Cheatsheet Probability and Statistics

Marc-Olivier Jufer
mjufer@ethz.ch

August 12, 2025

1 Mathematical framework

1.1 Probability space

Def. 1.1. The set Ω is called the **sample space**. An element $\omega \in \Omega$ is called an **outcome** or **elementary experiment**.

Ex. 1.1. *Throw of a die :* $\Omega = \{1, 2, 3, 4, 5, 6\}$

Def. 1.2. A **sigma-algebra** is a subset $\mathcal{F} \subset \mathcal{P}(\Omega)$ satisfying the following properties :

P1. $\Omega \in \mathcal{F}$

P2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$: If A is an event, "not A" is also an event.

P3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$: if $A_1, A_2, \ldots$ are events, then "A_1 or A_2 or ..." is an event

Ex. 1.2. *Examples of sigma-algebras for* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *:*

- $\mathcal{F} = \{\emptyset, \{1, 2, 3, 4, 5, 6\}\}$
- $\mathcal{F} = \mathcal{P}(\Omega)$
- $\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4, 5, 6\}, \{1, 2, 3, 4, 5, 6\}\}$

Non examples of sigma-algebras for $\Omega = \{1, 2, 3, 4, 5, 6\}$ *:*

- $\mathcal{F} = \{\{1, 2, 3, 4, 5, 6\}\}$ *: **P2** is not satisfied*
- $\mathcal{F} = \{\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \{1\}, \{2, 3, 4, 5, 6\}, \Omega\}$ *: **P3** is not satisfied*

Def. 1.3. Let Ω a sample space and $\mathcal{F}$ a sigma-algebra. A **probability measure** on $(\Omega, \mathcal{F})$ is a map

$$\mathbb{P} : \mathcal{F} \to [0, 1], \quad A \mapsto \mathbb{P}[A]$$

that satisfies the properties

P1. $\mathbb{P}[\Omega] = 1$

P2. (countable additivity) $\mathbb{P}[A] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$ if $A = \bigcup_{i=1}^{\infty} A_i$ (disjoint union)

Int. A probability measure is a map that associates to each event a number in $[0, 1]$

Ex. 1.3. *For* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *and* $\mathcal{F} = \mathcal{P}(\Omega)$*, the mapping* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *defined by*

$$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \frac{|A|}{6}$$

is a probability measure on $(\Omega, \mathcal{F})$.

Def. 1.4. Let Ω a sample space, $\mathcal{F}$ a sigma-algebra and $\mathbb{P}$ a probability measure. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

Int. To construct a probabilistic model, we give

- a sample space Ω : all the possible outcomes of the experiment
- a sigma-algebra $\mathcal{F} \subset \mathcal{P}(\Omega)$: the set of events
- a probability measure $\mathbb{P}$: gives a number in $[0, 1]$ to every event

Def. 1.5. Let $\omega \in \Omega$ (a possible outcome). Let A be an event. We say the event A **occurs** (**does not occur**) (for ω) if $\omega \in A$ ($\omega \notin A$).

1.2 Examples of probability spaces

Def. 1.6. Let Ω be a finite sample space. The **Laplace model** on Ω is the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is defined by

$$\forall A \in \mathcal{F} \quad P[A] = \frac{|A|}{|\Omega|}$$

1.3 Properties of Events

Prop 1.1. (Consequences of definition 1.2). Let $\mathcal{F}$ be a sigma-algebra on Ω. We have

P4. $\emptyset \in \mathcal{F}$

P5. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$

P6. $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$

P7. $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$

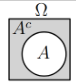
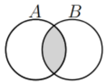
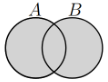
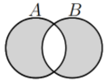
Event	Graphical representation	Probab. interpretation
A^c		A does **not** occur
$A \cap B$		A **and** B occur
$A \cup B$		A **or** B occurs
$A \Delta B$		one and only one of A or B occurs

Figure 1: Representation of set operations

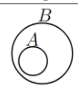
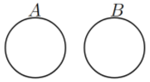
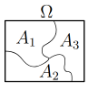
Relation	Graphical representation	Probab. interpretation
$A \subset B$		If A occurs, then B occurs
$A \cap B = \emptyset$		A and B cannot occur at the same time
$\Omega = A_1 \cup A_2 \cup A_3$ with A_1, A_2, A_3 pairwise disjoint		for each outcome ω, one and only one of the events A_1, A_2, A_3 is satisfied.

Figure 2: Representation of set relations

1.4 Properties of probability measures

Prop 1.2. (Consequences of definition 1.3). Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$.

P3. We have $\mathbb{P}[\emptyset] = 0$

P4. (**additivity**) Let $k \geq 1$, let $A_1, \ldots, A_k$ be k pairwise disjoint events, then

$$\mathbb{P}[A_1 \cup \ldots \cup A_k] = \mathbb{P}[A_1] + \ldots + \mathbb{P}[A_k]$$

P5. Let A be an event, then

$$\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$$

P6. If A and B are two events (not necessarily disjoint), then

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

Prop 1.3. (**Monotonicity**). Let $A, B \in \mathcal{F}$, then

$$A \subset B \Rightarrow \mathbb{P}[A] \leq \mathbb{P}[B]$$

Prop 1.4. (**Union bound**).Let $A_1, A_2, \ldots$ be a sequence of events (not necessarily disjoint), then we have

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] \leq \sum_{i=1}^{\infty} \mathbb{P}[A_i]$$

Union bound also applies to a finite collection of events.

Prop 1.5. Let (A_n) be an increasing sequence of events (i.e. $\forall n \; A_n \subset A_{n+1}$). Then

$$\lim_{n\to\infty} \mathbb{P}[A_n] = \mathbb{P}\left[\bigcup_{n=1} \infty A_n\right].$$ **increasing limit**

Let (B_n) be a decreasing sequence of events (i.e. $\forall n \; B_n \supset B_{n+1}$). Then

$$\lim_{n\to\infty} \mathbb{P}[B_n] = \mathbb{P}\left[\bigcap_{n=1} \infty B_n\right].$$ **decreasing limit**

[1] i.e. $\Omega = B_1 \cup \ldots \cup B_n$ and the events are pairwise disjoint.

