Basics of Statistical Inference with R

SUPPORTING MATERIALS

Antonio Calcagnì

DPSS, University of Padova
GNCS, INdAM

Doctoral School in Psychological Sciences
Fall 2022

Random experiments

A **random experiment** is an experiment whose outcomes cannot be determined in advance. Whereas the set of all possible outcomes (**sample space** Ω) can distinctly be determined (there is no fuzziness in this step), what is affected by uncertainty is the occurrence of an **event** of the sample space.

The most typical example is the experiment where a (fair) coin is tossed a number of times (e.g., three times). In this case,

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

where THT means that the first toss is tail, the second is head, and the the third is tail.

Random experiments

Subsets $A_1, \ldots, A_K$ of Ω are called **events**. For instance, the event that the first toss is tail is $A = \{THH, THT, TTH, TTT\}$. An event A *is said to occur* if an element $a \in A$ (e.g., $a = \{THH\}$) is the outcome of the experiment.

Since events are sets, we can use the **standard set operations** to perform computation on random events.

Given two events A_k, A_h ($k \neq h$):

- $A_k \cup A_h$ (union: the event that A or B or both occur)
- $A_k \cap A_h$ (intersection: the event that A and B both occur)
- A^c (complement: the event that A does not occur)
- $A_k \cap A_h = \emptyset$ (disjoint event)

Random experiments

The **probability** $\mathbb{P}$ of an event A is a measure such that

P1 $\mathbb{P}(A_k) \in [0, 1]$

P2 $\mathbb{P}(\bigcup_{k=1}^{K} A_k) = \sum_{k=1}^{K} \mathbb{P}(A_k) = 1$

P1 states that $\mathbb{P}(A_k) = 0$ indicates that A_k does not occur certainly whereas P2 gives a calculus for the total probability of disjoint events.

Random experiments

There are two ways to assign probability at least:

- **classic**: the probability of A is given by computing the elementary events which have been occurred during an experiment

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

- **frequentist**: the probability of A is given as the limiting frequency after a sequence of n independent attempts

$$\mathbb{P}(A) = \lim_{n \to \infty} \frac{f_A}{n}$$

where f_A is the empirical frequency for the event A

Random experiments

It should be noted that the *classic* approach to computing probabilities is performed **before** the random experiment is done whereas the *frequentist* (or *empirical*) approach is performed **after** the experiment is done and it requires that the experiment can be repeated infinitely many times.

For example, the probability of the event $A = \{THH, THT, TTH, TTT\}$ is

$$\mathbb{P}(A) = |A| \ / \ |\Omega| = 4/2^3 = 1/2$$

according to the classic approach to probability.

Random experiments

Consider two events A_k and A_h ($k \neq h$). Then, by fixing one of the two terms (e.g., A_h) we may ask whether knowing A_h changes the probability of A_k:

$$\mathbb{P}(A_k|A_h) = \frac{\mathbb{P}(A_k \cap A_h)}{\mathbb{P}(A_h)}$$

This is known as **conditional probability**. When $\mathbb{P}(A_k \cap A_h) = \emptyset$ then $\mathbb{P}(A_k|A_h) = \mathbb{P}(A_k)$ and knowing $\mathbb{P}(A_h)$ does not affect the probability of $\mathbb{P}(A_k)$.

For instance, suppose we toss a fair coin three times. Let A_h be the event that the total number of heads is two and let A_k be the event that the first toss is heads. Then $\mathbb{P}(A_k|A_h) = (2/8)/(3/8) = 2/3 = 0.67$.

Random experiments

The conditional probability also provides a calculus for the joint probability of $A_k \cap A_h$:

$$\mathbb{P}(A_k \cap A_h) = \mathbb{P}(A_k|A_h)\mathbb{P}(A_h)$$

which can be generalized for a sequence of events:

$$\mathbb{P}(A_1 \cap \ldots \cap A_K) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1) \cdots \mathbb{P}(A_k|A_{k-1} \cap \ldots \cap A_1) \cdots \mathbb{P}(A_K|A_{K-1} \cap \ldots \cap A_1)$$

Two events A_k and A_h ($k \neq h$) are said to be **independent** when $\mathbb{P}(A_k|A_h) = \mathbb{P}(A_k)$ or $\mathbb{P}(A_k \cap A_h) = \mathbb{P}(A_k)\mathbb{P}(A_h)$. Independence models the *lack of information between events* and it is often a model assumption.

In the general case of independence:

$$\mathbb{P}(A_1 \cap \ldots \cap A_K) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_k) \cdots \mathbb{P}(A_K)$$

Random variables

Very often defining probability spaces for some interesting empirical phenomena is difficult. Sometimes it is also unnecessary as we are only interested in particular outcomes of the random experiment. To this end, random variables can circumvent these issues by introducing parametric classes of probabilistic models.

Random variables

Non-formal definition: A **random variable** X is a function that maps subsets of the sample space Ω (or subsets of the event space $\mathcal{A}$, the σ-algebra associated to Ω) to real numbers.

The **support** of X - i.e. $\sup(X)$ - is the set of values that X may assume. For discrete random variables, $\sup(X)$ is countable finite (e.g., discrete subset of real numbers). For real random variables, $\sup(X)$ is infinite.

Random variables can be **univariate** (e.g., $\sup(X) \subset \mathbb{R}$), **bivariate** (e.g., $\sup(X) \subset \mathbb{R} \times \mathbb{R}$), or more generally **multivariate** (e.g., $\sup(X) \subset \mathbb{R} \times \overset{J-1}{\ldots} \times \mathbb{R}$).

