Probability–the Science of Uncertainty and Data

by Fabián Kozynski

Probability

Probability models and axioms

Definition (Sample space) A sample space Ω is the set of all possible outcomes. The set's elements must be mutually exclusive, collectively exhaustive and at the right granularity.

Definition (Event) An event is a subset of the sample space. Probability is assigned to events.

Definition (Probability axioms) A probability law $\mathbb{P}$ assigns probabilities to events and satisfies the following axioms:

Nonnegativity $\mathbb{P}(A) \geq 0$ for all events A.

Normalization $\mathbb{P}(\Omega) = 1$.

(Countable) additivity For every sequence of events $A_1, A_2, \dots$ such that $A_i \cap A_j = \varnothing$: $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.

Corollaries (Consequences of the axioms)

- $\mathbb{P}(\varnothing) = 0$.
- For any finite collection of disjoint events $A_1, \dots, A_n$, $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$.
- $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.
- $\mathbb{P}(A) \leq 1$.
- If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Example (Discrete uniform law) Assume Ω is finite and consists of n equally likely elements. Also, assume that $A \subset \Omega$ with k elements. Then $\mathbb{P}(A) = \frac{k}{n}$.

Conditioning and Bayes' rule

Definition (Conditional probability) Given that event B has occurred and that $P(B) > 0$, the probability that A occurs is

$$\mathbb{P}(A|B) \triangleq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark (Conditional probabilities properties) They are the same as ordinary probabilities. Assuming $\mathbb{P}(B) > 0$:

- $\mathbb{P}(A|B) \geq 0$.
- $\mathbb{P}(\Omega|B) = 1$
- $\mathbb{P}(B|B) = 1$.
- If $A \cap C = \varnothing$, $\mathbb{P}(A \cup C|B) = \mathbb{P}(A|B) + \mathbb{P}(C|B)$.

Proposition (Multiplication rule)

$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$.

Theorem (Total probability theorem) Given a partition $\{A_1, A_2, \dots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and for every event B, we have

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B|A_i).$$

Theorem (Bayes' rule) Given a partition $\{A_1, A_2, \dots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and if $\mathbb{P}(A_i) > 0$ for all i, then for every event B, the conditional probabilities $\mathbb{P}(A_i|B)$ can be obtained from the conditional probabilities $\mathbb{P}(B|A_i)$ and the initial probabilities $\mathbb{P}(A_i)$ as follows:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}.$$

Independence

Definition (Independence of events) Two events are independent if occurrence of one provides no information about the other. We say that A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Equivalently, as long as $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

$$\mathbb{P}(B|A) = \mathbb{P}(B) \qquad \mathbb{P}(A|B) = \mathbb{P}(A).$$

Remarks

- The definition of independence is symmetric with respect to A and B.
- The product definition applies even if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

Corollary If A and B are independent, then A and B^c are independent. Similarly for A^c and B, or for A^c and B^c.

Definition (Conditional independence) We say that A and B are independent conditioned on C, where $\mathbb{P}(C) > 0$, if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C).$$

Definition (Independence of a collection of events) We say that events $A_1, A_2, \dots, A_n$ are independent if for every collection of distinct indices $i_1, i_2, \dots, i_k$, we have

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Counting

This section deals with finite sets with uniform probability law. In this case, to calculate $\mathbb{P}(A)$, we need to count the number of elements in A and in Ω.

Remark (Basic counting principle) For a selection that can be done in r stages, with n_i choices at each stage i, the number of possible selections is $n_1 \cdot n_2 \cdots n_r$.

Definition (Permutations) The number of permutations (orderings) of n different elements is

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$

Definition (Combinations) Given a set of n elements, the number of subsets with exactly k elements is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Definition (Partitions) We are given an n–element set and nonnegative integers $n_1, n_2, \dots, n_r$, whose sum is equal to n. The number of partitions of the set into r disjoint subsets, with the i^{th} subset containing exactly n_i elements, is equal to

$$\binom{n}{n_1, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

Remark This is the same as counting how to assign n distinct elements to r people, giving each person i exactly n_i elements.

Discrete random variables

Probability mass function and expectation

Definition (Random variable) A random variable X is a function of the sample space Ω into the real numbers (or $\mathbb{R}^n$). Its range can be discrete or continuous.

Definition (Probability mass funtion (PMF)) The probability law of a discrete random variable X is called its PMF. It is defined as

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}).$$

Properties

$p_X(x) \geq 0,\ \forall x.$

$\sum_x p_X(x) = 1.$

Example (Bernoulli random variable) A Bernoulli random variable X with parameter $0 \leq p \leq 1$ ($X \sim \mathrm{Ber}(p)$) takes the following values:

$$X = \begin{cases} 1 & \text{w.p. } p, \\ 0 & \text{w.p. } 1-p. \end{cases}$$

An indicator random variable of an event ($I_A = 1$ if A occurs) is an example of a Bernoulli random variable.

Example (Discrete uniform random variable) A Discrete uniform random variable X between a and b with $a \leq b$ ($X \sim \mathrm{Uni}[a,b]$) takes any of the values in $\{a, a+1, \dots, b\}$ with probability $\frac{1}{b-a+1}$.

Example (Binomial random variable) A Binomial random variable X with parameters n (natural number) and $0 \leq p \leq 1$ ($X \sim \mathrm{Bin}(n,p)$) takes values in the set $\{0, 1, \dots, n\}$ with probabilities $p_X(i) = \binom{n}{i} p^i (1-p)^{n-i}$.

It represents the number of successes in n independent trials where each trial has a probability of success p. Therefore, it can also be seen as the sum of n independent Bernoulli random variables, each with parameter p.

Example (Geometric random variable) A Geometric random variable X with parameter $0 \leq p \leq 1$ ($X \sim \mathrm{Geo}(p)$) takes values in the set $\{1, 2, \dots\}$ with probabilities $p_X(i) = (1-p)^{i-1}p$.

It represents the number of independent trials until (and including) the first success, when the probability of success in each trial is p.

Definition (Expectation/mean of a random variable) The expectation of a discrete random variable is defined as

$$\mathbb{E}[X] \triangleq \sum_x x p_X(x).$$

assuming $\sum_x |x| p_X(x) < \infty$.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- If $X = c$ then $\mathbb{E}[X] = c$.

Example Expected value of know r.v.

- If $X \sim \mathrm{Ber}(p)$ then $\mathbb{E}[X] = p$.
- If $X = I_A$ then $\mathbb{E}[X] = \mathbb{P}(A)$.
- If $X \sim \mathrm{Uni}[a,b]$ then $\mathbb{E}[X] = \frac{a+b}{2}$.
- If $X \sim \mathrm{Bin}(n,p)$ then $\mathbb{E}[X] = np$.
- If $X \sim \mathrm{Geo}(p)$ then $\mathbb{E}[X] = \frac{1}{p}$.

Theorem (Expected value rule) Given a random variable X and a function $g : \mathbb{R} \to \mathbb{R}$, we construct the random variable $Y = g(X)$. Then

$$\sum_y y p_Y(y) = \mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_x g(x) p_X(x).$$

Remark (PMF of $Y = g(X)$) The PMF of $Y = g(X)$ is $p_Y(y) = \sum_{x : g(x) = y} p_X(x)$.

Remark In general $g(\mathbb{E}[X]) \neq \mathbb{E}[g(X)]$. They are equal if $g(x) = ax + b$.

Variance, conditioning on an event, multiple r.v.

Definition (Variance of a random variable) Given a random variable X with $\mu = \mathbb{E}[X]$, its variance is a measure of the spread of the random variable and is defined as

$$\mathrm{Var}(X) \triangleq \mathbb{E}\left[(X - \mu)^2\right] = \sum_x (x - \mu)^2 p_X(x).$$

Definition (Standard deviation)

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Properties (Properties of the variance)

- $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$, for all $a \in \mathbb{R}$.
- $\mathrm{Var}(X + b) = \mathrm{Var}(X)$, for all $b \in \mathbb{R}$.
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Example (Variance of known r.v.)

- If $X \sim \mathrm{Ber}(p)$, then $\mathrm{Var}(X) = p(1-p)$.
- If $X \sim \mathrm{Uni}[a,b]$, then $\mathrm{Var}(X) = \frac{(b-a)(b-a+2)}{12}$.
- If $X \sim \mathrm{Bin}(n,p)$, then $\mathrm{Var}(X) = np(1-p)$.
- If $X \sim \mathrm{Geo}(p)$, then $\mathrm{Var}(X) = \frac{1-p}{p^2}$

Proposition (Conditional PMF and expectation, given an event) Given the event A, with $\mathbb{P}(A) > 0$, we have the following

- $p_{X|A}(x) = \mathbb{P}(X = x|A)$.
- If A is a subset of the range of X, then: $p_{X|A}(x) \triangleq p_{X|\{X \in A\}}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} p_X(x), & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$
- $\sum_x p_{X|A}(x) = 1$.
- $\mathbb{E}[X|A] = \sum_x x p_{X|A}(x)$.
- $\mathbb{E}[g(X)|A] = \sum_x g(x) p_{X|A}(x)$.

Proposition (Total expectation rule) Given a partition of disjoint events $A_1, \dots, A_n$ such that $\sum_i \mathbb{P}(A_i) = 1$, and $\mathbb{P}(A_i) > 0$,

$$\mathbb{E}[X] = \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \cdots + \mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Memorylessness of the geometric random variable) When we condition a geometric random variable X on the event $X > n$ we have memorylessness, meaning that the "remaining time" $X - n$, given that $X > n$, is also geometric with the same parameter. Formally,

$$p_{X-n|X>n}(i) = p_X(i).$$

Definition (Joint PMF) The joint PMF of random variables $X_1, X_2, \dots, X_n$ is $p_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$.

Properties (Properties of joint PMF)

- $\sum_{x_1} \cdots \sum_{x_n} p_{X_1, \dots, X_n}(x_1, \dots, x_n) = 1$.
- $p_{X_1}(x_1) = \sum_{x_2} \cdots \sum_{x_n} p_{X_1, \dots, X_n}(x_1, x_2, \dots, x_n)$.
- $p_{X_2, \dots, X_n}(x_2, \dots, x_n) = \sum_{x_1} p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$.

Definition (Functions of multiple r.v.) If $Z = g(X_1, \dots, X_n)$, where $g : \mathbb{R}^n \to \mathbb{R}$, then $p_Z(z) = \mathbb{P}(g(X_1, \dots, X_n) = z)$.