1.5 Conditional probabilities

Def. 1.7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Let A, B be two events with $\mathbb{P}[B] > 0$. The **conditional probability of A given B** is defined by

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Ex. 1.4. *We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ corresponding to the throw of one die. Let $A = \{1,2,3\}$ and $B = \{2,4,6\}$. Then*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

Prop 1.6. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Let B be an event with positive probability. Then $\mathbb{P}[\,.\,|B]$ is a probability measure on Ω.

Prop 1.7. (**Formula of total probability**). Let $B_1, \ldots, B_n$ be a partition[1] of the sample space Ω with $\mathbb{P}[B_i] > 0$ for every $i \leq i \leq n$. Then one has

$$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|B_i]\,\mathbb{P}[B_i]$$

Prop 1.8. (**Bayes formula**). Let $B_1, \ldots, B_n \in \mathcal{F}$ be a partition of Ω with $\mathbb{P}[B_i] > 0 \; \forall i$. For every event A with $\mathbb{P}[A] > 0$ we have

$$\forall i = 1, \ldots, n \quad \mathbb{P}[B_i|A] = \frac{\mathbb{P}[A|B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}[A|B_j]\,\mathbb{P}[B_j]}$$

Ex. 1.5. *Test to detect a disease which concerns about* 1/10000 *of the population. The test gives the right answer* 99% *of the time. If a patient has a positive test, what is the probability that he is actually sick ?*
We modeled the situation as $\Omega = \{0,1\} \times \{0,1\}$. $\mathcal{F} = \mathcal{P}(\Omega)$ and an outcome is $\omega = (\omega_1, \omega_2)$, where ω_1 is 1 if the patient is sick and ω_2 is 1 if the test is positive. Let $S = \{(1,0),(1,1)\}$ be the event that the patient is sick and $T = \{(0,1),(1,1)\}$ the event that the test is positive.
From the hypotheses, we have

$$\mathbb{P}[S] = \frac{1}{10000}, \quad \mathbb{P}[T|S] = \frac{99}{100}, \quad \mathbb{P}[T|S^c] = \frac{1}{100}$$

By applying the Bayes formula to the partition $\Omega = S \cup S^c$, we obtain

$$\mathbb{P}[S|T] = \frac{\mathbb{P}[T|S]\,\mathbb{P}[S]}{\mathbb{P}[T|S]\,\mathbb{P}[S] + \mathbb{P}[T|S^c]\,\mathbb{P}[S^c]} \simeq 0.0098$$

1.6 Independence

Def. 1.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events A and B are said to be **independent** if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$$

A is independent of B iff A is independent of B^c.
If $\mathbb{P}[A] \in \{0,1\}$, then A is independent of every event.
If A is independent with itself (i.e. $\mathbb{P}[A \cap A] = \mathbb{P}[A]^2$), then $\mathbb{P}[A] \in \{0,1\}$.

Prop 1.9. Let $A, B \in \mathcal{F}$ be two events with $\mathbb{P}[A], \mathbb{P}[B] > 0$. Then the following are equivalent :

i. $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$: A and B are independent

ii. $\mathbb{P}[A|B] = \mathbb{P}[A]$: the occurrence of B has no influence on A

iii. $\mathbb{P}[B|A] = \mathbb{P}[B]$: the occurrence of A has no influence on B

Def. 1.9. Let I be an arbitrary set of indices. A collection of events $(A_i)_{i \in I}$ is said to be **independent** if

$$\forall J \subset I finite \quad \mathbb{P}\left[\bigcap_{j \in J} A_j\right] = \prod_{j \in J} \mathbb{P}[A_j]$$

Int. Three events A, B and C are independent if the following 4 equations are satisfied :

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$$

$$\mathbb{P}[A \cap C] = \mathbb{P}[A]\,\mathbb{P}[C]$$

$$\mathbb{P}[B \cap C] = \mathbb{P}[B]\,\mathbb{P}[C]$$

$$\mathbb{P}[A \cap B \cap C] = \mathbb{P}[A]\,\mathbb{P}[B]\,\mathbb{P}[C]$$

2 Random variables and distribution functions

2.1 Abstract definition

Def. 2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random variable** (r.v.) is a map $X : \Omega \to \mathbb{R}$ s.t.

$$\forall a \in \mathbb{R} \quad \{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}$$

Ex. 2.1. *We throw a fair die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$ and we consider the Laplace model $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose we gamble on the outcome in such a way that our profit is -1 if the outcome is 1, 2 or 3 ; 0 if the outcome is 4 and 2 if the outcome is 5 or 6. Our profit can be represented by the mapping X defined by*

$$\forall \omega \in \Omega \quad X(\omega) = \begin{cases} -1 & \text{if } \omega = 1, 2, 3, \\ 0 & \text{if } \omega = 4, \\ 2 & \text{if } \omega = 5, 6 \end{cases}$$

Since $\mathcal{F} = \mathcal{P}(\Omega)$, we have $\{\omega : X(\omega) \leq a\} \in \mathcal{F}$ for every a. Therefore, X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.

Def. 2.2. When events are defined in terms of random variable, we omit the dependence in ω. E.g. for $a \leq b$ we write

$$\{X \leq a\} = \{\omega \in \Omega : X(\omega) \leq a\}$$

$$\{a < X \leq b\} = \{\omega \in \Omega : a < X(\omega) < b\}$$

$$\{X \in \mathbb{Z}\} = \{\omega \in \Omega : X(\omega) \in \mathbb{Z}\}$$

When consider the probability of events as above, we omit the brackets

$$\mathbb{P}[X \leq a] = \mathbb{P}[\{X \leq a\}] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\}]$$

2.2 Distribution function

Def. 2.3. Let X be a random variable on a prob. space $(\Omega, \mathcal{F}, \mathbb{P})$. The **distribution function of X** is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$\forall a \in \mathbb{R} \quad F_X(a) = \mathbb{P}[X \leq a]$$

Ex. 2.2. *Same example with the die. Let X be the random variable defined as above. For $a \in \mathbb{R}$ we have*

$$F_X(a) = \begin{cases} 0 & \text{if } a < -1, \\ 1/2 & \text{if } -1 \leq a < 0, \\ 2/3 & \text{if } 0 \leq a < 2, \\ 1 & \text{if } a \geq 2 \end{cases}$$