Note: the adjective *random* indicates that we are dealing with random experiments (the function X is not random per-sé).

Random variables

Example: Fair coin tossed $n = 3$ times

- $\Omega = \{ttt, ttc, tct, ctt, ccc, cct, ctc, tcc\}$, $|\Omega| = 2^n$
- $\mathbb{P}$ defined according to the classic assignment: $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$
- $X \overset{\text{def}}{=}$ "number of heads", $\sup(X) = \{0, 1, 2, 3\}$

$$\begin{aligned} X = 0 &\iff \{ccc\} \\ X = 1 &\iff \{ctc, tcc, cct\} \\ X = 2 &\iff \{ttc, tct, ctt\} \\ X = 3 &\iff \{ttt\} \end{aligned}$$

Random variables

Example: Fair coin tossed $n = 3$ times

- $\Omega = \{ttt, ttc, tct, ctt, ccc, cct, ctc, tcc\}$, $|\Omega| = 2^n$
- $\mathbb{P}$ defined according to the classic assignment: $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$
- $X \stackrel{\text{def}}{=}$ "number of heads", $\quad \sup(X) = \{0, 1, 2, 3\}$

$$
\begin{array}{lcll}
X = 0 & \iff & \{ccc\} & \mathbb{P}(X = 0) = 1/8 \\
X = 1 & \iff & \{ctc, tcc, cct\} & \mathbb{P}(X = 1) = 3/8 \\
X = 2 & \iff & \{ttc, tct, ctt\} & \mathbb{P}(X = 2) = 3/8 \\
X = 3 & \iff & \{ttt\} & \mathbb{P}(X = 3) = 1/8
\end{array}
$$

Note: $\sup(X)$ can be considered the new sample space over which $\mathbb{P}$ assigns probabilities.

Random variables

The probabilities $\mathbb{P}(X = x)$ induced by a random variable give rise to the **distribution function** F_X. Depending on if X is discrete or continuous, probability distribution can be discrete or continuous too.

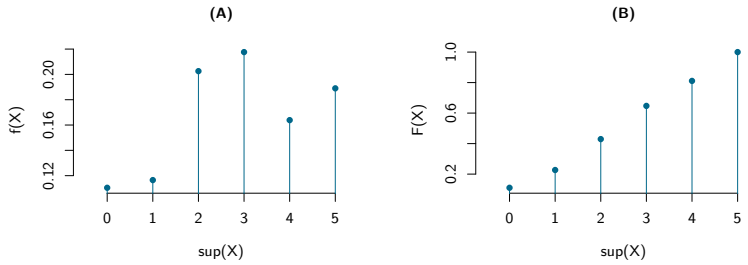
F_X defines the way probabilities can be computed by means of random variables:

$$
\begin{aligned}
F_X(X = x) &= \mathbb{P}(X \leq x) \quad x \in \sup(X) \\
F_X(X \in [a, b]) &= \mathbb{P}(a \leq X \leq b) \quad x \in \sup(X)
\end{aligned}
$$

From F_X we can derive continuous or discrete **density functions** $f_X(X = x)$ or $f_X(X \in [a, b])$. In general, $f_X(x)$ satisfies the axioms of probabilities:

- $f_X(x \in [x_0, x_0 + \epsilon]) = \int_{x_0}^{x_0+\epsilon} f_X(x)\ dx \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$

Random variables

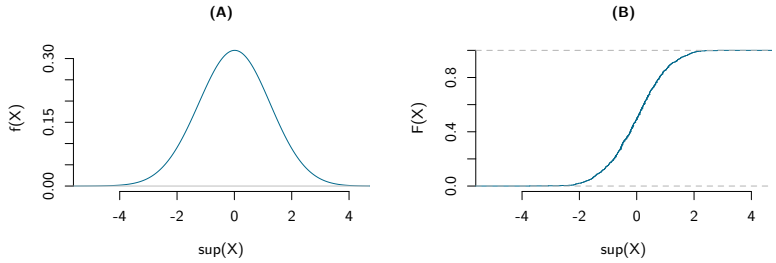


(A) Discrete density function (aka, *probability mass function*)
(B) Discrete distribution function (aka, *cumulative probability mass function*)

$f(X = x) \geq 0$

$\sum_{x \in \sup(X)} f(X = x) = 1$

Random variables



(A) Continuous density function
(B) Continuous distribution function (aka, *cumulative density function*)

$f(x_0 \leq X \leq x_0 + \epsilon) = \int_{x_0}^{x_0+\epsilon} f(x)dx \geq 0 \qquad (\epsilon > 0)$

$\int_{-\infty}^{+\infty} f(x)dx = 1$

$f(x_0) = 0$

Random variables

Random variables allows for using the same probabilistic models to represent different random experiments.

For instance, the Binomial random model can be used to formalize the experiments of drawing marbles from an urn or the experiment of purchasing a given product from a finite set of choices. Similarly, the Normal random model can be used to represent the measurement error of a physical as well as psychophysic experiments.

Random variables

The distribution function F_X can be parameterized by some reals $\boldsymbol{\theta}$ called **parameters** that modify the way it assigns probabilities. The mathematical domain where the parameters lie is called **parameter space**.

In general $X \sim F_X(x;\boldsymbol{\theta})$ is used to signify that X has distribution function F_X parameterized by $\boldsymbol{\theta}$.