Proposition (Expected value rule for multiple r.v.) Given $g : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}[g(X_1, \dots, X_n)] = \sum_{x_1, \dots, x_n} g(x_1, \dots, x_n) p_{X_1, \dots, X_n}(x_1, \dots, x_n).$$

Properties (Linearity of expectations)

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
- $\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]$.

Conditioning on a random variable, independence

Definition (Conditional PMF given another random variable) Given discrete random variables X, Y and y such that $p_Y(y) > 0$ we define

$$p_{X|Y}(x|y) \triangleq \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

Proposition (Multiplication rule) Given jointly discrete random variables X, Y, and whenever the conditional probabilities are defined,

$$p_{X,Y}(x,y) = p_X(x) p_{Y|X}(y|x) = p_Y(y) p_{X|Y}(x|y).$$

Definition (Conditional expectation) Given discrete random variables X, Y and y such that $p_Y(y) > 0$ we define

$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \sum_x g(x) p_{X|Y}(x|y).$$

Theorem (Total probability and expectation theorems) If $p_Y(y) > 0$, then

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x|y),$$

$$\mathbb{E}[X] = \sum_y p_Y(y) \mathbb{E}[X|Y = y].$$

Definition (Independence of a random variable and an event) A discrete random variable X and an event A are independent if $\mathbb{P}(X = x \text{ and } A) = p_X(x)\mathbb{P}(A)$, for all x.

Definition (Independence of two random variables) Two discrete random variables X and Y are independent if $p_{X,Y}(x,y) = p_X(x) p_Y(y)$ for all x, y.

Remark (Independence of a collection of random variables) A collection $X_1, X_2, \dots, X_n$ of random variables are independent if

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n),\ \forall x_1, \dots, x_n.$$

Remark (Independence and expectation) In general, $\mathbb{E}[g(X,Y)] \neq g(\mathbb{E}[X], \mathbb{E}[Y])$. An exception is for linear functions: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Proposition (Expectation of product of independent r.v.) If X and Y are discrete independent random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables) IF X and Y are discrete independent random variables,

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

Continuous random variables

PDF, Expectation, Variance, CDF

Definition (Probability density function (PDF)) A probability density function of a r.v. X is a non-negative real valued function f_X that satisfies the following

- $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$.
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x$ for some random variable X.

Definition (Continuous random variable) A random variable X is continuous if its probability law can be described by a PDF f_X.

Remark Continuous random variables satisfy:

- For small $\delta > 0$, $\mathbb{P}(a \leq X \leq a + \delta) \approx f_X(a)\delta$.
- $\mathbb{P}(X = a) = 0,\ \forall a \in \mathbb{R}$.

Definition (Expectation of a continuous random variable) The expectation of a continuous random variable is

$$\mathbb{E}[X] \triangleq \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x.$$

assuming $\int_{-\infty}^{\infty} |x| f_X(x)\mathrm{d}x < \infty$.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\mathrm{d}x$.
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.

Definition (Variance of a continuous random variable) Given a continuous random variable X with $\mu = \mathbb{E}[X]$, its variance is

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mu)^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\mathrm{d}x.$$

It has the same properties as the variance of a discrete random variable.

Example (Uniform continuous random variable) A Uniform continuous random variable X between a and b, with $a < b$, ($X \sim \mathrm{Uni}(a,b)$) has PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

We have $\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Example (Exponential random variable) An Exponential random variable X with parameter $\lambda > 0$ ($X \sim Exp(\lambda)$) has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We have $E[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

Definition (Cumulative Distribution Function (CDF)) The CDF of a random variable X is $F_X(x) = \mathbb{P}(X \leq x)$.
In particular, for a continuous random variable, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(x) \mathrm{d}x,$$

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

Properties (Properties of CDF)

- If $y \geq x$, then $F_X(y) \geq F_X(x)$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.

Definition (Normal/Gaussian random variable) A Normal random variable X with mean μ and variance $\sigma^2 > 0$ ($X \sim \mathcal{N}(\mu, \sigma^2)$) has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

We have $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Remark (Standard Normal) The standard Normal is $\mathcal{N}(0,1)$.

Proposition (Linearity of Gaussians) Given $X \sim \mathcal{N}(\mu, \sigma^2)$, and if $a \neq 0$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
Using this $Y = \frac{X - \mu}{\sigma}$ is a standard gaussian.

Conditioning on an event, and multiple continuous r.v.

Definition (Conditional PDF given an event) Given a continuous random variable X and event A with $P(A) > 0$, we define the conditional PDF as the function that satisfies

$$\mathbb{P}(X \in B | A) = \int_B f_{X|A}(x) \mathrm{d}x.$$

Definition (Conditional PDF given $X \in A$) Given a continuous random variable X and an $A \subset \mathbb{R}$, with $P(A) > 0$:

$$f_{X|X \in A}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} f_X(x), & x \in A, \\ 0, & x \notin A. \end{cases}$$

Definition (Conditional expectation) Given a continuous random variable X and an event A, with $P(A) > 0$:

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} f_{X|A}(x) \mathrm{d}x.$$

Definition (Memorylessness of the exponential random variable)
When we condition an exponential random variable X on the event $X > t$ we have memorylessness, meaning that the "remaining time" $X - t$ given that $X > t$ is also geometric with the same parameter i.e.,

$$\mathbb{P}(X - t > x | X > t) = \mathbb{P}(X > x).$$

Theorem (Total probability and expectation theorems) Given a partition of the space into disjoint events $A_1, A_2, \ldots, A_n$ such that $\sum_i \mathbb{P}(A_i) = 1$ we have the following:

$$F_X(x) = \mathbb{P}(A_1) F_{X|A_1}(x) + \cdots + \mathbb{P}(A_n) F_{X|A_n}(x),$$

$$f_X(x) = \mathbb{P}(A_1) f_{X|A_1}(x) + \cdots + \mathbb{P}(A_n) f_{X|A_n}(x),$$

$$\mathbb{E}[X] = \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \cdots + \mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Jointly continuous random variables) A pair (collection) of random variables is jointly continuous if there exists a joint PDF $f_{X,Y}$ that describes them, that is, for every set $B \subset \mathbb{R}^n$

$$\mathbb{P}((X,Y) \in B) = \iint_B f_{X,Y}(x,y) \mathrm{d}x\mathrm{d}y.$$

Properties (Properties of joint PDFs)

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathrm{d}y$.
- $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{X,Y}(u,v) \mathrm{d}v \right] \mathrm{d}u$.
- $f_{X,Y}(x) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \, \partial y}$.

Example (Uniform joint PDF on a set S) Let $S \subset \mathbb{R}^2$ with area $s > 0$, then the random variable (X,Y) is uniform over S if it has PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{s}, & (x,y) \in S, \\ 0, & (x,y) \notin S. \end{cases}$$

Conditioning on a random variable, independence, Bayes' rule

Definition (Conditional PDF given another random variable)
Given jointly continuous random variables X, Y and a value y such that $f_Y(y) > 0$, we define the conditional PDF as

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Additionally we define $\mathbb{P}(X \in A | Y = y) \int_A f_{X|Y}(x|y) \mathrm{d}x$.

Proposition (Multiplication rule) Given jointly continuous random variables X, Y, whenever possible we have

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x) = f_Y(y) f_{X|Y}(x|y).$$

Definition (Conditional expectation) Given jointly continuous random variables X, Y, and y such that $f_Y(y) > 0$, we define the conditional expected value as

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) \mathrm{d}x.$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) \mathrm{d}x.$$

Theorem (Total probability and total expectation theorems)

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y) \mathrm{d}y,$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} f_Y(y) \mathbb{E}[X|Y = y] \mathrm{d}y.$$

Definition (Independence) Jointly continuous random variables X, Y are independent if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all x, y.

Proposition (Expectation of product of independent r.v.) If X and Y are independent continuous random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent,
$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables)
If X and Y are independent continuous random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Proposition (Bayes' rule summary)

- For X, Y discrete: $p_{X|Y}(x|y) = \frac{p_X(x) p_{Y|X}(y|x)}{p_Y(y)}$.
- For X, Y continuous: $f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)}$.
- For X discrete, Y continuous: $p_{X|Y}(x|y) = \frac{p_X(x) f_{Y|X}(y|x)}{f_Y(y)}$.
- For X continuous, Y discrete: $f_{X|Y}(x|y) = \frac{f_X(x) p_{Y|X}(y|x)}{p_Y(y)}$.

Derived distributions

Proposition (Discrete case) Given a discrete random variable X and a function g, the r.v. $Y = g(X)$ has PMF

$$p_Y(y) = \sum_{x: g(x) = y} p_X(x).$$

Remark (Linear function of discrete random variable) If $g(x) = ax + b$, then $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

Proposition (Linear function of continuous r.v.) Given a continuous random variable X and $Y = aX + b$, with $a \neq 0$, we have

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Corollary (Linear function of normal r.v.) If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, with $a \neq 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Example (General function of a continuous r.v.) If X is a continuous random variable and g is any function, to obtain the pdf of $Y = g(X)$ we follow the two-step procedure:

1. Find the CDF of Y: $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$.
2. Differentiate the CDF of Y to obtain the PDF: $f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}$.

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{\mathrm{d}h}{\mathrm{d}y}(y) \right|.$$

where $h = g^{-1}$ in the interval where g is monotonic.

Sums of independent r.v., covariance and correlation

Proposition (Discrete case) Let X, Y be discrete independent random variables and $Z = X + Y$, then the PMF of Z is

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x).$$

Proposition (Continuous case) Let X, Y be continuous independent random variables and $Z = X + Y$, then the PDF of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \mathrm{d}x.$$

Proposition (Sum of independent normal r.v.) Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ independent. Then
$Z = X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Definition (Covariance) We define the covariance of random variables X, Y as

$$\text{Cov}(X,Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Properties (Properties of covariance)

- If X, Y are independent, then $\text{Cov}(X,Y) = 0$.
- $\text{Cov}(X,X) = \text{Var}(X)$.
- $\text{Cov}(aX + b, Y) = a \, \text{Cov}(X,Y)$.
- $\text{Cov}(X, Y + Z) = \text{Cov}(X,Y) + \text{Cov}(X,Z)$.
- $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Proposition (Variance of a sum of r.v.)

$$\text{Var}(X_1 + \cdots + X_n) = \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).$$

Definition (Correlation coefficient) We define the correlation coefficient of random variables X, Y, with $\sigma_X, \sigma_Y > 0$, as

$$\rho(X,Y) \triangleq \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Properties (Properties of the correlation coefficient)

- $-1 \leq \rho \leq 1$.
- If X, Y are independent, then $\rho = 0$.
- $|\rho| = 1$ if and only if $X - \mathbb{E}[X] = c(Y - \mathbb{E}[Y])$.
- $\rho(aX + b, Y) = \text{sign}(a)\rho(X,Y)$.