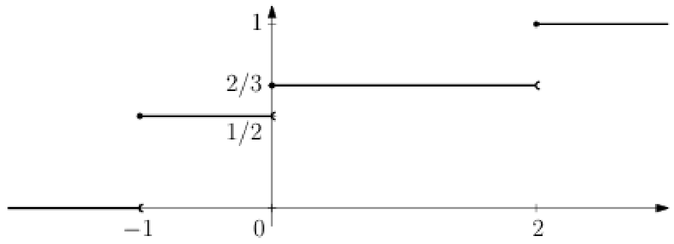


Figure 3: Graph of the distribution function F_X

Prop 2.1. (**Basic identity**). Let $a < b$ be two real numbers. Then

$$\mathbb{P}[a < X \leq b] = F(b) - F(a)$$

Prop 2.2. Let X be a r.v. on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The distribution function $F = F_X : \mathbb{R} \to [0, 1]$ of X satisfies the following properties :

i. F is nondecreasing

ii. F is right continuous[2]

iii. $\lim_{a \to -\infty} F(a) = 0$ and $\lim_{a \to \infty} F(a) = 1$

[2]i.e. $F(a) = \lim_{h \downarrow 0} F(a + h)$ for every $a \in \mathbb{R}$

2.3 Independence

Def. 2.4. Let $X_1, \ldots, X_n$ be n random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that they are **independent** if $\forall x_1, \ldots, x_n \in \mathbb{R} \quad \mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n] = \mathbb{P}[X_1 \leq x_1] \ldots \mathbb{P}[X_n \leq x_n]$.
One can show that $X_1, \ldots, X_n$ are independent iff $\forall I_1 \subset \mathbb{R}, \ldots, I_n \subset \mathbb{R}$ intervals $\{X_1 \in I_1\}, \ldots, \{X_n \in I_n\}$ are independent.

Prop 2.3. (**Grouping**). Let $X_1, \ldots, X_n$ be n independent r.v. Let $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ be some indices and $\phi_1, \ldots, \phi_k$ some functions. Then $Y_1 = \phi_1(X_1, \ldots, X_{i_1}), Y_2 = \phi_2(X_{i_1+1}, \ldots, X_{i_2}), \ldots, Y_k = \phi_k(X_{i_{k-1}+1}, \ldots, X_{i_k})$ are independent.

Def. 2.5. An infinite sequence $X_1, X_2, \ldots$ of random variables is said to be

- **independent** if $X_1, \ldots, X_n$ are independent for every n
- **independent and identically distributed** (i.i.d) if they are independent and they have the same distribution function, i.e. $\forall i, j \quad F_{X_i} = F_{X_j}$.

2.4 Transformation of random variables

We can create r.v. from other r.v. on the same probability space. For example, consider $Z_1 = \exp(X_1), Z_2 = X_1 + X_2$. Not to forget : r.v. are maps $\Omega \to \mathbb{R}$.
We can work with r.v. as if they were real numbers with the following notation :

Def. 2.6. If X is a r.v. and $\phi : \mathbb{R} \to \mathbb{R}$, then we write

$$\phi(X) := \phi \circ X$$

to to $\phi(X)$ a new mapping $\Omega \to \mathbb{R}$.
We also consider function of several variables. If $X_1, \ldots, X_n$ are n r.v. and $\phi : \mathbb{R}^n \to \mathbb{R}$, then we write

$$\phi(X_1, \ldots, X_n) := \phi \circ (X_1, \ldots, X_n)$$

2.5 Construction of random variables

Def. 2.7. Let $p \in [0, 1]$. A r.v. X is said to be a **Bernoulli r.v. with parameter p** if

$$\mathbb{P}[X = 0] = 1 - p \quad \text{and} \quad \mathbb{P}[X = 1] = p$$

In this case, we write $X \sim \text{Ber}(p)$.

Prop 2.4. (**Existence theorem of Kolmogorov**). There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an infinite sequence of r.v. $X_1, X_2, \ldots$ (on this probability space) that is an iid sequence of Bernoulli r.v. with parameter 1/2.

Prop 2.5. A r.v. U is said to be **uniform r.v. in** $[0,1]$ if its distribution function is equal to

$$F_U(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

In this csae, we write $U \sim \mathcal{U}([0,1])$.

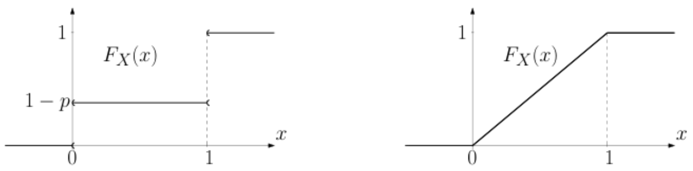


Figure 4: Left: distribution function of a Bernoulli r.v. with parameter p. Right: distribution function of a uniform r.v. in $[0,1]$.

Prop 2.6. The mapping $Y : \Omega \to [0,1]$ defined by $Y(\omega) = \sum_{n=1}^{\infty} 2^{-n} X_n(\omega)$ is a uniform r.v. in $[0,1]$.

Def. 2.8. The **generalized inverse** of F[3] is the mapping $F^{-1} : (0,1) \to \mathbb{R}$ defined by

$$\forall \alpha \in (0,1) \quad F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \ge \alpha\}$$

Int. By definition of the infimum and using right continuity of F, we have $\forall x \in \mathbb{R}$ and $\forall \alpha \in (0,1)$

$$(F^{-1}(\alpha) \le x) \iff (\alpha \le F(x))$$

Prop 2.7. (**Inverse transform sampling**). Let $F : \mathbb{R} \to [0,1]$.[4] Let U be a uniform r.v. in $[0,1]$. Then the r.v. $X = F^{-1}(U)$ has distribution $F_X = F$.

Prop 2.8. Let $F_1, F_2, \dots$ be a sequence of functions $\mathbb{R} \to [0,1]$.[5] Then there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of independent r.v. $X_1, X_2, \dots$ on this probability space s.t.

