta|m) = p(y|\theta,m) \, p(\theta|m) \; . \end{gathered} \tag{3}$$

Sources:
- original work

Metadata: ID: P89 | shortcut: jl-lfnprior | author: JoramSoch | date: 2020-05-05, 04:21.

5.1.7 Posterior distribution

Definition: Consider measured data y and some unknown latent parameters θ. The distribution of θ conditional 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 (→ Definition I/1.4.4) is called the posterior density:

$$p(\theta|y,m) = \mathcal{D}(\theta;\phi) \; . \tag{2}$$

Sources:
- original work

Metadata: ID: D32 | shortcut: post | author: JoramSoch | date: 2020-03-03, 16:43.

5.1.8 Posterior density is proportional to joint likelihood

Theorem: In a full probability model (→ Definition I/5.1.4) m describing measured data y using model parameters θ, the posterior density (→ Definition I/5.1.7) over the model parameters is proportional to the joint likelihood (→ Definition I/5.1.5):

$$p(\theta|y,m) \propto p(y,\theta|m) \; . \tag{1}$$

Proof: In a full probability model (→ Definition I/5.1.4), the posterior distribution (→ Definition I/5.1.7) can be expressed using Bayes' theorem (→ Proof 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 (→ Definition I/1.2.4) 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:

$$p(\theta|y,m) \propto p(y,\theta|m) \; . \tag{4}$$

Sources:
- original work

Metadata: ID: P90 | shortcut: post-jl | author: JoramSoch | date: 2020-05-05, 04:46.

5.1.9 Marginal likelihood

Definition: Let there be a generative model (→ Definition I/5.1.1) m describing measured data y using model parameters θ and a prior distribution (→ Definition I/5.1.3) on θ. Then, the marginal probability (→ Definition I/1.2.3) density function (→ Definition I/1.4.4) of y across the parameter space Θ is called the marginal likelihood:

$$p(y|m) = \int_{\Theta} p(y|\theta,m)\, p(\theta|m)\, \mathrm{d}\theta \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D33 | shortcut: ml | author: JoramSoch | date: 2020-03-03, 16:49.

5.1.10 Marginal likelihood is integral of joint likelihood

Theorem: In a full probability model (→ Definition I/5.1.4) m describing measured data y using model parameters θ, the marginal likelihood (→ Definition I/5.1.9) is the integral of the joint likelihood (→ Definition I/5.1.5) across the parameter space Θ:

$$p(y|m) = \int_{\Theta} p(y,\theta|m)\, \mathrm{d}\theta \; . \tag{1}$$

Proof: In a full probability model (→ Definition I/5.1.4), the marginal likelihood (→ Definition I/5.1.9) is defined as the marginal probability (→ Definition I/1.2.3) of the data y, given only the model m:

$$p(y|m) \; . \tag{2}$$

Using the law of marginal probabililty (→ Definition I/1.2.3), this can be obtained by integrating over the product of likelihood function (→ Definition I/5.1.2) and prior density (→ Definition I/5.1.3):

$$p(y|m) = \int_{\Theta} p(y|\theta,m)\, p(\theta|m)\, \mathrm{d}\theta \; . \tag{3}$$

Applying the law of conditional probability (→ Definition I/1.2.4) to the integrand, we have:

$$p(y|m) = \int_{\Theta} p(y,\theta|m)\, \mathrm{d}\theta \; . \tag{4}$$

Sources:
- original work

Metadata: ID: P91 | shortcut: ml-jl | author: JoramSoch | date: 2020-05-05, 04:59.

5.2 Prior distributions

5.2.1 Flat vs. hard vs. soft

Definition: Let $p(\theta|m)$ be a prior distribution (→ Definition I/5.1.3) for the parameter θ of a generative model (→ Definition I/5.1.1) m. Then,

- the distribution is called a "flat prior", if its precision (→ Definition "prec") is zero or variance (→ Definition I/1.6.1) is infinite;
- the distribution is called a "hard prior", if its precision (→ Definition "prec") is infinite or variance (→ Definition I/1.6.1) is zero;
- the distribution is called a "soft prior", if its precision (→ Definition "prec") and variance (→ Definition I/1.6.1) 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.

Metadata: ID: D116 | shortcut: prior-flat | author: JoramSoch | date: 2020-12-02, 17:04.

5.2.2 Uniform vs. non-uniform

Definition: Let $p(\theta|m)$ be a prior distribution (→ Definition I/5.1.3) for the parameter θ of a generative model (→ Definition I/5.1.1) m where θ belongs to the parameter space Θ. Then,

- the distribution is called a "uniform prior", if its density (→ Definition I/1.4.4) or mass (→ Definition I/1.4.1) is constant over Θ;
- the distribution is called a "non-uniform prior", if its density (→ Definition I/1.4.4) or mass (→ Definition I/1.4.1) 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.

Metadata: ID: D117 | shortcut: prior-uni | author: JoramSoch | date: 2020-12-02, 17:21.

5.2.3 Informative vs. non-informative

Definition: Let $p(\theta|m)$ be a prior distribution (→ Definition I/5.1.3) for the parameter θ of a generative model (→ Definition 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 (→ Proof I/5.1.8);
- the distribution is called a "non-informative prior", if it does not influence (→ Proof I/5.1.8) the posterior hyperparameters (→ Definition 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.

Metadata: ID: D118 | shortcut: prior-inf | author: JoramSoch | date: 2020-12-02, 17:28.

5.2.4 Empirical vs. non-empirical

Definition: Let $p(\theta|m)$ be a prior distribution (→ Definition I/5.1.3) for the parameter θ of a generative model (→ Definition I/5.1.1) m. Then,
- the distribution is called an "empirical prior", if it has been derived from empirical data (→ Proof I/5.1.8);
- 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.

Metadata: ID: D119 | shortcut: prior-emp | author: JoramSoch | date: 2020-12-02, 17:37.

5.2.5 Conjugate vs. non-conjugate

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|m)$. Then,
- the prior distribution (→ Definition I/5.1.3) is called "conjugate", if it, when combined with the likelihood function (→ Definition I/5.1.2), leads to a posterior distribution (→ Definition I/5.1.7) that belongs to the same family of probability distributions (→ Definition I/1.3.1);
- 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.

Metadata: ID: D120 | shortcut: prior-conj | author: JoramSoch | date: 2020-12-02, 17:55.

5.2.6 Maximum entropy priors

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|\lambda, m)$ using prior hyperparameters (→ Definition I/5.1.3) λ. Then, the prior distribution is called a "maximum entropy prior", if

1) when θ is a discrete random variable (→ Definition I/1.1.7), it maximizes the entropy (→ Definition I/2.1.1) of the prior probability mass function (→ Definition I/1.4.1):

$$\lambda_{\mathrm{maxent}} = \operatorname*{arg\,max}_{\lambda} \mathrm{H}\left[p(\theta|\lambda,m)\right] \; ; \tag{1}$$

2) when θ is a continuous random variable (→ Definition I/1.1.7), it maximizes the differential entropy (→ Definition I/2.2.1) of the prior probability density function (→ Definition I/1.4.4):

$$\lambda_{\mathrm{maxent}} = \operatorname*{arg\,max}_{\lambda} \mathrm{h}\left[p(\theta|\lambda,m)\right] \; . \tag{2}$$

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.

Metadata: ID: D121 | shortcut: prior-maxent | author: JoramSoch | date: 2020-12-02, 18:13.

5.2.7 Empirical Bayes priors

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta,m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|\lambda,m)$ using prior hyperparameters (→ Definition I/5.1.3) λ. Let $p(y|\lambda,m)$ be the marginal likelihood (→ Definition I/5.1.9) when integrating the parameters out of the joint likelihood (→ Proof I/5.1.10). Then, the prior distribution is called an "Empirical Bayes 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.

Metadata: ID: D122 | shortcut: prior-eb | author: JoramSoch | date: 2020-12-02, 18:19.

5.2.8 Reference priors

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta,m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|\lambda,m)$ using prior hyperparameters (→ Definition I/5.1.3) λ. Let $p(\theta|y,\lambda,m)$ be the posterior distribution (→ Definition I/5.1.7) that is proportional to the the joint likelihood (→ Proof I/5.1.8). Then, the prior distribution is called a "reference prior", if it maximizes the expected (→ Definition I/1.5.1) Kullback-Leibler divergence (→ Definition I/2.5.1) of the posterior distribution relative to the prior distribution:

$$\lambda_{\mathrm{ref}} = \operatorname*{arg\,max}_{\lambda} \left\langle \mathrm{KL}\left[p(\theta|y,\lambda,m) \,||\, p(\theta|\lambda,m)\right] \right\rangle \; . \tag{1}$$

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.

Metadata: ID: D123 | shortcut: prior-ref | author: JoramSoch | date: 2020-12-02, 18:26.

5.3 Bayesian inference

5.3.1 Bayes' theorem

Theorem: Let A and B be two arbitrary statements about random variables (→ Definition I/1.1.3), 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 (→ Definition I/1.2.4) is defined as the ratio of joint probability (→ Definition I/1.2.2), i.e. the probability of both statements being true, and marginal probability (→ Definition I/1.2.3), 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.

Metadata: ID: P4 | shortcut: bayes-th | author: JoramSoch | date: 2019-09-27, 16:24.

5.3.2 Bayes' rule

Theorem: Let A_1, A_2 and B be arbitrary statements about random variables (→ Definition I/1.1.3) where A_1 and A_2 are mutually exclusive. Then, Bayes' rule states that the posterior odds (→ Definition "post-odd") are equal to the Bayes factor (→ Definition IV/3.4.1) times the prior odds (→ Definition "prior-odd"), 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 (→ Proof I/5.3.1), the conditional probabilities (→ Definition I/1.2.4) 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.

Metadata: ID: P12 | shortcut: bayes-rule | author: JoramSoch | date: 2020-01-06, 20:55.

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 (→ Definition I/1.1.3). 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.

Metadata: ID: D88 | shortcut: duni | author: JoramSoch | date: 2020-07-28, 04:05.

1.1.2 Probability mass function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a discrete uniform distribution (→ Definition II/1.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the probability mass function (→ Definition I/1.4.1) 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 (→ Definition 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 (→ Definition I/1.4.1) is

$$\sum_{x=a}^{b} f_X(x) = 1 \; , \tag{4}$$

we have

$$f_X(x) = \frac{1}{n} = \frac{1}{b-a+1} \; . \tag{5}$$

Sources:
- original work

Metadata: ID: P140 | shortcut: duni-pmf | author: JoramSoch | date: 2020-07-28, 04:57.

1.1.3 Cumulative distribution function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a discrete uniform distribution (→ Definition II/1.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the cumulative distribution function (→ Definition I/1.4.8) of X is

$$F_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if} \; x < a \\ \frac{\lfloor x \rfloor - a + 1}{b-a+1} \; , & \text{if} \; a \leq x \leq b \\ 1 \; , & \text{if} \; x > b \; . \end{array} \right. \tag{2}$$

Proof: The probability mass function of the discrete uniform distribution (→ Proof 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 (→ Definition I/1.4.8) 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

$$F_X(x) \overset{(3)}{=} \frac{\lfloor x \rfloor - a + 1}{n}, \; \text{if} \; a \leq x \leq b \; . \tag{6}$$

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.

Sources:
- original work

Metadata: ID: P141 | shortcut: duni-cdf | author: JoramSoch | date: 2020-07-28, 05:34.

1.1.4 Quantile function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a discrete uniform distribution (→ Definition II/1.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the quantile function (→ Definition I/1.4.13) 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 (→ Proof 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 $Q_X(p)$ is defined as (→ Definition I/1.4.13) 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 (→ Proof II/1.1.3) multiples of $1/n$ with $n = b - a + 1$, the quantile function (→ Definition I/1.4.13) is only defined for such values. First, we have $Q_X(p) = -\infty$, if $p = 0$. Second, since the cumulative probability increases step-wise (→ Proof 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:

$$\begin{split} Q_X(p) &= a + p \cdot (b - a + 1) - 1 \\ &= a + pb - pa + p - 1 \\ &= a(1-p) + (b+1)p - 1 \; . \end{split} \tag{7}$$

Sources:
- original work

Metadata: ID: P142 | shortcut: duni-qf | author: JoramSoch | date: 2020-07-28, 06:17.

1.2 Bernoulli distribution

1.2.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.2.1) p and $X = 0$ with probability (→ Definition I/1.2.1) $q = 1 - p$.

Sources:
- Wikipedia (2020): "Bernoulli distribution"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-03-22; URL: https://en.wikipedia.org/wiki/Bernoulli_distribution.

Metadata: ID: D44 | shortcut: bern | author: JoramSoch | date: 2020-03-22, 17:40.

1.2.2 Probability mass function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a Bernoulli distribution (→ Definition II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the probability mass function (→ Definition I/1.4.1) 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 (→ Definition II/1.2.1).

Sources:
- original work

Metadata: ID: P96 | shortcut: bern-pmf | author: JoramSoch | date: 2020-05-11, 22:10.

1.2.3 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a Bernoulli distribution (→ Definition II/1.2.1):

$$X \sim \mathrm{Bern}(p) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = p \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.5.1) 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 (→ Proof 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.

Metadata: ID: P22 | shortcut: bern-mean | author: JoramSoch | date: 2020-01-16, 10:58.

1.3 Binomial distribution

1.3.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.2.6) trials, where each trial has two possible outcomes (→ Definition 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.

Metadata: ID: D45 | shortcut: bin | author: JoramSoch | date: 2020-03-22, 17:52.

1.3.2 Probability mass function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a binomial distribution (→ Definition II/1.3.1):

$$X \sim \mathrm{Bin}(n,p) \; . \tag{1}$$

Then, the probability mass function (→ Definition I/1.4.1) of X is

$$f_X(x) = \binom{n}{x} \, p^x \, (1-p)^{n-x} \; . \tag{2}$$

Proof: A binomial variable is defined as (→ Definition II/1.3.1) the number of successes observed in n independent (→ Definition I/1.2.6) trials, where each trial has two possible outcomes (→ Definition II/1.2.1) (success/failure) and the probability (→ Definition I/1.2.1) 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.

Sources:
- original work

Metadata: ID: P97 | shortcut: bin-pmf | author: JoramSoch | date: 2020-05-11, 22:35.

1.3.3 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a binomial distribution (→ Definition II/1.3.1):

$$X \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = np \; . \tag{2}$$

Proof: By definition, a binomial random variable (→ Definition II/1.3.1) is the sum of n independent and identical Bernoulli trials (→ Definition 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 (→ Proof I/1.5.4), this is equal to

$$\begin{split} \mathrm{E}(X) &= \mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n) \\ &= \sum_{i=1}^{n} \mathrm{E}(X_i) \; . \end{split} \tag{4}$$

With the expected value of the Bernoulli distribution (→ Proof 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.

Metadata: ID: P23 | shortcut: bin-mean | author: JoramSoch | date: 2020-01-16, 11:06.

1.4 Poisson distribution

1.4.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.4.1) 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.

Metadata: ID: D62 | shortcut: poiss | author: JoramSoch | date: 2020-05-25, 23:34.

1.4.2 Probability mass function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a Poisson distribution (→ Definition II/1.4.1):

$$X \sim \mathrm{Poiss}(\lambda) \; . \tag{1}$$

Then, the probability mass function (→ Definition I/1.4.1) 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 (→ Definition II/1.4.1).

Sources:
- original work

Metadata: ID: P102 | shortcut: poiss-pmf | author: JoramSoch | date: 2020-05-14, 20:39.

1.4.3 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a Poisson distribution (→ Definition II/1.4.1):

$$X \sim \mathrm{Poiss}(\lambda) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = \lambda \; . \tag{2}$$

Proof: The expected value of a discrete random variable (→ Definition I/1.5.1) 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 (→ Proof II/1.4.2), 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.

Metadata: ID: P151 | shortcut: poiss-mean | author: JoramSoch | date: 2020-08-19, 06:09.

2 Multivariate discrete distributions

2.1 Categorical distribution

2.1.1 Definition

Definition: Let X be a random vector (→ Definition I/1.1.4). 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 (→ Definition I/1.2.1) 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.

Metadata: ID: D46 | shortcut: cat | author: JoramSoch | date: 2020-03-22, 18:09.

2.1.2 Probability mass function

Theorem: Let X be a random vector (→ Definition I/1.1.4) following a categorical distribution (→ Definition II/2.1.1):

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the probability mass function (→ Definition I/1.4.1) 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 (→ Definition II/2.1.1).

Sources:
- original work

Metadata: ID: P98 | shortcut: cat-pmf | author: JoramSoch | date: 2020-05-11, 22:58.

2.1.3 Mean

Theorem: Let X be a random vector (→ Definition I/1.1.4) following a categorical distribution (→ Definition II/2.1.1):

$$X \sim \mathrm{Cat}([p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = [p_1, \ldots, p_k] \; . \tag{2}$$

Proof: If we conceive the outcome of a categorical distribution (→ Definition 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 $\mathrm{Pr}(X = e_1) = p_1$, ..., $\mathrm{Pr}(X = e_k) = p_k$. Consequently, the expected value (→ Definition I/1.5.1) is

$$\begin{split}
\mathrm{E}(X) &= \sum_{x \in \mathcal{X}} x \cdot \mathrm{Pr}(X = x) \\
&= \sum_{i=1}^{k} e_i \cdot \mathrm{Pr}(X = e_i) \\
&= \sum_{i=1}^{k} e_i \cdot p_i \\
&= [p_1, \ldots, p_k] \; .
\end{split} \tag{3}$$

Sources:
- original work

Metadata: ID: P24 | shortcut: cat-mean | author: JoramSoch | date: 2020-01-16, 11:17.

2.2 Multinomial distribution

2.2.1 Definition

Definition: Let X be a random vector (→ Definition I/1.1.4). 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 (→ Definition I/1.2.6) trials, where each trial has k possible outcomes (→ Definition 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.

Metadata: ID: D47 | shortcut: mult | author: JoramSoch | date: 2020-03-22, 17:52.

2.2.2 Probability mass function

Theorem: Let X be a random vector (→ Definition I/1.1.4) following a multinomial distribution (→ Definition II/2.2.1):

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the probability mass function (→ Definition I/1.4.1) 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 is defined as (→ Definition II/2.2.1) a vector of the numbers of observations belonging to k distinct categories in n independent (→ Definition I/1.2.6) trials, where each trial has k possible outcomes (→ Definition II/2.1.1) and the category probabilities (→ Definition I/1.2.1) are identical across trials.
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_k observations for category k. This number is equal to the number of possibilities in which x_1 category 1 objects, ..., x_k category k objects 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_k observations for category k, 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.

Sources:
- original work

Metadata: ID: P99 | shortcut: mult-pmf | author: JoramSoch | date: 2020-05-11, 23:30.

2.2.3 Mean

Theorem: Let X be a random vector (→ Definition I/1.1.4) following a multinomial distribution (→ Definition II/2.2.1):

$$X \sim \mathrm{Mult}(n, [p_1, \ldots, p_k]) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = [np_1, \ldots, np_k] \; . \tag{2}$$

Proof: By definition, a multinomial random variable (→ Definition II/2.2.1) is the sum of n independent and identical categorical trials (→ Definition 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 (→ Proof I/1.5.4), 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 (→ Proof II/2.1.3), we have:

$$\mathrm{E}(X) = \sum_{i=1}^{n} [p_1, \ldots, p_k] = n \cdot [p_1, \ldots, p_k] = [np_1, \ldots, np_k] \; . \tag{5}$$

Sources:
- original work

Metadata: ID: P25 | shortcut: mult-mean | author: JoramSoch | date: 2020-01-16, 11:26.

3 Univariate continuous distributions

3.1 Continuous uniform distribution

3.1.1 Definition

Definition: Let X be a continuous random variable (→ Definition I/1.1.3). 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).

Metadata: ID: D3 | shortcut: cuni | author: JoramSoch | date: 2020-01-27, 14:05.

3.1.2 Probability density function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) of X is

$$f_X(x) = \left\{ \begin{array}{rl} \frac{1}{b-a} \; , & \text{if} \; a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{array} \right. \tag{2}$$

Proof: A continuous uniform variable is defined as (→ Definition II/3.1.1) having a constant probability density between minimum a and maximum b. Therefore,

$$\begin{split} f_X(x) &\propto 1 \quad \text{for all} \quad x \in [a,b] \quad \text{and} \\ f_X(x) &= 0, \quad \text{if} \quad x < a \quad \text{or} \quad x > b \; . \end{split} \tag{3}$$

To ensure that $f_X(x)$ is a proper probability density function (→ Definition I/1.4.4), 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:

$$
\begin{aligned}
\int_a^b 1 \, \mathrm{d}x &= c(a,b) \\
[x]_a^b &= c(a,b) \\
c(a,b) &= b-a \; .
\end{aligned} \tag{6}
$$

Sources:
- original work

Metadata: ID: P37 | shortcut: cuni-pdf | author: JoramSoch | date: 2020-01-31, 15:41.

3.1.3 Cumulative distribution function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition II/3.1.1):

$$
X \sim \mathcal{U}(a,b) \; . \tag{1}
$$

Then, the cumulative distribution function (→ Definition I/1.4.8) of X is

$$
F_X(x) = \left\{
\begin{array}{rl}
0 \; , & \text{if} \; x < a \\
\frac{x-a}{b-a} \; , & \text{if} \; a \leq x \leq b \\
1 \; , & \text{if} \; x > b \; .
\end{array}
\right. \tag{2}
$$

Proof: The probability density function of the continuous uniform distribution (→ Proof II/3.1.2) is:

$$
\mathcal{U}(x; a, b) = \left\{
\begin{array}{rl}
\frac{1}{b-a} \; , & \text{if} \; a \leq x \leq b \\
0 \; , & \text{otherwise} \; .
\end{array}
\right. \tag{3}
$$

Thus, the cumulative distribution function (→ Definition I/1.4.8) 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)

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{a} \mathcal{U}(z; a, b) \, \mathrm{d}z + \int_{a}^{x} \mathcal{U}(z; a, b) \, \mathrm{d}z \\
&= \int_{-\infty}^{a} 0 \, \mathrm{d}z + \int_{a}^{x} \frac{1}{b-a} \, \mathrm{d}z \\
&= 0 + \frac{1}{b-a} [z]_a^x \\
&= \frac{x-a}{b-a} \; .
\end{aligned} \tag{6}
$$

Finally, if $x > b$, we have

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{b} \mathcal{U}(z;a,b)\,\mathrm{d}z + \int_{b}^{x} \mathcal{U}(z;a,b)\,\mathrm{d}z \\ &= F_X(b) + \int_{b}^{x} 0\,\mathrm{d}z \\ &= \frac{b-a}{b-a} + 0 \\ &= 1 \; . \end{aligned} \tag{7}$$

This completes the proof.

Sources:
- original work

Metadata: ID: P38 | shortcut: cuni-cdf | author: JoramSoch | date: 2020-01-02, 18:05.

3.1.4 Quantile function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition II/3.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the quantile function (→ Definition I/1.4.13) 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 (→ Proof II/3.1.3) 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 $Q_X(p)$ is defined as (→ Definition I/1.4.13) 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 (→ Proof I/1.4.14)

$$Q_X(p) = F_X^{-1}(x) \; . \tag{5}$$

This can be derived by rearranging equation (3):

$$\begin{aligned} p &= \frac{x-a}{b-a} \\ x &= p(b-a) + a \\ x &= bp + a(1-p) \; . \end{aligned} \tag{6}$$

Sources:
- original work

Metadata: ID: P39 | shortcut: cuni-qf | author: JoramSoch | date: 2020-01-02, 18:27.

3.1.5 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = \frac{1}{2}(a+b) \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.5.1) 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 (→ Proof II/3.1.2), this becomes:

$$\begin{split} \mathrm{E}(X) &= \int_a^b x \cdot \frac{1}{b-a} \, \mathrm{d}x \\ &= \left[\frac{1}{2} \frac{x^2}{b-a} \right]_a^b \\ &= \frac{1}{2} \frac{b^2-a^2}{b-a} \\ &= \frac{1}{2} \frac{(b+a)(b-a)}{b-a} \\ &= \frac{1}{2}(a+b) \; . \end{split} \tag{4}$$

Sources:
- original work

Metadata: ID: P82 | shortcut: cuni-mean | author: JoramSoch | date: 2020-03-16, 16:12.

3.1.6 Median

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition II/3.1.1):

$$X \sim \mathcal{U}(a, b) \; . \tag{1}$$

Then, the median (→ Definition I/1.9.1) of X is

$$\mathrm{median}(X) = \frac{1}{2}(a+b) \; . \tag{2}$$

Proof: The median (→ Definition I/1.9.1) is the value at which the cumulative distribution function (→ Definition I/1.4.8) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the continuous uniform distribution (→ Proof II/3.1.3) 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 (→ Proof II/3.1.4) is

$$x = bp + a(1-p) \; . \tag{5}$$

Setting $p = 1/2$, we obtain:

$$\mathrm{median}(X) = b \cdot \frac{1}{2} + a \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{2}(a+b) \; . \tag{6}$$

Sources:
- original work

Metadata: ID: P83 | shortcut: cuni-med | author: JoramSoch | date: 2020-03-16, 16:19.

3.1.7 Mode

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a continuous uniform distribution (→ Definition II/3.1.1):

$$X \sim \mathcal{U}(a,b) \; . \tag{1}$$

Then, the mode (→ Definition I/1.9.2) of X is

$$\mathrm{mode}(X) \in [a,b] \; . \tag{2}$$

Proof: The mode (→ Definition I/1.9.2) is the value which maximizes the probability density function (→ Definition I/1.4.4):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the continuous uniform distribution (→ Proof II/3.1.2) is:

$$f_X(x) = \left\{ \begin{array}{rl} \frac{1}{b-a} \; , & \text{if } a \leq x \leq b \\ 0 \; , & \text{otherwise} \; . \end{array} \right. \tag{4}$$

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.

Sources:
- original work

Metadata: ID: P84 | shortcut: cuni-med | author: JoramSoch | date: 2020-03-16, 16:29.

3.2 Normal distribution

3.2.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D4 | shortcut: norm | author: JoramSoch | date: 2020-01-27, 14:15.

3.2.2 Standard normal distribution

Definition: Let X be a random variable (→ Definition I/1.1.3). Then, X is said to be standard normally distributed, if X follows a normal distribution (→ Definition 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.

Metadata: ID: D63 | shortcut: snorm | author: JoramSoch | date: 2020-05-26, 23:32.

3.2.3 Relation to standard normal distribution

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition 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 (→ Definition II/3.2.2) 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 (→ Proof I/1.4.9) as

$$F_Y(y) = \begin{cases} 0 \; , & \text{if } y < \min(\mathcal{Y}) \\ F_X(g^{-1}(y)) \; , & \text{if } y \in \mathcal{Y} \\ 1 \; , & \text{if } y > \max(\mathcal{Y}) \; . \end{cases} \tag{5}$$

The cumulative distribution function of the normally distributed (→ Proof II/3.2.9) 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 (→ Definition I/1.4.8) of the standard normal distribution (→ Definition II/3.2.2).

Sources:
- original work

Metadata: ID: P111 | shortcut: norm-snorm | author: JoramSoch | date: 2020-05-26, 23:01.

3.2.4 Relation to standard normal distribution

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition 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 (→ Definition II/3.2.2) 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 (→ Proof I/1.4.5) as

$$f_Y(y) = \left\{ \begin{array}{rl} f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{array} \right. \tag{5}$$

where $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$. With the probability density function of the normal distribution (→ Proof II/3.2.7), 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 (→ Definition I/1.4.4) of the standard normal distribution (→ Definition II/3.2.2).

Sources:
- original work

Metadata: ID: P176 | shortcut: norm-snorm2 | author: JoramSoch | date: 2020-10-15, 11:42.

3.2.5 Relation to standard normal distribution

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition 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 (→ Definition II/3.2.2) 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 (→ Proof II/4.1.5) 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 (→ Definition I/1.1.4) following a multivariate normal distribution (→ Definition 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:

$$Z \sim \mathcal{N}(0, 1) \; . \tag{6}$$

Sources:
- original work

Metadata: ID: P180 | shortcut: norm-snorm3 | author: JoramSoch | date: 2020-10-22, 06:34.