Conditional expectation and variance, sum of random number of r.v.

Definition (Conditional expectation as a random variable) Given random variables X, Y the conditional expectation $\mathbb{E}[X|Y]$ is the random variable that takes the value $\mathbb{E}[X|Y = y]$ whenever $Y = y$.

Theorem (Law of iterated expectations)

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Definition (Conditional variance as a random variable) Given random variables X, Y the conditional variance $\text{Var}(X|Y)$ is the random variable that takes the value $\text{Var}(X|Y = y)$ whenever $Y = y$.

Theorem (Law of total variance)

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y]).$$

Proposition (Sum of a random number of independent r.v.)
Let N be a nonnegative integer random variable.
Let $X, X_1, X_2, \ldots, X_N$ be i.i.d. random variables.
Let $Y = \sum_i X_i$. Then

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X],$$

$$\text{Var}(Y) = \mathbb{E}[N] \text{Var}(X) + (\mathbb{E}[X])^2 \text{Var}(N).$$

Convergence of random variables

Inequalities, convergence, and the Weak Law of Large Numbers

Theorem (Markov inequality) Given a random variable $X \geq 0$ and, for every $a > 0$ we have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Theorem (Chebyshev inequality) Given a random variable X with $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$, for every $\epsilon > 0$ we have

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Theorem (Weak Law of Large Number (WLLN)) Given a sequence of i.i.d. random variables $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, we define

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

for every $\epsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}(|M_n - \mu| \geq \epsilon) = 0.$$

Definition (Convergence in probability) A sequence of random variables $\{Y_i\}$ converges in probability to the random variable Y if

$$\lim_{n \to \infty} \mathbb{P}(|Y_i - Y| \geq \epsilon) = 0,$$

for every $\epsilon > 0$.

Properties (Properties of convergence in probability) If $X_n \to a$ and $Y_n \to b$ in probability, then

- $X_n + Y_n \to a + b$.
- If g is a continuous function, then $g(X_n) \to g(a)$.
- $\mathbb{E}[X_n]$ does not always converge to a.

The Central Limit Theorem

Theorem (Central Limit Theorem (CLT)) Given a sequence of independent random variables $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, we define

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Then, for every z, we have

$$\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z),$$

where $Z \sim \mathcal{N}(0,1)$.

Corollary (Normal approximation of a binomial) Let $X \sim Bin(n,p)$ with n large. Then S_n can be approximated by $Z \sim \mathcal{N}(np, np(1-p))$.

Remark (De Moivre-Laplace 1/2 approximation) Let $X \sim Bin$, then $\mathbb{P}(X = i) = \mathbb{P}\left(i - \frac{1}{2} \leq X \leq i + \frac{1}{2}\right)$ and we can use the CLT to approximate the PMF of X.

18.6501x Fundamentals of Statistics

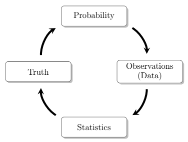
This is a cheat sheet for statistics based on the online course given by Prof. Philippe Rigollet. Compiled by Janus B. Advincula.

Last Updated December 18, 2019

Introduction to Statistics

What is Statistics?

Statistical view Data comes from a *random process*. The goal is to learn how this process works in order to make predictions or to understand what plays a role in it.



Statistics vs. Probability

Probability Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, in average 80 will be cured and at least 65 will be cured with 99.99% chances.

Statistics Observe that $\frac{78}{100}$ patients were cured. We (will be able to) conclude that we are 95% confident that for other studies, the drug will be effective on between 69.88% and 86.11% of patients.

Probability Redux

Let $X_1, \dots, X_n$ be i.i.d. random variables with $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Law of Large Numbers

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{\mathbb{P},a.s.} \mu.$$

Central Limit Theorem

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Equivalently,

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \sigma^2\right).$$

Hoeffding's Inequality Let n be a positive integer and $X, X_1, \dots X_n$ be i.i.d. random variables such that $\mathbb{E}[X] = \mu$ and $X \in [a,b]$ almost surely. Then,

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}} \quad \forall \epsilon > 0$$

The Gaussian Distribution

Because of the CLT, the Gaussian (a.k.a. normal) distribution is ubiquitous in statistics.

- $X \sim \mathcal{N}(\mu, \sigma^2)$
- $\mathbb{E}[X] = \mu$
- $\text{Var}(X) = \sigma^2 > 0$

Gaussian density (PDF)

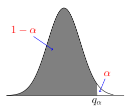
$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Useful Properties of Gaussian
It is invariant under *affine transformation*.

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any $a, b \in \mathbb{R}$,
$$aX + b \sim \mathcal{N}\left(a\mu + b, a^2\sigma^2\right).$$
- **Standardization:** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then
$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$
We can compute probabilities from the CDF of $Z \sim \mathcal{N}(0,1)$:
$$\mathbb{P}(u \leq X \leq v) = \mathbb{P}\left(\frac{u-\mu}{\sigma} \leq Z \leq \frac{v-\mu}{\sigma}\right)$$
- Symmetry: If $X \sim \mathcal{N}(0, \sigma^2)$, then $-X \sim \mathcal{N}(0, \sigma^2)$. If $x > 0$,
$$\mathbb{P}(|X| > x) = \mathbb{P}(X > x) + \mathbb{P}(-X > x) = 2\,\mathbb{P}(X > x)$$

Quantiles Let $\alpha \in (0,1)$. The quantile of order $1-\alpha$ of a random variable X is the number q_α such that

$$\mathbb{P}(X \leq q_\alpha) = 1 - \alpha.$$



Let F denote the CDF of X.

- $F(q_\alpha) = 1 - \alpha$
- If F is invertible, then $q_\alpha = F^{-1}(1-\alpha)$
- $\mathbb{P}(X > q_\alpha) = \alpha$
- If $X \sim \mathcal{N}(0,1)$, $\mathbb{P}(|X| > q_{\alpha/2}) = \alpha$

Three Types of Convergence

Almost Surely (a.s.) Convergence

$$T_n \xrightarrow[n\to\infty]{a.s.} T \iff \mathbb{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1$$

Convergence in Probability

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \iff \mathbb{P}(|T_n - T| \geq \epsilon) \xrightarrow[n\to\infty]{} 0 \quad \forall\epsilon > 0$$

Convergence in Distribution

$$T_n \xrightarrow[n\to\infty]{(d)} T \iff \mathbb{E}[f(T_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(T)]$$

for all continuous and bounded function f.

Properties

- If $(T_n)_{n\geq 1}$ converges a.s., then it also converges in probability, and the two limits are equal.
- If $(T_n)_{n\geq 1}$ converges in probability, then it also converges in distribution.
- Convergence in distribution implies convergence in probability if the limit has a density (e.g. Gaussian):
$$T_n \xrightarrow[n\to\infty]{(d)} T \quad \Rightarrow \quad \mathbb{P}(a \leq T_n \leq b) \xrightarrow[n\to\infty]{} \mathbb{P}(a \leq T \leq b)$$

Addition, Multiplication, Division

Assume

$$T_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}} T \quad \text{and} \quad U_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}} U.$$

- $T_n + U_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}} T + U$
- $T_n U_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}} TU$
- If, in addition, $U \neq 0$ a.s., then
$$\frac{T_n}{U_n} \xrightarrow[n\to\infty]{a.s./\mathbb{P}} \frac{T}{U}$$

Slutsky's Theorem

Let (X_n), (Y_n) be two sequences of random variables such that

$$(i)\ T_n \xrightarrow[n\to\infty]{(d)} T \quad \text{and} \quad (ii)\ U_n \xrightarrow[n\to\infty]{\mathbb{P}} u$$

where T is a random variable and u is a given real number. Then,

- $T_n + U_n \xrightarrow[n\to\infty]{(d)} T + u$
- $T_n U_n \xrightarrow[n\to\infty]{(d)} Tu$
- If, in addition, $u \neq 0$, then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{(d)} \frac{T}{u}$.

Continuous Mapping Theorem

If f is a continuous function, then

$$T_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}/(d)} T \quad \Rightarrow \quad f(T_n) \xrightarrow[n\to\infty]{a.s./\mathbb{P}/(d)} f(T).$$

Foundation of Inference

Statistical Model

Let the observed outcome of a statistical experiment be a *sample* $X_1, \dots, X_n$ of n i.i.d. random variables in some measurable space E (usually $E \subseteq \mathbb{R}$) and denote by $\mathbb{P}$ their common distribution. A *statistical model* associated to that statistical experiment is a *pair*

$$\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$$

where

- E is called *sample space*;
- $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is a family of probability measures on E;
- Θ is any set, called *parameter set*.

Parametric, Nonparametric and Semiparametric Models

- Usually, we will assume that the statistical model is **well-specified**, i.e., defined such that $\exists\theta$ such that $\mathbb{P} = \mathbb{P}_\theta$. This particular θ is called the **true parameter** and is unknown.
- We often assume that $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$. The model is called **parametric**.
- Sometimes we could have Θ be infinite dimensional, in which case the model is called **nonparametric**.
- If $\Theta = \Theta_1 \times \Theta_2$, where Θ_1 is finite dimensional and Θ_2 is infinite dimensional, then we have a **semiparametric** model. In these models, we only care to estimate the finite dimensional parameter and the infinite dimensional one is called **nuisance parameter**.

Identifiability

The parameter θ is called *identifiable* if and only if the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, i.e.,

$$\theta \neq \theta' \quad \implies \quad \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently,

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \quad \implies \quad \theta = \theta'.$$

Parameter Estimation

Statistic Any *measurable* function of the sample, e.g., $\bar{X}_n$, $\max_i X_i$, etc.