- for every i X_i has distribution function F_i (i.e. $\forall x \mathbb{P}[X_1 \le x] = F_i(x)$)
- $X_1, X_2, \dots$ are independent

[3] satifying prop. 2.2
[4] See footnote 3
[5] See footnote 3

3 Discrete and continuous r.v.

3.1 Discontinuity / continuity points of F

Prop 3.1. Let $X : \Omega \to \mathbb{R}$ be a r.v. with distribution function F. Then for every a in $\mathbb{R}$ we have

$$\mathbb{P}[X = a] = F(a) - F(a-)$$

where $F(a-) := \lim_{h \downarrow 0} F(a-h)$.

Int. Fix $a \in \mathbb{R}$

→ If F is not continuous at a point $a \in \mathbb{R}$, then the "jump size" $F(a) - F(a-)$ is equal to the probability that $X = a$

→ If F is continuous at a point $a \in \mathbb{R}$, then $\mathbb{P}[X = a] = 0$

3.2 Almost sure events

Def. 3.1. Let $A \in \mathcal{F}$ be an event. We say that A occurs **almost surely** (**a.s.**) if $\mathbb{P}[A] = 1$.

3.3 Discrete random variables

Def. 3.2. A r.v. $X : \Omega \to \mathbb{R}$ is said to be **discrete** if there exists some set $W \subset \mathbb{R}$ finite or countable s.t. $X \in W$ a.s..

Def. 3.3. Let X be a discrete r.v. taking some values in some finite or countable set $W \subset \mathbb{R}$. The **distribution of** X is the sequence of numbers $(p(x))_{x \in W}$ defined by

$$\forall x \in W \quad p(x) := \mathbb{P}[X = x]$$

Prop 3.2. The distribution $(p(x))_{x \in W}$ of a discrete r.v. satisfies $\sum_{x \in W} p(x) = 1$.

Prop 3.3. Let X be a discrete r.v. with values in a finite or countable set W almost surely, and distribution p. Then the distribution function of X is given by

$$\forall x \in \mathbb{R} \quad F_X(x) = \sum_{\substack{y \le x \\ y \in W}} p(y)$$

Int. $W = \{\text{positions of the jumps of } F_X\}$,
$p(x) =$ "height of the jump" at $x \in W$.

3.4 Examples of discrete random variables

The simplest (non constant) r.v. is the Bernoulli r.v. defined in definition 2.7.

Def. 3.4. Let $0 \le p \le 1$, let $n \in \mathbb{N}$. A r.v. X is said to be a **binomial r.v. with parameters** n **and** p if it takes values in $W = \{0, \dots, n\}$ and

$$\forall k \in \{0, \dots, n\} \quad \mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

In that case we write $X \sim \text{Bin}(n,p)$.

Prop 3.4. (**Sum of independent Bernoulli and binomail**. Let $0 \le p \le 1$, let $n \in \mathbb{N}$. Let $X_1, \dots, X_n$ be n independent Bernoulli r.v. with parameter p. Then

$$S_n := X_1 + \dots + X_n$$

is a binomial r.v. with parameter n and p.

Int. In particular, the distribution $\text{Bin}(1,p)$ is the same as the distribution $\text{Ber}(p)$. On can also check that if $X \sim \text{Bin}(m,p)$ and $Y \sim \text{Bin}(n,p)$ and X, Y are independent, then $X + Y \sim \text{Bin}(m+n, p)$.

Def. 3.5. Let $0 \le p \le 1$. A r.v. X is said to be a **geometric r.v. with parameter** p if it takes values in $W = \mathbb{N} \setminus \{0\}$ and

$$\forall k \in \mathbb{N} \quad \mathbb{P}[X = k] = (1-p)^{k-1} \cdot p$$

In this case, we write $X \sim \text{Geom}(p)$.

Prop 3.5. Let $X_1, X_2, \ldots$ be a sequence of infinitely many independent Bernoulli r.v. with parameter p. Then

$$T := \min\{n \geq 1 : X_n = 1\}$$

is a geometric r.v. with parameter p.

Prop 3.6. (**Absence of memory of the geometric distribution**). Let $T \sim \mathrm{Geom}(p)$ for some $0 < p < 1$. Then

$$\forall n \geq 0\, \forall k \geq 1 \quad \mathbb{P}[T \geq n+k | T > n] = \mathbb{P}[T \geq k]$$

Def. 3.6. Let $\lambda > 0$ be a positive real number. A r.v. X is said to be a **Poisson r.v. with parameter λ** if it takes values in $W = \mathbb{N}$ and

$$\forall k \in \mathbb{N} \quad \mathbb{P}[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

In this case, we write $X \sim \mathrm{Poisson}(\lambda)$.

Prop 3.7. (**Poisson approximation of the binomail**). Let $\lambda > 0$. For every $n \geq 1$, consider a r.v. $X_n \sim \mathrm{Bin}(n, \frac{\lambda}{n})$. Then

$$\forall k \in \mathbb{N} \quad \lim_{n \to \infty} \mathbb{P}[X_n = k] = \mathbb{P}[N = k]$$

where N is a Poisson r.v. with parameter λ.

3.5 Continous random variables

Def. 3.7. A r.v. $X : \Omega \to \mathbb{R}$ is said to be **continuous** if its distribution function F_X can be written as

$$F_X(a) = \int_{-\infty}^{a} f(x)\, dx \quad \text{for all } a \in \mathbb{R}$$

for some nonnegative function $f : \mathbb{R} \to \mathbb{R}_+$, called the **density** of X.

Int. $f(x)\, dx$ represents the probability that X takes a value in the infinitesimal interval $[x, x + dx]$.

Prop 3.8. The density f of a r.v. satisfies $\int_{-\infty}^{+\infty} f(x)\, dx = 1$.

Prop 3.9. Let X be a r.v. Assume the distribution function F_X is continuous and piecewise $\mathcal{C}^1$, i.e. that there exist $x_0 = -\infty < x_1 < \ldots < x_{n-1} < x_n = +\infty$ s.t. F_X is $\mathcal{C}^1$ on every interval $(x_i, x_{i+1}$. Then X is a continuous r.v. and a density f can be constructed by defining

$$\forall x \in (x_i, x_{i+1}) \quad f(x) = F_X'(x)$$

and setting arbitrary values at $x_1, \ldots, x_{n-1}$.