The class of parameterized distribution functions will be called **parametric probabilistic models**. Depending on the support of X we may then have discrete parametric models as well as continuous parametric models.

Random variables

Some univariate **discrete probabilistic models**

Model	Notation	sup(X)	$\boldsymbol{\theta}$	f_X
Bernoulli	$\mathcal{B}er(x;\pi)$	$\{0,1\}$	$\pi \in [0,1]$	$\pi^x(1-\pi)^{1-x}$
Binomial	$\mathcal{B}in(x;n,\pi)$	$\mathbb{N}_0$	$n \in \mathbb{N}$, $\pi \in [0,1]$	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$
Poisson	$\mathcal{P}oi(x;\lambda)$	$\mathbb{N}_0$	$\lambda \in \mathbb{R}^+$	$\frac{\lambda^x}{x!}\exp(-\lambda)$
Geometric	$\mathcal{G}(x;\pi)$	$\mathbb{N}$	$\pi \in [0,1]$	$\pi(1-x)^{x-1}$
Multinomial	$\mathcal{M}ulti(\mathbf{x};n,\boldsymbol{\pi})$	$\mathbb{N}_0$	$n \in \mathbb{N}$, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$, $\boldsymbol{\pi}^T\mathbf{1}_K = 1$	$\binom{n!}{x_1!\cdots x_K!}\pi_1^{x_1}\cdots\pi_K^{x_K}$

Random variables

Some univariate **continuous probabilistic models**

Model	Notation	sup(X)	$\boldsymbol{\theta}$	f_X
Normal	$\mathcal{N}(x;\mu,\sigma^2)$	$\mathbb{R}$	$\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$	$\left(\sigma\sqrt{2\pi}\right)^{-1}\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$
Uniform	$\mathcal{U}(x;\alpha,\beta)$	$[\alpha,\beta] \subset \mathbb{R}$	$\alpha \in \mathbb{R}$ $\beta \in \mathbb{R}$ $\alpha < \beta$	$\frac{1}{\beta-\alpha}$
Exponential	$\mathcal{E}xp(x;\lambda)$	$\mathbb{R}^+$	$\lambda \in \mathbb{R}$	$\lambda\exp(-\lambda x)$
Beta	$\mathcal{B}eta(x;\alpha,\beta)$	$[0,1] \subset \mathbb{R}$	$\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}^+$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$
Chi-square	$\chi^2(x;\nu)$	$\mathbb{R}^+$	$\nu \in \mathbb{N}$	$\left(2^{\nu/2}\Gamma(\nu/2)\right)^{-1}x^{\nu/2-1}\exp(-x/2)$

Random variables

When using random variables it is useful to consider various characteristics (e.g., position, dispersion, shape) that can be summarized numerically.

Expectation. It is denoted by $\mathbb{E}[X]$ and quantifies the mean value to which a sequence of random experiments is expected to converge:

$$\mathbb{E}[X] = \sum_{x \in \text{sup}(X)} x f_X(x) \quad \text{(discrete case)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{(continuous case)}$$

Random variables

Variance. It is denoted by $\mathbb{V}\text{ar}[X]$ and quantifies the dispersion of the outcomes of a sequence of random experiments:

$$\mathbb{V}\text{ar}[X] = \sum_{x \in \text{sup}(X)} (x - \mathbb{E}[X])^2 f_X(x) \quad \text{(discrete case)}$$

$$\mathbb{V}\text{ar}[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x)\, dx \quad \text{(continuous case)}$$

Random variables

For two (or more) random variables $X_1, \ldots, X_J$ an important characteristic to be calculated is the **covariance**, which summarizes the joint variability of the involved r.vs.

Given a pair X_h, X_k ($h \neq k$), we have:

$$\begin{aligned} \mathbb{C}\text{ov}[X_h, X_k] &= \mathbb{E}[(X_h - \mu_{X_h})(X_k - \mu_{X_k})] \\ &= \mathbb{E}[X_h X_k] - \mu_{X_h}\mu_{X_k} \end{aligned}$$

where in general $\mu_X = \mathbb{E}[X]$. The covariance offers a measure of linear association between X_h and X_k. In particular:

- $\mathbb{C}\text{ov}[X_h, X_k] > 0$ indicates that X_h and X_k are positively associated
- $\mathbb{C}\text{ov}[X_h, X_k] < 0$ indicates that X_h and X_k are negatively associated
- $\mathbb{C}\text{ov}[X_h, X_k] = 0$ indicates that X_h and X_k are not *linearly* associated

Random variables

Expectations for some important distributions

Model	Notation	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$
Bernoulli	$\mathcal{B}er(x; \pi)$	π	$\pi(1-\pi)$
Binomial	$\mathcal{B}in(x; n, \pi)$	$n\pi$	$n\pi(1-\pi)$
Poisson	$\mathcal{P}oi(x; \lambda)$	λ	λ
Geometric	$\mathcal{G}(x; \pi)$	$\frac{1}{\pi}$	$\frac{1-\pi}{\pi^2}$
Multinomial	$\mathcal{M}ulti(\mathbf{x}; n, \boldsymbol{\pi})$	$n\pi_1, \ldots, n\pi_J$	$n\pi_1(1-\pi_1), \ldots, n\pi_J(1-\pi_J)$