Estimator of θ Any statistic whose expression does not depend on θ

- An estimator $\hat{\theta}_n$ of θ is weakly (resp.strongly) **consistent** if
$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}\text{ (resp. }a.s.)} \theta \quad \text{(w.r.t. }\mathbb{P}).$$
- An estimator $\hat{\theta}_n$ of θ is **asymptotically normal** if
$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \sigma^2\right)$$

Bias of an Estimator

- **Bias** of an estimator of $\hat{\theta}_n$ of θ:
$$\text{bias}\left(\hat{\theta}_n\right) = \mathbb{E}\left[\hat{\theta}_n\right] - \theta$$
- If $\text{bias}\left(\hat{\theta}_n\right) = 0$, we say that $\hat{\theta}_n$ is **unbiased**.

Jensen's Inequality

- If the function $f(x)$ is convex,
$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$
- If the function $g(x)$ is concave,
$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X]).$$

Quadratic Risk

- We want estimators to have low bias and low variance at the same time.
- The **risk** (or **quadratic risk**) of an estimator $\hat{\theta}_n \in \mathbb{R}$ is
$$R\left(\hat{\theta}_n\right) = \mathbb{E}\left[\left|\hat{\theta}_n - \theta\right|^2\right] = \text{variance} + \text{bias}^2$$
- Low quadratic risk means that both bias and variance are small.

Confidence Intervals

Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model based on observations $X_1, \dots, X_n$, and assume $\Theta \subseteq \mathbb{R}$. Let $\alpha \in (0,1)$.

- Confidence interval (C.I.) of level $1-\alpha$ for θ: Any random (depending on $X_1, \dots, X_n$) interval $\mathcal{I}$ whose boundaries do not depend on θ and such that
$$\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall\theta \in \Theta.$$
- C.I. of asymptotic level $1-\alpha$ for θ: Any random interval $\mathcal{I}$ whose boundaries do not depend on θ and such that
$$\lim_{n\to\infty}\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall\theta \in \Theta.$$

Example We observe $R_1, \dots, R_n \overset{\text{iid}}{\sim} \text{Ber}(p)$ for some unknown $p \in (0,1)$.

- Statistical model: $\left(\{0,1\}, (\text{Ber}(p))_{p\in(0,1)}\right)$
- From CLT:
$$\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$
- It yields
$$\mathcal{I} = \left[\overline{R}_n - \frac{q_{\frac{\alpha}{2}}\sqrt{p(1-p)}}{\sqrt{n}}, \overline{R}_n + \frac{q_{\frac{\alpha}{2}}\sqrt{p(1-p)}}{\sqrt{n}}\right]$$
- But this is **not** a confidence interval because it depends on p!

Three solutions:

1. Conservative bound
2. Solving the (quadratic) equation for p
3. Plug-in

The Delta Method

Let $(Z_n)_{n\geq 1}$ be a sequence of random variables that satisfies

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \sigma^2\right)$$

for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ (the sequence $(Z_n)_{n\geq 1}$ is said to be **asymptotically normal around** θ). Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at the point θ. Then,

- $(g(Z_n))_{n\geq 1}$ is also asymptotically normal around $g(\theta)$.
- More precisely,
$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, (g'(\theta))^2\sigma^2\right).$$

Introduction to Hypothesis Testing

Statistical Formulation Consider a sample $X_1, \dots, X_n$ of i.i.d. random variables and a statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$. Let Θ_0 and Θ_1 be disjoint subsets of Θ.

Consider the two hypotheses:

- $H_0 : \theta \in \Theta_0$
- $H_1 : \theta \in \Theta_1$

H_0 is the **null hypothesis** and H_1 is the **alternative hypothesis**.

Asymmetry in the hypotheses H_0 and H_1 do not play a symmetric role: the data is only used to try to disprove H_0. Lack of evidence does not mean that H_0 is true.

A test is a statistic $\psi \in \{0,1\}$ such that:

- If $\psi = 0$, H_0 is not rejected.
- If $\psi = 1$, H_0 is rejected.

Errors

- **Rejection region** of a test ψ:
$$R_\psi = \{x \in E^n : \psi(x) = 1\}.$$
- **Type 1 error** of a test ψ:
$$\begin{aligned}\alpha_\psi : \Theta_0 &\to \mathbb{R} \quad (\text{or } [0,1])\\ \theta &\mapsto \mathbb{P}_\theta[\psi = 1]\end{aligned}$$
- **Type 2 error** of a test ψ:
$$\begin{aligned}\beta_\psi : \Theta_1 &\to \mathbb{R}\\ \theta &\mapsto \mathbb{P}_\theta[\psi = 0]\end{aligned}$$
- **Power** of a test ψ:
$$\pi_\psi = \inf_{\theta\in\Theta_1}(1 - \beta_\psi(\theta))$$

Level, test statistic and rejection region

- A test ψ has level α if
$$\alpha_\psi(\theta) \leq \alpha, \quad \forall\theta \in \Theta_0.$$
- A test ψ has asymptotic level α if
$$\lim_{n\to\infty}\alpha_\psi(\theta) \leq \alpha, \quad \forall\theta \in \Theta_0.$$
- In general, a test has the form
$$\psi = \mathbb{1}\{T_n > c\}$$
for some statistic T_n and threshold $c \in \mathbb{R}$. T_n is called the **test statistic**. The rejection region is $R_\psi = \{T_n > c\}$.

p-value The (asymptotic) p-value of a test ψ_α is the smallest (asymptotic) level α at which ψ_α rejects H_0.

Methods of Estimation

Total Variation Distance

Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that there exists $\theta^* \in \Theta$ such that $X_1 \sim \mathbb{P}_{\theta^*}$.

Statistician's goal: Given $X_1, \dots, X_n$, find an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ such that $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^*}$ for the true parameter θ^*.

The **total variation distance** between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \max_{A\subset E}|\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|$$

Total Variation Distance between Discrete Measures Assume that E is discrete (i.e., finite or countable). The total variation distance between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2}\sum_{x\in E}|p_\theta(x) - p_{\theta'}(x)|$$

Total Variation Distance between Continuous Measures Assume that E is continuous. The total variation distance between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is

$$\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2}\int|f_\theta(x) - f_{\theta'}(x)|\,dx$$

Properties of Total Variation

- $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \text{TV}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ **symmetric**
- $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$, $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq 1$ **positive**
- If $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$, then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ **definite**
- $\text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq \text{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \text{TV}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ **triangle inequality**

These imply that the total variation is a **distance** between probability distributions.

Kullback-Leibler (KL) Divergence

The Kullback-Leibler (KL) divergence between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

$$\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases}\displaystyle\sum_{x\in E}p_\theta(x)\log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) & \text{if } E \text{ is discrete}\\ \displaystyle\int_E f_\theta(x)\log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right)dx & \text{if } E \text{ is continuous}\end{cases}$$

KL-divergence is also known as **relative entropy**.

Properties of KL-divergence

- $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq \text{KL}(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
- $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
- If $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$, then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ (definite)
- $\text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \nleq \text{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + \text{KL}(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$ in general

Maximum Likelihood Estimation

Likelihood, Discrete Case Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that E is discrete (i.e., finite or countable).

Definition The likelihood of the model is the map L_n (or just L) defined as

$$\begin{aligned}L_n : E^n \times \Theta &\to \mathbb{R}\\ (x_1, \dots, x_n; \theta) &\mapsto \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]\\ &= \prod_{i=1}^n \mathbb{P}_\theta[X_i = x_i]\end{aligned}$$

Likelihood, Continuous Case Let $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that all the $\mathbb{P}_\theta$ have density f_θ.

Definition The likelihood of the model is the map L defined as

$$L : E^n \times \Theta \to \mathbb{R}$$

$$(x_1, \dots, x_n; \theta) \mapsto \prod_{i=1}^n f_\theta(x_i)$$

Maximum Likelihood Estimator Let $X_1, \dots, X_n$ be an i.i.d. sample associated with a statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ and let L be the corresponding likelihood.

Definition The maximum likelihood estimator of θ is defined as

$$\hat{\theta}_n^{\text{MLE}} = \underset{\theta\in\Theta}{\text{argmax}}\, L(X_1, \dots, X_n, \theta),$$

provided it exists.

Log-likelihood Estimator In practice, we use the fact that

$$\hat{\theta}_n^{\text{MLE}} = \underset{\theta\in\Theta}{\text{argmax}}\, \log L(X_1, \dots, X_n, \theta),$$

Concave and Convex Functions

A twice-differentiable function $h : \Theta \subset \mathbb{R} \to \mathbb{R}$ is said to be **concave** if its second derivative satisfies

$$h''(\theta) \leq 0, \quad \forall\theta \in \Theta.$$

It is said to be **strictly concave** if the inequality is strict: $h''(\theta) < 0$. Moreover, h is said to be (strictly) **convex** if $-h$ is (strictly) concave, i.e. $h''(\theta) \geq 0$ ($h''(\theta) > 0$).

Multivariate Concave Functions More generally, for a multivariate function: $h : \Theta \subset \mathbb{R}^d \to \mathbb{R}$, $d \geq 2$, define the

- **gradient vector:**
$$\nabla h(\theta) = \begin{pmatrix}\frac{\partial h(\theta)}{\partial\theta_1}\\ \vdots\\ \frac{\partial h(\theta)}{\partial\theta_d}\end{pmatrix} \in \mathbb{R}^d$$
- **Hessian matrix:**
$$\mathbb{H}h(\theta) = \begin{pmatrix}\frac{\partial^2 h(\theta)}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 h(\theta)}{\partial\theta_1\partial\theta_d}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 h(\theta)}{\partial\theta_d\partial\theta_1} & \cdots & \frac{\partial^2 h(\theta)}{\partial\theta_d\partial\theta_d}\end{pmatrix} \in \mathbb{R}^{d\times d}$$

h is concave $\iff x^\intercal\mathbb{H}h(\theta)x \leq 0, \forall x \in \mathbb{R}^d, \theta \in \Theta$

h is strictly concave $\iff x^\intercal\mathbb{H}h(\theta)x < 0, \forall x \in \mathbb{R}^d, \theta \in \Theta$

Consistency of Maximum Likelihood Estimator Under mild regularity conditions, we have

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$$

Covariance In general, when $\theta \subset \mathbb{R}^d$, $d \geq 2$, its coordinates are not necessarily independent. The covariance between two random variables X and Y is

$$\begin{aligned}\text{Cov}(X,Y) &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

Properties

- $\text{Cov}(X,X) = \text{Var}(X)$
- $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
- If X and Y are independent, then $\text{Cov}(X,Y) = 0$

Covariance Matrix The covariance matrix of a random vector

$$X = \left(X^{(1)}, \dots, X^{(d)}\right)^\intercal \in \mathbb{R}^d$$

is given by

$$\Sigma = \text{Cov}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\intercal\right].$$

This is a matrix of size $d \times d$.