3.2.6 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

$$\begin{aligned} I_P^2 &\overset{(3)}{=} \left(\int_0^P \exp\left[-x^2\right] \, \mathrm{d}x \right) \left(\int_0^P \exp\left[-y^2\right] \, \mathrm{d}y \right) \\ &= \int_0^P \int_0^P \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \\ &= \iint_{S_P} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \end{aligned} \tag{6}$$

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

$$\iint_{C_1} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \le I_P^2 \le \iint_{C_2} \exp\left[-\left(x^2+y^2\right)\right] \, \mathrm{d}x \, \mathrm{d}y \tag{7}$$

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

$$\int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp\left[-r^2\right] \, r \, \mathrm{d}r \, \mathrm{d}\theta \le I_P^2 \le \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp\left[-r^2\right] \, r \, \mathrm{d}r \, \mathrm{d}\theta \; . \tag{8}$$

Solving the definite integrals yields:

$$
\begin{split}
\int_0^{\frac{\pi}{2}} \int_0^{r_1} \exp\left[-r^2\right] r \, \mathrm{d}r \, \mathrm{d}\theta \leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \int_0^{r_2} \exp\left[-r^2\right] r \, \mathrm{d}r \, \mathrm{d}\theta \\
\int_0^{\frac{\pi}{2}} \left[-\frac{1}{2}\exp\left[-r^2\right]\right]_0^{r_1} \mathrm{d}\theta \leq I_P^2 \leq \int_0^{\frac{\pi}{2}} \left[-\frac{1}{2}\exp\left[-r^2\right]\right]_0^{r_2} \mathrm{d}\theta \\
-\frac{1}{2}\int_0^{\frac{\pi}{2}} \left(\exp\left[-r_1^2\right]-1\right) \mathrm{d}\theta \leq I_P^2 \leq -\frac{1}{2}\int_0^{\frac{\pi}{2}} \left(\exp\left[-r_2^2\right]-1\right) \mathrm{d}\theta \\
-\frac{1}{2}\left[\left(\exp\left[-r_1^2\right]-1\right)\theta\right]_0^{\frac{\pi}{2}} \leq I_P^2 \leq -\frac{1}{2}\left[\left(\exp\left[-r_2^2\right]-1\right)\theta\right]_0^{\frac{\pi}{2}} \\
\frac{1}{2}\left(1-\exp\left[-r_1^2\right]\right)\frac{\pi}{2} \leq I_P^2 \leq \frac{1}{2}\left(1-\exp\left[-r_2^2\right]\right)\frac{\pi}{2} \\
\frac{\pi}{4}\left(1-\exp\left[-P^2\right]\right) \leq I_P^2 \leq \frac{\pi}{4}\left(1-\exp\left[-2P^2\right]\right)
\end{split} \tag{9}
$$

Calculating the limit for $P \to \infty$, we obtain

$$
\begin{split}
\lim_{P\to\infty} \frac{\pi}{4}\left(1-\exp\left[-P^2\right]\right) \leq \lim_{P\to\infty} I_P^2 \leq \lim_{P\to\infty} \frac{\pi}{4}\left(1-\exp\left[-2P^2\right]\right) \\
\frac{\pi}{4} \leq I^2 \leq \frac{\pi}{4} \; ,
\end{split} \tag{10}
$$

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:

$$
\begin{split}
\int_{-\infty}^{+\infty} \exp\left[-x^2\right] \mathrm{d}x &= 2\int_0^{\infty} \exp\left[-x^2\right] \mathrm{d}x \\
&\overset{(11)}{=} 2\,\frac{\sqrt{\pi}}{2} \\
&= \sqrt{\pi} \; .
\end{split} \tag{12}
$$

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.

Metadata: ID: P196 | shortcut: norm-gi | author: JoramSoch | date: 2020-11-25, 04:47.

3.2.7 Probability density function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/3.2.1).

Sources:
- original work

Metadata: ID: P33 | shortcut: norm-pdf | author: JoramSoch | date: 2020-01-27, 15:15.

3.2.8 Moment-generating function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the moment-generating function (→ Definition I/1.4.15) 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 (→ Proof II/3.2.7) 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 (→ Definition I/1.4.15) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] \; . \tag{4}$$

Using the expected value for continuous random variables (→ Definition I/1.5.1), 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

$$
\begin{split}
M_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{(-\infty-\mu)/(\sqrt{2}\sigma)}^{(+\infty-\mu)/(\sqrt{2}\sigma)} \exp\left[t\left(\sqrt{2}\sigma u + \mu\right) - \frac{1}{2}\left(\frac{\sqrt{2}\sigma u + \mu - \mu}{\sigma}\right)^2\right] \, \mathrm{d}\left(\sqrt{2}\sigma u + \mu\right) \\
&= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \exp\left[\left(\sqrt{2}\sigma u + \mu\right)t - u^2\right] \, \mathrm{d}u \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[\sqrt{2}\sigma u t - u^2\right] \, \mathrm{d}u \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u^2 - \sqrt{2}\sigma u t\right)\right] \, \mathrm{d}u \\
&= \frac{\exp(\mu t)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u - \frac{\sqrt{2}}{2}\sigma t\right)^2 + \frac{1}{2}\sigma^2 t^2\right] \, \mathrm{d}u \\
&= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-\left(u - \frac{\sqrt{2}}{2}\sigma t\right)^2\right] \, \mathrm{d}u
\end{split} \tag{6}
$$

Now substituting $v = u - \sqrt{2}/2\,\sigma t$, i.e. $u = v + \sqrt{2}/2\,\sigma t$, we have

$$
\begin{split}
M_X(t) &= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty-\sqrt{2}/2\,\sigma t}^{+\infty-\sqrt{2}/2\,\sigma t} \exp\left[-v^2\right] \, \mathrm{d}\left(v + \sqrt{2}/2\,\sigma t\right) \\
&= \frac{\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left[-v^2\right] \, \mathrm{d}v \; .
\end{split} \tag{7}
$$

With the Gaussian integral (→ Proof II/3.2.6)

$$
\int_{-\infty}^{+\infty} \exp\left[-x^2\right] \, \mathrm{d}x = \sqrt{\pi} \; , \tag{8}
$$

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.

Metadata: ID: P71 | shortcut: norm-mgf | author: JoramSoch | date: 2020-03-03, 11:29.

3.2.9 Cumulative distribution function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distributions (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Proof II/3.2.7) 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 (→ Definition I/1.4.8) 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.

Metadata: ID: P85 | shortcut: norm-cdf | author: JoramSoch | date: 2020-03-20, 01:33.

3.2.10 Cumulative distribution function without error function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Definition I/1.4.4) of the standard normal distribution (→ Definition II/3.2.2) and $n!!$ is a double factorial.

Proof:
1) First, consider the standard normal distribution (→ Definition II/3.2.2) $\mathcal{N}(0,1)$ which has the probability density function (→ Proof II/3.2.7)

$$
\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \; . \tag{3}
$$

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

$$
\begin{aligned}
T(x) &= \frac{1}{\sqrt{2\pi}} \cdot \sum_{i=1}^{\infty} \left(\frac{x^{2i-1}}{(2i-1)!!} \cdot e^{-\frac{1}{2}x^2} \right) + c \\
&= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c \\
&\overset{(3)}{=} \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + c .
\end{aligned} \tag{6}
$$

2) Next, let $\Phi(x)$ be the cumulative distribution function (→ Definition I/1.4.8) of the standard normal distribution (→ Definition II/3.2.2):

$$
\Phi(x) = \int_{-\infty}^{x} \varphi(x)\,\mathrm{d}x . \tag{7}
$$

It can be obtained by matching $T(0)$ to $\Phi(0)$ which is $1/2$, because the standard normal distribution is symmetric around zero:

$$\begin{aligned}
T(0) = \varphi(0) \cdot \sum_{i=1}^{\infty} \frac{0^{2i-1}}{(2i-1)!!} + c = \frac{1}{2} = \Phi(0) \\
\Leftrightarrow c = \frac{1}{2} \\
\Rightarrow \Phi(x) = \varphi(x) \cdot \sum_{i=1}^{\infty} \frac{x^{2i-1}}{(2i-1)!!} + \frac{1}{2} \, .
\end{aligned} \tag{8}$$

3) Finally, the cumulative distribution functions (→ Definition I/1.4.8) of the standard normal distribution (→ Definition II/3.2.2) and the general normal distribution (→ Definition II/3.2.1) are related to each other (→ Proof II/3.2.3) 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*, arXiv: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.

Metadata: ID: P86 | shortcut: norm-cdfwerf | author: JoramSoch | date: 2020-03-20, 04:26.

3.2.11 Quantile function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distributions (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \, . \tag{1}$$

Then, the quantile function (→ Definition I/1.4.13) 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 (→ Proof II/3.2.9) 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 (→ Proof I/1.4.14):

$$Q_X(p) = F_X^{-1}(x) \; . \tag{4}$$

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.

Metadata: ID: P87 | shortcut: norm-qf | author: JoramSoch | date: 2020-03-20, 04:47.

3.2.12 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = \mu \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.5.1) 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 (→ Proof II/3.2.7), this reads:

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x \; . \end{aligned} \tag{4}$$

Substituting $z = x - \mu$, we have:

$$
\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} (z+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \, \mathrm{d}(z+\mu) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (z+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \, \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \, \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2\right] \, \mathrm{d}z \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{+\infty} z \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \, \mathrm{d}z + \mu \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \, \mathrm{d}z \right) \; .
\end{aligned} \tag{5}
$$

The general antiderivatives are

$$
\begin{aligned}
\int x \cdot \exp\left[-ax^2\right] \mathrm{d}x &= -\frac{1}{2a} \cdot \exp\left[-ax^2\right] \\
\int \exp\left[-ax^2\right] \mathrm{d}x &= \frac{1}{2}\sqrt{\frac{\pi}{a}} \cdot \mathrm{erf}\left[\sqrt{a}x\right]
\end{aligned} \tag{6}
$$

where $\mathrm{erf}(x)$ is the error function. Using this, the integrals can be calculated as:

$$
\begin{aligned}
\mathrm{E}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \right]_{-\infty}^{+\infty} + \mu \left[\sqrt{\frac{\pi}{2}}\sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z\right] \right]_{-\infty}^{+\infty} \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\left[\lim_{z\to\infty} \left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \right) - \lim_{z\to-\infty} \left(-\sigma^2 \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot z^2\right] \right) \right] \right. \\
&\quad \left. + \mu \left[\lim_{z\to\infty} \left(\sqrt{\frac{\pi}{2}}\sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z\right] \right) - \lim_{z\to-\infty} \left(\sqrt{\frac{\pi}{2}}\sigma \cdot \mathrm{erf}\left[\frac{1}{\sqrt{2}\sigma} z\right] \right) \right] \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left([0-0] + \mu \left[\sqrt{\frac{\pi}{2}}\sigma - \left(-\sqrt{\frac{\pi}{2}}\sigma \right) \right] \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \mu \cdot 2\sqrt{\frac{\pi}{2}}\sigma \\
&= \mu \; .
\end{aligned} \tag{7}
$$

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.

Metadata: ID: P15 | shortcut: norm-mean | author: JoramSoch | date: 2020-01-09, 15:04.

3.2.13 Median

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the median (→ Definition I/1.9.1) of X is

$$\mathrm{median}(X) = \mu \; . \tag{2}$$

Proof: The median (→ Definition I/1.9.1) is the value at which the cumulative distribution function (→ Definition I/1.4.8) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the normal distribution (→ Proof II/3.2.9) 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:

$$\mathrm{median}(X) = \sqrt{2}\sigma \cdot \mathrm{erf}^{-1}(0) + \mu = \mu \; . \tag{6}$$

Sources:
- original work

Metadata: ID: P16 | shortcut: norm-med | author: JoramSoch | date: 2020-01-09, 15:33.

3.2.14 Mode

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the mode (→ Definition I/1.9.2) of X is

$$\mathrm{mode}(X) = \mu \; . \tag{2}$$

Proof: The mode (→ Definition I/1.9.2) is the value which maximizes the probability density function (→ Definition I/1.4.4):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the normal distribution (→ Proof II/3.2.7) 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):

$$\begin{split} f'_X(x) = 0 &= \frac{1}{\sqrt{2\pi}\sigma^3} \cdot (-x+\mu) \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\ 0 &= -x+\mu \\ x &= \mu \; . \end{split} \tag{7}$$

By plugging this value into the second deriative (6),

$$\begin{split} f''_X(\mu) &= -\frac{1}{\sqrt{2\pi}\sigma^3} \cdot \exp(0) + \frac{1}{\sqrt{2\pi}\sigma^5} \cdot (0)^2 \cdot \exp(0) \\ &= -\frac{1}{\sqrt{2\pi}\sigma^3} < 0 \; , \end{split} \tag{8}$$

we confirm that it is in fact a maximum which shows that

$$\mathrm{mode}(X) = \mu \; . \tag{9}$$

Sources:
- original work

Metadata: ID: P17 | shortcut: norm-mode | author: JoramSoch | date: 2020-01-09, 15:58.

3.2.15 Variance

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the variance (→ Definition I/1.6.1) of X is

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

Proof: The variance (→ Definition I/1.6.1) is the probability-weighted average of the squared deviation from the mean (→ Definition I/1.5.1):

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mathrm{E}(X))^2 \cdot f_{\mathrm{X}}(x) \, \mathrm{d}x \; . \tag{3}$$

With the expected value (→ Proof II/3.2.12) and probability density function (→ Proof II/3.2.7) of the normal distribution, this reads:

$$\begin{split} \mathrm{Var}(X) &= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] \mathrm{d}x \; . \end{split} \tag{4}$$

Substituting $z = x - \mu$, we have:

$$\begin{split} \mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty-\mu}^{+\infty-\mu} z^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}(z+\mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} z^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{z}{\sigma} \right)^2 \right] \mathrm{d}z \; . \end{split} \tag{5}$$

Now substituting $z = \sqrt{2}\sigma x$, we have:

$$\begin{split} \mathrm{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma x)^2 \cdot \exp\left[-\frac{1}{2} \left(\frac{\sqrt{2}\sigma x}{\sigma} \right)^2 \right] \mathrm{d}(\sqrt{2}\sigma x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot 2\sigma^2 \cdot \sqrt{2}\sigma \int_{-\infty}^{+\infty} x^2 \cdot \exp\left[-x^2 \right] \mathrm{d}x \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 \cdot e^{-x^2} \, \mathrm{d}x \; . \end{split} \tag{6}$$

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.

Metadata: ID: P18 | shortcut: norm-var | author: JoramSoch | date: 2020-01-09, 22:47.

3.2.16 Full width at half maximum

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}$$

Then, the full width at half maximum (→ Definition I/1.10.2) (FWHM) of X is

$$\mathrm{FWHM}(X) = 2\sqrt{2\ln 2}\sigma \; . \tag{2}$$

Proof: The probability density function of the normal distribution (→ Proof II/3.2.7) 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 (→ Proof II/3.2.14) 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 (→ Definition I/1.10.2)

$$f_X(x_{\mathrm{FWHM}}) = \frac{1}{2} f_{\mathrm{max}} \overset{(5)}{=} \frac{1}{2\sqrt{2\pi}\sigma} \; . \tag{6}$$

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 (→ Definition I/1.10.2) 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.

Metadata: ID: P152 | shortcut: norm-fwhm | author: JoramSoch | date: 2020-08-19, 06:39.

3.2.17 Differential entropy

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a normal distribution (→ Definition II/3.2.1):

$$
X \sim \mathcal{N}(\mu, \sigma^2) \; . \tag{1}
$$

Then, the differential entropy (→ Definition 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 (→ Definition 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 (→ Definition I/1.5.1)

$$\mathrm{h}(X) = -\mathrm{E}\left[\ln p(x)\right] \; . \tag{4}$$

With the probability density function of the normal distribution (→ Proof II/3.2.7), the differential entropy of X is:

$$\begin{split}
\mathrm{h}(X) &= -\mathrm{E}\left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\right)\right] \\
&= -\mathrm{E}\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\,\mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \mathrm{E}\left[(x-\mu)^2\right] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \sigma^2 \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}\ln(2\pi\sigma^2 e) \; .
\end{split} \tag{5}$$

Sources:

- Wang, Peng-Hua (2012): "Differential Entropy"; in: *National Taipei University*; URL: https://web.ntpu.edu.tw/~phwang/teaching/2012s/IT/slides/chap08.pdf.

Metadata: ID: P101 | shortcut: norm-dent | author: JoramSoch | date: 2020-05-14, 20:09.

3.2.18 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ Definition I/1.1.3). Assume two normal distributions (→ Definition II/3.2.1) P and Q specifying the probability distribution of X as

$$\begin{split}
P &: \; X \sim \mathcal{N}(\mu_1, \sigma_1^2) \\
Q &: \; X \sim \mathcal{N}(\mu_2, \sigma_2^2) \; .
\end{split} \tag{1}$$

Then, the Kullback-Leibler divergence (→ Definition 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 (→ Definition 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 (→ Definition II/3.2.1) in (1), yields

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \int_{-\infty}^{+\infty} \mathcal{N}(x;\mu_1,\sigma_1^2)\,\ln \frac{\mathcal{N}(x;\mu_1,\sigma_1^2)}{\mathcal{N}(x;\mu_2,\sigma_2^2)}\,\mathrm{d}x \\
&= \left\langle \ln \frac{\mathcal{N}(x;\mu_1,\sigma_1^2)}{\mathcal{N}(x;\mu_2,\sigma_2^2)} \right\rangle_{p(x)} \; .
\end{split} \tag{4}$$

Using the probability density function of the normal distribution (→ Proof II/3.2.7), this becomes:

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{2\pi}\sigma_1}\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]}{\frac{1}{\sqrt{2\pi}\sigma_2}\cdot\exp\left[-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]} \right\rangle_{p(x)} \\
&= \left\langle \ln \left(\sqrt{\frac{\sigma_2^2}{\sigma_1^2}} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 \right] \right) \right\rangle_{p(x)} \\
&= \left\langle \frac{1}{2}\ln\frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 \right\rangle_{p(x)} \\
&= \frac{1}{2}\left\langle -\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \ln\frac{\sigma_1^2}{\sigma_2^2} \right\rangle_{p(x)} \\
&= \frac{1}{2}\left\langle -\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{x^2-2\mu_2 x+\mu_2^2}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2} \right\rangle_{p(x)} \; .
\end{split} \tag{5}$$

Because trace function and expected value (→ Definition I/1.5.1) are both linear operators, the expectation can be moved inside the trace:

$$\begin{split}
\mathrm{KL}[P\,||\,Q] &= \frac{1}{2}\left[-\frac{\left\langle (x-\mu_1)^2 \right\rangle}{\sigma_1^2} + \frac{\left\langle x^2-2\mu_2 x+\mu_2^2 \right\rangle}{\sigma_2^2} - \left\langle \ln\frac{\sigma_1^2}{\sigma_2^2} \right\rangle \right] \\
&= \frac{1}{2}\left[-\frac{\left\langle (x-\mu_1)^2 \right\rangle}{\sigma_1^2} + \frac{\left\langle x^2 \right\rangle - \left\langle 2\mu_2 x \right\rangle + \left\langle \mu_2^2 \right\rangle}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2} \right] \; .
\end{split} \tag{6}$$

The first expectation corresponds to the variance (→ Definition I/1.6.1)

$$\left\langle (X-\mu)^2 \right\rangle = \mathrm{E}[(X-\mathrm{E}(X))^2] = \mathrm{Var}(X) \tag{7}$$

and the variance of a normally distributed random variable (→ Proof II/3.2.15) 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 (→ Proof II/3.2.8)

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \langle x \rangle = \mu \quad \text{and} \quad \left\langle x^2 \right\rangle = \mu^2 + \sigma^2 \; , \tag{9}$$

the Kullback-Leibler divergence in (6) becomes

$$\begin{split} \mathrm{KL}[P \,||\, Q] &= \frac{1}{2} \left[-\frac{\sigma_1^2}{\sigma_1^2} + \frac{\mu_1^2 + \sigma_1^2 - 2\mu_2\mu_1 + \mu_2^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} \right] \\ &= \frac{1}{2} \left[\frac{\mu_1^2 - 2\mu_1\mu_2 + \mu_2^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right] \\ &= \frac{1}{2} \left[\frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln \frac{\sigma_1^2}{\sigma_2^2} - 1 \right] \end{split} \tag{10}$$

which is equivalent to (2).

Sources:
- original work

Metadata: ID: P193 | shortcut: norm-kl | author: JoramSoch | date: 2020-11-19, 07:08.

3.3 Gamma distribution

3.3.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D7 | shortcut: gam | author: JoramSoch | date: 2020-02-08, 23:29.

3.3.2 Standard gamma distribution

Definition: Let X be a random variable (→ Definition I/1.1.3). Then, X is said to have a standard gamma distribution, if X follows a gamma distribution (→ Definition II/3.3.1) 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

Metadata: ID: D64 | shortcut: sgam | author: JoramSoch | date: 2020-05-26, 23:36.

3.3.3 Relation to standard gamma distribution

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1) 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 (→ Definition II/3.3.2) 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 (→ Proof I/1.4.9) 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 (→ Proof II/3.3.6) 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:

$$\begin{split} F_Y(y) &\overset{(5)}{=} F_X(g^{-1}(y)) \\ &\overset{(6)}{=} \int_{-\infty}^{y/b} \frac{b^a}{\Gamma(a)} t^{a-1} \exp[-bt] \, \mathrm{d}t \; . \end{split} \tag{7}$$

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 (→ Definition I/1.4.8) of the standard gamma distribution (→ Definition II/3.3.2).

Sources:
- original work

Metadata: ID: P112 | shortcut: gam-sgam | author: JoramSoch | date: 2020-05-26, 23:14.

3.3.4 Relation to standard gamma distribution

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1) 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 (→ Definition II/3.3.2) 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 (→ Proof I/1.4.5) as

$$
f_Y(y) = \left\{ \begin{array}{rl} f_X(g^{-1}(y)) \, \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \; , & \text{if } y \in \mathcal{Y} \\ 0 \; , & \text{if } y \notin \mathcal{Y} \end{array} \right. \tag{5}
$$

where $\mathcal{Y} = \{y = g(x) : x \in \mathcal{X}\}$. With the probability density function of the gamma distribution (→ Proof II/3.3.5), 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 (→ Definition I/1.4.4) of the standard gamma distribution (→ Definition II/3.3.2).

Sources:
- original work

Metadata: ID: P177 | shortcut: gam-sgam2 | author: JoramSoch | date: 2020-10-15, 12:04.

3.3.5 Probability density function

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$
X \sim \mathrm{Gam}(a, b) \; . \tag{1}
$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/3.3.1).

Sources:
- original work

Metadata: ID: P45 | shortcut: gam-pdf | author: JoramSoch | date: 2020-02-08, 23:41.

3.3.6 Cumulative distribution function

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$
X \sim \mathrm{Gam}(a, b) \; . \tag{1}
$$

Then, the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Proof II/3.3.5) is:

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \; . \tag{3}$$

Thus, the cumulative distribution function (→ Definition I/1.4.8) is:

$$\begin{split} F_X(x) &= \int_0^x \mathrm{Gam}(z; a, b) \, \mathrm{d}z \\ &= \int_0^x \frac{b^a}{\Gamma(a)} z^{a-1} \exp[-bz] \, \mathrm{d}z \\ &= \frac{b^a}{\Gamma(a)} \int_0^x z^{a-1} \exp[-bz] \, \mathrm{d}z \; . \end{split} \tag{4}$$

Substituting $t = bz$, i.e. $z = t/b$, this becomes:

$$\begin{split} F_X(x) &= \frac{b^a}{\Gamma(a)} \int_{b \cdot 0}^{bx} \left(\frac{t}{b} \right)^{a-1} \exp\left[-b \left(\frac{t}{b} \right) \right] \mathrm{d}\left(\frac{t}{b} \right) \\ &= \frac{b^a}{\Gamma(a)} \cdot \frac{1}{b^{a-1}} \cdot \frac{1}{b} \int_0^{bx} t^{a-1} \exp[-t] \, \mathrm{d}t \\ &= \frac{1}{\Gamma(a)} \int_0^{bx} t^{a-1} \exp[-t] \, \mathrm{d}t \; . \end{split} \tag{5}$$

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.

Metadata: ID: P178 | shortcut: gam-cdf | author: JoramSoch | date: 2020-10-15, 12:34.

3.3.7 Quantile function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the quantile function (→ Definition I/1.4.13) 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 (→ Proof II/3.3.6) 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 $Q_X(p)$ is defined as (→ Definition I/1.4.13) 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 (→ Proof I/1.4.14)

$$Q_X(p) = F_X^{-1}(x) \; . \tag{5}$$

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.

Metadata: ID: P194 | shortcut: gam-qf | author: JoramSoch | date: 2020-11-19, 07:31.

3.3.8 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$X \sim \text{Gam}(a, b) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\text{E}(X) = \frac{a}{b} \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.5.1) 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 (→ Proof II/3.3.5), this reads:

$$\begin{split} \mathrm{E}(X) &= \int_0^\infty x \cdot \frac{b^a}{\Gamma(a)} x^{a-1} \exp[-bx] \, \mathrm{d}x \\ &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] \, \mathrm{d}x \\ &= \int_0^\infty \frac{1}{b} \cdot \frac{b^{a+1}}{\Gamma(a)} x^{(a+1)-1} \exp[-bx] \, \mathrm{d}x \; . \end{split} \tag{4}$$

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 (→ Proof II/3.3.5), 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.

Metadata: ID: P108 | shortcut: gam-mean | author: JoramSoch | date: 2020-05-19, 06:54.

3.3.9 Variance

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the variance (→ Definition I/1.6.1) of X is

$$\mathrm{Var}(X) = \frac{a}{b^2} \; . \tag{2}$$

Proof: The variance (→ Definition I/1.6.1) can be expressed in terms of expected values (→ Proof I/1.6.2) as

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

The expected value of a gamma random variable (→ Proof II/3.3.8) is

$$\mathrm{E}(X) = \frac{a}{b} \; . \tag{4}$$

With the probability density function of the gamma distribution (→ Proof II/3.3.5), 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 (→ Proof II/3.3.5), we get

$$\begin{split} \mathrm{E}(X^2) &= \frac{a\,(a+1)}{b^2} \int_0^\infty \mathrm{Gam}(x; a+2, b) \, \mathrm{d}x \\ &= \frac{a^2+a}{b^2} \; . \end{split} \tag{7}$$

Plugging (7) and (4) into (3), the variance of a gamma random variable finally becomes

$$\begin{split} \mathrm{Var}(X) &= \frac{a^2+a}{b^2} - \left(\frac{a}{b}\right)^2 \\ &= \frac{a}{b^2} \; . \end{split} \tag{8}$$

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.

Metadata: ID: P109 | shortcut: gam-var | author: JoramSoch | date: 2020-05-19, 07:20.

3.3.10 Logarithmic expectation

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Gam}(a, b) \; . \tag{1}$$

Then, the expectation (→ Definition I/1.5.1) 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 (→ Proof II/3.3.5) 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

$$\begin{split} f_Y(y) \, \mathrm{d}y &= \frac{1}{\Gamma(a)} \, (e^y)^a \, \exp[-e^y] \, \mathrm{d}y \\ &= \frac{1}{\Gamma(a)} \exp\left[ay - e^y\right] \, \mathrm{d}y \; . \end{split} \tag{5}$$

Because $f_Y(y)$ integrates to one, we have

$$\begin{split} 1 &= \int_{\mathbb{R}} f_Y(y) \, \mathrm{d}y \\ 1 &= \int_{\mathbb{R}} \frac{1}{\Gamma(a)} \exp\left[ay - e^y\right] \, \mathrm{d}y \\ \Gamma(a) &= \int_{\mathbb{R}} \exp\left[ay - e^y\right] \, \mathrm{d}y \; . \end{split} \tag{6}$$

Note that the integrand in (6) is differentiable with respect to a:

$$\begin{split} \frac{\mathrm{d}}{\mathrm{d}a} \exp\left[ay - e^y\right] \, \mathrm{d}y &= y \exp\left[ay - e^y\right] \, \mathrm{d}y \\ &\overset{(5)}{=} \Gamma(a) \, y \, f_Y(y) \, \mathrm{d}y \; . \end{split} \tag{7}$$

Now we can calculate the expected value of $Y = \ln(X)$:

$$\begin{split} \mathrm{E}(Y) &= \int_{\mathbb{R}} y \, f_Y(y) \, \mathrm{d}y \\ &\overset{(7)}{=} \frac{1}{\Gamma(a)} \int_{\mathbb{R}} \frac{\mathrm{d}}{\mathrm{d}a} \exp\left[ay - e^y\right] \, \mathrm{d}y \\ &= \frac{1}{\Gamma(a)} \frac{\mathrm{d}}{\mathrm{d}a} \int_{\mathbb{R}} \exp\left[ay - e^y\right] \, \mathrm{d}y \\ &\overset{(6)}{=} \frac{1}{\Gamma(a)} \frac{\mathrm{d}}{\mathrm{d}a} \Gamma(a) \\ &= \frac{\Gamma'(a)}{\Gamma(a)} \; . \end{split} \tag{8}$$

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 (→ Proof II/3.3.3) on a gamma-distributed (→ Definition II/3.3.1) random variable (→ Definition I/1.1.3),

$$X \sim \mathrm{Gam}(a,1) \quad \Rightarrow \quad \frac{1}{b} X \sim \mathrm{Gam}(a,b) \; , \tag{12}$$

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.

Metadata: ID: P110 | shortcut: gam-logmean | author: JoramSoch | date: 2020-05-25, 21:28.

3.3.11 Expectation of x ln x

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a gamma distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Gam}(a,b) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) 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 (→ Definition I/1.5.1), the law of the unconscious statistician (→ Proof I/1.5.8) and the probability density function of the gamma distribution (→ Proof II/3.3.5), 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 (→ Proof II/3.3.10) 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 (→ Proof II/3.3.10)

$$
\mathrm{E}(\ln Y) = \psi(a+1) - \ln(b) \tag{5}
$$

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 CrossValidated*, retrieved on 2020-10-15; URL: https://stats.stackexchange.com/questions/457357/what-is-the-expected-value-of-x-logx-of-the-gamma-distribution.

Metadata: ID: P179 | shortcut: gam-xlogx | author: JoramSoch | date: 2020-10-15, 13:02.