3.6 Examples of continuous random variables

Def. 3.8. A continuous r.v. X is said to be **uniform in $[a, b]$** if its density is equal to

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b], \\ 0 & x \notin [a, b] \end{cases}$$

In this case, we write $X \sim \mathcal{U}([a, b])$.

Def. 3.9. A continuous r.v. T is said to be **exponential with parameter $\lambda > 0$** if its density is equal to

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0 \end{cases}$$

In this case, we write $T \sim \mathrm{Exp}(\lambda)$.

Def. 3.10. A continuous r.v. X is said to be **normal with parameters m and $\sigma^2 > 0$** if its density is equal to

$$f_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-m)^2}{2\sigma^2}}$$

In this case, we write $X \sim \mathcal{N}(m, \sigma^2)$.

4 Expectation

4.1 Expectation for general r.v.

Def. 4.1. Let $X : \Omega \to \mathbb{R}_+$ be a r.v. with nonnegative values. The **expectation** of X is defined as

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x)) dx$$

Prop 4.1. Let X be a nonnegative r.v. Then we have $\mathbb{E}[X] \geq 0$, with equality iff $X = 0$ almost surely.

Def. 4.2. Let X be a r.v. If $\mathbb{E}[|X|] < \infty$, then the expectation of X is defined by $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$, where X_+ and X_- are the positive and negative parts of X defined by $X_+(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \geq 0, \\ 0 & \text{if } X(\omega) < 0, \end{cases}$ and $X_-(\omega) = \begin{cases} -X(\omega) & \text{if } X(\omega) \leq 0, \\ 0 & \text{if } X(\omega) > 0. \end{cases}$

4.2 Expectation of a discrete r.v.

Prop 4.2. Let $X : \Omega \to \mathbb{R}$ be a discrete r.v. with values in W (finite or countable) almost surely. We have

$$\mathbb{E}[X] = \sum_{x \in W} x \cdot \mathbb{P}[X = x]$$

provided the sum is well defined.

Prop 4.3. Let $X : \Omega \to \mathbb{R}$ be a discrete r.v. with values in W (finite or countable) almost surely. For every $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[\phi(X)] = \sum_{x \in W} \phi(x) \cdot \mathbb{P}[X = x]$$

provided the sum is well defined.

4.3 Expectation of a continuous r.v.

Prop 4.4. Let X be a contiuous r.v. with density f. Then we have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

provided the integral is well defined.

Prop 4.5. Let X be a continuous r.v. with density f. Let $\phi : \mathbb{R} \to \mathbb{R}$ be s.t. $\phi(X)$ is a r.v. Then we have

$$\mathbb{E}[\phi(X)] = \int_{-\infty}^{\infty} \phi(x) f(x) dx$$

provided the integral is well defined.

4.4 Calculus

Prop 4.6. (**Linearity of the expectation**). Let $X, Y : \Omega \to \mathbb{R}$ be r.v.'s, let $\lambda \in \mathbb{R}$. Provided the expectations are well defined, we have

1. $\mathbb{E}[\lambda \cdot X] = \lambda \cdot \mathbb{E}[X]$
2. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Prop 4.7. Let X, Y be two r.v. If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

4.5 Tailsum formulas

Prop 4.8. (**Tailsum formula for nonnegative r.v.'s**). Let X be a r.v., s.t. $X \geq 0$ almost surely. Then we have $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]\,dx$.

Prop 4.9. (**Tailsum formula for discrete r.v.'s**). Let X be a discrete r.v. taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$. Then $\mathbb{E}[X] = \sum_{n=1}^\infty \mathbb{P}[X \geq n]$.

4.6 Characterizations via expecations

Prop 4.10. Let X be a r.v. Let $f : \mathbb{R} \to \mathbb{R}$ s.t. $\int_{-\infty}^{+\infty} f(x)dx = 1$. then the following are equivalent :

i. X is continuous with density f,

ii. For every function $\phi : \mathbb{R} \to \mathbb{R}$ piecwise continuous, bounded : $\mathbb{E}[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)f(x)dx$

Prop 4.11. Let X, Y be two discrete r.v.'s. Then the following are equivalent

i. X, Y are independent

ii. For every $\phi : \mathbb{R} \to \mathbb{R}$, $\psi : \mathbb{R} \to \mathbb{R}$ piecewise contiuous, bounded : $\mathbb{E}[\phi(X)\psi(Y)] = \mathbb{E}[\phi(X)]\,\mathbb{E}[\psi(Y)]$.

Prop 4.12. Let $X_1, \ldots, X_n$ be n r.v.'s. Then the following are equivalent

i. $X_1, \ldots, X_n$ are independent

ii. For every $\phi_1 : \mathbb{R} \to \mathbb{R}, \ldots, \phi_n : \mathbb{R} \to \mathbb{R}$ piecewise continuous, bounded : $\mathbb{E}[\phi_1(X_1) \cdots \phi_n(X_n)] = \mathbb{E}[\phi_1(X_1)] \cdots \mathbb{E}[\phi_n(X_n)]$.

4.7 Inequalities

Prop 4.13. (**Monotonicity**). Let X, Y be two r.v.'s s.t. $X \leq Y$ a.s. Then $\mathbb{E}[X] \leq \mathbb{E}[Y]$, provided the two expectations are well defined.

Prop 4.14. (**Markov's inequality**). Let X be a nonnegative r.v. Then for every $a > 0$, we have

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

Prop 4.15. (**Jensen's inequality**). Let X be a r.v. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathbb{E}[\phi(X)]$ and $\mathbb{E}[X]$ are well defined, then

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$$

4.8 Variance

Def. 4.3. Let X be a variable s.t. $\mathbb{E}\left[X^2\right] < \infty$. The **variance of X** is defined by

$$\sigma_X^2 = \mathbb{E}\left[(X - m)^2\right], \quad \text{where } m = \mathbb{E}[X]$$

The square root σ_X of the variance is called the **standard deviation of X**.

Prop 4.16. Let X be a r.v. s.t. $\mathbb{E}\left[X^2\right] < \infty$. Then for every $a \geq 0$ we have

$$\mathbb{P}[|X - m| \geq a] \leq \frac{\sigma_X^2}{a^2}, \quad \text{where } m = \mathbb{E}[X]$$

Prop 4.17. (**Basic properties of the variance**).