If $X \in \mathbb{R}^d$ and A, B are matrices:

$$\text{Cov}(AX + B) = \text{Cov}(AX) = A\,\text{Cov}(X)A^\intercal = A\Sigma_X A^\intercal$$

The Multivariate Gaussian Distribution If $(X,T)^\intercal$ is a Gaussian vector then its PDF depends on 5 parameters:

$$\mathbb{E}[X], \text{Var}(X), \mathbb{E}[Y], \text{Var}(Y), \text{ and } \text{Cov}(X,Y).$$

A Gaussian vector $X \in \mathbb{R}^d$ is completely determined by its expected value and covariance matrix Σ:

$$X \sim \mathcal{N}_d(\mu, \Sigma).$$

It has PDF over $\mathbb{R}^d$ given by:

$$f(x) = \frac{1}{((2\pi)^d\det(\Sigma))^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\right)$$

The Multivariate CLT Let $X_1, \dots, X_n \in \mathbb{R}^d$ be independent copies of a random vector X such that $\mathbb{E}[X] = \mu$, $\text{Cov}(X) = \Sigma$, then

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma)$$

Multivariate Delta Method Let $(T_n)_{n\geq 1}$ sequence of random vectors in $\mathbb{R}^d$ such that

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma),$$

for some $\theta \in \mathbb{R}^d$ and some covariance $\Sigma \in \mathbb{R}^{d\times d}$. Let $g : \mathbb{R}^d \to \mathbb{R}^k$ ($k \geq 1$) be continuously differentiable at θ. Then,

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \nabla g(\theta)^\intercal\Sigma\,\nabla g(\theta)\right),$$

where $\nabla g(\theta) = \frac{\partial g(\theta)}{\partial\theta} = \left(\frac{\partial g_j}{\partial\theta_i}\right)_{\substack{1\leq i\leq d\\ 1\leq j\leq k}} \in \mathbb{R}^{d\times k}$

Fisher Information

Define the log-likelihood for one observation as

$$\ell(\theta) = \log L_1(X, \theta), \quad \theta \in \Theta \subset \mathbb{R}^d.$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the Fisher information of the statistical model is defined as

$$I(\theta) = \mathbb{E}\left[\nabla\ell(\theta)\nabla\ell(\theta)^\intercal\right] - \mathbb{E}[\nabla\ell(\theta)]\,\mathbb{E}[\nabla\ell(\theta)]^\intercal = -\mathbb{E}[\mathbb{H}\ell(\theta)].$$

If $\Theta \subset \mathbb{R}$, we get

$$I(\theta) = \text{Var}\left[\ell'(\theta)\right] = -\mathbb{E}\left[\ell''(\theta)\right].$$

Asymptotic Normality of the MLE

Theorem Let $\theta^* \in \Theta$ (the true parameter). Assume the following:

1. The parameter is identifiable.
2. For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ.
3. θ^* is not on the boundary of Θ.
4. $I(\theta)$ is invertible in a neighborhood of θ^*.
5. A few more technical conditions.

Then, $\hat{\theta}_n^{\text{MLE}}$ satisfies

- $\hat{\theta}_n^{\text{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$ w.r.t. $\mathbb{P}_{\theta^*}$;
- $\sqrt{n}\left(\hat{\theta}_n^{\text{MLE}} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, I^{-1}(\theta^*)\right)$ w.r.t. $\mathbb{P}_{\theta^*}$.

The Method of Moments

Moments

Let $X_1, \dots, X_n$ be an i.i.d. sample associated with a statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$. Assume that $E \subseteq \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$.

Population Moments Let $m_k(\theta) = \mathbb{E}_\theta\left[X_1^k\right]$, $1 \leq k \leq d$.

Empirical Moments Let $\hat{m}_k = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^n X_i^k$, $1 \leq k \leq d$.

From LLN,

$$\hat{m}_k \xrightarrow[n\to\infty]{\mathbb{P}/a.s.} m_k(\theta)$$

More compactly, we say that the whole vector converges:

$$(\hat{m}_1, \dots, \hat{m}_d) \xrightarrow[n\to\infty]{\mathbb{P}/a.s.} (m_1(\theta), \dots, m_d(\theta))$$

Moments Estimator

Let

$$\begin{aligned}M : \Theta &\to \mathbb{R}^d\\ \theta &\mapsto M(\theta) = (m_1(\theta), \dots, m_d(\theta))\end{aligned}$$

Assume M is one-to-one:

$$\theta = M^{-1}(m_1(\theta), \dots, m_d(\theta))$$

Moments estimator of θ:

$$\widehat{\theta}_n^{\text{MM}} = M^{-1}(\widehat{m}_1, \dots, \widehat{m}_d)$$

provided it exists.

Generalized Method of Moments

Applying the multivariate CLT and Delta method yields:

Theorem

$$\sqrt{n}\left(\widehat{\theta}_n^{\text{MM}} - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \Gamma(\theta)),$$

where $\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial\theta}M(\theta)\right]^\intercal\Sigma(\theta)\left[\frac{\partial M^{-1}}{\partial\theta}M(\theta)\right]$

MLE vs. Moment Estimator

- Comparison of the quadratic risks: In general, the MLE is more accurate.
- MLE still gives good results if the model is misspecified.
- Computational issues: Sometimes, the MLE is intractable but MM is easier (polynomial equations).

M-Estimation

- Let $X_1, \dots, X_n$ be i.i.d. with some unknown distribution $\mathbb{P}$ in some sample space E ($E \subseteq \mathbb{R}^d$ for some $d \geq 1$).
- No statistical model needs to be assumed (similar to ML).
- The goal is to estimate some parameter μ^* associated with $\mathbb{P}$, e.g. its mean, variance, median, other quantiles, the true parameter in some statistical model, etc.
- We want to find a function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values for the unknown μ^*, such that
$$Q(\mu) := \mathbb{E}[\rho(X_1, \mu)]$$
achieves its minimum at $\mu = \mu^*$.

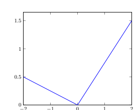
Examples (1)

- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x,\mu) = (x-\mu)^2$, for all $x, \mu \in \mathbb{R}$: $\mu^* = \mathbb{E}[X]$.
- If $E = \mathcal{M} = \mathbb{R}^d$ and $\rho(x,\mu) = \|x-\mu\|_2^2$, for all $x, \mu \in \mathbb{R}^d$: $\mu^* = \mathbb{E}[X] \in \mathbb{R}^d$.
- If $E = \mathcal{M} = \mathbb{R}$ and $\rho(x,\mu) = |x-\mu|$, for all $x, \mu \in \mathbb{R}$: μ^* is a **median** of $\mathbb{P}$.

Example (2) If $E = \mathcal{M} = \mathbb{R}$, $\alpha \in (0,1)$ is fixed and $\rho(x,\mu) = C_\alpha(x-\mu)$, for all $x, \mu \in \mathbb{R}$: μ^* is a α-quantile of $\mathbb{P}$.

Check Function

$$C_\alpha = \begin{cases}-(1-\alpha)x & \text{if } x < 0\\ \alpha x & \text{if } x \geq 0.\end{cases}$$



MLE is an M-estimator Assume that $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$ is a statistical model associated with the data.

Theorem Let $\mathcal{M} = \Theta$ and $\rho(x,\theta) = -\log L_1(x,\theta)$, provided the likelihood is positive everywhere. Then,

$$\mu^* = \theta^*,$$

where $\mathbb{P} = \mathbb{P}_{\theta^*}$ (i.e., θ^* is the true value of the parameter).

Statistical Analysis

- Define $\hat{\mu}_n$ as a minimizer of
$$Q_n(\mu) := \frac{1}{n}\sum_{i=1}^n\rho(X_i, \mu).$$
- Let $J(\mu) = \frac{\partial^2 Q(\mu)}{\partial\mu\partial\mu^\intercal}$.
- Under some regularity conditions, $J(\mu) = \mathbb{E}\left[\frac{\partial^2\rho(X_1,\mu)}{\partial\mu\partial\mu^\intercal}\right]$
- Let $K(\mu) = \text{Cov}\left(\frac{\partial\rho(X_1,\mu)}{\partial\mu}\right)$
- Remark: In the log-likelihood case,
$$J(\theta) = K(\theta) = I(\theta) \qquad \text{(Fisher information)}$$

Asymptotic Normality Let $\mu^* \in \mathcal{M}$ (the true parameter). Assume the following:

1. μ^* is the only minimizer of the function Q,
2. $J(\mu)$ is invertible for all $\mu \in \mathcal{M}$,
3. A few more technical conditions.