3.3.12 Kullback-Leibler divergence

Theorem: Let X be a random variable (→ Definition I/1.1.3). Assume two gamma distributions (→ Definition II/3.3.1) 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 (→ Definition 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 (→ Definition 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 (→ Definition II/3.3.1) in (1), yields

$$\begin{split} \mathrm{KL}[P\,||\,Q] &= \int_{-\infty}^{+\infty} \mathrm{Gam}(x; a_1, b_1) \, \ln \frac{\mathrm{Gam}(x; a_1, b_1)}{\mathrm{Gam}(x; a_2, b_2)} \, \mathrm{d}x \\ &= \left\langle \ln \frac{\mathrm{Gam}(x; a_1, b_1)}{\mathrm{Gam}(x; a_2, b_2)} \right\rangle_{p(x)} \; . \end{split} \tag{4}$$

Using the probability density function of the gamma distribution (→ Proof II/3.3.5), this becomes:

$$\begin{split} \mathrm{KL}[P\,||\,Q] &= \left\langle \ln \frac{\frac{{b_1}^{a_1}}{\Gamma(a_1)} x^{a_1-1} \exp[-b_1 x]}{\frac{{b_2}^{a_2}}{\Gamma(a_2)} x^{a_2-1} \exp[-b_2 x]} \right\rangle_{p(x)} \\ &= \left\langle \ln \left(\frac{{b_1}^{a_1}}{{b_2}^{a_2}} \cdot \frac{\Gamma(a_2)}{\Gamma(a_1)} \cdot x^{a_1-a_2} \cdot \exp[-(b_1 - b_2) x] \right) \right\rangle_{p(x)} \\ &= \left\langle a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \ln x - (b_1 - b_2) \cdot x \right\rangle_{p(x)} \; . \end{split} \tag{5}$$

Using the mean of the gamma distribution (→ Proof II/3.3.8) and the expected value of a logarithmized gamma variate (→ Proof II/3.3.10)

$$\begin{split} x \sim \mathrm{Gam}(a, b) \quad \Rightarrow \quad &\langle x \rangle = \frac{a}{b} \quad \text{and} \\ &\langle \ln x \rangle = \psi(a) - \ln(b) \; , \end{split} \tag{6}$$

the Kullback-Leibler divergence from (5) becomes:

$$\begin{split} \mathrm{KL}[P\,||\,Q] &= a_1 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot (\psi(a_1) - \ln(b_1)) - (b_1 - b_2) \cdot \frac{a_1}{b_1} \\ &= a_2 \cdot \ln b_1 - a_2 \cdot \ln b_2 - \ln \Gamma(a_1) + \ln \Gamma(a_2) + (a_1 - a_2) \cdot \psi(a_1) - (b_1 - b_2) \cdot \frac{a_1}{b_1} \; . \end{split} \tag{7}$$

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.

Metadata: ID: P93 | shortcut: gam-kl | author: JoramSoch | date: 2020-05-05, 08:41.

3.4 Exponential distribution

3.4.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D8 | shortcut: exp | author: JoramSoch | date: 2020-02-08, 23:48.

3.4.2 Special case of gamma distribution

Theorem: The exponential distribution (→ Definition II/3.4.1) is a special case of the gamma distribution (→ Definition II/3.3.1) with shape $a = 1$ and rate $b = \lambda$.

Proof: The probability density function of the gamma distribution (→ Proof II/3.3.5) 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

$$\begin{aligned} \mathrm{Gam}(x; 1, \lambda) &= \frac{\lambda^1}{\Gamma(1)} x^{1-1} \exp[-\lambda x] \\ &= \frac{x^0}{\Gamma(1)} \lambda \exp[-\lambda x] \\ &= \lambda \exp[-\lambda x] \end{aligned} \tag{2}$$

which is equivalent to the probability density function of the exponential distribution (→ Proof II/3.4.3).

Sources:
- original work

Metadata: ID: P69 | shortcut: exp-gam | author: JoramSoch | date: 2020-03-02, 20:49.

3.4.3 Probability density function

Theorem: Let X be a non-negative random variable (→ Definition I/1.1.3) following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) of X is

$$f_X(x) = \lambda \exp[-\lambda x] \; . \tag{2}$$

Proof: This follows directly from the definition of the exponential distribution (→ Definition II/3.4.1).

Sources:
- original work

Metadata: ID: P46 | shortcut: exp-pdf | author: JoramSoch | date: 2020-02-08, 23:53.

3.4.4 Cumulative distribution function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Proof II/3.4.3) 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 (→ Definition I/1.4.8) 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):

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{0} \mathrm{Exp}(z;\lambda)\,\mathrm{d}z + \int_{0}^{x} \mathrm{Exp}(z;\lambda)\,\mathrm{d}z \\
&= \int_{-\infty}^{0} 0\,\mathrm{d}z + \int_{0}^{x} \lambda \exp[-\lambda z]\,\mathrm{d}z \\
&= 0 + \lambda \left[-\frac{1}{\lambda} \exp[-\lambda z] \right]_{0}^{x} \\
&= \lambda \left[\left(-\frac{1}{\lambda} \exp[-\lambda x] \right) - \left(-\frac{1}{\lambda} \exp[-\lambda \cdot 0] \right) \right] \\
&= 1 - \exp[-\lambda x] \; .
\end{aligned} \tag{6}
$$

Sources:
- original work

Metadata: ID: P48 | shortcut: exp-cdf | author: JoramSoch | date: 2020-02-11, 14:48.

3.4.5 Quantile function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following an exponential distribution (→ Definition II/3.4.1):

$$
X \sim \mathrm{Exp}(\lambda) \; . \tag{1}
$$

Then, the quantile function (→ Definition I/1.4.13) 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 (→ Proof II/3.4.4) 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 $Q_X(p)$ is defined as (→ Definition I/1.4.13) 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 (→ Proof I/1.4.14)

$$
Q_X(p) = F_X^{-1}(x) \; . \tag{5}
$$

This can be derived by rearranging equation (3):

$$
\begin{aligned}
p &= 1 - \exp[-\lambda x] \\
\exp[-\lambda x] &= 1 - p \\
-\lambda x &= \ln(1-p) \\
x &= -\frac{\ln(1-p)}{\lambda} \; .
\end{aligned} \tag{6}
$$

Sources:
- original work

Metadata: ID: P50 | shortcut: exp-qf | author: JoramSoch | date: 2020-02-12, 15:48.

3.4.6 Mean

Theorem: Let X be a random variable (→ Definition I/1.1.3) following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = \frac{1}{\lambda} \; . \tag{2}$$

Proof: The expected value (→ Definition I/1.5.1) 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 (→ Proof II/3.4.3), 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-deriative

$$\int x \cdot \exp(-\lambda x) \, \mathrm{d}x = \left(-\frac{1}{\lambda} x - \frac{1}{\lambda^2} \right) \exp(-\lambda x) \; , \tag{5}$$

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.

Metadata: ID: P47 | shortcut: exp-mean | author: JoramSoch | date: 2020-02-10, 21:57.

3.4.7 Median

Theorem: Let X be a random variable (→ Definition I/1.1.3) following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the median (→ Definition I/1.9.1) of X is

$$\mathrm{median}(X) = \frac{\ln 2}{\lambda} \; . \tag{2}$$

Proof: The median (→ Definition I/1.9.1) is the value at which the cumulative distribution function (→ Definition I/1.4.8) is 1/2:

$$F_X(\mathrm{median}(X)) = \frac{1}{2} \; . \tag{3}$$

The cumulative distribution function of the exponential distribution (→ Proof II/3.4.4) 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:

$$\mathrm{median}(X) = -\frac{\ln(1-\frac{1}{2})}{\lambda} = \frac{\ln 2}{\lambda} \; . \tag{6}$$

Sources:
- original work

Metadata: ID: P49 | shortcut: exp-med | author: JoramSoch | date: 2020-02-11, 15:03.

3.4.8 Mode

Theorem: Let X be a random variable (→ Definition I/1.1.3) following an exponential distribution (→ Definition II/3.4.1):

$$X \sim \mathrm{Exp}(\lambda) \; . \tag{1}$$

Then, the mode (→ Definition I/1.9.2) of X is

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

Proof: The mode (→ Definition I/1.9.2) is the value which maximizes the probability density function (→ Definition I/1.4.4):

$$\mathrm{mode}(X) = \operatorname*{arg\,max}_x f_X(x) \; . \tag{3}$$

The probability density function of the exponential distribution (→ Proof II/3.4.3) is:

$$f_X(x) = \left\{ \begin{array}{rl} 0 \; , & \text{if } x < 0 \\ \lambda \exp[-\lambda x] \; , & \text{if } x \geq 0 \; . \end{array} \right. \tag{4}$$

Since

$$\lim_{x \rightarrow 0} f_X(x) = \infty \tag{5}$$

and

$$f_X(x) < \infty \quad \text{for any} \quad x \neq 0 \; , \tag{6}$$

it follows that

$$\mathrm{mode}(X) = 0 \; . \tag{7}$$

Sources:
- original work

Metadata: ID: P51 | shortcut: exp-mode | author: JoramSoch | date: 2020-02-12, 15:53.

3.5 Chi-square distribution

3.5.1 Definition

Definition: Let $X_1, ..., X_k$ be independent (→ Definition I/1.2.6) random variables (→ Definition I/1.1.3) where each of them is following a standard normal distribution (→ Definition II/3.2.2):

$$X_i \sim \mathcal{N}(0,1) \; . \tag{1}$$

Then, the sum of their squares follows a chi-square 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-square distribution (→ Proof II/3.5.3) with k degress 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.

Metadata: ID: D100 | shortcut: chi2 | author: kjpetrykowski | date: 2020-10-13, 01:20.

3.5.2 Special case of gamma distribution

Theorem: The chi-square distribution (→ Definition II/3.5.1) with k degrees of freedom is a special case of the gamma distribution (→ Definition II/3.3.1) 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 (→ Proof II/3.3.5) for $x > 0$, where α is the shape parameter and β is the rate paramete, 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-square distribution (→ Proof II/3.5.3).

Sources:
- original work

Metadata: ID: P174 | shortcut: chi2-gam | author: kjpetrykowski | date: 2020-10-12, 22:15.

3.5.3 Probability density function

Theorem: Let Y be a random variable (→ Definition I/1.1.3) following a chi-square distribution (→ Definition II/3.5.1):

$$Y \sim \chi^2(k) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/3.5.1) with k degrees of freedom is defined as the sum of k squared standard normal random variables (→ Definition II/3.2.2):

$$X_1, \ldots, X_k \sim \mathcal{N}(0,1) \quad \Rightarrow \quad Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k) \; . \tag{3}$$

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 (→ Definition I/1.4.4) and cumulative distribution function (→ Definition I/1.4.8) of Y. Because the PDF is the first derivative of the CDF (→ Proof I/1.4.7), 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 (→ Definition I/1.4.8) 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 (→ Definition I/1.4.4) of the standard normal distribution (→ Definition II/3.2.2) 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 (→ Definition "norm-pdf"), equation (6) can be developed as follows:

$$\begin{aligned} f_Y(y) \, \mathrm{d}y &= \int_V \prod_{i=1}^{k} \left(\frac{1}{\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2} x_i^2 \right] \, \mathrm{d}x_i \right) \\ &= \int_V \frac{\exp\left[-\frac{1}{2} \left(x_1^2 + \ldots + x_k^2 \right) \right]}{(2\pi)^{k/2}} \, \mathrm{d}x_1 \, \ldots \, \mathrm{d}x_k \\ &= \frac{1}{(2\pi)^{k/2}} \int_V \exp\left[-\frac{y}{2} \right] \, \mathrm{d}x_1 \, \ldots \, \mathrm{d}x_k \; . \end{aligned} \tag{7}$$

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:

$$
\begin{aligned}
f_Y(y)\,\mathrm{d}y &= \frac{\exp\left[-y/2\right]}{(2\pi)^{k/2}} \cdot A\,\mathrm{d}r \\
&= \frac{\exp\left[-y/2\right]}{(2\pi)^{k/2}} \cdot 2r^{k-1}\,\frac{\pi^{k/2}}{\Gamma(k/2)} \cdot \frac{\mathrm{d}y}{2y^{1/2}} \\
&= \frac{1}{2^{k/2}\,\Gamma(k/2)} \cdot \frac{2\sqrt{y}^{k-1}}{2\sqrt{y}} \cdot \exp\left[-y/2\right]\,\mathrm{d}y \\
&= \frac{1}{2^{k/2}\,\Gamma(k/2)} \cdot y^{\frac{k}{2}-1} \cdot \exp\left[-\frac{y}{2}\right]\,\mathrm{d}y \; .
\end{aligned} \tag{11}
$$

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.

Metadata: ID: P197 | shortcut: chi2-pdf | author: JoramSoch | date: 2020-11-25, 05:56.

3.5.4 Moments

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a chi-square distribution (→ Definition II/3.5.1):

$$
X \sim \chi^2(k) \; . \tag{1}
$$

If $m > -k/2$, then $E(X^m)$ exists and is equal to:

$$
\mathrm{E}(X^m) = \frac{2^m \Gamma\left(\frac{k}{2}+m\right)}{\Gamma\left(\frac{k}{2}\right)} \; . \tag{2}
$$

Proof: Combining the definition of the m-th raw moment (→ Definition I/1.12.3) with the probability density function of the chi-square distribution (→ Proof II/3.5.3), we have:

$$
\mathrm{E}(X^m) = \int_0^\infty \frac{1}{\Gamma\left(\frac{k}{2}\right) 2^{k/2}}\, x^{(k/2)+m-1}\, e^{-x/2} \mathrm{d}x \; . \tag{3}
$$

Now define a new variable $u = x/2$. As a result, we obtain:

$$
\mathrm{E}(X^m) = \int_0^\infty \frac{1}{\Gamma\left(\frac{k}{2}\right) 2^{(k/2)-1}}\, 2^{(k/2)+m-1}\, u^{(k/2)+m-1}\, e^{-u} \mathrm{d}u \; . \tag{4}
$$

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 (→ Definition I/1.12.1) of a chi-square distribution (→ Definition II/3.5.1) exist and the m-th raw moment is given by the foregoing equation.

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.

Metadata: ID: P175 | shortcut: chi2-mom | author: kjpetrykowski | date: 2020-10-13, 01:30.

3.6 Beta distribution

3.6.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D53 | shortcut: beta | author: JoramSoch | date: 2020-05-10, 20:29.

3.6.2 Probability density function

Theorem: Let X be a random variable (→ Definition I/1.1.3) following a beta distribution (→ Definition II/3.6.1):

$$X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/3.6.1).

Sources:
- original work

Metadata: ID: P94 | shortcut: beta-pdf | author: JoramSoch | date: 2020-05-05, 21:03.

3.6.3 Moment-generating function

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a beta distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}$$

Then, the moment-generating function (→ Definition I/1.4.15) 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 (→ Proof II/3.6.2) is

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \tag{3}$$

and the moment-generating function (→ Definition I/1.4.15) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX} \right] \; . \tag{4}$$

Using the expected value for continuous random variables (→ Definition I/1.5.1), the moment-generating function of X therefore is

$$\begin{split} M_X(t) &= \int_0^1 \exp[tx] \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \\ &= \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 e^{tx} \, x^{\alpha-1} \, (1-x)^{\beta-1} \, \mathrm{d}x \; . \end{split} \tag{5}$$

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.

Metadata: ID: P198 | shortcut: beta-mgf | author: JoramSoch | date: 2020-11-25, 06:55.

3.6.4 Cumulative distribution function

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a beta distribution (→ Definition II/3.3.1):

$$X \sim \mathrm{Bet}(\alpha, \beta) \; . \tag{1}$$

Then, the cumulative distribution function (→ Definition I/1.4.8) 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 (→ Proof II/3.6.2) is:

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \, x^{\alpha-1} \, (1-x)^{\beta-1} \; . \tag{3}$$

Thus, the cumulative distribution function (→ Definition I/1.4.8) is:

$$\begin{aligned} F_X(x) &= \int_0^x \mathrm{Bet}(z; \alpha, \beta) \, \mathrm{d}z \\ &= \int_0^x \frac{1}{\mathrm{B}(\alpha, \beta)} \, z^{\alpha-1} \, (1-z)^{\beta-1} \, \mathrm{d}z \\ &= \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^x z^{\alpha-1} \, (1-z)^{\beta-1} \, \mathrm{d}z \; . \end{aligned} \tag{4}$$

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.

Metadata: ID: P195 | shortcut: beta-cdf | author: JoramSoch | date: 2020-11-19, 08:01.

3.7 Wald distribution

3.7.1 Definition

Definition: Let X be a random variable (→ Definition I/1.1.3). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D95 | shortcut: wald | author: tomfaulkenberry | date: 2020-09-04, 12:00.

3.7.2 Probability density function

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a Wald distribution (→ Definition II/3.7.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/3.7.1).

Sources:
- original work

Metadata: ID: P162 | shortcut: wald-pdf | author: tomfaulkenberry | date: 2020-09-04, 12:00.

3.7.3 Moment-generating function

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a Wald distribution (→ Definition II/3.7.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) . \tag{1}$$

Then, the moment-generating function (→ Definition I/1.4.15) 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 (→ Proof II/3.7.2) 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 (→ Definition I/1.4.15) is defined as

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] . \tag{4}$$

Using the definition of expected value for continuous random variables (→ Definition I/1.5.1), 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](https://dlmf.nist.gov/10.25), 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](https://dlmf.nist.gov/10.39.2) 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](https://dlmf.nist.gov/10.32.10) 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

$$\begin{aligned} M_X(t) &= \frac{\alpha}{\sqrt{2\pi}} \cdot e^{\alpha\gamma} \cdot 2\left(\frac{\gamma^2 - 2t}{\alpha^2}\right)^{1/4} \cdot \sqrt{\frac{\pi}{2\sqrt{\alpha^2(\gamma^2 - 2t)}}} \cdot \exp\left[-\sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\ &= \frac{\alpha}{\sqrt{2} \cdot \sqrt{\pi}} \cdot e^{\alpha\gamma} \cdot 2 \cdot \frac{(\gamma^2 - 2t)^{1/4}}{\sqrt{\alpha}} \cdot \frac{\sqrt{\pi}}{\sqrt{2} \cdot \sqrt{\alpha} \cdot (\gamma^2 - 2t)^{1/4}} \cdot \exp\left[-\sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\ &= e^{\alpha\gamma} \cdot \exp\left[-\sqrt{\alpha^2(\gamma^2 - 2t)}\right] \\ &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \; . \end{aligned} \tag{11}$$

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.

Metadata: ID: P168 | shortcut: wald-mgf | author: tomfaulkenberry | date: 2020-09-13, 12:00.

3.7.4 Mean

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a Wald distribution (→ Definition II/3.7.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}$$

Then, the mean or expected value (→ Definition I/1.5.1) of X is

$$\mathrm{E}(X) = \frac{\alpha}{\gamma} \; . \tag{2}$$

Proof: The mean or expected value $\mathrm{E}(X)$ is the first moment (→ Definition I/1.12.1) of X, so we can use (→ Proof I/1.12.2) the moment-generating function of the Wald distribution (→ Proof II/3.7.3) 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

$$\begin{split} M_X'(t) &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}\left(\alpha^2(\gamma^2 - 2t)\right)^{-1/2} \cdot -2\alpha^2 \\ &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2t)}} \; . \end{split} \tag{5}$$

Evaluating (5) at $t = 0$ gives the desired result:

$$\begin{split} M_X'(0) &= \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2(0))}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2(\gamma^2 - 2(0))}} \\ &= \exp\left[\alpha\gamma - \sqrt{\alpha^2 \cdot \gamma^2}\right] \cdot \frac{\alpha^2}{\sqrt{\alpha^2 \cdot \gamma^2}} \\ &= \exp[0] \cdot \frac{\alpha^2}{\alpha\gamma} \\ &= \frac{\alpha}{\gamma} \; . \end{split} \tag{6}$$

Sources:
- original work

Metadata: ID: P169 | shortcut: wald-mean | author: tomfaulkenberry | date: 2020-09-13, 12:00.

3.7.5 Variance

Theorem: Let X be a positive random variable (→ Definition I/1.1.3) following a Wald distribution (→ Definition II/3.7.1):

$$X \sim \mathrm{Wald}(\gamma, \alpha) \; . \tag{1}$$

Then, the variance (→ Definition I/1.6.1) 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 (→ Proof I/1.6.2):

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2. \tag{3}$$

We then use the moment-generating function of the Wald distribution (→ Proof II/3.7.3) 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:

$$\begin{aligned}
M_X''(t) &= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1/2} \cdot -\frac{1}{2}\left(\alpha^2(\gamma^2 - 2t)\right)^{-1/2} \cdot -2\alpha^2 \\
&+ \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot -\frac{1}{2}(\gamma^2 - 2t)^{-3/2} \cdot -2 \\
&= \alpha^2 \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-1} \\
&+ \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \cdot (\gamma^2 - 2t)^{-3/2} \\
&= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2t)}\right] \left[\frac{\alpha}{\gamma^2 - 2t} + \frac{1}{\sqrt{(\gamma^2 - 2t)^3}}\right] \; .
\end{aligned} \tag{7}$$

Applying (4) yields

$$\begin{aligned}
\mathrm{E}(X^2) &= M_X''(0) \\
&= \alpha \cdot \exp\left[\alpha\gamma - \sqrt{\alpha^2(\gamma^2 - 2(0))}\right] \left[\frac{\alpha}{\gamma^2 - 2(0)} + \frac{1}{\sqrt{(\gamma^2 - 2(0))^3}}\right] \\
&= \alpha \cdot \exp\left[\alpha\gamma - \alpha\gamma\right] \cdot \left[\frac{\alpha}{\gamma^2} + \frac{1}{\gamma^3}\right] \\
&= \frac{\alpha^2}{\gamma^2} + \frac{\alpha}{\gamma^3} \; .
\end{aligned} \tag{8}$$

Since the mean of a Wald distribution (→ Proof II/3.7.4) 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).

Sources:
- original work

Metadata: ID: P170 | shortcut: wald-var | author: tomfaulkenberry | date: 2020-09-13, 12:00.

4 Multivariate continuous distributions

4.1 Multivariate normal distribution

4.1.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ Definition I/1.1.4). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D1 | shortcut: mvn | author: JoramSoch | date: 2020-01-22, 05:20.

4.1.2 Probability density function

Theorem: Let X be a random vector (→ Definition I/1.1.4) following a multivariate normal distribution (→ Definition II/4.1.1):

$$X \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/4.1.1).

Sources:
- original work

Metadata: ID: P34 | shortcut: mvn-pdf | author: JoramSoch | date: 2020-01-27, 15:23.

4.1.3 Differential entropy

Theorem: Let x follow a multivariate normal distribution (→ Definition II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the differential entropy (→ Definition 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 (→ Definition 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 (→ Definition I/1.5.1)

$$\mathrm{h}(X) = - \mathrm{E}\left[\ln p(x) \right] \; . \tag{4}$$

With the probability density function of the multivariate normal distribution (→ Proof II/4.1.2), 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

Metadata: ID: P100 | shortcut: mvn-dent | author: JoramSoch | date: 2020-05-14, 19:49.

4.1.4 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (→ Definition I/1.1.4). Assume two multivariate normal distributions (→ Definition II/4.1.1) P and Q specifying the probability distribution of x as

$$\begin{aligned} P:\; x &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ Q:\; x &\sim \mathcal{N}(\mu_2, \Sigma_2) \; . \end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence (→ Definition I/2.5.1) of P from Q is given by

$$\mathrm{KL}[P\,||\,Q] = \frac{1}{2} \left[(\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - n \right] \; . \tag{2}$$

Proof: The KL divergence for a continuous random variable (→ Definition 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 (→ Definition II/4.1.1) in (1), yields

$$\begin{aligned} \mathrm{KL}[P\,||\,Q] &= \int_{\mathbb{R}^n} \mathcal{N}(x; \mu_1, \Sigma_1) \, \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} \, \mathrm{d}x \\ &= \left\langle \ln \frac{\mathcal{N}(x; \mu_1, \Sigma_1)}{\mathcal{N}(x; \mu_2, \Sigma_2)} \right\rangle_{p(x)} \; . \end{aligned} \tag{4}$$

Using the probability density function of the multivariate normal distribution (→ Proof II/4.1.2), this becomes:

$$\begin{aligned} \mathrm{KL}[P\,||\,Q] &= \left\langle \ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_1|}} \cdot \exp\left[-\frac{1}{2}(x-\mu_1)^\mathrm{T} \Sigma_1^{-1} (x-\mu_1) \right]}{\frac{1}{\sqrt{(2\pi)^n |\Sigma_2|}} \cdot \exp\left[-\frac{1}{2}(x-\mu_2)^\mathrm{T} \Sigma_2^{-1} (x-\mu_2) \right]} \right\rangle_{p(x)} \\ &= \left\langle \frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^\mathrm{T} \Sigma_1^{-1} (x-\mu_1) + \frac{1}{2}(x-\mu_2)^\mathrm{T} \Sigma_2^{-1} (x-\mu_2) \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - (x-\mu_1)^\mathrm{T} \Sigma_1^{-1} (x-\mu_1) + (x-\mu_2)^\mathrm{T} \Sigma_2^{-1} (x-\mu_2) \right\rangle_{p(x)} \; . \end{aligned} \tag{5}$$

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:

$$\begin{aligned} \mathrm{KL}[P\,||\,Q] &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1} (x-\mu_1)(x-\mu_1)^\mathrm{T} \right] + \mathrm{tr}\left[\Sigma_2^{-1} (x-\mu_2)(x-\mu_2)^\mathrm{T} \right] \right\rangle_{p(x)} \\ &= \frac{1}{2} \left\langle \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1} (x-\mu_1)(x-\mu_1)^\mathrm{T} \right] + \mathrm{tr}\left[\Sigma_2^{-1} \left(x x^\mathrm{T} - 2\mu_2 x^\mathrm{T} + \mu_2 \mu_2^\mathrm{T} \right) \right] \right\rangle_{p(x)} \; . \end{aligned} \tag{6}$$

Because trace function and expected value (→ Definition I/1.5.1) are both linear operators, the expectation can be moved inside the trace:

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}\left\langle (x-\mu_1)(x-\mu_1)^\mathrm{T}\right\rangle_{p(x)}\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left\langle xx^\mathrm{T} - 2\mu_2 x^\mathrm{T} + \mu_2\mu_2^\mathrm{T}\right\rangle_{p(x)}\right]\right) \\
&= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}\left\langle (x-\mu_1)(x-\mu_1)^\mathrm{T}\right\rangle_{p(x)}\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(\left\langle xx^\mathrm{T}\right\rangle_{p(x)} - \left\langle 2\mu_2 x^\mathrm{T}\right\rangle_{p(x)} + \left\langle \mu_2\mu_2^\mathrm{T}\right\rangle_{p(x)}\right)\right]\right)
\end{aligned} \tag{7}
$$

Using the expectation of a linear form for the multivariate normal distribution (→ Proof II/4.1.5)

$$
x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \langle Ax \rangle = A\mu \tag{8}
$$

and the expectation of a quadratic form for the multivariate normal distribution (→ Proof I/1.5.7)

$$
x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad \left\langle x^\mathrm{T} A x \right\rangle = \mu^\mathrm{T} A \mu + \mathrm{tr}(A\Sigma) \; , \tag{9}
$$

the Kullback-Leibler divergence from (7) becomes:

$$
\begin{aligned}
\mathrm{KL}[P\,||\,Q] &= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[\Sigma_1^{-1}\Sigma_1\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(\Sigma_1 + \mu_1\mu_1^\mathrm{T} - 2\mu_2\mu_1^\mathrm{T} + \mu_2\mu_2^\mathrm{T}\right)\right]\right) \\
&= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{tr}\left[I_n\right] + \mathrm{tr}\left[\Sigma_2^{-1}\Sigma_1\right] + \mathrm{tr}\left[\Sigma_2^{-1}\left(\mu_1\mu_1^\mathrm{T} - 2\mu_2\mu_1^\mathrm{T} + \mu_2\mu_2^\mathrm{T}\right)\right]\right) \\
&= \frac{1}{2}\left(\ln\frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}\left[\Sigma_2^{-1}\Sigma_1\right] + \mathrm{tr}\left[\mu_1^\mathrm{T}\Sigma_2^{-1}\mu_1 - 2\mu_1^\mathrm{T}\Sigma_2^{-1}\mu_2 + \mu_2^\mathrm{T}\Sigma_2^{-1}\mu_2\right]\right) \\
&= \frac{1}{2}\left[\ln\frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{tr}\left[\Sigma_2^{-1}\Sigma_1\right] + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right] \; .
\end{aligned} \tag{10}
$$

Finally, rearranging the terms, we get:

$$
\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \; . \tag{11}
$$

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.

Metadata: ID: P92 | shortcut: mvn-kl | author: JoramSoch | date: 2020-05-05, 06:57.

4.1.5 Linear transformation

Theorem: Let x follow a multivariate normal distribution (→ Definition 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 (→ Definition I/1.4.15) 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

$$\begin{aligned} M_y(t) &\overset{(2)}{=} \mathbb{E}\left(\exp\left[t^\mathrm{T}(Ax+b)\right]\right) \\ &= \mathbb{E}\left(\exp\left[t^\mathrm{T} Ax\right] \cdot \exp\left[t^\mathrm{T} b\right]\right) \\ &= \exp\left[t^\mathrm{T} b\right] \cdot \mathbb{E}\left(\exp\left[t^\mathrm{T} Ax\right]\right) \\ &\overset{(3)}{=} \exp\left[t^\mathrm{T} b\right] \cdot M_x(At) \; . \end{aligned} \tag{4}$$

The moment-generating function of the multivariate normal distribution (→ Proof "mvn-mgf") 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

$$\begin{aligned} M_y(t) &\overset{(4)}{=} \exp\left[t^\mathrm{T} b\right] \cdot M_x(At) \\ &\overset{(5)}{=} \exp\left[t^\mathrm{T} b\right] \cdot \exp\left[t^\mathrm{T} A\mu + \frac{1}{2} t^\mathrm{T} A\Sigma A^\mathrm{T} t\right] \\ &= \exp\left[t^\mathrm{T}\left(A\mu + b\right) + \frac{1}{2} t^\mathrm{T} A\Sigma A^\mathrm{T} t\right] \; . \end{aligned} \tag{6}$$

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.

Metadata: ID: P1 | shortcut: mvn-ltt | author: JoramSoch | date: 2019-08-27, 12:14.

4.1.6 Marginal distributions

Theorem: Let x follow a multivariate normal distribution (→ Definition II/4.1.1):

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the marginal distribution (→ Definition I/1.3.3) 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 μ_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 (→ Proof II/4.1.5) 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 $S\Sigma S^{\mathrm{T}} = \Sigma_s$.

Sources:
- original work

Metadata: ID: P35 | shortcut: mvn-marg | author: JoramSoch | date: 2020-01-29, 15:12.