1. Let X be a r.v. with $\mathbb{E}\left[X^2\right] < \infty$. Then $\sigma_X^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

2. Let X be a r.v. with $\mathbb{E}\left[X^2\right] < \infty$, let $\lambda \in \mathbb{R}$. Then $\sigma_{\lambda X}^2 = \lambda^2 \cdot \sigma_X^2$.

3. Let $X_1, \ldots, X_n$ be n pairwise independent r.v.'s and $S = X_1 + \ldots + X_n$. Then $\sigma_S^2 = \sigma_{X_1}^2 + \ldots + \sigma_{X_n}^2$.

4.9 Covariance

Def. 4.4. Let X, Y be two r.v.'s. Assume that $\mathbb{E}\left[X^2\right] < \infty$ and $\mathbb{E}\left[Y^2\right] < \infty$ (finite second moment). We define the **covariance between X and Y** as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

Int. With X and Y independent : $\mathrm{Cov}(X, Y) = 0$.

5 Joint distribution

5.1 Discrete joint distributions

Def. 5.1. Let $X_1, \ldots, X_n$ be n discrete r.v.'s with $X_i \in W_i$ almost surely, for some $W_i \subset \mathbb{R}$ finite or countable. The **joint distribution** of $(X_1, \ldots, X_n)$ is the collection $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$ defined by

$$p(x_1, \ldots, x_n) = \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]$$

Prop 5.1. The joint distribution of some r.v.'s $X_1, \ldots, X_n$ satisfies $\sum_{x_1 \in W_1, \ldots, x_n \in W_n} p(x_1, \ldots, x_n) = 1$.

Prop 5.2. Let $n \geq 1$ and $\phi : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary function. Let $X_1, \ldots, X_n$ be n discrete r.v.'s on $(\Omega, \mathcal{F}, \mathbb{P})$ with respective values in some finite or countable sets $W_1, \ldots, W_n$ a.s. Then $Z = \phi(X_1, \ldots, X_n)$ is a discrete r.v. with values in $W = \phi(W_1 \times \ldots \times W_n)$ a.s. and with distribution given by

$$\forall z \in W \quad \mathbb{P}[Z = z] = \sum_{\substack{x_1 \in W, \ldots, x_n \in W_n \\ \phi(x_1, \ldots, x_n) = z}} \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]$$

Prop 5.3. (**Marginal distributions**). Let $X_1, \ldots, X_n$ be n discrete r.v.'s with joint distribution $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$. For every i, we have $\forall z \in W_i \mathbb{P}[X_1 = z] = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} p(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$

Prop 5.4. (**Expectation of the image**). Let $X_1, \ldots, X_n$ be n discrete r.v.'s with joint distribution $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$. Let $\phi : \mathbb{R}^n \to \mathbb{R}$, then

$$\mathbb{E}[\phi(X_1, \ldots, X_n)] = \sum_{x_1, \ldots, x_n} \phi(x_1, \ldots, x_n)p(x_1, \ldots, x_n)$$

whenever the sum is well-defined.

Prop 5.5. (**Independence**). Let $X_1, \ldots, X_n$ be n discrete r.v.'s with joint distribution $p = p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$. The following are equivalent

i. $X_1, \ldots, X_n$ are independent

ii. $p(x_1, \ldots, x_n) = \mathbb{P}[X_1 = x_1] \cdots \mathbb{P}[X_n = x_n]$ for every $x_i \in W_i, \ldots, x_n \in W_n$

5.2 Continuous joint distribution

Def. 5.2. Let $n \geq 1$, some r.v.'s $X_n, \ldots, X_n : \Omega \to \mathbb{R}$ have a **continuous joint distribution** if there exists a function $f : \mathbb{R}^n \to \mathbb{R}_+$ s.t. $\mathbb{P}[X_1 \leq a_1, \ldots, X_n \leq a_n]$

$$= \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f(x_1, \ldots, x_n) dx_n \cdots dx_1$$

for every $a_1, \ldots, a_n \in \mathbb{R}$. A function f as above is called a **joint density of (X, Y)**.

Int. $f(x_1, \ldots, x_n)dx_1 \cdots dx_n$ represents the probability that the random vector $(X_1, \ldots, X_n)$ lies in the small region $[x_1, x_1 + dx_1] \times \ldots \times [x_n, x_n + dx_n]$.

Prop 5.6. (**Expectation of the image**). Let $\phi : \mathbb{R}^n \to \mathbb{R}$. If $X_1, \ldots, X_n$ have joint density f, then the expectation of the r.v. $Z = \phi(X_1, \ldots, X_n)$ can be calculated by the formula $\mathbb{E}[\phi(X_1, \ldots, X_n)]$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(x_1, \ldots, x_n) \cdot f(x_1, \ldots, x_n) dx_n \cdots dx_1$$

5.3 Marginal densities

Prop 5.7. Let $X_1, \ldots, X_n$ be n r.v.'s with a joint density $f = f_{X_1, \ldots, X_n}$. Then for every i, X_i is a continuous r.v. with density f_i given by $f_i(z)$

$$= \int_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathbb{R}^{n-1}} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$$

$$dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

Prop 5.8. (**Independence for continuous r.v.'s**). Let $X_1, \ldots, X_n$ be n continuous r.v.'s with respective densities $f_1, \ldots, f_n$. The following are equivalent

i. $X_1, \ldots, X_n$ are independent

ii. $X_1, \ldots, X_n$ are jointly continuous with joint density $f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$

6 Asymptotic results

For this section, fix a probability space $(\Omega, \mathcal{F}, \mathbb{P}$ and an infinite sequence of i.i.d. r.v.'s $X_1, X_2, \ldots$. For every n, consider the partial sum $S_n = X_1 + \ldots + X_n$.

Def. 6.1. The r.v. defined by $\frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$ (when n is large) is called the **empirical average**.

6.1 Law of large numbers

Prop 6.1. Assume that $\mathbb{E}[|X_1|] < \infty$. Defining $m = \mathbb{E}[X_1]$ we have $\lim_{n \to \infty} \frac{X_1 + \ldots + X_n}{n} = m$ a.s.