Then, $\hat{\mu}_n$ satisfies

- $\hat{\mu}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu^*$
- $\sqrt{n}(\hat{\mu}_n - \mu^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, J(\mu^*)^{-1}K(\mu^*)J(\mu^*)^{-1}\right)$

Hypothesis Testing

Parametric Hypothesis Testing

Hypotheses

$$H_0 : \Delta_c = \Delta_d \quad \text{vs.} \quad H_1 : \Delta_d > \Delta_c$$

Since the data is Gaussian by assumption, we don't need the CLT.

$$\overline{X}_n \sim \mathcal{N}\left(\Delta_d, \frac{\sigma_d^2}{n}\right) \quad \text{and} \quad \overline{Y}_m \sim \mathcal{N}\left(\Delta_c, \frac{\sigma_c^2}{m}\right)$$

Then,

$$\frac{\overline{X}_n - \overline{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\sigma_d^2}{n} + \frac{\sigma_c^2}{m}}} \sim \mathcal{N}(0,1)$$

Asymptotic test Assume that $m = cn$ and $n \to \infty$

Using Slutsky's theorem, we also have

$$\frac{\overline{X}_n - \overline{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\widehat{\sigma}_d^2}{n} + \frac{\widehat{\sigma}_c^2}{m}}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

where $\widehat{\sigma}_d^2 = \frac{1}{n-1}\sum_{i=1}^n\left(X_i - \overline{X}_n\right)^2$ and $\widehat{\sigma}_c^2 = \frac{1}{m-1}\sum_{i=1}^m\left(Y_i - \overline{Y}_m\right)^2$

We get the following test at asymptotic level α:

$$R_\psi = \left\{\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\widehat{\sigma}_d^2}{n} + \frac{\widehat{\sigma}_c^2}{m}}} > q_\alpha\right\}$$

The χ^2 Distribution

Definition For a positive integer d, the χ^2 distribution with d degrees of freedom is the law of the random variable $Z_1^2 + \cdots + Z_d^2$, where $Z_1, \dots, Z_d \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

Properties If $V \sim \chi_k^2$, then

- $\mathbb{E}[V] = \mathbb{E}\left[Z_1^2\right] + \cdots + \mathbb{E}\left[Z_d^2\right] = d$
- $\text{Var}(V) = \text{Var}(Z_1^2) + \cdots + \text{Var}(Z_d^2) = 2d$

Sample Variance $S_n = \frac{1}{n}\sum_{i=1}^n\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\overline{X}_n\right)^2$

Cochran's Theorem If $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

- $\overline{X}_n \perp\!\!\!\perp S_n$, for all n.
- $\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$

We often prefer the unbiased estimator of σ^2:

$$\widetilde{S}_n = \frac{1}{n-1}\sum_{i=1}^n\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}S_n$$

Student's T Distribution

Definition For a positive integer d, the Student's T distribution with d degrees of freedom (denoted by t_d) is the law of the random variable $\frac{Z}{\sqrt{V/d}}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi_d^2$ and $Z \perp\!\!\!\perp V$.

Student's T test (one-sample, two-sided)

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We want to test:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0$$

Test statistic:

$$T_n = \sqrt{n}\frac{\overline{X}_n}{\sqrt{\widetilde{S}_n}} = \frac{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}{\sqrt{\frac{\widetilde{S}_n}{\sigma^2}}}$$

Since $\sqrt{n}\frac{\overline{X}_n}{\sigma} \sim \mathcal{N}(0,1)$ (under H_0) and $\frac{\widetilde{S}_n}{\sigma^2} \sim \frac{\chi_{n-1}^2}{n-1}$ are independent by Cochran's theorem, we have

$$T_n \sim t_{n-1}.$$

Student's test with (non-asymptotic) level $\alpha \in (0,1)$:

$$\psi_\alpha = \mathbb{1}\left\{|T_n| > q_{\frac{\alpha}{2}}\right\},$$

where $q_{\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$-quantile of t_{n-1}.

Student's T test (one-sample, one-sided)

$$H_0: \mu \le \mu_0 \quad \text{vs.} \quad H_1: \mu > \mu_0$$

Test statistic:

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{\sqrt{\widehat{S}_n}} \sim t_{n-1} \qquad (\text{under } H_0)$$

Student's test with (non-asymptotic) level $\alpha \in (0,1)$:

$$\psi_\alpha = \mathbb{1}\{T_n > q_\alpha\}$$

where q_α is the $(1-\alpha)$-quantile of t_{n-1}.

Two-sample T-test

$$\frac{\overline{X}_n - \overline{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\hat\sigma_d^2}{n} + \frac{\hat\sigma_c^2}{m}}} \sim t_N$$

Welch-Satterthwaite formula

$$N = \frac{\left(\frac{\hat\sigma_d^2}{n} + \frac{\hat\sigma_c^2}{m}\right)^2}{\frac{\hat\sigma_d^4}{n^2(n-1)} + \frac{\hat\sigma_c^4}{m^2(m-1)}} \ge \min(n,m)$$

Wald's Test

A test based on the MLE Consider an i.i.d. sample $X_1, \dots, X_n$ with statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$, where $\Theta \subseteq \mathbb{R}^d$ $(d \ge 1)$ and let $\theta_0 \in \Theta$ be fixed and given. θ^* is the true parameter.

Consider the following hypotheses:

$$H_0: \theta^* = \theta_0 \quad \text{vs.} \quad H_1: \theta^* \neq \theta_0$$

Let $\widehat{\theta}_n^{\text{MLE}}$ be the MLE. Assume the MLE technical conditions are satisfied.

If H_0 is true, then

$$\sqrt{n}\, I\left(\widehat{\theta}^{\text{MLE}}\right)^{\frac{1}{2}}\left(\widehat{\theta}_n^{\text{MLE}} - \theta_0\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, \mathbb{I}_d\right)$$

Wald's test

$$T_n := n\left(\widehat{\theta}_n^{\text{MLE}} - \theta_0\right)^\intercal I\left(\widehat{\theta}_n^{\text{MLE}}\right)\left(\widehat{\theta}_n^{\text{MLE}} - \theta_0\right) \xrightarrow[n\to\infty]{(d)} \chi_d^2$$

Wald's test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ_d^2.

Wald's Test in 1 dimension In one dimension, Wald's test coincides with the two-sided test based on the asymptotic normality of the MLE.

Likelihood Ratio Test

Basic Form of the Likelihood Ratio Test Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathbb{P}_{\theta^*}$, and consider the associated statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\mathbb{R}^d}\right)$. Suppose that $\mathbb{P}_\theta$ is a discrete probability distribution with pmf given by p_θ.

In its most basic form, the likelihood ratio test can be used to decide between two hypotheses of the following form:

$$H_0: \theta^* = \theta_0 \quad \text{vs.} \quad H_1: \theta^* = \theta_1$$

Likelihood function

$$L_n : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$$

$$(x_1, \dots, x_n; \theta) \mapsto \prod_{i=1}^n p_\theta(x_i)$$

The likelihood ratio test in this set-up is of the form

$$\psi_C = \mathbb{1}\left(\frac{L_n(x_1, \dots, x_n; \theta_1)}{L_n(x_1, \dots, x_n; \theta_0)} > C\right)$$

where C is a threshold to be specified.

A test based on the log-likelihood Consider an i.i.d. sample $X_1, \dots, X_n$ with statistical model $\left(E, (\mathbb{P}_\theta)_{\theta\in\Theta}\right)$, where $\Theta \subseteq \mathbb{R}^d$ $(d \ge 1)$. Suppose the null hypothesis has the form

$$H_0: (\theta_{r+1}, \dots, \theta_d) = \left(\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}\right),$$

for some fixed and given numbers $\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}$.

Let

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \ell_n(\theta) \qquad \text{(MLE)}$$

and

$$\widehat{\theta}_n^c = \underset{\theta\in\Theta_0}{\operatorname{argmax}}\, \ell_n(\theta) \qquad \textit{(constrained MLE)}$$

where $\Theta_0 = \left\{\theta \in \Theta : (\theta_{r+1}, \dots, \theta_d) = \left(\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)}\right)\right\}$

Test statistic:

$$T_n = 2\left(\ell_n\left(\widehat{\theta}_n\right) - \ell_n\left(\widehat{\theta}_n^c\right)\right).$$

Wilk's Theorem Assume H_0 is true and the MLE technical conditions are satisfied. Then,

$$T_n \xrightarrow[n\to\infty]{(d)} \chi_{d-r}^2$$

Likelihood ratio test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\},$$

where q_α is the $(1-\alpha)$-quantile of χ_{d-r}^2.

Goodness of Fit Tests

Let X be a r.v. We want to know if the hypothesized distribution is a good fit for the data.

Key characteristic of Goodness of Fit tests: no parametric modeling.

Discrete distribution Let $E = \{a_1, \dots, a_K\}$ be a finite space and $(\mathbb{P}_\mathbf{p})_{\mathbf{p}\in\Delta_K}$ be the family of all probability distributions on E.

- $\Delta_K = \left\{\mathbf{p} = (p_1, \dots, p_K) \in (0,1)^K : \sum_{j=1}^K p_j = 1\right\}$
- For $\mathbf{p} \in \Delta_K$ and $X \sim \mathbb{P}_\mathbf{p}$,
$$\mathbb{P}_\mathbf{p}[X = a_j] = p_j, \quad j = 1, \dots, K.$$

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathbb{P}_\mathbf{p}$, for some unknown $\mathbf{p} \in \Delta_K$, and let $\mathbf{p}^0 \in \Delta_K$ be fixed.

We want to test:

$$H_0: \mathbf{p} = \mathbf{p}^0 \quad \text{vs.} \quad H_1: \mathbf{p} \neq \mathbf{p}^0$$

with asymptotic level $\alpha \in (0,1)$.

The Probability Simplex in K Dimensions The probability simplex in $\mathbb{R}^K$, denoted by Δ_K, is the set of all vectors $\mathbf{p} = [p_1, \dots, p_K]^\intercal$ such that

$$\mathbf{p} \cdot \mathbf{1} = \mathbf{p}^\intercal \mathbf{1} = 1, \quad p_i \ge 0 \quad \text{for all } K$$

where $\mathbf{1}$ denotes the vector $\mathbf{1} = (1, \dots, 1)^\intercal$

Categorical Likelihood

- Likelihood of the model:
$$L_n(X_1, \dots, X_n; \mathbf{p}) = p_1^{N_1} p_2^{N_2} \dots p_K^{N_K}$$
where $N_j = \#\{i = 1, \dots, n : X_i = a_j\}$.
- Let $\widehat{\mathbf{p}}$ be the MLE:
$$\widehat{\mathbf{p}}_j = \frac{N_j}{n}, \quad j = 1, \dots, K.$$
$\widehat{\mathbf{p}}$ maximizes $\log L_n(X_1, \dots, X_n, \mathbf{p})$ *under the constraint.*

χ^2 test If H_0 is true, then $\sqrt{n}\left(\widehat{\mathbf{p}} - \mathbf{p}^0\right)$ is asymptotically normal, and the following holds:

Theorem Under H_0:

$$T_n = n\sum_{j=1}^n \frac{\left(\widehat{\mathbf{p}}_j - \mathbf{p}_j^0\right)^2}{\mathbf{p}_j^0} \xrightarrow[n\to\infty]{(d)} \chi_{K-1}^2$$

CDF and empirical CDF Let $X_1, \dots, X_n$ be i.i.d. real random variables. The CDF of X_1 is defined as

$$F(t) = \mathbb{P}[X_1 \le t], \quad \forall t \in \mathbb{R}.$$

It completely characterizes the distribution of X_1.