Random variables

Expectations for some important distributions

Model	Notation	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$
Normal	$\mathcal{N}(x;\mu,\sigma^2)$	μ	σ^2
Uniform	$\mathcal{U}(x;\alpha,\beta)$	$\frac{1}{2}(\alpha+\beta)$	$\frac{1}{12}(\beta-\alpha)^2$
Exponential	$\mathcal{E}xp(x;\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Beta	$\mathcal{B}eta(x;\alpha,\beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Chi-square	$\chi^2(x;\nu)$	ν	2ν

Random variables

Some important properties for expectations

$\mathbb{E}[\alpha] = \alpha$

$\mathbb{E}[X_h + X_k] = \mathbb{E}[X_h] + \mathbb{E}[X_k]$

$\mathbb{E}[\beta X_h] = \beta\mathbb{E}[X_h]$

$\mathbb{E}[\alpha + \beta X_h] = \alpha + \beta\mathbb{E}[X_h]$

$\mathbb{V}\text{ar}[\alpha] = 0$

$\mathbb{V}\text{ar}[X_h + X_k] = \mathbb{V}\text{ar}[X_h] + \mathbb{V}\text{ar}[X_k] + 2\mathbb{C}\text{ov}[X_h, X_k]$

$\mathbb{V}\text{ar}[\beta X_h] = \beta^2\mathbb{V}\text{ar}[X_h]$

$\mathbb{V}\text{ar}[\alpha + \beta X_h] = \beta^2\mathbb{V}\text{ar}[X_h]$

Random variables

Often a random experiment is described by more than one random variable (**random vectors**).

Given a random vector $X = (X_1, \ldots, X_J)$ the **joint probability distribution** is defined as

$$F_{X_1,\ldots,X_J}(X_1 = x_1, \ldots, X_J = x_J) = \mathbb{P}(X_1 \leq x_1, \ldots, X_J \leq x_J)$$

The **marginal probability distribution** is obtained by integration (continuous case) or summation (discrete case). For example, in the continuous case ($J = 2$):

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_2$$

Random variables

The **conditional probability distribution** is defined as follows ($J = 2$):

$$f_{X_1|X_2}(x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

If $X_1 \perp\!\!\!\perp X_2$ (**independence**), then:

$$f_{X_1|X_2}(x_1) = f_{X_1}$$

or alternatively

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$$

Random variables

In the multivariate context, **conditional expectations** can be obtained as well:

$$\mathbb{E}[X_1|X_2 = x_2] = \int_{-\infty}^{\infty} x_1 \; f_{X_1|X_2}(x_1) \; dx_1$$

$$\mathbb{V}\text{ar}[X_1|X_2 = x_2] = \mathbb{E}\left[(X_1 - \mathbb{E}[X_1|X_2])^2 \middle| X_2 = x_2\right]$$

Random variables

Similarly to the univariate case, there are several parametric probabilistic models for the multivariate case. For example, the most relevant model for the continuous case is the **multivariate Normal model** $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_j \\ \vdots \\ \mu_J \end{bmatrix}, \begin{bmatrix} \sigma^2_{11} & \dots & \sigma_{1j} & \dots & \sigma_{1J} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{j1} & \dots & \sigma^2_{jj} & \dots & \sigma_{jJ} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{J1} & \dots & \sigma_{Jj} & \dots & \sigma^2_{JJ} \end{bmatrix} \right)$$

with $\boldsymbol{\mu}_{J\times 1}$ being the vector of the means and $\boldsymbol{\Sigma}_{J\times J}$ the **covariance matrix**.

Random variables

Random variables $X_1, \dots, X_J$ are said to be **independent and identically distributed** (iid) iff:

$$f_{X_1,\dots,X_J}(x_1, \dots, x_J) = f_{X_1}(x_1) \cdots f_{X_J}(x_J)$$

$$f_{X_1} = f_{X_2} = \dots = f_{X_J}$$

Independent and identically distributed random variables constitute the building block of simple **random samples**. Moreover, they are at the base of **limit theorems**, which are important for statistical inference.

(Weak) Law of Large Numbers

Let $X_1, \dots, X_n$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\text{ar}[X_i] = \sigma^2$ for each term of the sequence and let

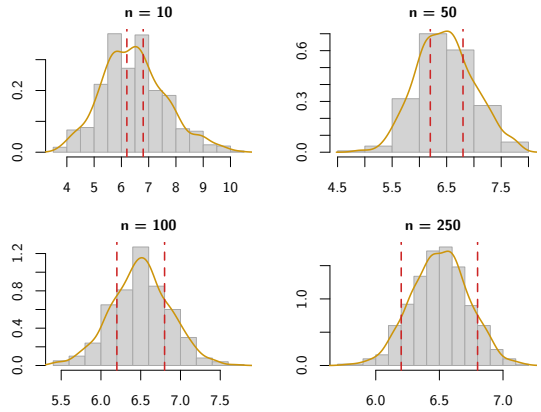
$$\overline{X} = \frac{1}{n} \sum_i X_i$$

be the mean of the random sequence. Then given any positive number ϵ (no matter how small) we have:

$$\boxed{\lim_{n\to\infty} \mathbb{P}(\mu - \epsilon < \overline{X}_n < \mu + \epsilon) = 1}$$

In other words, the random variable $\overline{X}$ is close to μ for large n.