4.1.7 Conditional distributions

Theorem: Let x follow a multivariate normal distribution (→ Definition II/4.1.1)

$$x \sim \mathcal{N}(\mu, \Sigma) \; . \tag{1}$$

Then, the conditional distribution (→ Definition I/1.3.4) 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 (→ Definition I/1.5.1) and covariance (→ Definition I/1.7.1) 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 (→ Definition I/1.3.2) of x_1 and x_2 is:

$$x_1, x_2 \sim \mathcal{N}(\mu, \Sigma) \; . \tag{6}$$

Moreover, the marginal distribution (→ Definition I/1.3.3) of x_2 follows from (→ Proof II/4.1.6) (1) and (4) as

$$x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) \; . \tag{7}$$

According to the law of conditional probability (→ Definition I/1.2.4), 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 (→ Proof II/4.1.2), 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:

$$\begin{split} p(x_1|x_2) = &\frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\exp\left[-\frac{1}{2}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^\mathrm{T} \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\right. \\ &\qquad \left. + \frac{1}{2}\left(x_2-\mu_2\right)^\mathrm{T}\Sigma_{22}^{-1}\left(x_2-\mu_2\right)\right] \; . \end{split} \tag{12}$$

Multiplying out within the exponent of (12), we have

$$\begin{split} p(x_1|x_2) = &\frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ &\exp\left[-\frac{1}{2}\left((x_1-\mu_1)^\mathrm{T}\Sigma^{11}(x_1-\mu_1) + 2(x_1-\mu_1)^\mathrm{T}\Sigma^{12}(x_2-\mu_2) + (x_2-\mu_2)^\mathrm{T}\Sigma^{22}(x_2-\mu_2)\right)\right. \\ &\qquad \left. + \frac{1}{2}(x_2-\mu_2)^\mathrm{T}\Sigma_{22}^{-1}(x_2-\mu_2)\right] \end{split} \tag{13}$$

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:

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ & \exp\left[-\frac{1}{2}\big((x_1 - \mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(x_1 - \mu_1) - \right. \\ & \qquad 2(x_1 - \mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) + \\ & \qquad (x_2 - \mu_2)^{\mathrm{T}}\left[\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\right](x_2 - \mu_2)\big) \\ & \left. + \frac{1}{2}\big((x_2 - \mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\big) \right]. \end{aligned} \tag{16}$$

Eliminating some terms, we have:

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ & \exp\left[-\frac{1}{2}\big((x_1 - \mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(x_1 - \mu_1) - \right. \\ & \qquad 2(x_1 - \mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) + \\ & \qquad \left. (x_2 - \mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)\big) \right]. \end{aligned} \tag{17}$$

Rearranging the terms, we have

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2} \cdot \right. \\ & \left. \left[(x_1 - \mu_1) - \Sigma_{12}^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\right]^{\mathrm{T}} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[(x_1 - \mu_1) - \Sigma_{12}^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\right] \right] \\ =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\left[-\frac{1}{2} \cdot \right. \\ & \left. \left[x_1 - \big(\mu_1 + \Sigma_{12}^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\big)\right]^{\mathrm{T}} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - \big(\mu_1 + \Sigma_{12}^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\big)\right] \right] \end{aligned} \tag{18}$$

where we have used the fact that $\Sigma_{21}^{\mathrm{T}} = \Sigma_{12}$, because Σ is a covariance matrix.

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

$$\begin{split} p(x_1|x_2) = & \frac{1}{\sqrt{(2\pi)^{n_1}|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|}} \cdot \exp\left[-\frac{1}{2} \cdot \right. \\ & \left. \left[x_1 - \left(\mu_1 + \Sigma_{12}^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\right)\right]^{\mathrm{T}} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - \left(\mu_1 + \Sigma_{12}^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)\right)\right] \right] \end{split} \tag{21}$$

which is the probability density function of a multivariate normal distribution (→ Proof II/4.1.2)

$$p(x_1|x_2) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{1|2}) \tag{22}$$

with the mean $\mu_{1|2}$ and variance $\Sigma_{1|2}$ 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.

Metadata: ID: P88 | shortcut: mvn-cond | author: JoramSoch | date: 2020-03-20, 08:44.

4.2 Normal-gamma distribution

4.2.1 Definition

Definition: Let X be an $n \times 1$ random vector (→ Definition I/1.1.4) and let Y be a positive random variable (→ Definition I/1.1.3). Then, X and Y are said to follow a normal-gamma distribution

$$X, Y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; , \tag{1}$$

if and only if their joint probability (→ Definition I/1.2.2) density function (→ Definition I/1.4.4) is given by

$$f_{X,Y}(x, y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{2}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is the probability density function of the multivariate normal distribution (→ Proof II/4.1.2) with mean μ and covariance Σ and $\mathrm{Gam}(x; a, b)$ is the probability density function of the gamma distribution (→ Proof II/3.3.5) with shape a and rate b. The $n \times n$ matrix Λ is referred to as the precision matrix (→ Definition I/1.7.8) 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.

Metadata: ID: D5 | shortcut: ng | author: JoramSoch | date: 2020-01-27, 14:28.

4.2.2 Probability density function

Theorem: Let x and y follow a normal-gamma distribution (→ Definition II/4.2.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then, the joint probability (→ Definition I/1.2.2) density function (→ Definition I/1.4.4) 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 probability density of the normal-gamma distribution is defined as (→ Definition II/4.2.1) as the product of a multivariate normal distribution (→ Definition II/4.1.1) over x conditional on y and a univariate gamma distribution (→ Definition II/3.3.1) over y:

$$p(x,y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{3}$$

With the probability density function of the multivariate normal distribution (→ Proof II/4.1.2) and the probability density function of the gamma distribution (→ Proof II/3.3.5), this becomes:

$$p(x,y) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \cdot \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right] \; . \tag{4}$$

Using the relation $|yA| = y^n|A|$ for an $n \times n$ matrix A and rearranging the terms, we have:

$$p(x,y) = \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \frac{b^a}{\Gamma(a)} \cdot y^{a+\frac{n}{2}-1} \exp\left[-\frac{y}{2}\left((x-\mu)^\mathrm{T}\Lambda(x-\mu) + 2b\right)\right] \; . \tag{5}$$

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.

Metadata: ID: P44 | shortcut: ng-pdf | author: JoramSoch | date: 2020-02-07, 20:44.

4.2.3 Kullback-Leibler divergence

Theorem: Let x be an $n \times 1$ random vector (→ Definition I/1.1.4) and let y be a positive random variable (→ Definition I/1.1.3). Assume two normal-gamma distributions (→ Definition II/4.2.1) P and Q specifying the joint distribution of x and y as

$$
\begin{split}
P: \; (x,y) &\sim \mathrm{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\
Q: \; (x,y) &\sim \mathrm{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) \; .
\end{split} \tag{1}
$$

Then, the Kullback-Leibler divergence (→ Definition I/2.5.1) of P from Q is given by

$$
\begin{split}
\mathrm{KL}[P\,||\,Q] = &\frac{1}{2}\frac{a_1}{b_1}\left[(\mu_2-\mu_1)^T\Lambda_2(\mu_2-\mu_1)\right] + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \\
&+ a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \; .
\end{split} \tag{2}
$$

Proof: The probability density function of the normal-gamma distribution (→ Proof II/4.2.2) 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 (→ Proof II/4.1.4) is

$$
\mathrm{KL}[P\,||\,Q] = \frac{1}{2}\left[(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - n\right] \tag{4}
$$

and the Kullback-Leibler divergence of the univariate gamma distribution (→ Proof II/3.3.12) 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 (→ Definition 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 (→ Definition II/4.2.1) 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 (→ Definition I/1.2.4), 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

$$\begin{split} &\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} \\ &= \left\langle \frac{1}{2} \left[(\mu_2-\mu_1)^T (y\Lambda_2)(\mu_2-\mu_1) + \mathrm{tr}\left((y\Lambda_2)(y\Lambda_1)^{-1} \right) - \ln \frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - n \right] \right\rangle_{p(y)} \\ &= \left\langle \frac{y}{2} (\mu_2-\mu_1)^T \Lambda_2 (\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \right\rangle_{p(y)} \end{split} \tag{9}$$

and using the relation (→ Proof II/3.3.8) $y \sim \mathrm{Gam}(a,b) \Rightarrow \langle y \rangle = a/b$, we have

$$\left\langle \mathrm{KL}[p(x|y)\,||\,q(x|y)] \right\rangle_{p(y)} = \frac{1}{2} \frac{a_1}{b_1} (\mu_2-\mu_1)^T \Lambda_2 (\mu_2-\mu_1) + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{n}{2} \; . \tag{10}$$

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.

Metadata: ID: P6 | shortcut: ng-kl | author: JoramSoch | date: 2019-12-06, 09:35.

4.2.4 Marginal distributions

Theorem: Let x and y follow a normal-gamma distribution (→ Definition II/4.2.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then, the marginal distribution (→ Definition I/1.3.3) of y is a gamma distribution (→ Definition II/3.3.1)

$$y \sim \mathrm{Gam}(a, b) \tag{2}$$

and the marginal distribution (→ Definition I/1.3.3) of x is a multivariate t-distribution (→ Definition "mvt")

$$x \sim \mathrm{t}\left(\mu, \left(\frac{a}{b} \Lambda \right)^{-1}, 2a \right) \; . \tag{3}$$

Proof: The probability density function of the normal-gamma distribution (→ Proof II/4.2.2) is given by

$$\begin{split} p(x,y) &= p(x|y) \cdot p(y) \\ p(x|y) &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \\ p(y) &= \mathrm{Gam}(y; a, b) \; . \end{split} \tag{4}$$

Using the law of marginal probability (→ Definition I/1.2.3), 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 (→ Proof II/3.3.5) with shape parameter a and rate parameter b.

Using the law of marginal probability (→ Definition I/1.2.3), 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) \\
&= \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 (→ Proof "mvt-pdf") with mean vector μ, shape matrix $\left(\frac{a}{b}\Lambda\right)^{-1}$ and $2a$ degrees of freedom.

Sources:
- original work

Metadata: ID: P36 | shortcut: ng-marg | author: JoramSoch | date: 2020-01-29, 21:42.

4.2.5 Conditional distributions

Theorem: Let x and y follow a normal-gamma distribution (→ Definition II/4.2.1):

$$x, y \sim \mathrm{NG}(\mu, \Lambda, a, b) \; . \tag{1}$$

Then,
1) the conditional distribution (→ Definition I/1.3.4) of x given y is a multivariate normal distribution (→ Definition II/4.1.1)

$$x|y \sim \mathcal{N}(\mu, (y\Lambda)^{-1}) \; ; \tag{2}$$

2) the conditional distribution (→ Definition I/1.3.4) of a subset vector x_1, given the complement vector x_2 and y, is also a multivariate normal distribution (→ Definition 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 (→ Definition I/1.5.1) and covariance (→ Definition I/1.7.1)

$$\begin{aligned} \mu_{1|2}(y) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2}(y) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} \end{aligned} \tag{4}$$

where μ_1, μ_2 and Σ_{11}, Σ_{12}, Σ_{22}, Σ_{21} are block-wise components (→ Proof II/4.1.7) of μ and $\Sigma(y) = (y\Lambda)^{-1}$;
3) the conditional distribution (→ Definition I/1.3.4) of y given x is a gamma distribution (→ Definition II/3.3.1)

$$y|x \sim \mathrm{Gam}\left(a + \frac{n}{2}, b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right) \tag{5}$$

where n is the dimensionality of x.

Proof:
1) This follows from the definition of the normal-gamma distribution (→ Definition II/4.2.1):

$$\begin{aligned} p(x, y) &= p(x|y) \cdot p(y) \\ &= \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \; . \end{aligned} \tag{6}$$

2) This follows from (2) and the conditional distributions of the multivariate normal distribution (→ Proof II/4.1.7):

$$\begin{aligned} & x \sim \mathcal{N}(\mu, \Sigma) \\ \Rightarrow \; & x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2}) \\ & \mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ & \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \; . \end{aligned} \tag{7}$$

3) The conditional density of y given x follows from Bayes' theorem (→ Proof I/5.3.1) as

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)} \; . \tag{8}$$

The conditional distribution (→ Definition I/1.3.4) of x given y is a multivariate normal distribution (→ Proof II/4.2.2)

$$p(x|y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) = \sqrt{\frac{|y\Lambda|}{(2\pi)^n}} \exp\left[-\frac{1}{2}(x-\mu)^\mathrm{T}(y\Lambda)(x-\mu)\right] \; , \tag{9}$$

the marginal distribution (→ Definition I/1.3.3) of y is a gamma distribution (→ Proof II/4.2.4)

$$p(y) = \mathrm{Gam}(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp\left[-by\right] \tag{10}$$

and the marginal distribution (→ Definition I/1.3.3) of x is a multivariate t-distribution (→ Proof II/4.2.4)

$$\begin{split} p(x) &= \mathrm{t}\left(x; \mu, \left(\frac{a}{b}\Lambda\right)^{-1}, 2a\right) \\ &= \sqrt{\frac{\left|\frac{a}{b}\,\Lambda\right|}{(2a\,\pi)^n}} \cdot \frac{\Gamma\left(\frac{2a+n}{2}\right)}{\Gamma\left(\frac{2a}{2}\right)} \cdot \left(1 + \frac{1}{2a}(x-\mu)^\mathrm{T}\left(\frac{a}{b}\Lambda\right)(x-\mu)\right)^{-\frac{2a+n}{2}} \\ &= \sqrt{\frac{|\Lambda|}{(2\pi)^n}} \cdot \frac{\Gamma\left(a+\frac{n}{2}\right)}{\Gamma(a)} \cdot b^a \cdot \left(b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right)^{-\left(a+\frac{n}{2}\right)} \; . \end{split} \tag{11}$$

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 (→ Proof II/3.3.5) 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

$$p(y|x) = \mathrm{Gam}\left(y; a+\frac{n}{2}, b + \frac{1}{2}(x-\mu)^\mathrm{T}\Lambda(x-\mu)\right) \; . \tag{14}$$

Sources:

- original work

Metadata: ID: P146 | shortcut: ng-cond | author: JoramSoch | date: 2020-08-05, 06:54.

4.3 Dirichlet distribution

4.3.1 Definition

Definition: Let X be a random vector (→ Definition I/1.1.4). 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 (→ Definition I/1.4.4) 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.

Metadata: ID: D54 | shortcut: dir | author: JoramSoch | date: 2020-05-10, 20:36.

4.3.2 Probability density function

Theorem: Let X be a random vector (→ Definition I/1.1.4) following a Dirichlet distribution (→ Definition II/4.3.1):

$$X \sim \mathrm{Dir}(\alpha) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/4.3.1).

Sources:
- original work

Metadata: ID: P95 | shortcut: dir-pdf | author: JoramSoch | date: 2020-05-05, 21:22.

4.3.3 Exceedance probabilities

Theorem: Let $r = [r_1, \ldots, r_k]$ be a random vector (→ Definition I/1.1.4) following a Dirichlet distribution (→ Definition II/4.3.1) with concentration parameters $\alpha = [\alpha_1, \ldots, \alpha_k]$:

$$r \sim \mathrm{Dir}(\alpha) \; . \tag{1}$$

1) If $k = 2$, then the exceedance probability (→ Definition I/1.2.5) 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 (→ Definition I/1.2.5) 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 (→ Definition II/4.3.1), the exceedance probability (→ Definition I/1.2.5) for a particular r_i is defined as:

$$\begin{split} \varphi_i &= p\left(\forall j \in \left\{ 1, \ldots, k \,\middle|\, j \neq i \right\} : \, r_i > r_j | \alpha \right) \\ &= p\left(\bigwedge_{j \neq i} r_i > r_j \,\middle|\, \alpha \right) \; . \end{split} \tag{4}$$

The probability density function of the Dirichlet distribution (→ Proof II/4.3.2) 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 (→ Definition II/4.3.1).

1) If $k = 2$, the probability density function of the Dirichlet distribution (→ Proof II/4.3.2) 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 (→ Proof II/3.6.2)

$$p(r_1) = \frac{r_1^{\alpha_1 - 1} \, (1 - r_1)^{\alpha_2 - 1}}{\mathrm{B}(\alpha_1, \alpha_2)} \tag{8}$$

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 (→ Proof II/3.6.4), 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 (→ Proof "gam-dir") 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 (→ Proof II/3.3.5) 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 (→ Definition II/3.3.1)

$$q_1 \sim \mathrm{Gam}(\alpha_1, 1), \; \ldots, \; q_k \sim \mathrm{Gam}(\alpha_k, 1), \; q_s = \sum_{j=1}^{k} q_j \tag{16}$$

and the Dirichlet random vector (→ Definition II/4.3.1)

$$r = (r_1, \ldots, r_k) = \left(\frac{q_1}{q_s}, \ldots, \frac{q_k}{q_s} \right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k) \; . \tag{17}$$

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 (→ Proof II/3.3.6) 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 (**??**) and (**??**), 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 (→ Definition I/1.2.3), 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 (**??**) and (**??**), 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 (**??**) and (**??**), 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 (→ Definition I/1.2.5) for one element from a Dirichlet-distributed (→ Definition II/4.3.1) random vector (→ Definition I/1.1.4) is an integral from zero to infinity where the first term in the integrand conforms to a product of gamma (→ Definition II/3.3.1) cumulative distribution functions (→ Definition I/1.4.8) and the second term is a gamma (→ Definition II/3.3.1) probability density function (→ Definition I/1.4.4).

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.

Metadata: ID: P181 | shortcut: dir-ep | author: JoramSoch | date: 2020-10-22, 08:04.

5 Matrix-variate continuous distributions

5.1 Matrix-normal distribution

5.1.1 Definition

Definition: Let X be an $n \times p$ random matrix (→ Definition I/1.1.5). Then, X is said to be matrix-normally distributed with mean M, covariance (→ Definition I/1.7.5) across rows U and covariance (→ Definition I/1.7.5) across columns V

$$X \sim \mathcal{MN}(M, U, V) \; , \tag{1}$$

if and only if its probability density function (→ Definition I/1.4.4) 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.

Metadata: ID: D6 | shortcut: matn | author: JoramSoch | date: 2020-01-27, 14:37.

5.1.2 Probability density function

Theorem: Let X be a random matrix (→ Definition I/1.1.5) following a matrix-normal distribution (→ Definition II/5.1.1):

$$X \sim \mathcal{MN}(M, U, V) \; . \tag{1}$$

Then, the probability density function (→ Definition I/1.4.4) 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 (→ Definition II/5.1.1).

Sources:
- original work

Metadata: ID: P70 | shortcut: matn-pdf | author: JoramSoch | date: 2020-03-02, 21:03.

5.1.3 Equivalence to multivariate normal distribution

Theorem: The matrix X is matrix-normally distributed (→ Definition II/5.1.1)

$$X \sim \mathcal{MN}(M, U, V) \; , \tag{1}$$

if and only if $\mathrm{vec}(X)$ is multivariate normally distributed (→ Definition II/4.1.1)

$$\mathrm{vec}(X) \sim \mathcal{MN}(\mathrm{vec}(M), V \otimes U) \tag{2}$$

where $\mathrm{vec}(X)$ is the vectorization operator and $\otimes$ is the Kronecker product.

Proof: The probability density function of the matrix-normal distribution (→ Proof II/5.1.2) 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:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2}\mathrm{vec}(X-M)^\mathrm{T}\, (V \otimes U)^{-1}\, \mathrm{vec}(X-M)\right] \; . \tag{7}$$

Using the vectorization property $\mathrm{vec}(A+B) = \mathrm{vec}(A) + \mathrm{vec}(B)$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V|^n |U|^p}} \cdot \exp\left[-\frac{1}{2}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]^\mathrm{T}\, (V \otimes U)^{-1}\, \left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]\right] \; . \tag{8}$$

Using the Kronecker-determinant relation $|A \otimes B| = |A|^m |B|^n$, we have:

$$\mathcal{MN}(X; M, U, V) = \frac{1}{\sqrt{(2\pi)^{np} |V \otimes U|}} \cdot \exp\left[-\frac{1}{2}\left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]^\mathrm{T}\, (V \otimes U)^{-1}\, \left[\mathrm{vec}(X) - \mathrm{vec}(M)\right]\right] \; . \tag{9}$$

This is the probability density function of the multivariate normal distribution (→ Proof II/4.1.2) 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 (→ Definition I/1.4.4) are identical, it is proven that the associated probability distributions (→ Definition I/1.3.1) 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.

Metadata: ID: P26 | shortcut: matn-mvn | author: JoramSoch | date: 2020-01-20, 21:09.

5.1.4 Transposition

Theorem: Let X be a random matrix (→ Definition I/1.1.5) following a matrix-normal distribution (→ Definition 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 (→ Proof II/5.1.2) 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 (→ Proof II/5.1.2) with mean M^T, covariance across rows V and covariance across columns U.

Sources:

- original work

Metadata: ID: P144 | shortcut: matn-trans | author: JoramSoch | date: 2020-08-03, 22:21.

5.1.5 Linear transformation

Theorem: Let X be an $n \times p$ random matrix (→ Definition I/1.1.5) following a matrix-normal distribution (→ Definition 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 (→ Proof II/5.1.3),

$$X \sim \mathcal{MN}(M, U, V) \quad \Leftrightarrow \quad \mathrm{vec}(X) \sim \mathcal{N}(\mathrm{vec}(M), V \otimes U) \; , \tag{3}$$

and the linear transformation theorem for the multivariate normal distribution (→ Proof II/4.1.5) 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

$$\begin{split} \mathrm{vec}(Y) &= \mathrm{vec}(AXB + C) \\ &= \mathrm{vec}(AXB) + \mathrm{vec}(C) \\ &= (B^\mathrm{T} \otimes A)\mathrm{vec}(X) + \mathrm{vec}(C) \; . \end{split} \tag{5}$$

Using (3) and (4), we have

$$\begin{split} \mathrm{vec}(Y) &\sim \mathcal{N}((B^\mathrm{T} \otimes A)\mathrm{vec}(M) + \mathrm{vec}(C), (B^\mathrm{T} \otimes A)(V \otimes U)(B^\mathrm{T} \otimes A)^\mathrm{T}) \\ &= \mathcal{N}(\mathrm{vec}(AMB) + \mathrm{vec}(C), (B^\mathrm{T}V \otimes AU)(B^\mathrm{T} \otimes A)^\mathrm{T}) \\ &= \mathcal{N}(\mathrm{vec}(AMB + C), B^\mathrm{T}VB \otimes AUA^\mathrm{T}) \; . \end{split} \tag{6}$$

Using (3), we finally have:

$$Y \sim \mathcal{MN}(AMB + C, AUA^\mathrm{T}, B^\mathrm{T}VB) \; . \tag{7}$$

Sources:
- original work

Metadata: ID: P145 | shortcut: matn-ltt | author: JoramSoch | date: 2020-08-03, 22:24.

5.2 Wishart distribution

5.2.1 Definition

Definition: Let X be an $n \times p$ matrix following a matrix-normal distribution (→ Definition 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.

Metadata: ID: D43 | shortcut: wish | author: JoramSoch | date: 2020-03-22, 17:15.

Chapter III

Statistical Models

1 Univariate normal data

1.1 Multiple linear regression

1.1.1 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 (→ Definition 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.

Metadata: ID: D36 | shortcut: mlr | author: JoramSoch | date: 2020-03-21, 20:09.

1.1.2 Ordinary least squares

Theorem: Given a linear regression model (→ Definition III/1.1.1) 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 (→ Definition III/1.1.6) 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.

Metadata: ID: P2 | shortcut: mlr-ols | author: JoramSoch | date: 2019-09-27, 07:18.

1.1.3 Ordinary least squares

Theorem: Given a linear regression model (→ Definition III/1.1.1) 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 (→ Definition III/1.1.6) are given by

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; . \tag{2}$$

Proof: The residual sum of squares (→ Definition III/1.1.6) is defined as

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \varepsilon_i = \varepsilon^\mathrm{T} \varepsilon = (y - X\beta)^\mathrm{T} (y - X\beta) \tag{3}$$

which can be developed into

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= y^\mathrm{T} y - y^\mathrm{T} X \beta - \beta^\mathrm{T} X^\mathrm{T} y + \beta^\mathrm{T} X^\mathrm{T} X \beta \\
&= y^\mathrm{T} y - 2 \beta^\mathrm{T} X^\mathrm{T} y + \beta^\mathrm{T} X^\mathrm{T} X \beta \; .
\end{aligned} \tag{4}
$$

The derivative of $\mathrm{RSS}(\beta)$ with respect to β is

$$
\frac{\mathrm{dRSS}(\beta)}{\mathrm{d}\beta} = -2 X^\mathrm{T} y + 2 X^\mathrm{T} X \beta \tag{5}
$$

and setting this deriative to zero, we obtain:

$$
\begin{aligned}
\frac{\mathrm{dRSS}(\hat{\beta})}{\mathrm{d}\beta} &= 0 \\
0 &= -2 X^\mathrm{T} y + 2 X^\mathrm{T} X \hat{\beta} \\
X^\mathrm{T} X \hat{\beta} &= X^\mathrm{T} y \\
\hat{\beta} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; .
\end{aligned} \tag{6}
$$

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.

Metadata: ID: P40 | shortcut: mlr-ols2 | author: JoramSoch | date: 2020-02-03, 18:43.

1.1.4 Total sum of squares

Definition: Let there be a multiple linear regression with independent observations (→ Definition III/1.1.1) 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.

Metadata: ID: D37 | shortcut: tss | author: JoramSoch | date: 2020-03-21, 21:44.

1.1.5 Explained sum of squares

Definition: Let there be a multiple linear regression with independent observations (→ Definition III/1.1.1) 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 (→ Proof III/1.1.2).

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.

Metadata: ID: D38 | shortcut: ess | author: JoramSoch | date: 2020-03-21, 21:57.

1.1.6 Residual sum of squares

Definition: Let there be a multiple linear regression with independent observations (→ Definition III/1.1.1) 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 (→ Proof III/1.1.2).

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.

Metadata: ID: D39 | shortcut: rss | author: JoramSoch | date: 2020-03-21, 22:03.

1.1.7 Total, explained and residual sum of squares

Theorem: Assume a linear regression model (→ Definition III/1.1.1) 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 (→ Definition III/1.1.4), ESS is the explained sum of squares (→ Definition III/1.1.5) and RSS is the residual sum of squares (→ Definition III/1.1.6).

Proof: The total sum of squares (→ Definition III/1.1.4) 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 (→ Proof III/1.1.2), 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 (→ Definition III/1.1.5) and residual sum of squares (→ Definition III/1.1.6), 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.

Metadata: ID: P76 | shortcut: mlr-pss | author: JoramSoch | date: 2020-03-09, 22:18.

1.1.8 Estimation matrix

Definition: In multiple linear regression (→ Definition III/1.1.1), the estimation matrix is the matrix E that results in ordinary least squares (→ Proof III/1.1.2) or weighted least squares (→ Proof III/1.1.13) parameter estimates when right-multiplied with the measured data:

$$Ey = \hat{\beta} \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D81 | shortcut: emat | author: JoramSoch | date: 2020-07-22, 05:17.

1.1.9 Projection matrix

Definition: In multiple linear regression (→ Definition III/1.1.1), the projection matrix is the matrix P that results in the fitted signal explained by estimated parameters (→ Definition III/1.1.8) 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.

Metadata: ID: D82 | shortcut: pmat | author: JoramSoch | date: 2020-07-22, 05:25.

1.1.10 Residual-forming matrix

Definition: In multiple linear regression (→ Definition III/1.1.1), the residual-forming matrix is the matrix R that results in the vector of residuals left over by estimated parameters (→ Definition III/1.1.8) 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.

Metadata: ID: D83 | shortcut: rfmat | author: JoramSoch | date: 2020-07-22, 05:35.

1.1.11 Estimation, projection and residual-forming matrix

Theorem: Assume a linear regression model (→ Definition III/1.1.1) 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 (→ Proof III/1.1.2). 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 (→ Definition III/1.1.8), projection matrix (→ Definition III/1.1.9) and residual-forming matrix (→ Definition III/1.1.10) 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 (→ Definition III/1.1.6)

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta} \left[(y - X\beta)^\mathrm{T} (y - X\beta) \right] \tag{4}$$

and the solution to this (→ Proof III/1.1.2) is given by

$$\begin{aligned} \hat{\beta} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \\ &\overset{(3)}{=} Ey \; . \end{aligned} \tag{5}$$

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.

Metadata: ID: P75 | shortcut: mlr-mat | author: JoramSoch | date: 2020-03-09, 21:18.

1.1.12 Idempotence of projection and residual-forming matrix

Theorem: The projection matrix (→ Definition III/1.1.9) and the residual-forming matrix (→ Definition III/1.1.10) 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 (→ Proof III/1.1.11)

$$
P = X(X^\mathrm{T} X)^{-1} X^\mathrm{T} \; , \tag{2}
$$

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 (→ Proof III/1.1.11)

$$R = I_n - X(X^\mathrm{T} X)^{-1} X^\mathrm{T} = I_n - P \; , \tag{4}$$

such that

$$\begin{split} R^2 &= (I_n - P)(I_n - P) \\ &= I_n - P - P + P^2 \\ &\overset{(3)}{=} I_n - 2P + P \\ &= I_n - P \\ &\overset{(4)}{=} R \; . \end{split} \tag{5}$$

Sources:

- Wikipedia (2020): "Projection matrix"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-07-22; URL: https://en.wikipedia.org/wiki/Projection_matrix#Properties.

Metadata: ID: P135 | shortcut: mlr-idem | author: JoramSoch | date: 2020-07-22, 06:28.

1.1.13 Weighted least squares

Theorem: Given a linear regression model (→ Definition III/1.1.1) 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 (→ Definition III/1.1.6) 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

$$W V W^\mathrm{T} = I_n \; . \tag{3}$$

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:

$$W V W = I_n \quad \Leftrightarrow \quad V = W^{-1} W^{-1} \quad \Leftrightarrow \quad V^{-1} = W W \quad \Leftrightarrow \quad W = V^{-1/2} \; . \tag{4}$$

Left-multiplying the linear regression equation (1) with W, the linear transformation theorem (→ Proof II/4.1.5) implies that

$$W y = W X\beta + W\varepsilon, \; W\varepsilon \sim \mathcal{N}(0, \sigma^2 W V W^T) \; . \tag{5}$$

Applying (3), we see that (5) is actually a linear regression model (→ Definition III/1.1.1) 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 (→ Proof III/1.1.2) 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.