6.2 Monte-Carlo integration

6.3 Convergence in distribution

Def. 6.2. Let $(X_n)_{n \in \mathbb{N}}$ and X be some r.v.'s. We write $X_n \approx X$ as $n \to \infty$, if for every $x \in \mathbb{R}$: $\lim n \to \infty \mathbb{P}[X_n \leq x] = \mathbb{P}[X \leq x]$

6.4 Central limit theorem

Prop 6.2. (**Central limit theorem**). Assume that $\mathbb{E}\left[X_1^2\right]$ is well defined and finite. Defining $m = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$, we have

$$\mathbb{P}\left[\frac{S_n - n \cdot m}{\sqrt{\sigma^2 n}} \leq a\right] \underset{n \to \infty}{\longrightarrow} \Phi(a) = \frac{1}{\sqrt{2\phi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

for every $a \in \mathbb{R}$.

Statistics

1 Basic concepts of an estimator

We wish to estimate an unknown parameter θ based on a sample $X_1, X_2, \ldots, X_n$.

1.1 Definition of an estimator

Def. 1.1. An **estimator** is a r.v. $T : \Omega \to \mathbb{R}$ of the form

$$T = t(X_1, X_2, \ldots, X_n)$$

where $t : \mathbb{R}^n \to \mathbb{R}$ is a measurable function. Inserting the observed data

$$x_1, x_2, \ldots, x_n \text{ with } x_i = X_i(\omega))$$

yields the **estimate** $t(x_1, \ldots, x_n)$ for θ.

Def. 1.2.

1. **Last observation estimator** : $T^{(1)} = X_n$
2. **Sample mean estimator** : $T^{(2)} = \frac{1}{n} \sum_{i=1}^{n} X_i$

1.2 Bias and mean squared error

The estimator T is a random variable which distribution (under $\mathbb{P}_\theta$) depends on the unknown parameter θ.

Def. 1.3. An estimator T is called **unbiased** for θ if, for all $\theta \in \Theta$:

$$\mathbb{E}_\theta[T] = \theta$$

Def. 1.4. For $\theta \in \Theta$, the **bias** of an estimator T is defined as

$$\mathrm{Bias}_\theta(T) = \mathbb{E}_\theta[T] - \theta$$

The **mean squared error** (MSE) is defined as

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta\left[(T - \theta)^2\right]$$

For an unbiased estimator, the MSE equals the variance :

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta[T] + (\mathrm{Bias}_\theta(T))^2$$

1.3 Maximum likelihood estimation (MLE)

Def. 1.5. For the observed sample $(x_1, \ldots, x_n)$, the **likelihood** function is defined by

$$L(x_1, \ldots, x_n; \theta) = \begin{cases} \prod_{i=1}^{n} p_{X_i}(x_i; \theta), & \text{if } X_i \text{ are discrete} \\ \prod_{i=1}^{n} f_{X_i}(x_i; \theta), & \text{if } X_i \text{ are continuous} \end{cases}$$

Def. 1.6. The **maximum likelihood estimator** of θ is defined as

$$\hat{\theta}(x_1, \ldots, x_n) \in \arg\max_{\theta\in\Theta} L(x_1, \ldots, x_n; \theta)$$

In practice, one maximises the log-likelihood function $l(\theta; x_1, \ldots, x_n) = \log L(x_1, \ldots, x_n; \theta)$, and then obtains the estimator by replacing the data with the random variables :

$$T_{ML} = t_{ML}(X_1, \ldots, X_n)$$

1.3.1 Application of the method

The maximum likelihood method is a way to systematically determine an estimator.

1. Find the joint density / distribution of the random variables
2. Determine the **log-likelihood function** from it : $f(\theta) := \ln(L(x_1, \ldots, x_n; \theta))$
3. Differentiate $f(\theta)$ with respect to θ
4. Find the zero(s) of $f'(\theta)$
5. Show that $f''(\theta) < 0$ or use another argument to demonstrate that a maximum has been found (possibly check boundary points)

1.4 Models with multiple parameters

Consider the parameter space $\Theta \subset \mathbb{R}^m$, where m is the number of parameters. The stochastic model is given by a family of probability measures $(P_\theta)_{\theta in \Theta}$, and our goal is to estimate the vector

$$\theta = (\theta_1, \theta_2, \ldots, \theta_m).$$

All previous definitions extend to this setting.

2 Confidence intervals

2.1 Definition

Def. 2.1. Let $\alpha \in [0, 1]$. A **confidence interval** for θ with confidence level $1 - \alpha$ is a random interval $I = [A, B]$, with endpoints $A = a(X_1, \ldots, X_n)$, $B = b(X_1, \ldots, X_n)$, where $a, b : \mathbb{R}^n \to \mathbb{R}$, s.t. for all $\theta \in \Theta$

$$P_\theta[A \le \theta \le B] \ge 1 - \alpha$$

2.2 Distribution statements

Def. 2.2. A continuous r.v. X is said to be **chi-squared distributed** with m degrees of freedom if its density is given by

$$f_X(y) = \frac{1}{2^{m/2}\Gamma(m/2)} y^{\frac{m}{2}-1} e^{-y/2}, \quad y \ge 0$$

Where $\Gamma(v) = \int_0^\infty t^{v-1}e^{-t}dt$ and for $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$
We write $X \sim \chi^2_m$.

Prop 2.1. (**Sum of squares theorem**). If $X_1, X_2, \ldots, X_m$ are iid $\sim N(0, 1)$, then

$$Y = \sum_{i=1}^m X_i^2 \sim \chi^2_m$$

Def. 2.3. A continuous r.v. X is said to be **t-distributed** with m degrees of freedom if its density is given by

$$f_X(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad x \in \mathbb{R}$$

We write $X \sim t_m$.

Prop 2.2. Let X and Y be independent r.v. with $X \sim N(0, 1)$ and $Y \sim \chi^2_m$. Then the quotient

$$Z := \frac{X}{\sqrt{Y/m}}$$

is t-distributed with m degrees of freedom.

2.3 Normal model with unknown variance and mean

Def. 2.4. **Sample mean** : $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$
Sample variance : $S^2 = \frac{1}{n-1}\sum_{i+1}^n (X_i - \bar{X}_n)^2$

Prop 2.3. If $X_1, \ldots, X_n$ are iid $\sim N(m, \sigma^2)$, then $\bar{X}_n$ and S^2 are independent.