(Weak) law of large numbers



Notes:
$X_i \sim \chi^2(n, \lambda = 6.5)$, $n = (10, 50, 100, 250)$. Dotted red lines indicate the set $I_{\epsilon,\lambda} = [\lambda \pm \epsilon]$, $\epsilon = 0.3$.
As n increases, $I_{\epsilon,\lambda}$ gets larger:
$\mathbb{P}(\overline{X}_{n=10} \in I_{\epsilon,\lambda}) = 0.218$, $\mathbb{P}(\overline{X}_{n=50} \in I_{\epsilon,\lambda}) = 0.422$, $\mathbb{P}(\overline{X}_{n=100} \in I_{\epsilon,\lambda}) = 0.586$,
$\mathbb{P}(\overline{X}_{n=250} \in I_{\epsilon,\lambda}) = 0.814$

Central limit theorem

Let $X_1, \ldots, X_n$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\text{ar}[X_i] = \sigma^2$ for each term of the sequence and let

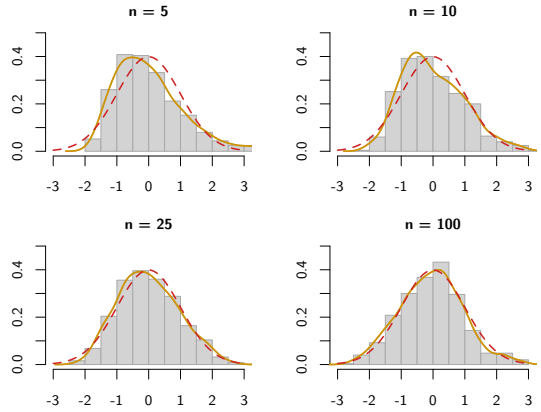
$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

be the **standardized random variable** with $\mathbb{E}[Z] = 0$ e $\mathbb{V}\text{ar}[Z] = 1$. Then for $x \in \mathbb{R}$ we have:

$$\boxed{\lim_{n\to\infty} \mathbb{P}(Z_n \leq x) = \mathbb{P}(Z \leq x)} \quad \text{with} \quad Z \sim \mathcal{N}(0,1)$$

In other words, the random variable Z_n has a distribution that is approximately standardized Normal (no matter how $X_1, \ldots, X_n$ are distributed).

Central limit theorem



Notes:
$X_i \sim \mathcal{E}xp(n, \lambda = 1)$, $n = (5, 10, 25, 100)$. Dotted red curves indicate the standardized Normal distribution.
As n increases, the distribution of Z_n approximates the standardized Normal distribution.

Statistical inference

A **statistical model** can generally be defined as a triplet

$$\mathcal{M} = \{F_Y(y;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p, y \in \mathcal{Y}\}$$

where

- $F_Y(y;\boldsymbol{\theta})$ is a parametric probabilistic model
- $\boldsymbol{\Theta}$ is the parametric space for $\boldsymbol{\theta}$
- $\mathcal{Y}$ is the sample space, i.e. the space where $\sup(Y)$ is defined

Examples:

- *Normal model*:
  $p = 2$, $\boldsymbol{\theta} = \{\mu, \sigma^2\} \in \boldsymbol{\Theta} \subset \mathbb{R} \times \mathbb{R}^+$, $\mathcal{Y} \subseteq \mathbb{R}$, and $F_Y(y;\boldsymbol{\theta}) = \mathcal{N}(y;\mu,\sigma^2)$
- *Bernoulli model*:
  $p = 1$, $\boldsymbol{\theta} = \pi \in [0,1]$, $\mathcal{Y} \subseteq \{0,1\}$, $F_Y(y;\boldsymbol{\theta}) = \mathcal{B}in(y;\pi)$

Statistical inference

In general, we have two ways for dealing with a statistical model $\mathcal{M}$:

(i) **Top-down approach**: the observer knows in advance the elements of the model - i.e. $\boldsymbol{\theta}$, $\mathcal{Y}$, and $F_Y(y;\boldsymbol{\theta})$ - with the purpose of simulating new instances/samples $\{y_1,\ldots,y_n\}$ from $\mathcal{M}$. For instance, this approach can be used to assess the inner-working mechanisms of $\mathcal{M}$.

(ii) **Bottom-up approach**: the observer has a set of instances/observations $\mathbf{y}=\{y_1,\ldots,y_n\}$ but nothing is known about $\mathcal{M}$ in advance. Then, the purpose here is to infer the most plausible model $\mathcal{M}^0 \in \{\mathcal{M}_1,\ldots,\mathcal{M}_K\}$ which has generated the observed sample $\mathbf{y}$.

Statistical inference

Example

With the goal of determining the level of a cognitive ability μ^0 in a non-clinical population, a sample of observations $\mathbf{y}_{n\times 1}$ has been collected by means of a cognitive test. From a statistical point-of-view, we need to determine which of the models $F(y;\mu_1),\ldots,F(y;\mu_k),\ldots,F(y;\mu_K)$ is the most plausible for μ^0 given $\mathbf{y}$.

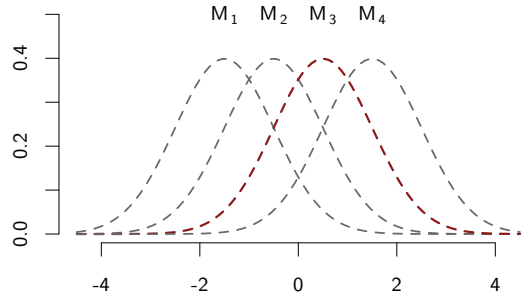
By previous knowledge about μ, we can set $\mathcal{Y}=\mathbb{R}$ and

$$F(\mathbf{y};\mu)=\mathcal{N}(\mathbf{y};\mu,\sigma^2=1)$$

Then, the goal becomes that of estimating $\mu^0 \in \mathbb{R}$ given $\mathbf{y}$, which implies selecting the most plausible model from the set $F(\mathbf{y};\mu_1),\ldots,F(\mathbf{y};\mu_k),\ldots,F(\mathbf{y};\mu_K)$.