Metadata: ID: P77 | shortcut: mlr-wls | author: JoramSoch | date: 2020-03-11, 11:22.

1.1.14 Weighted least squares

Theorem: Given a linear regression model (→ Definition III/1.1.1) 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 (→ Definition III/1.1.6) 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 (→ Proof II/4.1.5) 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 (→ Definition III/1.1.1) 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 (→ Definition III/1.1.6) as

$$\mathrm{wRSS}(\beta) = \sum_{i=1}^{n} (W\varepsilon)_i = (W\varepsilon)^\mathrm{T}(W\varepsilon) = (Wy - WX\beta)^\mathrm{T}(Wy - WX\beta) \tag{7}$$

which can be developed into

$$\begin{split} \mathrm{wRSS}(\beta) &= y^\mathrm{T} W^\mathrm{T} W y - y^\mathrm{T} W^\mathrm{T} W X \beta - \beta^\mathrm{T} X^\mathrm{T} W^\mathrm{T} W y + \beta^\mathrm{T} X^\mathrm{T} W^\mathrm{T} W X \beta \\ &= y^\mathrm{T} W W y - 2 \beta^\mathrm{T} X^\mathrm{T} W W y + \beta^\mathrm{T} X^\mathrm{T} W W X \beta \\ &\overset{(4)}{=} y^\mathrm{T} V^{-1} y - 2 \beta^\mathrm{T} X^\mathrm{T} V^{-1} y + \beta^\mathrm{T} X^\mathrm{T} V^{-1} X \beta \; . \end{split} \tag{8}$$

The derivative of $\mathrm{wRSS}(\beta)$ with respect to β is

$$\frac{\mathrm{dwRSS}(\beta)}{\mathrm{d}\beta} = -2 X^\mathrm{T} V^{-1} y + 2 X^\mathrm{T} V^{-1} X \beta \tag{9}$$

and setting this deriative to zero, we obtain:

$$\begin{split} \frac{\mathrm{dwRSS}(\hat{\beta})}{\mathrm{d}\beta} &= 0 \\ 0 &= -2 X^\mathrm{T} V^{-1} y + 2 X^\mathrm{T} V^{-1} X \hat{\beta} \\ X^\mathrm{T} V^{-1} X \hat{\beta} &= X^\mathrm{T} V^{-1} y \\ \hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \; . \end{split} \tag{10}$$

Since the quadratic form $y^\mathrm{T} V^{-1} y$ in (8) is positive, $\hat{\beta}$ minimizes $\mathrm{wRSS}(\beta)$.

Sources:
- original work

Metadata: ID: P136 | shortcut: mlr-wls2 | author: JoramSoch | date: 2020-07-22, 06:48.

1.1.15 Maximum likelihood estimation

Theorem: Given a linear regression model (→ Definition III/1.1.1) with correlated observations

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \; , \tag{1}$$

the maximum likelihood estimates (→ Definition I/4.1.3) of β and σ^2 are given by

$$\begin{split} \hat{\beta} &= (X^\mathrm{T} V^{-1} X)^{-1} X^\mathrm{T} V^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n} (y - X\hat{\beta})^\mathrm{T} (y - X\hat{\beta}) \; . \end{split} \tag{2}$$

Proof: With the probability density function of the multivariate normal distribution (→ Proof II/4.1.2), the linear regression equation (1) implies the following likelihood function (→ Definition 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 (→ Definition 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

$$
\begin{aligned}
\frac{\mathrm{dLL}(\beta,\sigma^2)}{\mathrm{d}\beta} &= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(-\frac{1}{2\sigma^2}\left(y^\mathrm{T} P y - 2\beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta\right)\right) \\
&= \frac{1}{2\sigma^2}\frac{\mathrm{d}}{\mathrm{d}\beta}\left(2\beta^\mathrm{T} X^\mathrm{T} P y - \beta^\mathrm{T} X^\mathrm{T} P X \beta\right) \\
&= \frac{1}{2\sigma^2}\left(2X^\mathrm{T} P y - 2X^\mathrm{T} P X \beta\right) \\
&= \frac{1}{\sigma^2}\left(X^\mathrm{T} P y - X^\mathrm{T} P X \beta\right)
\end{aligned} \tag{6}
$$

and setting this derivative to zero gives the MLE for β:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\beta},\sigma^2)}{\mathrm{d}\beta} &= 0 \\
0 &= \frac{1}{\sigma^2}\left(X^\mathrm{T} P y - X^\mathrm{T} P X \hat{\beta}\right) \\
0 &= X^\mathrm{T} P y - X^\mathrm{T} P X \hat{\beta} \\
X^\mathrm{T} P X \hat{\beta} &= X^\mathrm{T} P y \\
\hat{\beta} &= \left(X^\mathrm{T} P X\right)^{-1} X^\mathrm{T} P y
\end{aligned} \tag{7}
$$

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.

Sources:
- original work

Metadata: ID: P78 | shortcut: mlr-mle | author: JoramSoch | date: 2020-03-11, 12:27.

1.2 Bayesian linear regression

1.2.1 Conjugate prior distribution

Theorem: Let

$$
y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}
$$

be a linear regression model (→ Definition III/1.1.1) 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 (→ Definition I/5.2.5) for this model is a normal-gamma distribution (→ Definition II/4.2.1)

$$
p(\beta, \tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1}) \cdot \mathrm{Gam}(\tau; a_0, b_0) \tag{2}
$$

where $\tau = 1/\sigma^2$ is the inverse noise variance or noise precision.

Proof: By definition, a conjugate prior (→ Definition I/5.2.5) is a prior distribution (→ Definition I/5.1.3) that, when combined with the likelihood function (→ Definition I/5.1.2), leads to a posterior

distribution (→ Definition I/5.1.7) that belongs to the same family of probability distributions (→ Definition I/1.3.1). 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 (→ Definition 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 (→ Definition 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 (→ Definition II/4.2.1) 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 (→ Proof II/4.2.2)

$$p(\beta,\tau) = \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2} (\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0) \right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \tau^{a_0 - 1} \exp[-b_0 \tau] \tag{10}$$

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.

Metadata: ID: P9 | shortcut: blr-prior | author: JoramSoch | date: 2020-01-03, 15:26.

1.2.2 Posterior distribution

Theorem: Let

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

be a linear regression model (→ Definition III/1.1.1) 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 (→ Proof III/1.2.1) 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 (→ Definition I/5.1.7) is also a normal-gamma distribution (→ Definition II/4.2.1)

$$p(\beta, \tau|y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1}) \cdot \mathrm{Gam}(\tau; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ Definition 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 (→ Proof I/5.3.1), the posterior distribution (→ Definition 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 (→ Proof I/5.1.8) 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 (→ Definition 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 (→ Definition I/1.7.8) $P = V^{-1}$.

Combining the likelihood function (→ Definition I/5.1.2) (8) with the prior distribution (→ Definition I/5.1.3) (2), the joint likelihood (→ Definition I/5.1.5) of the model is given by

$$\begin{aligned} p(y,\beta,\tau) &= p(y|\beta,\tau)\, p(\beta,\tau) \\ &= \sqrt{\frac{|\tau P|}{(2\pi)^n}} \exp\left[-\frac{\tau}{2}(y - X\beta)^\mathrm{T} P (y - X\beta)\right] \cdot \\ &\quad \sqrt{\frac{|\tau \Lambda_0|}{(2\pi)^p}} \exp\left[-\frac{\tau}{2}(\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right] \cdot \\ &\quad \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0-1} \exp[-b_0 \tau]\; . \end{aligned} \tag{9}$$

Collecting identical variables gives:

$$\begin{aligned} p(y,\beta,\tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P||\Lambda_0|}\, \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left((y - X\beta)^\mathrm{T} P (y - X\beta) + (\beta - \mu_0)^\mathrm{T} \Lambda_0 (\beta - \mu_0)\right)\right]\; . \end{aligned} \tag{10}$$

Expanding the products in the exponent gives:

$$\begin{aligned} p(y,\beta,\tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P||\Lambda_0|}\, \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left(y^\mathrm{T} P y - y^\mathrm{T} P X \beta - \beta^\mathrm{T} X^\mathrm{T} P y + \beta^\mathrm{T} X^\mathrm{T} P X \beta + \right.\right. \\ &\qquad \left.\left. \beta^\mathrm{T} \Lambda_0 \beta - \beta^\mathrm{T} \Lambda_0 \mu_0 - \mu_0^\mathrm{T} \Lambda_0 \beta + \mu_0^\mathrm{T} \Lambda_0 \mu_0\right)\right]\; . \end{aligned} \tag{11}$$

Completing the square over β, we finally have

$$\begin{aligned} p(y,\beta,\tau) = &\sqrt{\frac{\tau^{n+p}}{(2\pi)^{n+p}} |P||\Lambda_0|}\, \frac{{b_0}^{a_0}}{\Gamma(a_0)}\, \tau^{a_0-1} \exp[-b_0\tau] \cdot \\ &\exp\left[-\frac{\tau}{2}\left((\beta - \mu_n)^\mathrm{T} \Lambda_n (\beta - \mu_n) + (y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n)\right)\right] \end{aligned} \tag{12}$$

with the posterior hyperparameters (→ Definition I/5.1.7)

$$\begin{aligned} \mu_n &= \Lambda_n^{-1}(X^\mathrm{T} P y + \Lambda_0 \mu_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0\; . \end{aligned} \tag{13}$$

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 (→ Definition I/5.1.7)

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(y^\mathrm{T} P y + \mu_0^\mathrm{T} \Lambda_0 \mu_0 - \mu_n^\mathrm{T} \Lambda_n \mu_n) \; . \end{aligned} \tag{15}$$

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 (→ Definition I/1.2.2) posterior distribution (→ Definition 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.

Metadata: ID: P10 | shortcut: blr-post | author: JoramSoch | date: 2020-01-03, 17:53.

1.2.3 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 (→ Definition III/1.1.1) 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 (→ Proof III/1.2.1) 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 (→ Definition IV/3.1.1) 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 (→ Definition 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 the law of marginal probability (→ Definition I/1.2.3), the model evidence (→ Definition I/5.1.9) 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 (→ Definition I/1.2.4), the integrand is equivalent to the joint likelihood (→ Definition 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 (→ Definition 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 (→ Proof III/1.2.2) $p(\beta,\tau|y)$, the joint likelihood $p(y,\beta,\tau)$ is obtained as

$$
\begin{split}
p(y,\beta,\tau) = \; &\sqrt{\frac{\tau^n |P|}{(2\pi)^n}} \sqrt{\frac{\tau^p |\Lambda_0|}{(2\pi)^p}} \, \frac{{b_0}^{a_0}}{\Gamma(a_0)} \, \tau^{a_0-1} \exp[-b_0 \tau] \cdot \\
&\exp\left[-\frac{\tau}{2}\left((\beta - \mu_n)^T \Lambda_n (\beta - \mu_n) + (y^T P y + \mu_0^T \Lambda_0 \mu_0 - \mu_n^T \Lambda_n \mu_n)\right)\right] \; .
\end{split} \tag{9}
$$

Using the probability density function of the multivariate normal distribution (→ Proof II/4.1.2), 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:

$$\begin{split}\int p(y,\beta,\tau)\,\mathrm{d}\beta = \sqrt{\frac{\tau^n|P|}{(2\pi)^n}}\sqrt{\frac{|\Lambda_0|}{|\Lambda_n|}}\,\frac{{b_0}^{a_0}}{\Gamma(a_0)}\,\tau^{a_0-1}\exp[-b_0\tau]\cdot \\ \exp\left[-\frac{\tau}{2}(y^TPy+\mu_0^T\Lambda_0\mu_0-\mu_n^T\Lambda_n\mu_n)\right] \; .\end{split} \tag{11}$$

Using the probability density function of the gamma distribution (→ Proof II/3.3.5), 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 (→ Definition IV/3.1.1) 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.

Metadata: ID: P11 | shortcut: blr-lme | author: JoramSoch | date: 2020-01-03, 22:05.

1.2.4 Posterior probability of alternative hypothesis

Theorem: Let there be a linear regression model (→ Definition III/1.1.1) with normally distributed (→ Definition II/4.1.1) errors:

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and assume a normal-gamma (→ Definition II/4.2.1) prior distribution (→ Definition 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 (→ Definition I/5.1.7) probability (→ Definition I/1.2.1) of the alternative hypothesis (→ Definition "h1")

$$\mathrm{H}_1 : \, c^\mathrm{T}\beta > 0 \tag{3}$$

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 (→ Definition "con"), $\mathrm{T}(x;\nu)$ is the cumulative distribution function (→ Definition I/1.4.8) of the t-distribution (→ Definition "t") with ν degrees of freedom (→ Definition "dof") and μ, Σ and ν can be obtained from the posterior hyperparameters (→ Definition I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (→ Proof III/1.2.2) is given by a normal-gamma distribution (→ Definition II/4.2.1) 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 (→ Definition I/5.1.7)

$$\begin{aligned}
\mu_n &= \Lambda_n^{-1}(X^\mathrm{T}Py + \Lambda_0\mu_0) \\
\Lambda_n &= X^\mathrm{T}PX + \Lambda_0 \\
a_n &= a_0 + \frac{n}{2} \\
b_n &= b_0 + \frac{1}{2}(y^\mathrm{T}Py + \mu_0^\mathrm{T}\Lambda_0\mu_0 - \mu_n^\mathrm{T}\Lambda_n\mu_n) \; .
\end{aligned} \tag{6}$$

The marginal distribution of a normal-gamma distribution is a multivariate t-distribution (→ Proof II/4.2.4), such that the marginal (→ Definition I/1.3.3) posterior (→ Definition I/5.1.7) distribution of β is

$$p(\beta|y) = \mathrm{t}(\beta; \mu, \Sigma, \nu) \tag{7}$$

with the posterior hyperparameters (→ Definition 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 (→ Proof "mvt-ltt"), γ also follows a multivariate t-distribution (→ Definition "mvt"):

$$p(\gamma|y) = \mathrm{t}(\gamma; c^\mathrm{T}\mu, c^\mathrm{T}\Sigma\, c, \nu) \; . \tag{9}$$

Because c^T is a $1 \times p$ vector, γ is a scalar and actually has a non-central scaled t-distribution (→ Definition "ncst"). Therefore, the posterior probability of H_1 can be calculated using a one-dimensional integral:

$$\begin{aligned}
\Pr\left(\mathrm{H}_1|y\right) &= p(\gamma > 0|y) \\
&= \int_0^{+\infty} p(\gamma|y)\,\mathrm{d}\gamma \\
&= 1 - \int_{-\infty}^0 p(\gamma|y)\,\mathrm{d}\gamma \\
&= 1 - \mathrm{T}_\mathrm{ncst}(0; c^\mathrm{T}\mu, c^\mathrm{T}\Sigma\, c, \nu) \; .
\end{aligned} \tag{10}$$

Using the relation between non-central scaled t-distribution and standard t-distribution (→ Proof "ncst-t"), 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.

Metadata: ID: P133 | shortcut: blr-pp | author: JoramSoch | date: 2020-07-17, 17:03.

1.2.5 Posterior credibility region excluding null hypothesis

Theorem: Let there be a linear regression model (→ Definition III/1.1.1) with normally distributed (→ Definition II/4.1.1) errors:

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

and assume a normal-gamma (→ Definition II/4.2.1) prior distribution (→ Definition 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 (→ Definition I/5.1.7) credibility region (→ Definition "cr") that does not contain the omnibus null hypothesis (→ Definition "h0")

$$\mathrm{H}_0: \; C^\mathrm{T}\beta = 0 \tag{3}$$

is given by the credibility level (→ Definition "cr")

$$(1 - \alpha) = \mathrm{F}\left(\left[\mu^\mathrm{T} C (C^\mathrm{T} \Sigma \, C)^{-1} C^\mathrm{T} \mu\right] / q; q, \nu\right) \tag{4}$$

where C is a $p \times q$ contrast matrix (→ Definition "con"), $\mathrm{F}(x; v, w)$ is the cumulative distribution function (→ Definition I/1.4.8) of the F-distribution (→ Definition "f") with v numerator degrees of freedom (→ Definition "dof") w denominator degrees of freedom (→ Definition "dof") and μ, Σ and ν can be obtained from the posterior hyperparameters (→ Definition I/5.1.7) of Bayesian linear regression.

Proof: The posterior distribution for Bayesian linear regression (→ Proof III/1.2.2) is given by a normal-gamma distribution (→ Definition II/4.2.1) 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 (→ Definition 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 (→ Proof II/4.2.4), such that the marginal (→ Definition I/1.3.3) posterior (→ Definition I/5.1.7) distribution of β is

$$
p(\beta|y) = \mathrm{t}(\beta; \mu, \Sigma, \nu) \tag{7}
$$

with the posterior hyperparameters (→ Definition 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 (→ Proof “mvt-ltt”), γ also follows a multivariate t-distribution (→ Definition “mvt”):

$$
p(\gamma|y) = \mathrm{t}(\gamma; C^\mathrm{T} \mu, C^\mathrm{T} \Sigma \, C, \nu) \; . \tag{9}
$$

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 (→ Proof “mvt-f”), 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.

Metadata: ID: P134 | shortcut: blr-pcr | author: JoramSoch | date: 2020-07-17, 17:41.

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 (→ Definition 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.

Metadata: ID: D40 | shortcut: glm | author: JoramSoch | date: 2020-03-21, 22:24.

2.1.2 Ordinary least squares

Theorem: Given a general linear model (→ Definition 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 (→ Definition "ols") 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 (→ Definition "ols") (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 (→ Definition I/1.5.1) 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:

$$\begin{split}
X^\mathrm{T} \hat{E} &= 0 \\
X^\mathrm{T} \left(Y - X\hat{B} \right) &= 0 \\
X^\mathrm{T} Y - X^\mathrm{T} X \hat{B} &= 0 \\
X^\mathrm{T} X \hat{B} &= X^\mathrm{T} Y \\
\hat{B} &= (X^\mathrm{T} X)^{-1} X^\mathrm{T} Y \; .
\end{split} \tag{6}$$

Sources:
- original work

Metadata: ID: P106 | shortcut: glm-ols | author: JoramSoch | date: 2020-05-19, 06:02.

2.1.3 Weighted least squares

Theorem: Given a general linear model (→ Definition III/2.1.1) with correlated observations

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; , \tag{1}$$

the weighted least sqaures (→ Definition "wls") 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:

$$W W = V^{-1} \quad \Leftrightarrow \quad W = V^{-1/2} \; . \tag{4}$$

Left-multiplying the linear regression equation (1) with W, the linear transformation theorem (→ Proof II/5.1.5) 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 (→ Definition 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 (→ Proof III/2.1.2) giving

$$\begin{split}
\hat{B} &= (\tilde{X}^\mathrm{T}\tilde{X})^{-1}\tilde{X}^\mathrm{T}\tilde{Y} \\
&= \left((WX)^\mathrm{T}WX\right)^{-1}(WX)^\mathrm{T}WY \\
&= \left(X^\mathrm{T}W^\mathrm{T}WX\right)^{-1}X^\mathrm{T}W^\mathrm{T}WY \\
&= \left(X^\mathrm{T}WWX\right)^{-1}X^\mathrm{T}WWY \\
&\overset{(4)}{=} \left(X^\mathrm{T}V^{-1}X\right)^{-1}X^\mathrm{T}V^{-1}Y
\end{split} \tag{7}$$

which corresponds to the weighted least squares solution (2).

Sources:
- original work

Metadata: ID: P107 | shortcut: glm-wls | author: JoramSoch | date: 2020-05-19, 06:27.

2.1.4 Maximum likelihood estimation

Theorem: Given a general linear model (→ Definition III/2.1.1) with matrix-normally distributed (→ Definition II/5.1.1) errors

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \; , \tag{1}$$

maximum likelihood estimates (→ Definition 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 (→ Definition 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 (→ Definition I/4.1.2)

$$
\begin{aligned}
\mathrm{LL}(B,\Sigma) &= \log p(Y|B,\Sigma) \\
&= -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{v}{2}\log|V| \\
&\quad - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\right] \ .
\end{aligned} \tag{4}
$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$
\begin{aligned}
\mathrm{LL}(B,\Sigma) = &-\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{v}{2}\log(|V|) \\
&-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB - B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB\right)\right] \ .
\end{aligned} \tag{5}
$$

The derivative of the log-likelihood function (5) with respect to B is

$$
\begin{aligned}
\frac{\mathrm{dLL}(B,\Sigma)}{\mathrm{d}B} &= \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB - B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB\right)\right]\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\operatorname{tr}\left[-2\Sigma^{-1}Y^{\mathrm{T}}PXB\right]\right) + \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}B^{\mathrm{T}}X^{\mathrm{T}}PXB\right]\right) \\
&= -\frac{1}{2}\left(-2X^{\mathrm{T}}PY\Sigma^{-1}\right) - \frac{1}{2}\left(X^{\mathrm{T}}PXB\Sigma^{-1} + (X^{\mathrm{T}}PX)^{\mathrm{T}}B(\Sigma^{-1})^{\mathrm{T}}\right) \\
&= X^{\mathrm{T}}PY\Sigma^{-1} - X^{\mathrm{T}}PXB\Sigma^{-1}
\end{aligned} \tag{6}
$$

and setting this derivative to zero gives the MLE for B:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\Sigma)}{\mathrm{d}B} &= 0 \\
0 &= X^{\mathrm{T}}PY\Sigma^{-1} - X^{\mathrm{T}}PX\hat{B}\Sigma^{-1} \\
0 &= X^{\mathrm{T}}PY - X^{\mathrm{T}}PX\hat{B} \\
X^{\mathrm{T}}PX\hat{B} &= X^{\mathrm{T}}PY \\
\hat{B} &= \left(X^{\mathrm{T}}PX\right)^{-1}X^{\mathrm{T}}PY
\end{aligned} \tag{7}
$$

The derivative of the log-likelihood function (4) at $\hat{B}$ with respect to Σ is

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\Sigma)}{\mathrm{d}\Sigma} &= \frac{\mathrm{d}}{\mathrm{d}\Sigma}\left(-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}(Y-X\hat{B})^{\mathrm{T}}V^{-1}(Y-X\hat{B})\right]\right) \\
&= -\frac{n}{2}\left(\Sigma^{-1}\right)^{\mathrm{T}} + \frac{1}{2}\left(\Sigma^{-1}(Y-X\hat{B})^{\mathrm{T}}V^{-1}(Y-X\hat{B})\,\Sigma^{-1}\right)^{\mathrm{T}} \\
&= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(Y-X\hat{B})^{\mathrm{T}}V^{-1}(Y-X\hat{B})\,\Sigma^{-1}
\end{aligned} \tag{8}
$$

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.

Sources:
- original work

Metadata: ID: P7 | shortcut: glm-mle | author: JoramSoch | date: 2019-12-06, 10:40.

2.2 Multivariate Bayesian linear regression

2.2.1 Conjugate prior distribution

Theorem: Let

$$
Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}
$$

be a general linear model (→ Definition 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 (→ Definition II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (→ Definition II/5.1.1) Σ.
Then, the conjugate prior (→ Definition I/5.2.5) for this model is a normal-Wishart distribution (→ Definition "nw")

$$
p(B, T) = \mathcal{MN}(B; M_0, \Lambda_0^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_0^{-1}, \nu_0) \tag{2}
$$

where $T = \Sigma^{-1}$ is the inverse noise covariance (→ Definition I/1.7.5) or noise precision matrix (→ Definition I/1.7.8).

Proof: By definition, a conjugate prior (→ Definition I/5.2.5) is a prior distribution (→ Definition I/5.1.3) that, when combined with the likelihood function (→ Definition I/5.1.2), leads to a posterior distribution (→ Definition I/5.1.7) that belongs to the same family of probability distributions (→ Definition I/1.3.1). 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 (→ Definition 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

$$p(Y|B,T) = \mathcal{MN}(Y; XB, P, T^{-1}) = \sqrt{\frac{|T|^n|P|^v}{(2\pi)^{nv}}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(T(Y-XB)^\mathrm{T}P(Y-XB)\right)\right] \tag{4}$$

using the $v \times v$ precision matrix (→ Definition I/1.7.8) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (→ Definition I/1.7.8) $P = V^{-1}$.

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 β, 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 (→ Definition 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 (→ Definition "nw") 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 (→ Proof "nw-pdf")

$$p(B,T) = \sqrt{\frac{|T|^p|\Lambda_0|^v}{(2\pi)^{pv}}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(T(B-M_0)^\mathrm{T}\Lambda_0(B-M_0)\right)\right] \cdot \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \tag{10}$$

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.

Metadata: ID: P159 | shortcut: mblr-prior | author: JoramSoch | date: 2020-09-03, 07:33.

2.2.2 Posterior distribution

Theorem: Let

$$Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}$$

be a general linear model (→ Definition 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 (→ Definition II/5.1.1) V as well as unknown $p \times v$ regression coefficients B and unknown $v \times v$ noise covariance (→ Definition II/5.1.1) Σ. Moreover, assume a normal-Wishart prior distribution (→ Proof III/2.2.1) 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 (→ Definition I/5.1.7) is also a normal-Wishart distribution (→ Definition "nw")

$$p(B,T|Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \tag{3}$$

and the posterior hyperparameters (→ Definition I/5.1.7) are given by

$$\begin{aligned} M_n &= \Lambda_n^{-1}(X^\mathrm{T}PY + \Lambda_0 M_0) \\ \Lambda_n &= X^\mathrm{T}PX + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^\mathrm{T}PY + M_0^\mathrm{T}\Lambda_0 M_0 - M_n^\mathrm{T}\Lambda_n M_n \\ \nu_n &= \nu_0 + n \; . \end{aligned} \tag{4}$$

Proof: According to Bayes' theorem (→ Proof I/5.3.1), the posterior distribution (→ Definition 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 (→ Proof I/5.1.8) 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 (→ Definition 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 (→ Definition I/1.7.8) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (→ Definition I/1.7.8) $P = V^{-1}$.

Combining the likelihood function (→ Definition I/5.1.2) (8) with the prior distribution (→ Definition I/5.1.3) (2), the joint likelihood (→ Definition 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:

$$\begin{split}
p(Y,B,T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\Omega_0 T \right) \right] \cdot \\
&\exp\left[-\frac{1}{2} \mathrm{tr}\left(T \left[Y^\mathrm{T} P Y - Y^\mathrm{T} P X B - B^\mathrm{T} X^\mathrm{T} P Y + B^\mathrm{T} X^\mathrm{T} P X B + \right. \right. \right. \\
&\left. \left. \left. B^\mathrm{T} \Lambda_0 B - B^\mathrm{T} \Lambda_0 M_0 - M_0^\mathrm{T} \Lambda_0 B + M_0^\mathrm{T} \Lambda_0 \mu_0 \right] \right) \right] \; .
\end{split} \tag{11}$$

Completing the square over B, we finally have

$$
\begin{split}
p(Y,B,T) = \sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \cdot \\
\exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_n)^\mathrm{T}\Lambda_n(B-M_n) + (Y^\mathrm{T}PY + M_0^\mathrm{T}\Lambda_0 M_0 - M_n^\mathrm{T}\Lambda_n M_n)\right]\right)\right] \; .
\end{split} \tag{12}
$$

with the posterior hyperparameters (→ Definition I/5.1.7)

$$
\begin{split}
M_n &= \Lambda_n^{-1}(X^\mathrm{T}PY + \Lambda_0 M_0) \\
\Lambda_n &= X^\mathrm{T}PX + \Lambda_0 \; .
\end{split} \tag{13}
$$

Ergo, the joint likelihood is proportional to

$$
p(Y,B,T) \propto |T|^{p/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_n)^\mathrm{T}\Lambda_n(B-M_n)\right]\right)\right] \cdot |T|^{(\nu_n-v-1)/2} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_n T\right)\right] \tag{14}
$$

with the posterior hyperparameters (→ Definition I/5.1.7)

$$
\begin{split}
\Omega_n &= \Omega_0 + Y^\mathrm{T}PY + M_0^\mathrm{T}\Lambda_0 M_0 - M_n^\mathrm{T}\Lambda_n M_n \\
\nu_n &= \nu_0 + n \; .
\end{split} \tag{15}
$$

From the term in (14), we can isolate the posterior distribution over B given T:

$$
p(B|T,Y) = \mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \; . \tag{16}
$$

From the remaining term, we can isolate the posterior distribution over T:

$$
p(T|Y) = \mathcal{W}(T; \Omega_n^{-1}, \nu_n) \; . \tag{17}
$$

Together, (16) and (17) constitute the joint (→ Definition I/1.2.2) posterior distribution (→ Definition 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.

Metadata: ID: P160 | shortcut: mblr-post | author: JoramSoch | date: 2020-09-03, 08:37.

2.2.3 Log model evidence

Theorem: Let

$$
Y = XB + E, \; E \sim \mathcal{MN}(0, V, \Sigma) \tag{1}
$$

be a general linear model (→ Definition 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 (→ Definition II/5.1.1) V as well as unknown $p \times v$

regression coefficients B and unknown $v \times v$ noise covariance (→ Definition II/5.1.1) Σ. Moreover, assume a normal-Wishart prior distribution (→ Proof III/2.2.1) 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 log model evidence (→ Definition IV/3.1.1) for this model is

$$\begin{split} \log p(y|m) = &\frac{v}{2} \log |P| - \frac{nv}{2} \log(2\pi) + \frac{v}{2} \log |\Lambda_0| - \frac{v}{2} \log |\Lambda_n| + \\ &\frac{\nu_0}{2} \log \left| \frac{1}{2} \Omega_0 \right| - \frac{\nu_n}{2} \log \left| \frac{1}{2} \Omega_n \right| + \log \Gamma_v \left(\frac{\nu_n}{2} \right) - \log \Gamma_v \left(\frac{\nu_0}{2} \right) \end{split} \tag{3}$$

where the posterior hyperparameters (→ Definition I/5.1.7) are given by

$$\begin{split} M_n &= \Lambda_n^{-1} (X^\mathrm{T} P Y + \Lambda_0 M_0) \\ \Lambda_n &= X^\mathrm{T} P X + \Lambda_0 \\ \Omega_n &= \Omega_0 + Y^\mathrm{T} P Y + M_0^\mathrm{T} \Lambda_0 M_0 - M_n^\mathrm{T} \Lambda_n M_n \\ \nu_n &= \nu_0 + n \; . \end{split} \tag{4}$$

Proof: According to the law of marginal probability (→ Definition I/1.2.3), the model evidence (→ Definition I/5.1.9) 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 (→ Definition I/1.2.4), the integrand is equivalent to the joint likelihood (→ Definition 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 (→ Definition 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 (→ Definition I/1.7.8) $T = \Sigma^{-1}$ and the $n \times n$ precision matrix (→ Definition I/1.7.8) $P = V^{-1}$.