The **empirical CDF** of the sample $X_1, \dots, X_n$ is defined as

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \le t\} = \frac{\#\{i = 1, \dots, n : X_i \le t\}}{n}, \quad \forall t \in \mathbb{R}.$$

Consistency By the LLN, for all $t \in \mathbb{R}$,

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

Glivenko-Cantelli Theorem (Fundamental theorem of statistics)

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0$$

Asymptotic normality By the CLT, for all $t \in \mathbb{R}$,

$$\sqrt{n}\left(F_n(t) - F(t)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, F(t)\left(1 - F(t)\right)\right)$$

Donsker's Theorem If F is continuous, then

$$\sqrt{n}\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} \sup_{0\le t\le 1} |\mathbf{B}(t)|,$$

where $\mathbf{B}(t)$ is a Brownian bridge on $[0,1]$.

Kolmogorov-Smirnov Test

Let $T_n = \sup_{t\in\mathbb{R}} \sqrt{n}\,|F_n(t) - F(t)|$. By Donsker's theorem, if H_0 is true, then $T_n \xrightarrow[n\to\infty]{(d)} Z$, where Z has a known distribution (supremum of the absolute value of a Brownian bridge).

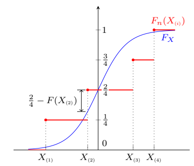
KS test with asymptotic level α:

$$\delta_\alpha^{\text{KS}} = \mathbb{1}\{T_n > q_\alpha\}$$

where q_α is the $(1-\alpha)$-quantile of Z.

Let $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ be the reordered sample. The expression for T_n reduces to

$$T_n = \sqrt{n}\max_{i=1,\dots,n}\left\{\max\left(\left|\frac{i-1}{n} - F^0\left(X_{(i)}\right)\right|, \left|\frac{i}{n} - F^0\left(X_{(i)}\right)\right|\right)\right\}.$$



Pivotal Distribution T_n is called a **pivotal statistic**: If H_0 is true, the distribution of T_n does not depend on the distribution of the X_i's.

Other Goodness of Fit Tests

Kolmogorov-Smirnov

$$d(F_n, F) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$$

Cramér-Von Mises

$$d^2(F_n, F) = \int_\mathbb{R} [F_n(t) - F(t)]^2\, dF(t) = \underset{X\sim F}{\mathbb{E}}\left[|F_n(X) - F(X)|^2\right]$$

Anderson-Darling

$$d^2(F_n, F) \int_\mathbb{R} \frac{[F_n(t) - F(t)]^2}{F(t)\left(1 - F(t)\right)}\, dF(t)$$

Kolmogorov-Lilliefors Test

We want to test if X has a Gaussian distribution with unknown parameters. In this case, Donsker's theorem is *no longer valid*. Instead, we compute the quantiles for the test statistic

$$\sup_{t\in\mathbb{R}}\left|F_n(t) - \Phi_{\hat\mu,\hat\sigma^2}(t)\right|$$

where $\hat\mu = \overline{X}_n$, $\hat\sigma^2 = S_n^2$ and $\Phi_{\hat\mu,\hat\sigma^2}(t)$ is the CDF of $\mathcal{N}(\hat\mu, \hat\sigma^2)$.

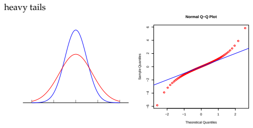
They do not depend on unknown parameters.

Quantile-Quantile (QQ) plots

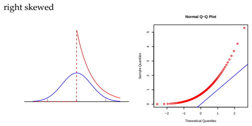
- Provide a visual way to perform goodness of fit tests.
- Not a formal test but quick and easy check to see if a distribution is plausible.
- Main idea: We want to check visually if the plot of F_n is close to that of F or, equivalently, if the plot of F_n^{-1} is close to F^{-1}.
- Check if the points
$$\left(F^{-1}(\tfrac{1}{n}), F_n^{-1}(\tfrac{1}{n})\right), \dots, \left(F^{-1}(\tfrac{n-1}{n}), F_n^{-1}(\tfrac{n-1}{n})\right)$$
are near the line $y = x$.
- F_n is not technically invertible but we define
$$F_n^{-1}(\tfrac{i}{n}) = X_i,$$
the i^{th} largest observation.

Four patterns

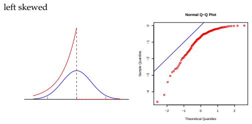
1. heavy tails



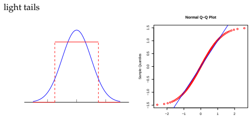
2. right skewed



3. left skewed



4. light tails



Bayesian Statistics

Introduction to Bayesian Statistics

Prior and Posterior

- Consider a probability distribution on a parameter space Θ with some PDF $\pi(\cdot)$: the **prior distribution**.
- Let $X_1, \dots, X_n$ be a sample of n random variables.
- Denote by $L_n(\cdot|\theta)$ the joint PDF of $X_1, \dots, X_n$ conditionally on θ, where $\theta \sim \pi$.
- **Remark:** $L_n(X_1, \dots, X_n|\theta)$ is the likelihood used in the frequentist approach.
- The conditional distribution of θ given $X_1, \dots, X_n$ is called the **posterior distribution**. Denote by $\pi(\cdot|X_1, \dots, X_n)$ its PDF.

Bayes' formula

$$\pi(\theta|X_1, \dots, X_n) \propto \pi(\theta) L_n(X_1, \dots, X_n|\theta), \quad \forall\theta \in \Theta$$

Bernoulli experiment with a Beta prior

- $p \sim \text{Beta}(a, a)$:
$$\pi(p) \propto p^{a-1}(1-p)^{a-1}, \quad p \in (0,1)$$
- Given p, $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$, so
$$L_n(X_1, \dots, X_n|p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$
- Hence,
$$\pi(p|X_1, \dots, X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}$$
- The posterior distribution is
$$\text{Beta}\left(a + \sum_{i=1}^n X_i, a + n - \sum_{i=1}^n X_i\right) \qquad \textbf{conjugate prior}$$

Non-informative Priors

- We can still use a Bayesian approach if we have no prior information about the parameter.
- Good candidate: $\pi(\theta) \propto 1$, i.e., constant PDF on Θ.
- If Θ is bounded, this is the uniform prior on Θ.
- If Θ is unbounded, this does not define a proper PDF on Θ.
- An **improper prior** on Θ is a measurable, non-negative function $\pi(\cdot)$ defined on Θ that is not integrable:
$$\int \pi(\theta)d\theta = \infty.$$
- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

Jeffreys Prior and Bayesian Confidence Interval

Jeffreys prior is an attempt to incorporate frequentist ideas of likelihood in the Bayesian framework, as well as an example of a **non-informative prior**:

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$

where $I(\theta)$ is the Fisher information matrix of the statistical model associated with $X_1, \dots, X_n$ in the frequentist approach (provided it exists).

Examples

- Bernoulli experiment: $\pi_J(\theta) \propto \dfrac{1}{\sqrt{p(1-p)}}$, $\quad p \in (0,1)$: the prior is $\text{Beta}(\frac{1}{2}, \frac{1}{2})$
- Gaussian experiment: $\pi_J(\theta) \propto 1, \theta \in \mathbb{R}$, is an improper prior

Jeffreys prior satisfies a **reparametrization invariance principle**: If η is a reparametrization of θ (i.e., $\eta = \phi(\theta)$ for some one-to-one map ϕ), then the PDF $\tilde\pi(\cdot)$ of η satisfies:

$$\tilde\pi(\eta) \propto \sqrt{\det \tilde{I}(\eta)},$$

where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ.

Bayesian confidence regions For $\alpha \in (0,1)$, a Bayesian confidence region with level α is a random subset $\mathcal{R}$ of the parameter space Θ, which depends on the sample $X_1, \dots, X_n$, such that

$$\mathbb{P}[\theta \in \mathcal{R}|X_1, \dots, X_n] = 1 - \alpha.$$

Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.

Bayesian confidence region and *confidence interval* are two **distinct** notions.

Bayesian estimation

- **Posterior mean:** $\widehat{\theta}^{(\pi)} = \int_\Theta \theta\pi(\theta|X_1, \dots, X_n)\, d\theta$
- **MAP (maximum a posteriori):** $\widehat{\theta}^{\text{MAP}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \pi(\theta|X_1, \dots, X_n)$

  It is the point that maximizes the posterior distribution, provided it is unique.

Linear Regression

Modeling Assumptions $(X_i, Y_i), i = 1, \dots, n$, are i.i.d. from some *unknown joint distribution* $\mathbb{P}$. $\mathbb{P}$ can be described entirely by (assuming all exist):

- either a joint PDF $h(x, y)$
- the marginal density of X, $h(x) = \int h(x,y)dy$ **and** the conditional density
$$h(y|x) = \frac{h(x,y)}{h(x)}$$

$h(y|x)$ answers all our questions. It contains all the information about Y given X.

Partial Modeling We can also describe the distribution only partially, e.g. using

- the expectation of Y: $\mathbb{E}[Y]$
- the conditional expectation of Y given $X = x$: $\mathbb{E}[X = x]$. The function
$$x \mapsto f(x) := \mathbb{E}[Y|X = x] = \int yh(y|x)dy$$
is called **regression function.**
- other possibilities:
  - the conditional median: $m(x)$ such that
  $$\int_{-\infty}^{m(x)} h(y|x)dy = \frac{1}{2}$$
  - conditional quantiles
  - conditional variance (not information about location)

Linear Regression We focus on modeling the regression function

$$f(x) = \mathbb{E}[Y|X = x].$$

Restrict to *simple* functions. The simplest is

$$f(x) = a + bx \qquad \text{linear (or affine) function}$$

Probabilistic Analysis Let X and Y be two r.v. (not neccessarily independent) with two moments and such that $\text{Var}(X) > 0$. The theoretical linear regression of Y on X is the line $x \mapsto a^* + b^*x$, where

$$(a^*, b^*) = \underset{(a,b)\in\mathbb{R}^2}{\operatorname{argmin}}\, \mathbb{E}\left[(Y - a - bX)^2\right]$$

which gives

$$b^* = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

$$a^* = \mathbb{E}[Y] - b^*\mathbb{E}[X] = \mathbb{E}[Y] - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}\mathbb{E}[X]$$

Noise The points are not exactly on the line $x \mapsto a^* + b^*x$ if $\text{Var}(Y|X = x) > 0$. The random variable $\varepsilon = Y - (a^* + b^*X)$ is called **noise** and satisfies

$$Y = a^* + b^*X + \varepsilon,$$

with $\mathbb{E}[\varepsilon] = 0$ and $\text{Cov}(X, \varepsilon) = 0$

Statistical Problem In practice, a^*, b^* need to be estimated from data.