Note: $\mathcal{N}(\mathbf{y};\mu,\sigma^2=1)$ is a *location model*.

Statistical inference



Notes:
$K=4$ plausible location models for the cognitive ability estimation.
The most plausible model given the data $\mathbf{y}$ is M_3 (red dotted curve).

Statistical inference

Determining $\mathcal{M}^0$ means making inference about the true but unknown parameter $\mu^0 \in \mathbb{R}$ of the true model $F^0(y;\boldsymbol{\theta})$. The procedure requires a **theory of statistical inference** which establishes the *correctness*, the *bias*, and the *uncertainty* of the estimates $\hat{\boldsymbol{\theta}}$.

A couple of approaches are available to this end: frequentist, Bayesian, information-theoretic based. Within the frequentist framework, the **maximum likelihood theory** is the most studied and most commonly used approach to statistical inference.

Statistical inference

A function of the data $t(\mathbf{y})$ is called **statistic** and, under some regularities, it summarizes the data information in the most optimal way. For example, the sample mean $\overline{y} = \frac{1}{n}\sum_i y_i$ is a statistic of the sample $\mathbf{y}$. As a statistic is computed over samples, which are in turn outcomes of r.vs., itself is a random variable $T(Y)$ with an own distribution as well. For example, the statistic $\overline{Y} = \sum_i Y_i$ is a random variable following the Normal distribution.

Statistical inference

Estimators $\hat{\theta}(Y)$ are statistics of the data and their outcomes $\hat{\theta}(\mathbf{y})$ or simply $\hat{\theta}$ are called **estimates**. For instance, in the location model $\mathcal{N}(y;\mu)$ the estimator for the parameter μ is $\hat{\mu} = \frac{1}{n}\sum_i y_i$. The probability distribution of $\hat{\theta}(Y)$ is called **sampling distribution** and provides information about $\hat{\theta}$. The variance of an estimator $\mathbb{V}\text{ar}\left[\hat{\theta}\right]$ (or $\sigma^2_{\hat{\theta}}$) provides important information about the uncertainty of the estimates $\hat{\theta}$.

Statistical inference

An estimator $\hat{\theta}(Y)$ for the parameter θ^0 is:

- **ubiased** iff

$$\text{B}(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta^0 = 0$$

This means that its average value over repeated samples is equal to the parameter being estimated

Statistical inference

An estimator $\hat{\theta}(Y)$ for the parameter θ^0 is:

- **efficient** iff its Mean Square Error (MSE)

$$\mathbb{E}\left[\left(\hat{\theta}(Y) - \theta^0\right)^2\right] = \mathbb{V}\text{ar}\left[\hat{\theta}\right] + \text{B}(\hat{\theta})^2$$

is as lower as possible.
For unbiased estimators, the efficiency of an estimator increases, therefore, as its sampling variance $\mathbb{V}\text{ar}\left[\hat{\theta}\right]$ declines.

Statistical inference

An estimator $\hat{\boldsymbol{\theta}}(Y)$ for the parameter θ^0 is:

- **consistent** if the bias and the sampling variance approach zero as n increases.

Statistical inference

Example

Consider a random sample $Y_1, \ldots, Y_n$ from a probabilistic model with parameters $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}\text{ar}[Y_i] = \sigma^2$.

Then, two estimators for μ are the following

$$\hat{\boldsymbol{\theta}}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad \hat{\boldsymbol{\theta}}_2 = Y_1$$

Statistical inference

Example

Both estimators are unbiased:

$$\begin{aligned} \text{B}\left(\hat{\boldsymbol{\theta}}_1\right) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] - \mu \\ &= \frac{1}{n}n\mu - \mu = 0 \end{aligned}$$

$$\begin{aligned} \text{B}\left(\hat{\boldsymbol{\theta}}_2\right) &= \mathbb{E}[Y_1] - \mu \\ &= \mu - \mu = 0 \end{aligned}$$

Statistical inference

Example

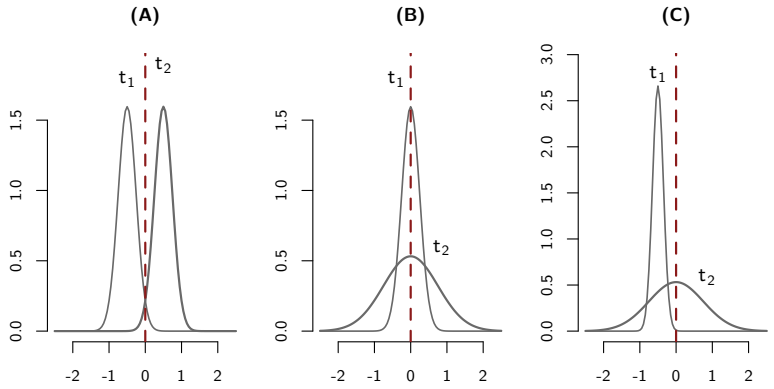
However, the second estimator is not as good as the first one:

$$\begin{aligned} \text{MSE}\left(\hat{\boldsymbol{\theta}}_1\right) &= \mathbb{V}\text{ar}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] + \text{B}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \\ &= \frac{1}{n^2}\mathbb{V}\text{ar}\left[\sum_{i=1}^{n} Y_i\right] + 0 \\ &= \frac{1}{n^2}n\sigma^2 + 0 = \frac{\sigma^2}{n} \end{aligned}$$

$$\begin{aligned} \text{MSE}\left(\hat{\boldsymbol{\theta}}_2\right) &= \mathbb{V}\text{ar}[Y_1] + \text{B}(Y_1)^2 \\ &= \sigma^2 + 0 = \sigma^2 \end{aligned}$$