When deriving the posterior distribution (→ Proof III/2.2.2) $p(B,T|Y)$, the joint likelihood $p(Y,B,T)$ is obtained as

$$
\begin{aligned}
p(Y,B,T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \cdot \\
&\exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[(B-M_n)^\mathrm{T}\Lambda_n(B-M_n) + (Y^\mathrm{T}PY + M_0^\mathrm{T}\Lambda_0 M_0 - M_n^\mathrm{T}\Lambda_n M_n)\right]\right)\right] \; .
\end{aligned} \tag{9}
$$

Using the probability density function of the matrix-normal distribution (→ Proof II/5.1.2), we can rewrite this as

$$
\begin{aligned}
p(Y,B,T) = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|T|^p |\Lambda_0|^v}{(2\pi)^{pv}}} \sqrt{\frac{(2\pi)^{pv}}{|T|^p |\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Omega_0 T\right)\right] \cdot \\
&\mathcal{MN}(B; M_n, \Lambda_n^{-1}, T^{-1}) \cdot \exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[Y^\mathrm{T}PY + M_0^\mathrm{T}\Lambda_0 M_0 - M_n^\mathrm{T}\Lambda_n M_n\right]\right)\right] \; .
\end{aligned} \tag{10}
$$

Now, B can be integrated out easily:

$$
\begin{aligned}
\int p(Y,B,T)\,\mathrm{d}B = &\sqrt{\frac{|T|^n |P|^v}{(2\pi)^{nv}}} \sqrt{\frac{|\Lambda_0|^v}{|\Lambda_n|^v}} \sqrt{\frac{|\Omega_0|^{\nu_0}}{2^{\nu_0 v}}} \frac{1}{\Gamma_v\left(\frac{\nu_0}{2}\right)} \cdot |T|^{(\nu_0-v-1)/2} \cdot \\
&\exp\left[-\frac{1}{2}\mathrm{tr}\left(T\left[\Omega_0 + Y^\mathrm{T}PY + M_0^\mathrm{T}\Lambda_0 M_0 - M_n^\mathrm{T}\Lambda_n M_n\right]\right)\right] \; .
\end{aligned} \tag{11}
$$

Using the probability density function of the Wishart distribution (→ Proof "wish-pdf"), 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 (→ Definition IV/3.1.1) of this model is given by

$$
\begin{aligned}
\log p(y|m) = &\frac{v}{2}\log|P| - \frac{nv}{2}\log(2\pi) + \frac{v}{2}\log|\Lambda_0| - \frac{v}{2}\log|\Lambda_n| + \\
&\frac{\nu_0}{2}\log\left|\frac{1}{2}\Omega_0\right| - \frac{\nu_n}{2}\log\left|\frac{1}{2}\Omega_n\right| + \log\Gamma_v\left(\frac{\nu_n}{2}\right) - \log\Gamma_v\left(\frac{\nu_0}{2}\right) \; .
\end{aligned} \tag{14}
$$

Sources:
- original work

Metadata: ID: P161 | shortcut: mblr-lme | author: JoramSoch | date: 2020-09-03, 09:23.

3 Poisson data

3.1 Poisson-distributed data

3.1.1 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 (→ Definition II/1.4.1) 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.

Metadata: ID: D41 | shortcut: poiss-data | author: JoramSoch | date: 2020-03-22, 22:50.

3.1.2 Maximum likelihood estimation

Theorem: Let there be a Poisson-distributed data (→ Definition III/3.1.1) 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 (→ Definition 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 (→ Proof "mean-sample")

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \; . \tag{3}$$

Proof: The likelihood function (→ Definition I/5.1.2) for each observation is given by the probability mass function of the Poisson distribution (→ Proof II/1.4.2)

$$p(y_i|\lambda) = \mathrm{Poiss}(y_i; \lambda) = \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \tag{4}$$

and because observations are independent (→ Definition I/1.2.6), 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 (→ Definition 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

$$
\begin{aligned}
\mathrm{LL}(\lambda) &= \sum_{i=1}^{n} \log \left[\frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \right] \\
&= \sum_{i=1}^{n} \left[y_i \cdot \log(\lambda) - \lambda - \log(y_i!) \right] \\
&= -\sum_{i=1}^{n} \lambda + \sum_{i=1}^{n} y_i \cdot \log(\lambda) - \sum_{i=1}^{n} \log(y_i!) \\
&= -n\lambda + \log(\lambda) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!)
\end{aligned} \tag{7}
$$

The derivatives of the log-likelihood with respect to λ are

$$
\begin{aligned}
\frac{\mathrm{dLL}(\lambda)}{\mathrm{d}\lambda} &= \frac{1}{\lambda} \sum_{i=1}^{n} y_i - n \\
\frac{\mathrm{d}^2\mathrm{LL}(\lambda)}{\mathrm{d}\lambda^2} &= -\frac{1}{\lambda^2} \sum_{i=1}^{n} y_i \; .
\end{aligned} \tag{8}
$$

Setting the first derivative to zero, we obtain:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{\lambda})}{\mathrm{d}\lambda} &= 0 \\
0 &= \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} y_i - n \\
\hat{\lambda} &= \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y} \; .
\end{aligned} \tag{9}
$$

Plugging this value into the second deriative, we confirm:

$$
\begin{aligned}
\frac{\mathrm{d}^2\mathrm{LL}(\hat{\lambda})}{\mathrm{d}\lambda^2} &= -\frac{1}{\bar{y}^2} \sum_{i=1}^{n} y_i \\
&= -\frac{n \cdot \bar{y}}{\bar{y}^2} \\
&= -\frac{n}{\bar{y}} < 0 \; .
\end{aligned} \tag{10}
$$

This demonstrates that the estimate $\hat{\lambda} = \bar{y}$ maximizes the likelihood $p(y|\lambda)$.

Sources:
- original work

Metadata: ID: P27 | shortcut: poiss-mle | author: JoramSoch | date: 2020-01-20, 21:53.

3.2 Poisson distribution with exposure values

3.2.1 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 (→ Definition II/1.4.1) 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/.

Metadata: ID: D42 | shortcut: poissexp | author: JoramSoch | date: 2020-03-22, 22:57.

3.2.2 Conjugate prior distribution

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ Definition III/3.2.1):

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}$$

Then, the conjugate prior (→ Definition I/5.2.5) for the model parameter λ is a gamma distribution (→ Definition II/3.3.1):

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Proof: With the probability mass function of the Poisson distribution (→ Proof II/1.4.2), the likelihood function (→ Definition 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 (→ Definition I/1.2.6), 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 (→ Proof "mean-sample") 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 (→ Proof II/3.3.5)

$$p(\lambda) = \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0-1} \exp[-b_0\lambda] \tag{9}$$

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/.

Metadata: ID: P41 | shortcut: poissexp-prior | author: JoramSoch | date: 2020-02-04, 14:11.

3.2.3 Posterior distribution

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ Definition III/3.2.1):

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}$$

Moreover, assume a gamma prior distribution (→ Proof III/3.2.2) over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Then, the posterior distribution (→ Definition I/5.1.7) is also a gamma distribution (→ Definition II/3.3.1)

$$p(\lambda|y) = \mathrm{Gam}(\lambda; a_n, b_n) \tag{3}$$

and the posterior hyperparameters (→ Definition 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 (→ Proof II/1.4.2), the likelihood function (→ Definition 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 (→ Definition I/1.2.6), 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 (→ Definition 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

$$
\begin{aligned}
p(y, \lambda) &= \prod_{i=1}^{n} \frac{{x_i}^{y_i}}{y_i!} \prod_{i=1}^{n} \lambda^{y_i} \prod_{i=1}^{n} \exp\left[-\lambda x_i\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{\sum_{i=1}^{n} y_i} \exp\left[-\lambda \sum_{i=1}^{n} x_i \right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \lambda^{n\bar{y}} \exp\left[-n\bar{x}\lambda\right] \cdot \frac{{b_0}^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp[-b_0 \lambda] \\
&= \prod_{i=1}^{n} \left(\frac{{x_i}^{y_i}}{y_i!} \right) \frac{{b_0}^{a_0}}{\Gamma(a_0)} \cdot \lambda^{a_0 + n\bar{y} - 1} \cdot \exp\left[-(b_0 + n\bar{x})\lambda\right]
\end{aligned} \tag{8}
$$

where $\bar{y}$ and $\bar{x}$ are the means (→ Proof “mean-sample”) 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 (→ Proof I/5.1.8):

$$p(\lambda|y) \propto p(y,\lambda) \; . \tag{10}$$

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 (→ Proof II/3.3.5):

$$p(\lambda|y) = \frac{{b_n}^{a_n}}{\Gamma(a_0)} \lambda^{a_n-1} \exp\left[-b_n\lambda\right] = \mathrm{Gam}(\lambda; a_n, b_n) \; . \tag{12}$$

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/.

Metadata: ID: P42 | shortcut: poissexp-post | author: JoramSoch | date: 2020-02-04, 14:42.

3.2.4 Log model evidence

Theorem: Consider data $y = \{y_1, \ldots, y_n\}$ following a Poisson distribution with exposure values (→ Definition III/3.2.1):

$$y_i \sim \mathrm{Poiss}(\lambda x_i), \quad i = 1, \ldots, n \; . \tag{1}$$

Moreover, assume a gamma prior distribution (→ Proof III/3.2.2) over the model parameter λ:

$$p(\lambda) = \mathrm{Gam}(\lambda; a_0, b_0) \; . \tag{2}$$

Then, the log model evidence (→ Definition IV/3.1.1) 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 (→ Definition I/5.1.7) are given by

$$\begin{split} a_n &= a_0 + n\bar{y} \\ a_n &= a_0 + n\bar{x} \; . \end{split} \tag{4}$$

Proof: With the probability mass function of the Poisson distribution (→ Proof II/1.4.2), the likelihood function (→ Definition 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 (→ Definition I/1.2.6), 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 (→ Definition 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 (→ Proof "mean-sample") 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 (→ Definition I/5.1.9):

$$p(y) = \int p(y,\lambda) \, \mathrm{d}\lambda \; . \tag{10}$$

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 (→ Proof II/3.3.5), 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 (→ Definition IV/3.1.1) is shown to be

$$
\begin{split}
\log p(y|m) = \sum_{i=1}^{n} y_i \log(x_i) - \sum_{i=1}^{n} \log y_i ! + \\
\log \Gamma(a_n) - \log \Gamma(a_0) + a_0 \log b_0 - a_n \log b_n \; .
\end{split}
\tag{13}
$$

Sources:
- original work

Metadata: ID: P43 | shortcut: poissexp-lme | author: JoramSoch | date: 2020-02-04, 15:12.

4 Probability 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 (→ Definition II/3.6.1) with shapes α and β:

$$y_i \sim \mathrm{Bet}(\alpha, \beta), \quad i = 1, \ldots, n \; . \tag{1}$$

Sources:
- original work

Metadata: ID: D77 | shortcut: beta-data | author: JoramSoch | date: 2020-06-28, 21:16.

4.1.2 Method of moments

Theorem: Let $y = \{y_1, \ldots, y_n\}$ be a set of observed counts independent and identically distributed (→ Definition "iid") according to a beta distribution (→ Definition II/3.6.1) 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 (→ Definition "mome") 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 (→ Proof "mean-sample") and $\bar{v}$ is the unbiased sample variance (→ Proof IV/1.1.3):

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \bar{v} &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \end{aligned} \tag{3}$$

Proof: Mean (→ Proof "beta-mean") and variance (→ Proof "beta-var") of the beta distribution (→ Definition II/3.6.1) 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 (→ Definition "mome") 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:

$$\begin{aligned}\bar{y}(\alpha+\beta) &= \alpha \\ \alpha\bar{y} + \beta\bar{y} &= \alpha \\ \beta\bar{y} &= \alpha - \alpha\bar{y} \\ \beta &= \frac{\alpha}{\bar{y}} - \alpha \\ \beta &= \alpha\left(\frac{1}{\bar{y}} - 1\right) \; .\end{aligned} \tag{6}$$

If we define $q = 1/\bar{y} - 1$ and plug (6) into the second equation, we have:

$$\begin{aligned}\bar{v} &= \frac{\alpha \cdot \alpha q}{(\alpha+\alpha q)^2(\alpha+\alpha q+1)} \\ &= \frac{\alpha^2 q}{(\alpha(1+q))^2(\alpha(1+q)+1)} \\ &= \frac{q}{(1+q)^2(\alpha(1+q)+1)} \\ &= \frac{q}{\alpha(1+q)^3+(1+q)^2} \\ q &= \bar{v}\left[\alpha(1+q)^3+(1+q)^2\right] \; .\end{aligned} \tag{7}$$

Noting that $1+q = 1/\bar{y}$ and $q = (1-\bar{y})/\bar{y}$, one obtains for α:

$$\begin{aligned}\frac{1-\bar{y}}{\bar{y}} &= \bar{v}\left[\frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2}\right] \\ \frac{1-\bar{y}}{\bar{y}\,\bar{v}} &= \frac{\alpha}{\bar{y}^3} + \frac{1}{\bar{y}^2} \\ \frac{\bar{y}^3(1-\bar{y})}{\bar{y}\,\bar{v}} &= \alpha + \bar{y} \\ \alpha &= \frac{\bar{y}^2(1-\bar{y})}{\bar{v}} - \bar{y} \\ &= \bar{y}\left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1\right) \; .\end{aligned} \tag{8}$$

Plugging this into equation (6), one obtains for β:

$$\begin{aligned}\beta &= \bar{y}\left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1\right) \cdot \left(\frac{1-\bar{y}}{\bar{y}}\right) \\ &= (1-\bar{y})\left(\frac{\bar{y}(1-\bar{y})}{\bar{v}} - 1\right) \; .\end{aligned} \tag{9}$$

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.

Metadata: ID: P28 | shortcut: beta-mom | author: JoramSoch | date: 2020-01-22, 02:53.

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 (→ Definition II/4.3.1) with concentration parameters $\alpha = [\alpha_1, \ldots, \alpha_k]$:

$$y_i \sim \mathrm{Dir}(\alpha), \quad i = 1, \ldots, n \; . \tag{2}$$

Sources:
- original work

Metadata: ID: D104 | shortcut: dir-data | author: JoramSoch | date: 2020-10-22, 05:06.

4.2.2 Maximum likelihood estimation

Theorem: Let there be a Dirichlet-distributed data (→ Definition 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 (→ Definition 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 (→ Definition I/5.1.2) for each observation is given by the probability density function of the Dirichlet distribution (→ Proof II/4.3.2)

$$p(y_i|\alpha) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} {y_{ij}}^{\alpha_j - 1} \tag{4}$$

and because observations are independent (→ Definition I/1.2.6), 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 (→ Definition 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

$$\begin{split} \mathrm{LL}(\alpha) &= \sum_{i=1}^{n} \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \sum_{i=1}^{n} \sum_{j=1}^{k} \log \Gamma(\alpha_j) + \sum_{i=1}^{n} \sum_{j=1}^{k} (\alpha_j - 1) \log y_{ij} \\ &= n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - n \sum_{j=1}^{k} \log \Gamma(\alpha_j) + n \sum_{j=1}^{k} (\alpha_j - 1) \, \frac{1}{n} \sum_{i=1}^{n} \log y_{ij} \\ &= n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - n \sum_{j=1}^{k} \log \Gamma(\alpha_j) + n \sum_{j=1}^{k} (\alpha_j - 1) \, \log \bar{y}_j \end{split} \tag{7}$$

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

$$\begin{split} \frac{\mathrm{dLL}(\alpha)}{\mathrm{d}\alpha_j} &= \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - n \sum_{j=1}^{k} \log \Gamma(\alpha_j) + n \sum_{j=1}^{k} (\alpha_j - 1) \, \log \bar{y}_j \right] \\ &= \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) \right] - \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n \log \Gamma(\alpha_j) \right] + \frac{\mathrm{d}}{\mathrm{d}\alpha_j} \left[n (\alpha_j - 1) \, \log \bar{y}_j \right] \\ &= n \psi\left(\sum_{j=1}^{k} \alpha_j\right) - n \psi(\alpha_j) + n \log \bar{y}_j \end{split} \tag{9}$$

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 (→ Definition II/4.3.1) belongs to the exponential family (→ Definition "dist-expfam"), 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.

Metadata: ID: P182 | shortcut: dir-mle | author: JoramSoch | date: 2020-10-22, 09:31.

5 Categorical data

5.1 Binomial observations

5.1.1 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, consititutes a set of binomial observations.

Sources:
- original work

Metadata: ID: D78 | shortcut: bin-data | author: JoramSoch | date: 2020-07-07, 07:04.

5.1.2 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 (→ Definition II/1.3.1):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Then, the conjugate prior (→ Definition I/5.2.5) for the model parameter p is a beta distribution (→ Definition II/3.6.1):

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Proof: With the probability mass function of the binomial distribution (→ Proof II/1.3.2), the likelihood function (→ Definition 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 (→ Proof II/3.6.2)

$$\mathrm{p}(p) = \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \tag{6}$$

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.

Metadata: ID: P29 | shortcut: bin-prior | author: JoramSoch | date: 2020-01-23, 23:38.

5.1.3 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 (→ Definition II/1.3.1):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Moreover, assume a beta prior distribution (→ Proof III/5.1.2) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Then, the posterior distribution (→ Definition I/5.1.7) is also a beta distribution (→ Definition II/3.6.1)

$$\mathrm{p}(p|y) = \mathrm{Bet}(p; \alpha_n, \beta_n) \; . \tag{3}$$

and the posterior hyperparameters (→ Definition 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 (→ Proof II/1.3.2), the likelihood function (→ Definition 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 (→ Definition I/5.1.5) of the model is given by

$$\begin{split} \mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \cdot frac1B(\alpha_0, \beta_0) \, p^{\alpha_0-1} \, (1-p)^{\beta_0-1} \\ &= \frac{1}{B(\alpha_0, \beta_0)} \binom{n}{y} \, p^{\alpha_0+y-1} \, (1-p)^{\beta_0+(n-y)-1} \; . \end{split} \tag{6}$$

Note that the posterior distribution is proportional to the joint likelihood (→ Proof I/5.1.8):

$$\mathrm{p}(p|y) \propto \mathrm{p}(y,p) \; . \tag{7}$$

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 (→ Proof II/3.6.2):

$$\mathrm{p}(p|y) = \frac{1}{B(\alpha_n, \beta_n)} \, p^{\alpha_n - 1} \, (1-p)^{\beta_n - 1} = \mathrm{Bet}(p; \alpha_n, \beta_n) \; . \tag{9}$$

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.

Metadata: ID: P30 | shortcut: bin-post | author: JoramSoch | date: 2020-01-24, 00:20.

5.1.4 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 (→ Definition II/1.3.1):

$$y \sim \mathrm{Bin}(n, p) \; . \tag{1}$$

Moreover, assume a beta prior distribution (→ Proof III/5.1.2) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Bet}(p; \alpha_0, \beta_0) \; . \tag{2}$$

Then, the log model evidence (→ Definition IV/3.1.1) 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 (→ Definition 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 (→ Proof II/1.3.2), the likelihood function (→ Definition 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 (→ Definition I/5.1.5) of the model is given by

$$\begin{split} \mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y} \, p^y \, (1-p)^{n-y} \cdot frac1B(\alpha_0, \beta_0) \, p^{\alpha_0 - 1} \, (1-p)^{\beta_0 - 1} \\ &= \binom{n}{y} \, \frac{1}{B(\alpha_0, \beta_0)} \, p^{\alpha_0 + y - 1} \, (1-p)^{\beta_0 + (n-y) - 1} \; . \end{split} \tag{6}$$

Note that the model evidence is the marginal density of the joint likelihood (→ Definition I/5.1.9):

$$\mathrm{p}(y) = \int \mathrm{p}(y,p) \, \mathrm{d}p \; . \tag{7}$$

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 (→ Proof II/3.6.2), 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 (→ Definition IV/3.1.1) (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

$$\begin{split} \log \mathrm{p}(y|m) = \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \\ + \log B(\alpha_n,\beta_n) - \log B(\alpha_0,\beta_0) \; . \end{split} \tag{13}$$

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.

Metadata: ID: P31 | shortcut: bin-lme | author: JoramSoch | date: 2020-01-24, 00:44.

5.2 Multinomial observations

5.2.1 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, $i = 1, \ldots, k$, consititutes a set of multinomial observations.

Sources:
- original work

Metadata: ID: D79 | shortcut: mult-data | author: JoramSoch | date: 2020-07-07, 07:12.

5.2.2 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 (→ Definition II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Then, the conjugate prior (→ Definition I/5.2.5) for the model parameter p is a Dirichlet distribution (→ Definition II/4.3.1):

$$\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}$$

Proof: With the probability mass function of the multinomial distribution (→ Proof II/2.2.2), the likelihood function (→ Definition 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 (→ Proof II/4.3.2)

$$\mathrm{p}(p) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \tag{6}$$

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

Metadata: ID: P79 | shortcut: mult-prior | author: JoramSoch | date: 2020-03-11, 14:15.

5.2.3 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 (→ Definition II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Moreover, assume a Dirichlet prior distribution (→ Proof III/5.2.2) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}$$

Then, the posterior distribution (→ Definition I/5.1.7) is also a Dirichlet distribution (→ Definition II/4.3.1)

$$\mathrm{p}(p|y) = \mathrm{Dir}(p; \alpha_n) \; . \tag{3}$$

and the posterior hyperparameters (→ Definition 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 (→ Proof II/2.2.2), the likelihood function (→ Definition 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 (→ Definition I/5.1.5) of the model is given by

$$\begin{aligned}
\mathrm{p}(y, p) &= \mathrm{p}(y|p)\,\mathrm{p}(p) \\
&= \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \\
&= \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}+y_j-1} \; .
\end{aligned} \tag{6}$$

Note that the posterior distribution is proportional to the joint likelihood (→ Proof I/5.1.8):

$$\mathrm{p}(p|y) \propto \mathrm{p}(y,p) \; . \tag{7}$$

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 (→ Proof II/4.3.2):

$$\mathrm{p}(p|y) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{nj})} \prod_{j=1}^{k} {p_j}^{\alpha_{nj}-1} = \mathrm{Dir}(p; \alpha_n) \; . \tag{9}$$

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

Metadata: ID: P80 | shortcut: mult-post | author: JoramSoch | date: 2020-03-11, 14:40.

5.2.4 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 (→ Definition II/2.2.1):

$$y \sim \mathrm{Mult}(n, p) \; . \tag{1}$$

Moreover, assume a Dirichlet prior distribution (→ Proof III/5.2.2) over the model parameter p:

$$\mathrm{p}(p) = \mathrm{Dir}(p; \alpha_0) \; . \tag{2}$$

Then, the log model evidence (→ Definition IV/3.1.1) for this model is

$$\begin{split} \log \mathrm{p}(y|m) = \log \Gamma(n+1) - \sum_{j=1}^{k} \log \Gamma(k_j+1) \\ + \log \Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right) - \log \Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right) \\ + \sum_{j=1}^{k} \log \Gamma(\alpha_{nj}) - \sum_{j=1}^{k} \log \Gamma(\alpha_{0j}) \; . \end{split} \tag{3}$$

and the posterior hyperparameters (→ Definition 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 (→ Proof II/2.2.2), the likelihood function (→ Definition 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 (→ Definition I/5.1.5) of the model is given by

$$\begin{split} \mathrm{p}(y,p) &= \mathrm{p}(y|p) \, \mathrm{p}(p) \\ &= \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} {p_j}^{y_j} \cdot \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}-1} \\ &= \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \prod_{j=1}^{k} {p_j}^{\alpha_{0j}+y_j-1} \; . \end{split} \tag{6}$$

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 (→ Proof II/4.3.2), p can now be integrated out easily

$$\begin{split} \mathrm{p}(y) &= \int \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{nj})} \prod_{j=1}^{k} {p_j}^{\alpha_{nj}-1} \, \mathrm{d}p \\ &= \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{nj})}{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)} \int \mathrm{Dir}(p; \alpha_n) \, \mathrm{d}p \\ &= \binom{n}{y_1, \ldots, y_k} \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_{0j}\right)}{\Gamma\left(\sum_{j=1}^{k} \alpha_{nj}\right)} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{nj})}{\prod_{j=1}^{k} \Gamma(\alpha_{0j})} \; , \end{split} \tag{9}$$

such that the log model evidence (→ Definition IV/3.1.1) (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

$$\begin{split} \log \mathrm{p}(y|m) = \log \Gamma(n+1) &- \sum_{j=1}^{k} \log \Gamma(k_j+1) \\ &+ \log \Gamma \left(\sum_{j=1}^{k} \alpha_{0j} \right) - \log \Gamma \left(\sum_{j=1}^{k} \alpha_{nj} \right) \\ &+ \sum_{j=1}^{k} \log \Gamma(\alpha_{nj}) - \sum_{j=1}^{k} \log \Gamma(\alpha_{0j}) \; . \end{split} \tag{13}$$

Sources:
- original work

Metadata: ID: P81 | shortcut: mult-lme | author: JoramSoch | date: 2020-03-11, 15:17.

5.3 Logistic regression

5.3.1 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 $y_i = 1$

$$l_i = \log_b \frac{\mathrm{Pr}(y_i = 1)}{\mathrm{Pr}(y_i = 0)} \tag{2}$$

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.

Metadata: ID: D76 | shortcut: logreg | author: JoramSoch | date: 2020-06-28, 20:51.

5.3.2 Probability and log-odds

Theorem: Assume a logistic regression model (→ Definition III/5.3.1)

$$l_i = x_i \beta + \varepsilon_i, \; i = 1, \ldots, n \tag{1}$$

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 (→ Definition I/5.1.2) consistent with (1), $p(y_i)$ are prior probabilities (→ Definition 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 (→ Proof I/5.3.1) and the law of marginal probability (→ Definition I/1.2.3), the posterior probabilities (→ Definition 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.

Metadata: ID: P105 | shortcut: logreg-pnlo | author: JoramSoch | date: 2020-05-19, 05:08.

5.3.3 Log-odds and probability

Theorem: Assume a logistic regression model (→ Definition III/5.3.1)

$$l_i = x_i\beta + \varepsilon_i, \; i = 1, \ldots, n \tag{1}$$

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.

Metadata: ID: P72 | shortcut: logreg-lonp | author: JoramSoch | date: 2020-03-03, 12:01.

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 (→ Definition III/1.1.1)

$$y = X\beta + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2 V) \tag{1}$$

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 (→ Definition I/4.1.3).

Sources:
- original work

Metadata: ID: D20 | shortcut: resvar | author: JoramSoch | date: 2020-02-25, 11:21.

1.1.2 Maximum likelihood estimator is biased

Theorem: Let $x = \{x_1, \ldots, x_n\}$ be a set of independent normally distributed (→ Definition II/3.2.1) observations with unknown mean (→ Definition I/1.5.1) μ and variance (→ Definition I/1.6.1) σ^2:

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

Then,
1) the maximum likelihood estimator (→ Definition I/4.1.3) of σ^2 is

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

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

2) and $\hat{\sigma}^2$ is a biased estimator (→ Definition "est-unb") of σ^2

$$\mathbb{E}\left[\hat{\sigma}^2\right] \neq \sigma^2 \; , \tag{4}$$

more precisely:

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-1}{n} \sigma^2 \; . \tag{5}$$

Proof:
1) This is equivalent to the maximum likelihood estimator for the univariate Gaussian with unknown variance (→ Proof "ug-mle") and a special case of the maximum likelihood estimator for multiple linear regression (→ Proof III/1.1.15) in which $y = x$, $X = 1_n$ and $\hat{\beta} = \bar{x}$:

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}(y - X\hat{\beta})^\mathrm{T}(y - X\hat{\beta}) \\
&= \frac{1}{n}(x - 1_n\bar{x})^\mathrm{T}(x - 1_n\bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \; .
\end{aligned} \tag{6}
$$

2) The expectation (→ Definition I/1.5.1) of the maximum likelihood estimator (→ Definition I/4.1.3) can be developed as follows:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - \mathbb{E}\left[\bar{x}^2\right]
\end{aligned} \tag{7}
$$

Due to the partition of variance into expected values (→ Proof I/1.6.2)

$$
\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \; , \tag{8}
$$

we have

$$
\begin{aligned}
\mathrm{Var}(x_i) &= \mathbb{E}(x_i^2) - \mathbb{E}(x_i)^2 \\
\mathrm{Var}(\bar{x}) &= \mathbb{E}(\bar{x}^2) - \mathbb{E}(\bar{x})^2 \; ,
\end{aligned} \tag{9}
$$

such that (7) becomes

$$
\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}(x_i) + \mathbb{E}(x_i)^2\right) - \left(\mathrm{Var}(\bar{x}) + \mathbb{E}(\bar{x})^2\right) \; . \tag{10}
$$

From (1), it follows that

$$\mathbb{E}(x_i) = \mu \quad \text{and} \quad \mathrm{Var}(x_i) = \sigma^2 \; . \tag{11}$$

The expectation (→ Definition I/1.5.1) of $\bar{x}$ given by (3) is

$$\begin{aligned} \mathbb{E}\left[\bar{x}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n x_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[x_i\right] \\ &\overset{(11)}{=} \frac{1}{n}\sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu \\ &= \mu \; . \end{aligned} \tag{12}$$

The variance of $\bar{x}$ given by (3) is

$$\begin{aligned} \mathrm{Var}\left[\bar{x}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n x_i\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left[x_i\right] \\ &\overset{(11)}{=} \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{1}{n^2} \cdot n \cdot \sigma^2 \\ &= \frac{1}{n}\sigma^2 \; . \end{aligned} \tag{13}$$

Plugging (11), (12) and (13) into (10), we have

$$\begin{aligned} \mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n}\sum_{i=1}^n \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\ \mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n} \cdot n \cdot \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu^2\right) \\ \mathbb{E}\left[\hat{\sigma}^2\right] &= \sigma^2 + \mu^2 - \frac{1}{n}\sigma^2 - \mu^2 \\ \mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{n-1}{n}\sigma^2 \end{aligned} \tag{14}$$

which proves the bias (→ Definition "est-unb") 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.