Least Squares The **least squares estimator** (LSE) of (a, b) is the minimizer of the sum of squared errors:

$$\sum_{i=1}^n (Y_i - a - bX_i)^2.$$

Then,

$$\hat{b} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}$$

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

Multivariate Regression

We have a vector of explanatory variables or **covariates**:

$$\mathbf{X}_i = \begin{pmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(p)} \end{pmatrix} \in \mathbb{R}^p.$$

The **response** or **dependent variable** is Y_i with

$$Y_i = \mathbf{X}_i^\intercal\boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, \dots, n$$

and $\boldsymbol{\beta}_1^*$ is called the **intercept**.

Least Squares Estimator The least squares estimator of $\boldsymbol{\beta}^*$ is the minimizer of the sum of squared errors

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - \mathbf{X}_i^\intercal\boldsymbol{\beta}\right)^2$$

LSE in Matrix Form

- Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\intercal \in \mathbb{R}^n$.
- Let $\mathbb{X}$ be the $n \times p$ matrix whose rows are $\mathbf{X}_1^\intercal, \dots, \mathbf{X}_n^\intercal$. $\mathbb{X}$ is called the **design matrix**.
- Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\intercal \in \mathbb{R}^n$, the unobserved noise. Then,
$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\beta}^* \text{ unknown.}$$
- The LSE $\widehat{\boldsymbol{\beta}}$ satisfies
$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2.$$

Closed Form Solution Assume that $\text{rank}(\mathbb{X}) = p$. Then,

$$\widehat{\boldsymbol{\beta}} = \left(\mathbb{X}^\intercal\mathbb{X}\right)^{-1}\mathbb{X}^\intercal\mathbf{Y}.$$

Geometric Interpretation of the LSE $\mathbb{X}\widehat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbb{X}$:

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = P\mathbf{Y},$$

where $P = \mathbb{X}\left(\mathbb{X}^\intercal\mathbb{X}\right)^{-1}\mathbb{X}^\intercal$.

Statistical Inference To make inference, we need more assumptions.

- The design matrix $\mathbb{X}$ is deterministic and $\text{rank}(\mathbb{X}) = p$.
- The model is **homoscedastic**: $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d.
- The noise vector $\boldsymbol{\varepsilon}$ is Gaussian:
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0, \sigma^2\mathbb{I}_n\right)$$
for some known or unknown $\sigma^2 > 0$.

Properties of LSE

- LSE = MSE
- Distribution of $\widehat{\boldsymbol{\beta}}$:
$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p\left(\boldsymbol{\beta}^*, \sigma^2\left(\mathbb{X}^\intercal\mathbb{X}\right)^{-1}\right)$$
- Quadratic Risk of $\widehat{\boldsymbol{\beta}}$:
$$\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\right] = \sigma^2\textbf{tr}\left(\left(\mathbb{X}^\intercal\mathbb{X}\right)^{-1}\right)$$
- Prediction Error:
$$\mathbb{E}\left[\|\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|_2^2\right] = \sigma^2(n - p)$$
- Unbiased estimator of σ^2:
$$\widehat{\sigma}^2 = \frac{\|\mathbf{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}\|_2^2}{n - p} = \frac{1}{n - p}\sum_{i=1}^n \widehat{\varepsilon}_i^2$$

Significance Tests

- Test whether the j^{th} explanatory variable is significant in the linear regression.
- $H_0: \beta_j = 0$ v.s. $H_1: \beta \neq 0$
- If γ_j $(\gamma_j > 0)$ is the j^{th} diagonal coefficient of $(\mathbb{X}^\intercal\mathbb{X})^{-1}$:
$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2\gamma_j}} \sim t_{n-p}$$
- Let $T_n^{(j)} = \dfrac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2\gamma_j}}$.
- Test with non-asymptotic level $\alpha \in (0,1)$:
$$R_{j,\alpha} = \left\{\left|T_n^{(j)}\right| > q_{\frac{\alpha}{2}}(t_{n-p})\right\}$$
where $q_{\frac{\alpha}{2}}(t_{n-p})$ is the $(1 - \frac{\alpha}{2})$-quantile of t_{n-p}.

Bonferroni's test Test whether a **group** of explanatory variables is significant in the linear regression.

- $H_0: \beta_j = 0\ \forall j \in S$ v.s. $H_1: \exists j \in S, \beta_j \neq 0$ where $S \subseteq \{1, \dots, p\}$.
- Bonferroni's test:
$$R_{S,\alpha} = \bigcup_{j\in S} R_{j,\frac{\alpha}{k}}, \quad \text{where } k = |S|$$

Generalized Linear Model

Generalization A generalized linear model (GLM) generalizes normal linear regression models in the following directions:

1. **Random component**: $Y|X = x \sim$ some distribution
2. **Regression function**:
$$g(\mu(x)) = x^\intercal\beta$$
where g is called **link function** and $\mu(x) = \mathbb{E}[Y|X = x]$ is the **regression function**.

Exponential Family

A family of distribution $\{\mathbb{P}_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ is said to be a k**-parameter exponential family** on $\mathbb{R}^q$, if there exist real-valued functions

- $\eta_1, \dots, \eta_k$ and $B(\theta)$
- $T_1, \dots, T_k$, and $h(y) \in \mathbb{R}^q$

such that the density function of $\mathbb{P}_\theta$ can be written as

$$f_\theta(y) = \exp\left[\sum_{i=1}^k \eta_i(\theta)T_i(y) - B(\theta)\right]h(y)$$

Examples of discrete distributions The following distributions form **discrete** exponential families of distributions with PMF:

- Bernoulli (p): $p^y(1-p)^{1-y}, y \in \{0,1\}$
- Poisson (λ): $\dfrac{\lambda^y}{y!}e^{-\lambda}, y = 0, 1, \dots$

Examples of continuous distributions The following distributions form **continuous** exponential families of distributions with PDF:

- Gamma (a, b): $\dfrac{1}{\Gamma(a)b^a}y^{a-1}e^{-\frac{y}{b}}$
- Inverse Gamma (α, β): $\dfrac{\beta^\alpha}{\Gamma(\alpha)}y^{-\alpha-1}e^{-\frac{\beta}{y}}$
- Inverse Gaussian (μ, σ^2): $\sqrt{\dfrac{\sigma^2}{2\pi y^3}}\exp\left(-\dfrac{\sigma^2(y-\mu)^2}{2\mu^2 y}\right)$

One-parameter Canonical Exponential Family

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

for some known functions $b(\theta)$ and $c(y, \phi)$.

- If ϕ is known, this is a one-parameter exponential family with θ being the **canonical parameter**.
- If ϕ is unknown, this may/may not be a two-parameter exponential family.
- ϕ is called **dispersion parameter**.

Expected value Note that

$$\ell(\theta) = \frac{Y\theta - b(\theta)}{\phi} + c(Y; \phi),$$

which leads to

$$\mathbb{E}[Y] = b'(\theta).$$

Variance

$$\text{Var}(Y) = b''(\theta) \cdot \phi$$

In GLM, we have $Y|X = x \sim$ distribution in exponential family. Then,

$$\mathbb{E}[Y|X = x] = f\left(X^\intercal\beta\right)$$

Link function β is the parameter of interest. A **link function** g relates the linear predictor $X^\intercal\beta$ to the mean parameter μ,

$$X^\intercal\beta = g(\mu) = g(\mu(X)).$$

g is required to be monotone increasing and differentiable

$$\mu = g^{-1}\left(X^\intercal\beta\right)$$

Canonical Link The function g that links the mean μ to the canonical parameter θ is called **canonical link**:

$$g(\mu) = \theta.$$

Since $\mu = b'(\theta)$, the canonical link is given by

$$g(\mu) = (b')^{-1}(\mu).$$

If $\phi > 0$, the canonical link function is strictly increasing.

Example Bernoulli distribution

$$p^y(1-p)^{1-y} = \exp\left(y\log\left(\frac{p}{1-p}\right) + \log(1-p)\right)$$
$$= \exp\left(y\theta - \log(1 + e^\theta)\right)$$

Hence, $\theta = \log\left(\dfrac{p}{1-p}\right)$ and $b(\theta) = \log\left(1 + e^\theta\right)$.

$$b'(\theta) = \frac{e^\theta}{1 + e^\theta} = \mu \iff \theta = \log\left(\frac{\mu}{1-\mu}\right)$$

The canonical link for the Bernoulli distribution is the **logit link**.

Model and Notation

Let $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, \dots, n$ be independent random pairs such that the conditional distribution of Y_i given $X_i = x_i$ has density in the canonical exponential family:

$$f_{\theta_i}(y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right]$$

Back to β: Given a link function g, note the following relationship between β and θ:

$$\theta_i = (b')^{-1}(\mu_i) = (b')^{-1}\left(g^{-1}(X_i^\intercal\beta)\right) \equiv h\left(X_i^\intercal\beta\right)$$

where h is defined as

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}.$$

If g is the *canonical link function*, h is the **identity** $g = (b')^{-1}$.

Log-likelihood The log-likelihood is given by

$$\ell_n(\mathbf{Y}, \mathbb{X}, \beta) = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} + \text{constant}$$
$$= \sum_i \frac{Y_i h\left(X_i^\intercal\beta\right) - b\left(h\left(X_i^\intercal\beta\right)\right)}{\phi} + \text{constant}$$

When we use the *canonical link function*, we obtain the expression

$$\ell_n(\mathbf{Y}, \mathbb{X}, \beta) = \sum_i \frac{Y_i X_i^\intercal\beta - b\left(X_i^\intercal\beta\right)}{\phi} + \text{constant}$$

Strict concavity The log-likelihood $\ell(\theta)$ is **strictly concave** (if $\text{rank}(\mathbb{X}) = p$) using the canonical function when $\phi > 0$. As a consequence, the maximum likelihood estimator is unique.

On the other hand, if another parametrization is used, the likelihood function may not be strictly concaving leading to *several local maxima*.

Recommended Resources

- Probability and Statistics (DeGroot and Schervish)
- Mathematical Statistics and Data Analysis (Rice)
- Fundamentals of Statistics [Lecture Slides] (`http://www.edx.org`)

Please share this cheatsheet with friends!