As $\text{MSE}\left(\hat{\boldsymbol{\theta}}_1\right) < \text{MSE}\left(\hat{\boldsymbol{\theta}}_2\right)$, the estimator $\hat{\boldsymbol{\theta}}_1$ should be preferred over $\hat{\boldsymbol{\theta}}_2$.

Statistical inference



Notes:
Sampling distributions for two estimators (dotted red line indicates the true parameter).
(A) - Biased estimators with same variance; **(B)** - Unbiased estimators with different variance;
(C) - Biased estimator (t_1) vs. unbiased estimator (t_2) with different variance.
Although t_2 is unbiased, t_1 would be preferred if bias could be removed.

Statistical modeling

Notation

The starting point of statistical modeling is the sample of observations $\mathbf{y} = (y_1, \ldots, y_n)$ which is a random realization of a set of random variables (*random vector*) $Y_1, \ldots, Y_n$. Most often, $Y_1, \ldots, Y_n$ are considered independent with identical distribution (iid) so that $\mathbf{y}$ is the outcome of a Bernoulli sampling schema.

The usual notation is then adopted to indicate the probabilistic model for a random variable $Y_i \sim \mathcal{F}(y; \boldsymbol{\theta})$ with $\mathcal{F}$ being a proper statistical distribution parameterized by $\boldsymbol{\theta}$. With a slight abuse of notation, the same will also be denoted by $\mathbf{y} \sim F(y; \boldsymbol{\theta})$.

Statistical modeling

Defining a statistical model

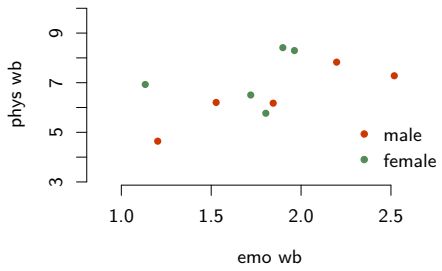
For a correct statistical model specification we will need to evaluate:

1. the probabilistic model $\mathcal{F}$ (aka, probabilistic distribution) underlying the response variable Y
2. the characteristics of $\mathcal{F}$ (e.g., expected value, variance, covariance) to be associated with the explanatory variables $\mathbf{X}$ through a parametric specification

Statistical modeling

Example 1

Subset of $n = 10$ data referring to the study of Physical well-being (`phys wb`) as a function of Emotional well-being (`emo wb`) and `gender`.



	phys wb	gender	emo wb
1	4.64	1	1.20
2	6.17	1	1.84
3	7.28	1	2.52
4	6.21	1	1.53
5	7.83	1	2.20
6	5.77	2	1.80
7	6.50	2	1.72
8	6.93	2	1.13
9	8.30	2	1.96
10	8.41	2	1.90

In **Example 1**, the response `phys wb` may be modeled using the Normal distribution $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{N}$ with parameters $\boldsymbol{\theta} = \{\mu, \sigma^2\} \in \mathbb{R} \times \mathbb{R}^+$, i.e.:

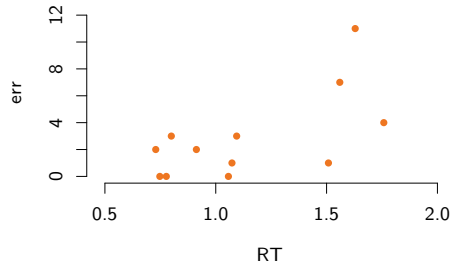
$$y_i \sim \mathcal{N}(y; \mu_i, \sigma^2)$$

with the explanatory variables `emo wb` ($\mathbf{x}_1$) and `gender` ($\mathbf{x}_2$) being associated to the *mean* of the model (systematic variation):

$$\mu_i = \mathbb{E}[Y_i] = \beta_0 + \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$$

The variance of the model σ^2 can be defined either as a function of the data or as a function of the mean and, consequently, as a function of the explanatory variables (as for GLMs).

Subset of $n = 12$ data referring to the number of errors in an experimental task (`err`) as a function of response times (`RT`).



	err	RT
1	11	1.63
2	0	1.06
3	0	0.75
4	4	1.76
5	3	1.09
6	7	1.56
7	1	1.07
8	2	0.91
9	0	0.78
10	2	0.73
11	3	0.80
12	1	1.51

In **Example 2** the response `err` may be modeled using the Poisson distribution $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{P}oi$ with parameters $\boldsymbol{\theta} = \lambda \in \mathbb{R}^+$, i.e.:

$$y_i \sim \mathcal{P}oi(y; \lambda_i)$$

with the explanatory variable `RT` ($\mathbf{x}$) being associated to the *mean* of the model:

$$\lambda_i = \mathbb{E}[Y_i] = \exp(\beta_0 + \mathbf{x}\beta)$$

In this particular case, the variance equals the mean, which is in turn a function of `RT`:

$$\mathbb{V}\text{ar}[Y_i] = \lambda_i$$

This is common in GLMs.