Metadata: ID: P61 | shortcut: resvar-bias | author: JoramSoch | date: 2020-02-24, 23:44.

1.1.3 Construction of unbiased estimator

Theorem: Let $x = \{x_1, \ldots, x_n\}$ be a set of independent normally distributed (→ Definition II/3.2.1) observations with unknown mean (→ Definition I/1.5.1) μ and variance (→ Definition I/1.6.1) σ^2:

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n \; . \tag{1}$$

An unbiased estimator (→ Definition "est-unb") of σ^2 is given by

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

Proof: It can be shown that (→ Proof IV/1.1.2) the maximum likelihood estimator (→ Definition I/4.1.3) of σ^2

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

is a biased estimator (→ Definition "est-unb") 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

$$\begin{aligned} \mathbb{E}\left[\frac{n}{n-1} \hat{\sigma}^2_{\text{MLE}}\right] &= \frac{n}{n-1} \mathbb{E}\left[\hat{\sigma}^2_{\text{MLE}}\right] \\ &\overset{(4)}{=} \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 \\ &= \sigma^2 \; , \end{aligned} \tag{5}$$

such that an unbiased estimator (→ Definition "est-unb") can be constructed as

$$\begin{aligned} \hat{\sigma}^2_{\text{unb}} &= \frac{n}{n-1} \hat{\sigma}^2_{\text{MLE}} \\ &\overset{(3)}{=} \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \; . \end{aligned} \tag{6}$$

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.

Metadata: ID: P62 | shortcut: resvar-unb | author: JoramSoch | date: 2020-02-25, 15:38.

1.2 R-squared

1.2.1 Definition

Definition: Let there be a linear regression model (→ Definition III/1.1.1) with independent (→ Definition I/1.2.6) 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 (→ Definition III/1.1.4)") that can be predicted from the independent variables X ("explained variance (→ Definition III/1.1.5)") 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.

Metadata: ID: D21 | shortcut: rsq | author: JoramSoch | date: 2020-02-25, 11:41.

1.2.2 Derivation of R² and adjusted R²

Theorem: Given a linear regression model (→ Definition III/1.1.1)

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with n independent observations and p independent variables,
1) the coefficient of determination (→ Definition 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 (→ Definition III/1.1.6) and total sum of squares (→ Definition III/1.1.4) are

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \text{TSS} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \;, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \end{aligned} \tag{4}$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares (→ Proof III/1.1.2) estimates.

Proof: The coefficient of determination R^2 is defined as (→ Definition IV/1.2.1) 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 (→ Definition III/1.1.5) as

$$\text{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \;, \tag{5}$$

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 (→ Proof III/1.1.7) 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 (→ Proof IV/1.1.3), 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 (→ Definition "dof").

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.

Metadata: ID: P8 | shortcut: rsq-der | author: JoramSoch | date: 2019-12-06, 11:19.

1.2.3 Relationship to maximum log-likelihood

Theorem: Given a linear regression model (→ Definition III/1.1.1) with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the coefficient of determination (→ Definition IV/1.2.1) can be expressed in terms of the maximum log-likelihood (→ Definition I/4.1.4) as

$$R^2 = 1 - \left(\exp[\Delta\mathrm{MLL}]\right)^{-2/n} \tag{2}$$

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 (→ Definition I/4.1.4) (MLL) of a linear regression model in terms of its residual sum of squares (→ Definition III/1.1.6) (RSS). The model in (1) implies the following log-likelihood function (→ Definition 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 (→ Proof III/1.1.15)

$$\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^\mathrm{T} (y - X\hat{\beta}) \tag{5}$$

and the residual sum of squares (→ Definition III/1.1.6) 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 (→ Definition 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 (→ Definition III/1.1.4):

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \; . \tag{10}$$

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.

Sources:
- original work

Metadata: ID: P14 | shortcut: rsq-mll | author: JoramSoch | date: 2020-01-08, 04:46.

1.3 Signal-to-noise ratio

1.3.1 Definition

Definition: Let there be a linear regression model (→ Definition III/1.1.1) with independent (→ Definition I/1.2.6) 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 (→ Proof III/1.1.15) $\hat{\beta}$ and residual variance (→ Definition 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.

Metadata: ID: D22 | shortcut: snr | author: JoramSoch | date: 2020-02-25, 12:01.

1.3.2 Relationship with R²

Theorem: Let there be a linear regression model (→ Definition III/1.1.1) with independent (→ Definition I/1.2.6) observations

$$
y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)
\tag{1}
$$

and parameter estimates (→ Definition "est") obtained with ordinary least squares (→ Proof III/1.1.2)

$$
\hat{\beta} = (X^\mathrm{T} X)^{-1} X^\mathrm{T} y \; .
\tag{2}
$$

Then, the signal-to noise ratio (→ Definition IV/1.3.1) can be expressed in terms of the coefficient of determination (→ Definition 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 (SNR) is defined as (→ Definition IV/1.3.1)

$$\mathrm{SNR} = \frac{\mathrm{Var}(X\hat{\beta})}{\hat{\sigma}^2} = \frac{\mathrm{Var}(\hat{y})}{\hat{\sigma}^2} \; . \tag{5}$$

Writing out the variances, 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 (→ Proof IV/1.1.2) (dividing by n) or the unbiased estimator fo the variance (→ Proof IV/1.1.3) (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 (→ Definition III/1.1.5), residual (→ Definition III/1.1.6) and total sum of squares (→ Definition III/1.1.4):

$$\mathrm{SNR} = \frac{\mathrm{ESS}}{\mathrm{RSS}} = \frac{\mathrm{ESS}/\mathrm{TSS}}{\mathrm{RSS}/\mathrm{TSS}} \; . \tag{8}$$

With the derivation of the coefficient of determination (→ Proof IV/1.2.2), this becomes

$$\mathrm{SNR} = \frac{R^2}{1 - R^2} \; . \tag{9}$$

Rearranging this equation for the coefficient of determination (→ Definition IV/1.2.1), we have

$$R^2 = \frac{\mathrm{SNR}}{1 + \mathrm{SNR}} \; , \tag{10}$$

Sources:
- original work

Metadata: ID: P63 | shortcut: snr-rsq | author: JoramSoch | date: 2020-02-26, 10:37.

2 Classical information criteria

2.1 Akaike information criterion

2.1.1 Definition

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (→ Definition I/4.1.3)

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta, m) \; . \tag{1}$$

Then, the Akaike information criterion (AIC) of this model is defined as

$$\mathrm{AIC}(m) = -2 \log p(y|\hat{\theta}, m) + 2\,p \tag{2}$$

where p is the number of free parameters estimated via (1).

Sources:

- Akaike H (1974): "A New Look at the Statistical Model Identification"; in: *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716-723; URL: https://ieeexplore.ieee.org/document/1100705; DOI: 10.1109/TAC.1974.1100705.

Metadata: ID: D23 | shortcut: aic | author: JoramSoch | date: 2020-02-25, 12:31.

2.2 Bayesian information criterion

2.2.1 Definition

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta, m)$ and maximum likelihood estimates (→ Definition I/4.1.3)

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta, m) \; . \tag{1}$$

Then, the Bayesian information criterion (BIC) of this model is defined as

$$\mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \tag{2}$$

where n is the number of data points and p is the number of free parameters estimated via (1).

Sources:

- Schwarz G (1978): "Estimating the Dimension of a Model"; in: *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464; URL: https://www.jstor.org/stable/2958889.

Metadata: ID: D24 | shortcut: bic | author: JoramSoch | date: 2020-02-25, 12:21.

2.2.2 Derivation

Theorem: Let $p(y|\theta, m)$ be the likelihood function (→ Definition I/5.1.2) of a generative model (→ Definition I/5.1.1) $m \in \mathcal{M}$ with model parameters $\theta \in \Theta$ describing measured data $y \in \mathbb{R}^n$.

Let $p(\theta|m)$ be a prior distribution (→ Definition I/5.1.3) on the model parameters. Assume that likelihood function and prior density are twice differentiable.
Then, as the number of data points goes to infinity, an approximation to the log-marginal likelihood (→ Definition I/5.1.9) $\log p(y|m)$, up to constant terms not depending on the model, is given by the Bayesian information criterion (→ Definition IV/2.2.1) (BIC) as

$$-2 \log p(y|m) \approx \mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \tag{1}$$

where $\hat{\theta}$ is the maximum likelihood estimator (→ Definition I/4.1.3) (MLE) of θ, n is the number of data points and p is the number of model parameters.

Proof: Let $\mathrm{LL}(\theta)$ be the log-likelihood function (→ Definition I/4.1.2)

$$\mathrm{LL}(\theta) = \log p(y|\theta, m) \tag{2}$$

and define the functions g and h as follows:

$$\begin{aligned} g(\theta) &= p(\theta|m) \\ h(\theta) &= \frac{1}{n} \, \mathrm{LL}(\theta) \; . \end{aligned} \tag{3}$$

Then, the marginal likelihood (→ Definition I/5.1.9) can be written as follows:

$$\begin{aligned} p(y|m) &= \int_{\Theta} p(y|\theta, m) \, p(\theta|m) \, \mathrm{d}\theta \\ &= \int_{\Theta} \exp\left[n \, h(\theta)\right] \, g(\theta) \, \mathrm{d}\theta \; . \end{aligned} \tag{4}$$

This is an integral suitable for Laplace approximation which states that

$$\int_{\Theta} \exp\left[n \, h(\theta)\right] \, g(\theta) \, \mathrm{d}\theta = \left(\sqrt{\frac{2\pi}{n}}\right)^p \exp\left[n \, h(\theta_0)\right] \left(g(\theta_0) \left|J(\theta_0)\right|^{-1/2} + O(1/n)\right) \tag{5}$$

where θ_0 is the value that maximizes $h(\theta)$ and $J(\theta_0)$ is the Hessian matrix evaluated at θ_0. In our case, we have $h(\theta) = 1/n \, \mathrm{LL}(\theta)$ such that θ_0 is the maximum likelihood estimator $\hat{\theta}$:

$$\hat{\theta} = \operatorname*{arg\,max}_{\theta} \mathrm{LL}(\theta) \; . \tag{6}$$

With this, (5) can be applied to (4) using (3) to give:

$$p(y|m) \approx \left(\sqrt{\frac{2\pi}{n}}\right)^p p(y|\hat{\theta}, m) \, p(\hat{\theta}|m) \, \left|J(\hat{\theta})\right|^{-1/2} \; . \tag{7}$$

Logarithmizing and multiplying with -2, we have:

$$-2 \log p(y|m) \approx -2 \, \mathrm{LL}(\hat{\theta}) + p \log n - p \log(2\pi) - 2 \log p(\hat{\theta}|m) + \log \left|J(\hat{\theta})\right| \; . \tag{8}$$

As $n \to \infty$, the last three terms are $O_p(1)$ and can therefore be ignored when comparing between models $\mathcal{M} = \{m_1, \ldots, m_M\}$ and using $p(y|m_j)$ to compute posterior model probabilies (→ Definition IV/3.5.1) $p(m_j|y)$. With that, the BIC is given as

$$\mathrm{BIC}(m) = -2 \log p(y|\hat{\theta}, m) + p \log n \; . \tag{9}$$

Sources:

- Claeskens G, Hjort NL (2008): "The Bayesian information criterion"; in: *Model Selection and Model Averaging*, ch. 3.2, pp. 78-81; URL: https://www.cambridge.org/core/books/model-selection-and-model-av E6F1EC77279D1223423BB64FC3A12C37; DOI: 10.1017/CBO9780511790485.

Metadata: ID: P32 | shortcut: bic-der | author: JoramSoch | date: 2020-01-26, 23:36.

2.3 Deviance information criterion

2.3.1 Definition

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta, m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|m)$. Together, likelihood function and prior distribution imply a posterior distribution (→ Definition I/5.1.7) $p(\theta|y, m)$. Define the posterior expected log-likelihood (→ Definition I/4.1.2) (PLL)

$$\mathrm{PLL}(m) = \left\langle \log p(y|\theta, m) \right\rangle_{\theta|y} \tag{1}$$

and the log-likelihood (→ Definition I/4.1.2) at the posterior expectation (LLP)

$$\mathrm{LLP}(m) = \log p(y| \left\langle \theta \right\rangle_{\theta|y}, m) \tag{2}$$

where $\langle \cdot \rangle_{\theta|y}$ denotes an expectation across the posterior distribution.
Then, the deviance information criterion (DIC) of the model is defined as

$$\mathrm{DIC}(m) = -2 \, \mathrm{LLP}(m) + 2 \, p_D \quad \text{or} \quad \mathrm{DIC}(m) = -2 \, \mathrm{PLL}(m) + p_D \tag{3}$$

where the "effective number of parameters" p_D is given by

$$p_D = -2 \, \mathrm{PLL}(m) + 2 \, \mathrm{LLP}(m) \; . \tag{4}$$

Sources:

- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002): "Bayesian measures of model complexity and fit"; in: *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 64, iss. 4, pp. 583-639; URL: https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353; DOI: 10.1111/1467-9868.00353.
- 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, eqs. 10-12; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D25 | shortcut: dic | author: JoramSoch | date: 2020-02-25, 12:46.

3 Bayesian model selection

3.1 Log model evidence

3.1.1 Definition

Definition: Let m be a generative model (→ Definition I/5.1.1) with likelihood function (→ Definition I/5.1.2) $p(y|\theta,m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|m)$. Then, the log model evidence (LME) of this model is defined as the logarithm of the marginal likelihood (→ Definition I/5.1.9):

$$\mathrm{LME}(m) = \log p(y|m) \; . \tag{1}$$

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. 13; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D26 | shortcut: lme | author: JoramSoch | date: 2020-02-25, 12:56.

3.1.2 Derivation

Theorem: Let $p(y|\theta,m)$ be a likelihood function (→ Definition I/5.1.2) of a generative model (→ Definition I/5.1.1) m for making inferences on model parameters θ given measured data y. Moreover, let $p(\theta|m)$ be a prior distribution (→ Definition I/5.1.3) on model parameters θ. Then, the log model evidence (→ Definition IV/3.1.1) (LME), also called marginal log-likelihood,

$$\mathrm{LME}(m) = \log p(y|m) \; , \tag{1}$$

can be expressed
1) as

$$\mathrm{LME}(m) = \log \int p(y|\theta,m)\, p(\theta|m)\, \mathrm{d}\theta \tag{2}$$

2) or

$$\mathrm{LME}(m) = \log p(y|\theta,m) + \log p(\theta|m) - \log p(\theta|y,m) \; . \tag{3}$$

Proof:
1) The first expression is a simple consequence of the law of marginal probability (→ Definition I/1.2.3) for continuous variables according to which

$$p(y|m) = \int p(y|\theta,m)\, p(\theta|m)\, \mathrm{d}\theta \tag{4}$$

which, when logarithmized, gives

$$\mathrm{LME}(m) = \log p(y|m) = \log \int p(y|\theta,m)\, p(\theta|m)\, \mathrm{d}\theta \; . \tag{5}$$

2) The second expression can be derived from Bayes' theorem (→ Proof I/5.3.1) which makes a statement about the posterior distribution (→ Definition I/5.1.7):

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \; . \tag{6}$$

Rearranging for $p(y|m)$ and logarithmizing, we have:

$$\begin{split} \mathrm{LME}(m) = \log p(y|m) &= \log \frac{p(y|\theta,m)\,p(\theta|m)}{p(\theta|y,m)} \\ &= \log p(y|\theta,m) + \log p(\theta|m) - \log p(\theta|y,m) \; . \end{split} \tag{7}$$

Sources:
- original work

Metadata: ID: P13 | shortcut: lme-der | author: JoramSoch | date: 2020-01-06, 21:27.

3.1.3 Partition into accuracy and complexity

Theorem: The log model evidence (→ Definition IV/3.1.1) can be partitioned into accuracy and complexity

$$\mathrm{LME}(m) = \mathrm{Acc}(m) - \mathrm{Com}(m) \tag{1}$$

where the accuracy term is the posterior (→ Definition I/5.1.7) expectation (→ Definition "mean-lotus") of the log-likelihood function (→ Definition I/4.1.2)

$$\mathrm{Acc}(m) = \langle \log p(y|\theta,m) \rangle_{p(\theta|y,m)} \tag{2}$$

and the complexity penalty is the Kullback-Leibler divergence (→ Definition I/2.5.1) of posterior (→ Definition I/5.1.7) from prior (→ Definition I/5.1.3)

$$\mathrm{Com}(m) = \mathrm{KL}\left[p(\theta|y,m)\,||\,p(\theta|m)\right] \; . \tag{3}$$

Proof: We consider Bayesian inference on data (→ Definition "data") y using model (→ Definition I/5.1.1) m with parameters θ. Then, Bayes' theorem (→ Proof I/5.3.1) makes a statement about the posterior distribution (→ Definition I/5.1.7), i.e. the probability of parameters, given the data and the model:

$$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)} \; . \tag{4}$$

Rearranging this for the model evidence (→ Proof IV/3.1.2), we have:

$$p(y|m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(\theta|y,m)} \; . \tag{5}$$

Logarthmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta,m) - \log \frac{p(\theta|y,m)}{p(\theta|m)} \; . \tag{6}$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y,m) \log p(y|\theta,m) \, \mathrm{d}\theta - \int p(\theta|y,m) \log \frac{p(\theta|y,m)}{p(\theta|m)} \, \mathrm{d}\theta \; . \tag{7}$$

By definition, the left-hand side is the log model evidence and the terms on the right-hand side correspond to the posterior expectation of the log-likelihood function and the Kullback-Leibler divergence of posterior from prior

$$\mathrm{LME}(m) = \left\langle \log p(y|\theta,m) \right\rangle_{p(\theta|y,m)} - \mathrm{KL}\left[p(\theta|y,m) \, || \, p(\theta|m) \right] \tag{8}$$

which proofs the partition given by (1).

Sources:
- Penny et al. (2007): "Bayesian Comparison of Spatially Regularised General Linear Models"; in: *Human Brain Mapping*, vol. 28, pp. 275–293; URL: https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.20327; DOI: 10.1002/hbm.20327.
- Soch et al. (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection"; in: *NeuroImage*, vol. 141, pp. 469–489; URL: https://www.sciencedirect.com/science/article/pii/S1053811916303615; DOI: 10.1016/j.neuroimage.2016.07.047.

Metadata: ID: P3 | shortcut: lme-anc | author: JoramSoch | date: 2019-09-27, 16:13.

3.1.4 Uniform-prior log model evidence

Definition: Assume a generative model (→ Definition I/5.1.1) m with likelihood function (→ Definition I/5.1.2) $p(y|\theta,m)$ and a uniform (→ Definition I/5.2.2) prior distribution (→ Definition I/5.1.3) $p_{\mathrm{uni}}(\theta|m)$. Then, the log model evidence (→ Definition IV/3.1.1) of this model is called "log model evidence with uniform prior" or "uniform-prior log model evidence" (upLME):

$$\mathrm{upLME}(m) = \log \int p(y|\theta,m) \, p_{\mathrm{uni}}(\theta|m) \, \mathrm{d}\theta \; . \tag{1}$$

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.

Metadata: ID: D113 | shortcut: uplme | author: JoramSoch | date: 2020-11-25, 07:28.

3.1.5 Cross-validated log model evidence

Definition: Let there be a data set (→ Definition "data") y with mutually exclusive and collectively exhaustive subsets $y_1, \ldots, y_S$. Assume a generative model (→ Definition I/5.1.1) m with model parameters θ implying a likelihood function (→ Definition I/5.1.2) $p(y|\theta,m)$ and a non-informative (→ Definition I/5.2.3) prior density (→ Definition I/5.1.3) $p_{\mathrm{ni}}(\theta|m)$.
Then, the cross-validated log model evidence of m is given by

$$\mathrm{cvLME}(m) = \sum_{i=1}^{S} \log \int p(y_i|\theta,m) \, p(\theta|y_{\neg i},m) \, \mathrm{d}\theta \tag{1}$$

where $y_{\neg i} = \bigcup_{j \neq i} y_j$ is the union of all data subsets except y_i and $p(\theta|y_{\neg i}, m)$ is the posterior distribution (→ Definition I/5.1.7) obtained from $y_{\neg i}$ when using the prior distribution (→ Definition I/5.1.3) $p_{\mathrm{ni}}(\theta|m)$:

$$p(\theta|y_{\neg i}, m) = \frac{p(y_{\neg i}|\theta, m)\, p_{\mathrm{ni}}(\theta|m)}{p(y_{\neg i}|m)} \; . \tag{2}$$

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, eqs. 13-15; URL: https://www.sciencedirect.com/science/article/pii/S1053811916303615; DOI: 10.1016/j.neuroimage
- 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. 6; URL: https://www.sciencedirect.com/science/article/pii/S105381191730527X; DOI: 10.1016/j.neuroimage.2017.06.056.
- 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, eqs. 14-15; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.
- Soch J (2018): "cvBMS and cvBMA: filling in the gaps"; in: *arXiv stat.ME*, arXiv:1807.01585; URL: https://arxiv.org/abs/1807.01585.

Metadata: ID: D111 | shortcut: cvlme | author: JoramSoch | date: 2020-11-19, 04:55.

3.1.6 Empirical Bayesian log model evidence

Definition: Let m be a generative model (→ Definition I/5.1.1) with model parameters θ and hyper-parameters λ implying the likelihood function (→ Definition I/5.1.2) $p(y|\theta, \lambda, m)$ and prior distribution (→ Definition I/5.1.3) $p(\theta|\lambda, m)$. Then, the Empirical Bayesian (→ Definition "eb") log model evidence (→ Definition IV/3.1.1) is the logarithm of the marginal likelihood (→ Definition I/5.1.9), maximized with respect to the hyper-parameters:

$$\mathrm{ebLME}(m) = \log p(y|\hat{\lambda}, m) \tag{1}$$

where

$$p(y|\lambda, m) = \int p(y|\theta, \lambda, m)\, (\theta|\lambda, m)\, \mathrm{d}\theta \tag{2}$$

and (→ Definition I/5.2.7)

$$\hat{\lambda} = \arg\max_{\lambda} \log p(y|\lambda, m) \; . \tag{3}$$

Sources:
- Wikipedia (2020): "Empirical Bayes method"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/Empirical_Bayes_method#Introduction.

Metadata: ID: D114 | shortcut: eblme | author: JoramSoch | date: 2020-11-25, 07:43.

3.1.7 Variational Bayesian log model evidence

Definition: Let m be a generative model (→ Definition I/5.1.1) with model parameters θ implying the likelihood function (→ Definition I/5.1.2) $p(y|\theta, m)$. Moreover, assume a prior distribution (→ Definition I/5.1.3) $p(\theta|m)$, a resulting posterior distribution (→ Definition I/5.1.7) $p(\theta|y, m)$ and an approximate (→ Definition "vb") posterior distribution (→ Definition I/5.1.7) $q(\theta)$. Then, the Variational Bayesian (→ Definition "vb") log model evidence (→ Definition IV/3.1.1) is the expectation of the log-likelihood function (→ Definition I/4.1.2) with respect to the approximate posterior, minus the Kullback-Leibler divergence (→ Definition I/2.5.1) between approximate posterior and true posterior distribution:

$$\mathrm{vbLME}(m) = \mathcal{L}\left[q(\theta)\right] - \mathrm{KL}\left[q(\theta)||p(\theta|y)\right] \tag{1}$$

where

$$\mathcal{L}\left[q(\theta)\right] = \int q(\theta) \log \frac{p(y, \theta|m)}{q(\theta)} \, \mathrm{d}\theta \tag{2}$$

and

$$\mathrm{KL}\left[q(\theta)||p(\theta|y)\right] = \int q(\theta) \log \frac{q(\theta)}{p(\theta|y, m)} \, \mathrm{d}\theta \; . \tag{3}$$

Sources:
- Wikipedia (2020): "Variational Bayesian methods"; in: *Wikipedia, the free encyclopedia*, retrieved on 2020-11-25; URL: https://en.wikipedia.org/wiki/Variational_Bayesian_methods#Evidence_lower_bound.
- Bishop CM (2006): "Variational Inference"; in: *Pattern Recognition for Machine Learning*, pp. 462-474, eqs. 10.2-10.4; URL: https://www.springer.com/gp/book/9780387310732.

Metadata: ID: D115 | shortcut: vblme | author: JoramSoch | date: 2020-11-25, 08:10.

3.2 Log family evidence

3.2.1 Definition

Definition: Let f be a family of M generative models (→ Definition I/5.1.1) $m_1, \ldots, m_M$, such that the following statement holds true:

$$f \Leftrightarrow m_1 \vee \ldots \vee m_M \; . \tag{1}$$

Then, the family evidence of f is the weighted average of the model evidences (→ Definition I/5.1.9) of $m_1, \ldots, m_M$ where the weights are the within-family prior model probabilities (→ Definition I/5.1.3)

$$p(y|f) = \sum_{i=1}^{M} p(y|m_i) \, p(m_i|f) \; . \tag{2}$$

The log family evidence is given by the logarithm of the family evidence:

$$\mathrm{LFE}(f) = \log p(y|f) = \log \sum_{i=1}^{M} p(y|m_i) \, p(m_i|f) \; . \tag{3}$$

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. 16; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D80 | shortcut: lfe | author: JoramSoch | date: 2020-07-13, 22:31.

3.2.2 Derivation

Theorem: Let f be a family of M generative models (→ Definition I/5.1.1) $m_1, \ldots, m_M$ with model evidences (→ Definition I/5.1.9) $p(y|m_1), \ldots, p(y|m_M)$. Then, the log family evidence (→ Definition IV/3.2.1)

$$\mathrm{LFE}(f) = \log p(y|f) \tag{1}$$

can be expressed as

$$\mathrm{LFE}(f) = \log \sum_{i=1}^{M} p(y|m_i)\, p(m_i|f) \tag{2}$$

where $p(m_i|f)$ are the within-family (→ Definition IV/3.2.1) prior (→ Definition I/5.1.3) model (→ Definition I/5.1.1) probabilities (→ Definition I/1.2.1).

Proof: We will assume "prior addivivity"

$$p(f) = \sum_{i=1}^{M} p(m_i) \tag{3}$$

and "posterior additivity" for family probabilities:

$$p(f|y) = \sum_{i=1}^{M} p(m_i|y) \tag{4}$$

Bayes' theorem (→ Proof I/5.3.1) for the family evidence (→ Definition IV/3.2.1) gives

$$p(y|f) = \frac{p(f|y)\, p(y)}{p(f)} \; . \tag{5}$$

Applying (3) and (4), we have

$$p(y|f) = \frac{\sum_{i=1}^{M} p(m_i|y)\, p(y)}{\sum_{i=1}^{M} p(m_i)} \; . \tag{6}$$

Bayes' theorem (→ Proof I/5.3.1) for the model evidence (→ Definition IV/3.2.1) gives

$$p(y|m_i) = \frac{p(m_i|y)\, p(y)}{p(m_i)} \tag{7}$$

which can be rearranged into

$$p(m_i|y)\, p(y) = p(y|m_i)\, p(m_i) \; . \tag{8}$$

Plugging (8) into (6), we have

$$
\begin{aligned}
p(y|f) &= \frac{\sum_{i=1}^{M} p(y|m_i)\, p(m_i)}{\sum_{i=1}^{M} p(m_i)} \\
&= \sum_{i=1}^{M} p(y|m_i) \cdot \frac{p(m_i)}{\sum_{i=1}^{M} p(m_i)} \\
&= \sum_{i=1}^{M} p(y|m_i) \cdot \frac{p(m_i, f)}{p(f)} \\
&= \sum_{i=1}^{M} p(y|m_i) \cdot p(m_i|f) \; .
\end{aligned} \tag{9}
$$

Equation (2) follows by logarithmizing both sides of (9).

Sources:
- original work

Metadata: ID: P132 | shortcut: lfe-der | author: JoramSoch | date: 2020-07-13, 22:58.

3.2.3 Calculation from log model evidences

Theorem: Let $m_1, \ldots, m_M$ be M statistical models with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1), \ldots, \mathrm{LME}(m_M)$ and belonging to F mutually exclusive model families $f_1, \ldots, f_F$. Then, the log family evidences (→ Definition IV/3.2.1) are given by:

$$
\mathrm{LFE}(f_j) = \log \sum_{m_i \in f_j} \left[\exp[\mathrm{LME}(m_i)] \cdot p(m_i|f_j) \right], \quad j = 1, \ldots, F, \tag{1}
$$

where $p(m_i|f_j)$ are within-family (→ Definition IV/3.2.1) prior (→ Definition I/5.1.3) model (→ Definition I/5.1.1) probabilities (→ Definition I/1.2.1).