2.4 Approximate confidence intervals

A general approximate approach is provided by the central limit theorem (CLT). Often, an estimator T is a function of a sum, say, $T = \frac{1}{n}\sum_{i=1}^n Y_i$. By the CLT for large n,

$$\sum_{i=1}^n Y_i \approx N\left(n\mathbb{E}\left[Y_i\right], n\mathrm{Var}[Y_i]\right)$$

which can be used to approximate the distribution T and hence to construct approximate confidence intervals.

3 Tests

3.1 Null and alternative hypotheses

Starting with a sample $X_1, \ldots, X_n$, consider a family of probability measures P_θ with $\theta \in \Theta$ that describes our possible models. The basic problem is to decide between two classes of models - namely, the null hypothesis and the alternative hypothesis.
One sets

- **Null hypothesis** $H_0 : \theta \in \Theta_0$
- **Alternative hypothesis** $H_A : \theta \in \Theta_A$

with $\Theta_0 \cap \Theta_A = \emptyset$ (default : $\Theta_A = \Theta \setminus \Theta_0$). When Θ_0 or Θ_A consists of a single value θ_0 or θ_A, they are called **simple**, otherwise they are called **composite**.

3.2 Tests and decisions

Def. 3.1. A **test** is a pair (T, K), where

- T is a statistic of the form $T = t(X_1, \ldots, X_n)$ (the **test statistic**) and
- $K \subset \mathbb{R}$ is a (deterministic) set, called the **critical region** (or **rejection region**).

A statistical test enables us to systematically accept or reject the null hypothesis. We first compute the test statistic $T(\omega) = t(X_1(\omega), \ldots, X_n(\omega))$ and then follow the decision rule : reject H_0 if $T(\omega) \in K$ and don't reject H_0 if $T(\omega) \notin K$.

There are two types of errors :

1. A **Type I error** occurs when the null hypothesis is wrongly rejected even though it is true. Probability : $P_\theta[T \in K]$, for $\theta \in \Theta_0$.
2. A **Type II error** occurs when the null hypothesis is not rejected even though it is false. Probability : $P_\theta[T \notin K] = 1 - P_\theta[T \in K]$ for $\theta \in \Theta_A$.

3.3 Significance level and power

Def. 3.2. Let $\alpha \in (0,1)$. A test (T,K) is said to have **significance level** α if for all $\theta \in \Theta_0$

$$P_\theta[T \in K] \leq \alpha$$

Int. The significance level is the probability you're willing to accept of making a wrong decision, specifically rejecting the null hypothesis when it's actually true.

Def. 3.3. The **power** of a test (T,K) is the function

$$\beta : \Theta_A \to [0,1], \quad \theta \mapsto \beta(\theta) := P_\theta[T \in K]$$

Int. The power is the "strength" of the test to avoid failing to detect an effect when one actually exists.

3.4 Construction of tests

Assume $\theta_0 \neq \theta_A$ are two fixed numbers. Assume both the null hypothesis and alternative hypothesis are simple, i.e. $H_0 : \theta = \theta_0 \quad H_A : \theta = \theta_A$. Assume r.v. $X_1, \dots, X_n$ are either jointly discrete or jointly continuous under both P_{θ_0} and P_{θ_A}. In particular, $L(x_1, \dots, x_n; \theta)$ is well-defined for $\theta = \theta_0$ and $\theta = \theta_A$.

Def. 3.4. For every $x_1, \dots, x_n$, the **likelihood ratio** is defined by

$$R(x_1, \dots, x_n) := \frac{L(x_1, \dots, x_n; \theta_A)}{L(x_1, \dots, x_n; \theta_0)}$$

By convention, if $L(x_1, \dots, x_n; \theta_0) = 0$, we set the ratio to $+\infty$.

Int. A large ratio indicates that the observations $x_1, \dots, x_n$ are far more likely under the alternative P_{θ_A} than under the null P_{θ_0}. Hence it makes sense to define the test statistic as $T := R(X_1, \dots, X_n)$ and the critical region as $K := (c, \infty)$, for some constant c.

Def. 3.5. Let $c \geq 0$. The **likelihood ratio test** with parameter c is the test (T,K) where $T = R(X_1, \dots, X_n)$ and $K = (c, \infty)$. It is optimal : no test have lower power while no greater significance level (**Neyman-Pearson Lemma**).

Def. 3.6. For composite hypotheses, the **generalized likelihood ratio** can be defined as

$$R(x_1, \dots, x_n) := \frac{\sup_{\theta \in \Theta_A} L(x_1, \dots, x_n; \theta)}{\sup_{\theta \in \Theta_0} L(x_1, \dots, x_n; \theta)}$$

and we choose $T := R(X_1, \dots, X_n)$ with $K = (c_0, \infty)$ where c_0 is chosen s.t. the test has the preassigned significance level.

3.5 The p-value

Let $X_1, \dots, X_n$ be a sample of size n. We wish to test a hypothesis $H_0 : \theta = \theta_0$ against an alternative $H_A : \theta \in \Theta_A$.

Def. 3.7. A **family of tests** $(T, (K_t)_{t \geq 0})$ is said to be **ordered** with respect to the test statistic T if for all $s, t \geq 0$

$$s \leq t \implies K_s \subset K_t$$

Typical examples are :

$K_t = (t, \infty)$ (right-tailed test), $K_t = (-\infty, -t)$ (left-tailed test),

or $\quad K_t = (-\infty, -t) \cup (t, \infty)$ (two-sided test).

Def. 3.8. Let $H_0 : \theta = \theta_0$ be a simple null hypothesis and let $(T, (K_t)_{t \geq 0})$ be an ordered family of tests. The **p-value** is defined as the r.v. p-value $= G(T)$, where the function $G : \mathbb{R}^+ \to [0,1]$ is given by

$$G(t) = P_\theta[T \in K_t]$$

Int. The p-value informs us which tests in our family would lead to rejection of H_0. If the observed p-value is p, then every test with significance level $a > p$ would reject H_0 and those with $a \leq p$ would not. The p-value doesn't depend on the alternative hypothesis.