Reconsider the location model used to estimate the cognitive ability μ^0 given the random sample $\mathbf{y}_{n\times 1}$ (see slide 37):

$$y_i \sim \mathcal{N}(y_i; \mu, \sigma^2 = 1)$$

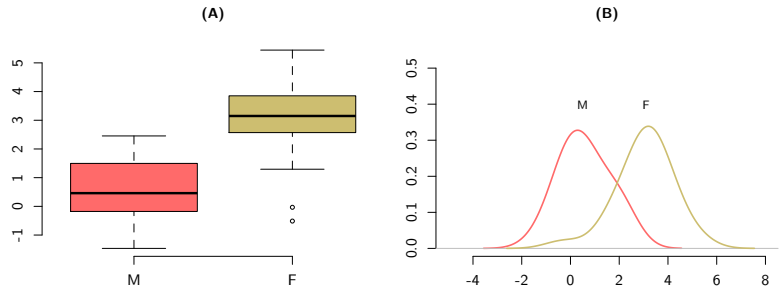
Then, we can extend the model to analyse whether the cognitive ability varies as a function of the categorical variable gender $\mathbf{z} \in \{0, 1\}^n$, which has the following levels $z_i = 0$ (male) and $z_i = 1$ (female).

This requires rewriting the mean of the model as a function of the new variable:

$$\mu_i = \beta_0 + z_i\beta_1$$

The result is still a location model but now it codifies two means, one for the male group when $z_i = 0$ ($\mu_i = \beta_0$) and the other one for the female group $z_i = 1$ ($\mu_i = \beta_0 + \beta_1$).

Statistical modeling
Defining a statistical model



Notes:

Linear model for the cognitive ability μ^0 as a function of gender.
In this example: $n = 100$ ($n_M = 50$), $\beta_0 = 0.5$, $\beta_1 = 2.6$, $\hat{\mu}_M = 0.5$, $\hat{\mu}_F = 3.1$.

(A) Observed data **y** plotted as a function of **z**
(B) Estimated densities $\hat{F}_Y(\mathbf{y}; \hat{\boldsymbol{\mu}})$ plotted as a function of **z**.

Statistical modeling
Basic notation

In general, y_i is called **response variable** whereas, depending on the context, $x_{i1}, \ldots, x_{iJ}$ are called **independent variables** (e.g., experimental settings) or simply **covariates** (e.g., social studies). The variables $\mathbf{x}_{i1}, \ldots, x_{iJ}$ are used as **predictors** of the responses y_i (asymmetric relation).

In this context, $x_{i1}, \ldots, x_{iJ}$ are considered *non stochastic* whereas y_i are thought as being random realizations from a random variable Y_i (only the response variables embed the stochasticity of the data collection process).

Statistical modeling
Basic notation

Depending on the context, the predictors $x_{i1}, \ldots, x_{iJ}$ can be **continuous** (i.e., $\mathbf{x}_i \in \mathbb{R}^J$) or **categorical** (i.e., $\mathbf{x}_i \in \mathcal{C} \subset \mathbb{N}^J$). In the last case, the elements of $\mathcal{C} = \{c_1, \ldots, c_k\}$ represent the **levels** that $\mathbf{x}_i$ may assume. For instance, if $k = 2$ the predictor is a dichotomous variable, otherwise when $k > 2$ the predictor is a polithomous variable.

The same applies for the response observations y_i. We may have **continuous** (i.e., $y_i \in \mathbb{R}$) or **positive continuous** responses (i.e., $y_i \in \mathbb{R}^+$) as well as categorical responses (i.e., $y_i \in \mathcal{C} \subset \mathbb{N}$) in the simplest form of **unordered categorical** responses or **ordered categorical** responses (i.e., $\ldots < c_{k-1} < c_k < c_{k+1} < \ldots$). In some circumstances, observations can be also collected in the form of **counts** (i.e., $y_i \in \mathbb{N}_0$) or **frequencies** (i.e., $y_i \in [0, 1]$).

Statistical modeling
Basic notation

As the stochasticity is embedded into y_i, the type of observation (e.g., continuous, categorical, counts) implies different random variable model Y_i. For instance, *continuous* responses may be modeled using a Normal random variable $Y_i \sim \mathcal{N}(y; \mu_i, \sigma^2)$, *dichotomous* responses may be modeled using a Bernoulli random variable $Y_i \sim \mathcal{B}er(\pi_i)$, *counts* may be modeled using a Poisson random variable $Y_i \sim \mathcal{P}oi(y; \lambda_i)$.

In order to infer the proper statistical model for a given response variable, we use **generalized linear models** (GLMs) which is a class of statistical models including many probabilistic models (e.g., Normal, Poisson, Gamma) for different response variables (e.g., continuous, counts, response times).

Response variable and covariates can be organized by means of a $n \times (J+1)$ matrix representation where n is the number of collected/sampled statistical units.

	Y	X_1	$\dots$	X_J
1	y_1	x_{11}	$\dots$	x_{1J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	x_{i1}	$\dots$	x_{iJ}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	$\dots$	x_{nJ}

Notes:

- We are interested in studying Y_i conditioned on $\mathbf{x}_i$ (asymmetric relation)
- For **non-grouped data**,
  the number of statistical units n equals the number of observations