Proof: Let us consider the (unlogarithmized) family evidence $p(y|f_j)$. According to the law of marginal probability (→ Definition I/1.2.3), this conditional probability is given by

$$
p(y|f_j) = \sum_{m_i \in f_j} \left[p(y|m_i, f_j) \cdot p(m_i|f_j) \right] \; . \tag{2}
$$

Because model families are mutually exclusive, it holds that $p(y|m_i, f_j) = p(y|m_i)$, such that

$$
p(y|f_j) = \sum_{m_i \in f_j} \left[p(y|m_i) \cdot p(m_i|f_j) \right] \; . \tag{3}
$$

Logarithmizing transforms the family evidence $p(y|f_j)$ into the log family evidence $\mathrm{LFE}(f_j)$:

$$
\mathrm{LFE}(f_j) = \log \sum_{m_i \in f_j} \left[p(y|m_i) \cdot p(m_i|f_j) \right] \; . \tag{4}
$$

The definition of the log model evidence (→ Definition IV/3.1.1)

$$\mathrm{LME}(m) = \log p(y|m) \tag{5}$$

can be exponentiated to then read

$$\exp\left[\mathrm{LME}(m)\right] = p(y|m) \tag{6}$$

and applying (6) to (4), we finally have:

$$\mathrm{LFE}(f_j) = \log \sum_{m_i \in f_j} \left[\exp[\mathrm{LME}(m_i)] \cdot p(m_i|f_j)\right] \; . \tag{7}$$

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. 16; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P65 | shortcut: lfe-lme | author: JoramSoch | date: 2020-02-27, 21:16.

3.3 Log Bayes factor

3.3.1 Definition

Definition: Let there be two generative models (→ Definition I/5.1.1) m_1 and m_2 which are mutually exclusive, but not necessarily collectively exhaustive:

$$\neg(m_1 \wedge m_2) \tag{1}$$

Then, the Bayes factor in favor of m_1 and against m_2 is the ratio of the model evidences (→ Definition I/5.1.9) of m_1 and m_2:

$$\mathrm{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{2}$$

The log Bayes factor is given by the logarithm of the Bayes factor:

$$\mathrm{LBF}_{12} = \log \mathrm{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{3}$$

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. 18; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D84 | shortcut: lbf | author: JoramSoch | date: 2020-07-22, 07:02.

3.3.2 Derivation

Theorem: Let there be two generative models (→ Definition I/5.1.1) m_1 and m_2 with model evidences (→ Definition I/5.1.9) $p(y|m_1)$ and $p(y|m_2)$. Then, the log Bayes factor (→ Definition IV/3.3.1)

$$\mathrm{LBF}_{12} = \log \mathrm{BF}_{12} \tag{1}$$

can be expressed as

$$\mathrm{LBF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{2}$$

Proof: The Bayes factor (→ Definition IV/3.4.1) is defined as the posterior (→ Definition I/5.1.7) odds ratio (→ Definition "odds") when both models (→ Definition I/5.1.1) are equally likely apriori (→ Definition I/5.1.3):

$$\mathrm{BF}_{12} = \frac{p(m_1|y)}{p(m_2|y)} \tag{3}$$

Plugging in the posterior odds ratio according to Bayes' rule (→ Proof I/5.3.2), we have

$$\mathrm{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} \; . \tag{4}$$

When both models are equally likely apriori, the prior (→ Definition I/5.1.3) odds ratio (→ Definition "odds") is one, such that

$$\mathrm{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{5}$$

Equation (2) follows by logarithmizing both sides of (5).

Sources:
- original work

Metadata: ID: P137 | shortcut: lbf-der | author: JoramSoch | date: 2020-07-22, 07:27.

3.3.3 Calculation from log model evidences

Theorem: Let m_1 and m_2 be two statistical models with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1)$ and $\mathrm{LME}(m_2)$. Then, the log Bayes factor (→ Definition IV/3.3.1) in favor of model m_1 and against model m_2 is the difference of the log model evidences:

$$\mathrm{LBF}_{12} = \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \tag{1}$$

Proof: The Bayes factor (→ Definition IV/3.4.1) is defined as the ratio of the model evidences (→ Definition I/5.1.9) of m_1 and m_2

$$\mathrm{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)} \tag{2}$$

and the log Bayes factor (→ Definition IV/3.3.1) is defined as the logarithm of the Bayes factor

$$\mathrm{LBF}_{12} = \log \mathrm{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{3}$$

With the definition of the log model evidence (→ Definition IV/3.1.1)

$$\mathrm{LME}(m) = \log p(y|m) \tag{4}$$

the log Bayes factor can be expressed as:
Resolving the logarithm and applying the definition of the log model evidence (→ Definition IV/3.1.1), we finally have:

$$\begin{split} \mathrm{LBF}_{12} &= \log p(y|m_1) - \log p(y|m_2) \\ &= \mathrm{LME}(m_1) - \mathrm{LME}(m_2) \; . \end{split} \tag{5}$$

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. 18; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P64 | shortcut: lbf-lme | author: JoramSoch | date: 2020-02-27, 20:51.

3.4 Bayes factor

3.4.1 Definition

Definition: Consider two competing generative models (→ Definition I/5.1.1) m_1 and m_2 for observed data y. Then the Bayes factor in favor m_1 over m_2 is the ratio of marginal likelihoods (→ Definition I/5.1.9) of m_1 and m_2:

$$\mathrm{BF}_{12} = \frac{p(y \mid m_1)}{p(y \mid m_2)}. \tag{1}$$

Note that by Bayes' theorem (→ Proof I/5.3.1), the ratio of posterior model probabilities (→ Definition IV/3.5.1) (i.e., the posterior model odds) can be written as

$$\frac{p(m_1 \mid y)}{p(m_2 \mid y)} = \frac{p(m_1)}{p(m_2)} \cdot \frac{p(y \mid m_1)}{p(y \mid m_2)}, \tag{2}$$

or equivalently by (1),

$$\frac{p(m_1 \mid y)}{p(m_2 \mid y)} = \frac{p(m_1)}{p(m_2)} \cdot \mathrm{BF}_{12}. \tag{3}$$

In other words, the Bayes factor can be viewed as the factor by which the prior model odds are updated (after observing data y) to posterior model odds – which is also expressed by Bayes' rule (→ Proof I/5.3.2).

Sources:

- Kass, Robert E. and Raftery, Adrian E. (1995): "Bayes Factors"; in: *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773-795; URL: https://dx.doi.org/10.1080/01621459.1995.10476572; DOI: 10.1080/01621459.1995.10476572.

Metadata: ID: D92 | shortcut: bf | author: tomfaulkenberry | date: 2020-08-26, 12:00.

3.4.2 Transitivity

Theorem: Consider three competing models (→ Definition I/5.1.1) m_1, m_2, and $m3$ for observed data y. Then the Bayes factor (→ Definition IV/3.4.1) for m_1 over m_3 can be written as:

$$\mathrm{BF}_{13} = \mathrm{BF}_{12} \cdot \mathrm{BF}_{23} . \tag{1}$$

Proof: By definition (→ Definition IV/3.4.1), the Bayes factor BF_{13} is the ratio of marginal likelihoods of data y over m_1 and m_3, respectively. That is,

$$\mathrm{BF}_{13} = \frac{p(y \mid m_1)}{p(y \mid m_3)} . \tag{2}$$

We can equivalently write

$$\begin{aligned}
\mathrm{BF}_{13} &\overset{(2)}{=} \frac{p(y \mid m_1)}{p(y \mid m_3)} \\
&= \frac{p(y \mid m_1)}{p(y \mid m_3)} \cdot \frac{p(y \mid m_2)}{p(y \mid m_2)} \\
&= \frac{p(y \mid m_1)}{p(y \mid m_2)} \cdot \frac{p(y \mid m_2)}{p(y \mid m_3)} \\
&\overset{(2)}{=} \mathrm{BF}_{12} \cdot \mathrm{BF}_{23} ,
\end{aligned} \tag{3}$$

which completes the proof of (1).

Sources:
- original work

Metadata: ID: P163 | shortcut: bf-trans | author: tomfaulkenberry | date: 2020-09-07, 12:00.

3.4.3 Computation using Savage-Dickey Density Ratio

Theorem: Consider two competing models (→ Definition I/5.1.1) on data y containing parameters δ and φ, namely $m_0 : \delta = \delta_0, \varphi$ and $m_1 : \delta, \varphi$. In this context, we say that δ is a parameter of interest, φ is a nuisance parameter (i.e., common to both models), and m_0 is a sharp point hypothesis nested within m_1. Suppose further that the prior for the nuisance parameter φ in m_0 is equal to the prior for φ in m_1 after conditioning on the restriction – that is, $p(\varphi \mid m_0) = p(\varphi \mid \delta = \delta_0, m_1)$. Then the Bayes factor (→ Definition IV/3.4.1) for m_0 over m_1 can be computed as:

$$\mathrm{BF}_{01} = \frac{p(\delta = \delta_0 \mid y, m_1)}{p(\delta = \delta_0 \mid m_1)} . \tag{1}$$

Proof: By definition (→ Definition IV/3.4.1), the Bayes factor BF_{01} is the ratio of marginal likelihoods of data y over m_0 and m_1, respectively. That is,

$$\mathrm{BF}_{01} = \frac{p(y \mid m_0)}{p(y \mid m_1)} . \tag{2}$$

The key idea in the proof is that we can use a "change of variables" technique to express BF_{01} entirely in terms of the "encompassing" model m_1. This proceeds by first unpacking the marginal likelihood (→ Definition I/5.1.9) for m_0 over the nuisance parameter φ and then using the fact that m_0 is a sharp hypothesis nested within m_1 to rewrite everything in terms of m_1. Specifically,

$$\begin{aligned} p(y \mid m_0) &= \int p(y \mid \varphi, m_0)\, p(\varphi \mid m_0)\, \mathrm{d}\varphi \\ &= \int p(y \mid \varphi, \delta = \delta_0, m_1)\, p(\varphi \mid \delta = \delta_0, m_1)\, \mathrm{d}\varphi \\ &= p(y \mid \delta = \delta_0, m_1) . \end{aligned} \tag{3}$$

By Bayes' theorem (→ Proof I/5.3.1), we can rewrite this last line as

$$p(y \mid \delta = \delta_0, m_1) = \frac{p(\delta = \delta_0 \mid y, m_1)\, p(y \mid m_1)}{p(\delta = \delta_0 \mid m_1)} . \tag{4}$$

Thus we have

$$\begin{aligned} \mathrm{BF}_{01} &\overset{(2)}{=} \frac{p(y \mid m_0)}{p(y \mid m_1)} \\ &= p(y \mid m_0) \cdot \frac{1}{p(y \mid m_1)} \\ &\overset{(3)}{=} p(y \mid \delta = \delta_0, m_1) \cdot \frac{1}{p(y \mid m_1)} \\ &\overset{(4)}{=} \frac{p(\delta = \delta_0 \mid y, m_1)\, p(y \mid m_1)}{p(\delta = \delta_0 \mid m_1)} \cdot \frac{1}{p(y \mid m_1)} \\ &= \frac{p(\delta = \delta_0 \mid y, m_1)}{p(\delta = \delta_0 \mid m_1)} , \end{aligned} \tag{5}$$

which completes the proof of (1).

Sources:

- Faulkenberry, Thomas J. (2019): "A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors"; in: *Communications for Statistical Applications and Methods*, vol. 26, no. 2, pp. 217-238; URL: https://dx.doi.org/10.29220/CSAM.2019.26.2.217; DOI: 10.29220/CSAM.2019.26.2.217.
- Penny, W.D. and Ridgway, G.R. (2013): "Efficient Posterior Probability Mapping Using Savage-Dickey Ratios"; in: *PLoS ONE*, vol. 8, iss. 3, art. e59655, eq. 16; URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059655; DOI: 10.1371/journal.pone.0059655.

Metadata: ID: P156 | shortcut: bf-sddr | author: tomfaulkenberry | date: 2020-08-26, 12:00.

3.4.4 Computation using Encompassing Prior Method

Theorem: Consider two models m_1 and m_e, where m_1 is nested within an encompassing model (→ Definition IV/3.4.5) m_e via an inequality constraint on some parameter θ, and θ is unconstrained under m_e. Then, the Bayes factor (→ Definition IV/3.4.1) is

$$\mathrm{BF}_{1e} = \frac{c}{d} = \frac{1/d}{1/c} \tag{1}$$

where $1/d$ and $1/c$ represent the proportions of the posterior and prior of the encompassing model, respectively, that are in agreement with the inequality constraint imposed by the nested model m_1.

Proof: Consider first that for any model m_1 on data y with parameter θ, Bayes' theorem (→ Proof I/5.3.1) implies

$$p(\theta \mid y, m_1) = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)}{p(y \mid m_1)} . \tag{2}$$

Rearranging equation (2) allows us to write the marginal likelihood (→ Definition I/5.1.9) for y under m_1 as

$$p(y \mid m_1) = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)}{p(\theta \mid y, m_1)} . \tag{3}$$

Taking the ratio of the marginal likelihoods for m_1 and the encompassing model (→ Definition IV/3.4.5) m_e yields the following Bayes factor (→ Definition IV/3.4.1):

$$\mathrm{BF}_{1e} = \frac{p(y \mid \theta, m_1) \cdot p(\theta \mid m_1)/p(\theta \mid y, m_1)}{p(y \mid \theta, m_e) \cdot p(\theta \mid m_e)/p(\theta \mid y, m_e)} . \tag{4}$$

Now, both the constrained model m_1 and the encompassing model (→ Definition IV/3.4.5) m_e contain the same parameter vector θ. Choose a specific value of θ, say θ', that exists in the support of both models m_1 and m_e (we can do this, because m_1 is nested within m_e). Then, for this parameter value θ', we have $p(y \mid \theta', m_1) = p(y \mid \theta', m_e)$, so the expression for the Bayes factor in equation (4) reduces to an expression involving only the priors and posteriors for θ' under m_1 and m_e:

$$\mathrm{BF}_{1e} = \frac{p(\theta' \mid m_1)/p(\theta' \mid y, m_1)}{p(\theta' \mid m_e)/p(\theta' \mid y, m_e)} . \tag{5}$$

Because m_1 is nested within m_e via an inequality constraint, the prior $p(\theta' \mid m_1)$ is simply a truncation of the encompassing prior $p(\theta' \mid m_e)$. Thus, we can express $p(\theta' \mid m_1)$ in terms of the encompassing prior $p(\theta' \mid m_e)$ by multiplying the encompassing prior by an indicator function over m_1 and then normalizing the resulting product. That is,

$$\begin{split} p(\theta' \mid m_1) &= \frac{p(\theta' \mid m_e) \cdot I_{\theta' \in m_1}}{\int p(\theta' \mid m_e) \cdot I_{\theta' \in m_1} \, \mathrm{d}\theta'} \\ &= \left(\frac{I_{\theta' \in m_1}}{\int p(\theta' \mid m_e) \cdot I_{\theta' \in m_1} \, \mathrm{d}\theta'} \right) \cdot p(\theta' \mid m_e) , \end{split} \tag{6}$$

where $I_{\theta' \in m_1}$ is an indicator function. For parameters $\theta' \in m_1$, this indicator function is identically equal to 1, so the expression in parentheses reduces to a constant, say c, allowing us to write the prior as

$$p(\theta' \mid m_1) = c \cdot p(\theta' \mid m_e). \tag{7}$$

By similar reasoning, we can write the posterior as

$$p(\theta' \mid y, m_1) = \left(\frac{I_{\theta' \in m_1}}{\int p(\theta' \mid y, m_e) \cdot I_{\theta' \in m_1} \, \mathrm{d}\theta'} \right) \cdot p(\theta' \mid y, m_e) = d \cdot p(\theta' \mid y, m_e). \tag{8}$$

Plugging (7) and (8) into (5), this gives us

$$\mathrm{BF}_{1e} = \frac{c \cdot p(\theta' \mid m_e)/d \cdot p(\theta' \mid y, m_e)}{p(\theta' \mid m_e)/p(\theta' \mid y, m_e)} = \frac{c}{d} = \frac{1/d}{1/c}, \tag{9}$$

which completes the proof. Note that by definition, $1/d$ represents the proportion of the posterior distribution for θ under the encompassing model (→ Definition IV/3.4.5) m_e that agrees with the constraints imposed by m_1. Similarly, $1/c$ represents the proportion of the prior distribution for θ under the encompassing model (→ Definition IV/3.4.5) m_e that agrees with the constraints imposed by m_1.

Sources:
- Klugkist, I., Kato, B., and Hoijtink, H. (2005): "Bayesian model selection using encompassing priors"; in: *Statistica Neerlandica*, vol. 59, no. 1., pp. 57-69; URL: https://dx.doi.org/10.1111/j.1467-9574.2005.00279.x; DOI: 10.1111/j.1467-9574.2005.00279.x.
- Faulkenberry, Thomas J. (2019): "A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors"; in: *Communications for Statistical Applications and Methods*, vol. 26, no. 2, pp. 217-238; URL: https://dx.doi.org/10.29220/CSAM.2019.26.2.217; DOI: 10.29220/CSAM.2019.26.2.217.

Metadata: ID: P157 | shortcut: bf-ep | author: tomfaulkenberry | date: 2020-09-02, 12:00.

3.4.5 Encompassing model

Definition: Consider a family f of generative models (→ Definition I/5.1.1) m on data y, where each $m \in f$ is defined by placing an inequality constraint on model parameter(s) θ (e.g., $m : \theta > 0$). Then the encompassing model m_e is constructed such that each m is nested within m_e and all inequality constraints on the parameter(s) θ are removed.

Sources:
- Klugkist, I., Kato, B., and Hoijtink, H. (2005): "Bayesian model selection using encompassing priors"; in: *Statistica Neerlandica*, vol. 59, no. 1, pp. 57-69; URL: https://dx.doi.org/10.1111/j.1467-9574.2005.00279.x; DOI: 10.1111/j.1467-9574.2005.00279.x.

Metadata: ID: D93 | shortcut: encm | author: tomfaulkenberry | date: 2020-09-02, 12:00.

3.5 Posterior model probability

3.5.1 Definition

Definition: Let $m_1, \ldots, m_M$ be M statistical models (→ Definition I/5.1.4) with model evidences (→ Definition I/5.1.9) $p(y|m_1), \ldots, p(y|m_M)$ and prior probabilities (→ Definition I/5.1.3) $p(m_1), \ldots, p(m_M)$.

Then, the conditional probability (→ Definition I/1.2.4) of model m_i, given the data y, is called the posterior probability (→ Definition I/5.1.7) of model m_i:

$$\mathrm{PP}(m_i) = p(m_i|y) \; . \tag{1}$$

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. 23; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: D87 | shortcut: pmp | author: JoramSoch | date: 2020-07-28, 03:30.

3.5.2 Derivation

Theorem: Let there be a set of generative models (→ Definition I/5.1.1) $m_1, \ldots, m_M$ with model evidences (→ Definition I/5.1.9) $p(y|m_1), \ldots, p(y|m_M)$ and prior probabilities (→ Definition I/5.1.3) $p(m_1), \ldots, p(m_M)$. Then, the posterior probability (→ Definition IV/3.5.1) of model m_i is given by

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{\sum_{j=1}^{M} p(y|m_j)\, p(m_j)}, \; i = 1, \ldots, M \; . \tag{1}$$

Proof: From Bayes' theorem (→ Proof I/5.3.1), the posterior model probability (→ Definition IV/3.5.1) of the i-th model can be derived as

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{p(y)} \; . \tag{2}$$

Using the law of marginal probability (→ Definition I/1.2.3), the denominator can be rewritten, such that

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{\sum_{j=1}^{M} p(y, m_j)} \; . \tag{3}$$

Finally, using the law of conditional probability (→ Definition I/1.2.4), we have

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{\sum_{j=1}^{M} p(y|m_j)\, p(m_j)} \; . \tag{4}$$

Sources:
- original work

Metadata: ID: P139 | shortcut: pmp-der | author: JoramSoch | date: 2020-07-28, 03:58.

3.5.3 Calculation from Bayes factors

Theorem: Let $m_0, m_1, \ldots, m_M$ be $M+1$ statistical models with model evidences (→ Definition IV/3.1.1) $p(y|m_0), p(y|m_1), \ldots, p(y|m_M)$. Then, the posterior model probabilities (→ Definition IV/3.5.1) of the models $m_1, \ldots, m_M$ are given by

$$p(m_i|y) = \frac{\mathrm{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \mathrm{BF}_{j,0} \cdot \alpha_j}, \quad i = 1, \ldots, M \; , \tag{1}$$

where $\mathrm{BF}_{i,0}$ is the Bayes factor (→ Definition IV/3.4.1) comparing model m_i with m_0 and α_i is the prior (→ Definition I/5.1.3) odds ratio (→ Definition "odds") of model m_i against m_0.

Proof: Define the Bayes factor (→ Definition IV/3.4.1) for m_i

$$\mathrm{BF}_{i,0} = \frac{p(y|m_i)}{p(y|m_0)} \tag{2}$$

and prior odds ratio of m_i against m_0

$$\alpha_i = \frac{p(m_i)}{p(m_0)} \; . \tag{3}$$

The posterior model probability (→ Proof IV/3.5.2) of m_i is given by

$$p(m_i|y) = \frac{p(y|m_i) \cdot p(m_i)}{\sum_{j=1}^M p(y|m_j) \cdot p(m_j)} \; . \tag{4}$$

Now applying (2) and (3) to (4), we have

$$\begin{split} p(m_i|y) &= \frac{\mathrm{BF}_{i,0} \, p(y|m_0) \cdot \alpha_i \, p(m_0)}{\sum_{j=1}^M \mathrm{BF}_{j,0} \, p(y|m_0) \cdot \alpha_j \, p(m_0)} \\ &= \frac{[p(y|m_0) \, p(m_0)] \, \mathrm{BF}_{i,0} \cdot \alpha_i}{[p(y|m_0) \, p(m_0)] \sum_{j=1}^M \mathrm{BF}_{j,0} \cdot \alpha_j} \; , \end{split} \tag{5}$$

such that

$$p(m_i|y) = \frac{\mathrm{BF}_{i,0} \cdot \alpha_i}{\sum_{j=1}^M \mathrm{BF}_{j,0} \cdot \alpha_j} \; . \tag{6}$$

Sources:

- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999): "Bayesian Model Averaging: A Tutorial"; in: *Statistical Science*, vol. 14, no. 4, pp. 382–417, eq. 9; URL: https://projecteuclid.org/euclid.ss/1009212519; DOI: 10.1214/ss/1009212519.

Metadata: ID: P74 | shortcut: pmp-bf | author: JoramSoch | date: 2020-03-03, 13:13.

3.5.4 Calculation from log Bayes factor

Theorem: Let m_1 and m_2 be two statistical models with the log Bayes factor (→ Definition IV/3.3.1) LBF_{12} in favor of model m_1 and against model m_2. Then, if both models are equally likely apriori (→ Definition I/5.1.3), the posterior model probability (→ Definition IV/3.5.1) of m_1 is

$$p(m_1|y) = \frac{\exp(\mathrm{LBF}_{12})}{\exp(\mathrm{LBF}_{12}) + 1} \; . \tag{1}$$

Proof: From Bayes' rule (→ Proof I/5.3.2), the posterior odds ratio (→ Definition "odds") is

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \cdot \frac{p(m_1)}{p(m_2)} \; . \tag{2}$$

When both models are equally likely apriori (→ Definition I/5.1.3), the prior odds ratio (→ Definition "odds") is one, such that

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \; . \tag{3}$$

Now the right-hand side corresponds to the Bayes factor (→ Definition IV/3.4.1), therefore

$$\frac{p(m_1|y)}{p(m_2|y)} = \mathrm{BF}_{12} \; . \tag{4}$$

Because the two posterior model probabilities (→ Definition IV/3.5.1) add up to 1, we have

$$\frac{p(m_1|y)}{1 - p(m_1|y)} = \mathrm{BF}_{12} \; . \tag{5}$$

Now rearranging for the posterior probability (→ Definition IV/3.5.1), this gives

$$p(m_1|y) = \frac{\mathrm{BF}_{12}}{\mathrm{BF}_{12} + 1} \; . \tag{6}$$

Because the log Bayes factor is the logarithm of the Bayes factor (→ Definition IV/3.3.1), we finally have

$$p(m_1|y) = \frac{\exp(\mathrm{LBF}_{12})}{\exp(\mathrm{LBF}_{12}) + 1} \; . \tag{7}$$

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. 21; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P73 | shortcut: pmp-lbf | author: JoramSoch | date: 2020-03-03, 12:27.

3.5.5 Calculation from log model evidences

Theorem: Let $m_1, \ldots, m_M$ be M statistical models with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1), \ldots, \mathrm{LME}(m_M)$. Then, the posterior model probabilities (→ Definition IV/3.5.1) are given by:

$$p(m_i|y) = \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)}, \quad i = 1, \ldots, M \; , \tag{1}$$

where $p(m_i)$ are prior (→ Definition I/5.1.3) model probabilities.

Proof: The posterior model probability (→ Proof IV/3.5.2) can be derived as

$$p(m_i|y) = \frac{p(y|m_i)\, p(m_i)}{\sum_{j=1}^{M} p(y|m_j)\, p(m_j)} \; . \tag{2}$$

The definition of the log model evidence (→ Definition IV/3.1.1)

$$\mathrm{LME}(m) = \log p(y|m) \tag{3}$$

can be exponentiated to then read

$$\exp\left[\mathrm{LME}(m)\right] = p(y|m) \tag{4}$$

and applying (4) to (2), we finally have:

$$p(m_i|y) = \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)} \; . \tag{5}$$

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. 23; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P66 | shortcut: pmp-lme | author: JoramSoch | date: 2020-02-27, 21:33.

3.6 Bayesian model averaging

3.6.1 Definition

Definition: Let $m_1, \ldots, m_M$ be M statistical models (→ Definition I/5.1.4) with posterior model probabilities (→ Definition IV/3.5.1) $p(m_1|y), \ldots, p(m_M|y)$ and posterior distributions (→ Definition I/5.1.7) $p(\theta|y, m_1), \ldots, p(\theta|y, m_M)$. Then, Bayesian model averaging (BMA) consists in finding the marginal (→ Definition I/1.3.3) posterior (→ Definition I/5.1.7) density (→ Definition I/1.4.4), conditional (→ Definition I/1.2.4) on the measured data y, but unconditional (→ Definition I/1.2.3) on the modelling approach m:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|y, m_i) \cdot p(m_i|y) \; . \tag{1}$$

Sources:

- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999): "Bayesian Model Averaging: A Tutorial"; in: *Statistical Science*, vol. 14, no. 4, pp. 382–417, eq. 1; URL: https://projecteuclid.org/euclid.ss/1009212519; DOI: 10.1214/ss/1009212519.

Metadata: ID: D89 | shortcut: bma | author: JoramSoch | date: 2020-08-03, 21:34.

3.6.2 Derivation

Theorem: Let $m_1, \ldots, m_M$ be M statistical models (→ Definition I/5.1.4) with posterior model probabilities (→ Definition IV/3.5.1) $p(m_1|y), \ldots, p(m_M|y)$ and posterior distributions (→ Definition I/5.1.7) $p(\theta|y, m_1), \ldots, p(\theta|y, m_M)$. Then, the marginal (→ Definition I/1.3.3) posterior (→ Definition I/5.1.7) density (→ Definition I/1.4.4), conditional (→ Definition I/1.2.4) on the measured data y, but unconditional (→ Definition I/1.2.3) on the modelling approach m, is given by:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|y, m_i) \cdot p(m_i|y) \; . \tag{1}$$

Proof: Using the law of marginal probability (→ Definition I/1.2.3), the probability distribution of the shared parameters θ conditional (→ Definition I/1.2.4) on the measured data y can be obtained by marginalizing (→ Definition I/1.2.3) over the discrete random variable (→ Definition I/1.1.3) model m:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta, m_i|y) \; . \tag{2}$$

Using the law of the conditional probability (→ Definition I/1.2.4), the summand can be expanded to give

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|y, m_i) \cdot p(m_i|y) \tag{3}$$

where $p(\theta|y, m_i)$ is the posterior distribution (→ Definition I/5.1.7) of the i-th model and $p(m_i|y)$ happens to be the posterior probability (→ Definition IV/3.5.1) of the i-th model.

Sources:
- original work

Metadata: ID: P143 | shortcut: bma-der | author: JoramSoch | date: 2020-08-03, 22:05.

3.6.3 Calculation from log model evidences

Theorem: Let $m_1, \ldots, m_M$ be M statistical models (→ Definition I/5.1.4) describing the same measured data y with log model evidences (→ Definition IV/3.1.1) $\mathrm{LME}(m_1), \ldots, \mathrm{LME}(m_M)$ and shared model parameters θ. Then, Bayesian model averaging (→ Definition IV/3.6.1) determines the following posterior distribution over θ:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|m_i, y) \cdot \frac{\exp[\mathrm{LME}(m_i)]\, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)]\, p(m_j)} \; , \tag{1}$$

where $p(\theta|m_i, y)$ is the posterior distributions over θ obtained using m_i.

Proof: According to the law of marginal probability (→ Definition I/1.2.3), the probability of the shared parameters θ conditional on the measured data y can be obtained (→ Proof IV/3.6.2) by marginalizing over the discrete variable model m:

$$p(\theta|y) = \sum_{i=1}^{M} p(\theta|m_i, y) \cdot p(m_i|y) \; , \tag{2}$$

where $p(m_i|y)$ is the posterior probability (→ Definition IV/3.5.1) of the i-th model. One can express posterior model probabilities in terms of log model evidences (→ Proof IV/3.5.5) as

$$p(m_i|y) = \frac{\exp[\mathrm{LME}(m_i)] \, p(m_i)}{\sum_{j=1}^{M} \exp[\mathrm{LME}(m_j)] \, p(m_j)} \tag{3}$$

and by plugging (3) into (2), one arrives at (1).

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. 25; URL: https://www.sciencedirect.com/science/article/pii/S0165027018301468; DOI: 10.1016/j.jneumeth.2018.05.017.

Metadata: ID: P67 | shortcut: bma-lme | author: JoramSoch | date: 2020-02-27, 21:58.