Arthur Charpentier

charpentier.arthur@uqam.ca

https://freakonometrics.github.io/

UQAM, 2019

# STT5100 - Applied Linear Models

## Probability & Statistics

# Agenda

- Introduction: Statistical Model
- **Probability**
- Usual notations, $\mathbb{P}$, $F$, $f$, $\mathbb{E}$, Var
- Usual distributions: discrete & continuous
- Conditional Distribution, Conditional Expectation, Mixtures
- Convergence, Approximation and Asymptotic Results
  - Law of Large Numbers (LLN)
  - Central Limit Theorem (CLT)
- **(Mathematical Statistics)**
- From descriptive statistics to mathematical statistics
- Sampling: mean and variance
- Confidence Interval
- Decision Theory and Testing Procedures

# Overview

| sample | | inference | | test |
|---|---|---|---|---|
| $\{x_1, \cdots, x_n\}$ | $\rightarrow$ | $\widehat{\theta}_n = \varphi(x_1, \cdots, x_n)$ | $\rightarrow$ | $H_0 : \theta_0 = \kappa$ |
| $\downarrow$ | | $\downarrow$ | | $\downarrow$ |
| probabilistic | | properties of | | distribution |
| model | | the estimator | | under $H_0$ of $T_n$ |
| $X_i$ i.i.d. | | $\mathbb{E}(\widehat{\theta}_n)$ | | confiance interval |
| distribution $F_{\theta_0}$ | | $\text{Var}(\widehat{\theta}_n)$ | | $\theta_0 \in [a, b]$ |
| with $F_{\theta_0} \in \{F_\theta, \theta \in \Theta\}$ | | (asymptotics or | | with 95% chance |
| | | finite distance) | | |

# Additional References

Abebe, Daniels & McKean (2001) Statistics and Data Analysis

Freedman (2009) Statistical Models: Theory and Practice. Cambridge University Press.

Grinstead & Snell (2015) Introduction to Probability

Hogg, McKean & Craig (2005) Introduction to Mathematical Statistics. Cambridge University Press.

Kerns (2010) Introduction to Probability and Statistics Using R.

# Probability Space

Assume that there is a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

- $\Omega$ is the fundamental space: $\Omega = \{\omega_i, i \in I\}$ is the set of all results from a random experiment.
- $\mathcal{A}$ is the $\sigma$-algebra of evevents, ie the set of all parts of $\Omega$.
- $\mathbb{P}$ is a probability measure on $\Omega$, i.e.
  - $\mathbb{P}(\Omega) = 1$
  - for any event $A$ in $\Omega$, $0 \leq \mathbb{P}(A) \leq 1$,
  - for any $A_1, \cdots, A_n$ mutually exclusive $(A_i \cap A_j = \emptyset)$,

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

A random variable $X$ is a function $\Omega \to \mathbb{R}$.

## Probability Space

One flip of a fair coin: the outcome is either heads or tails, $\Omega = \{\text{H}, \text{T}\}$, e.g. $\omega = \{H\} \in \Omega$.

The $\sigma$-algebra is $\mathcal{A} = \{\{\}, \{\text{H}\}, \{\text{T}\}, \{\text{H}, \text{T}\}\}$, or $\mathcal{F} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, \Omega\}$

There is a fifty percent chance of tossing heads and fifty percent for tails, $\mathbb{P}(\{\}) = 0$, $\mathbb{P}(\{\text{H}\}) = 0.5$ $\mathbb{P}(\{\text{T}\}) = 0.5$ and $\mathbb{P}(\{\text{H}, \text{T}\}) = 1$.

Consider a game where we gain 1 if the outcome is head, 0 otherwise. Let $X$ denote our financial income. $X$ is a random variable with values $\{0, 1\}$. $\mathbb{P}(X = 0) = 0.5$ and $\mathbb{P}(X = 1) = 0.5$ is the distribution of $X$ on $\{0, 1\}$.

# Probability Space

$n$ flip of a fair coin, the outcome is either heads or tails, each time, $\Omega = \{\text{H}, \text{T}\}^n$, e.g. $\omega = \{H, H, T, \cdots, T, H\} \in \Omega$.

The $\sigma$-algebra is $\mathcal{A} = \{\{\}, \{\text{H}\}, \{\text{T}\}, \{\text{H}, \text{H}\}\}, \{\text{H}, \text{T}\}, \{\text{T}, \text{H}\}\}, \cdots\}$.

There is a fifty percent chance of tossing heads and fifty percent for tails, $P(\omega) = 0$ if $\#\omega \neq n$, otherwise, probability is $1/2^n$,

$$\mathbb{P}(\{H, H, T, \cdots, T, H\}) = \frac{1}{2^n}$$

Consider a game where we gain 1 if the outcome is head, 0 otherwise. Let $X$ denote our financial income. $X$ is a random variable with values $\{0, 1, \cdots, n\}$ ($X$ is also the number of heads obtained out of $n$ draws). $P(X = 0) = 1/2^n$, $P(X = 1) = n/2^n$, etc, is the distribution of $X$ on $\{0, 1, \cdots, n\}$.

# Usual Functions

**Definition** Let $X$ denote a random variable, its cumulative distribution function (cdf) is

$$F(x) = \mathbb{P}(X \leq x), \text{ for all } x \in \mathbb{R}.$$

More formally, $F(x) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\})$, see c.d.f.

Observe that

- $F$ is an increasing function on $\mathbb{R}$ with values in $[0, 1]$,
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

$X$ and $Y$ are equal in distribution, denoted $X \overset{\mathcal{L}}{=} Y$ if for any $x$

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x) = F_Y(x).$$

The survival function is $\overline{F}(x) = 1 - F(x) = \mathbb{P}(X > x)$, see s.d.f.

In R, `pexp()` or `ppois()` return cdfs of exponential - $\mathcal{E}(1)$ - and Poisson distributions.

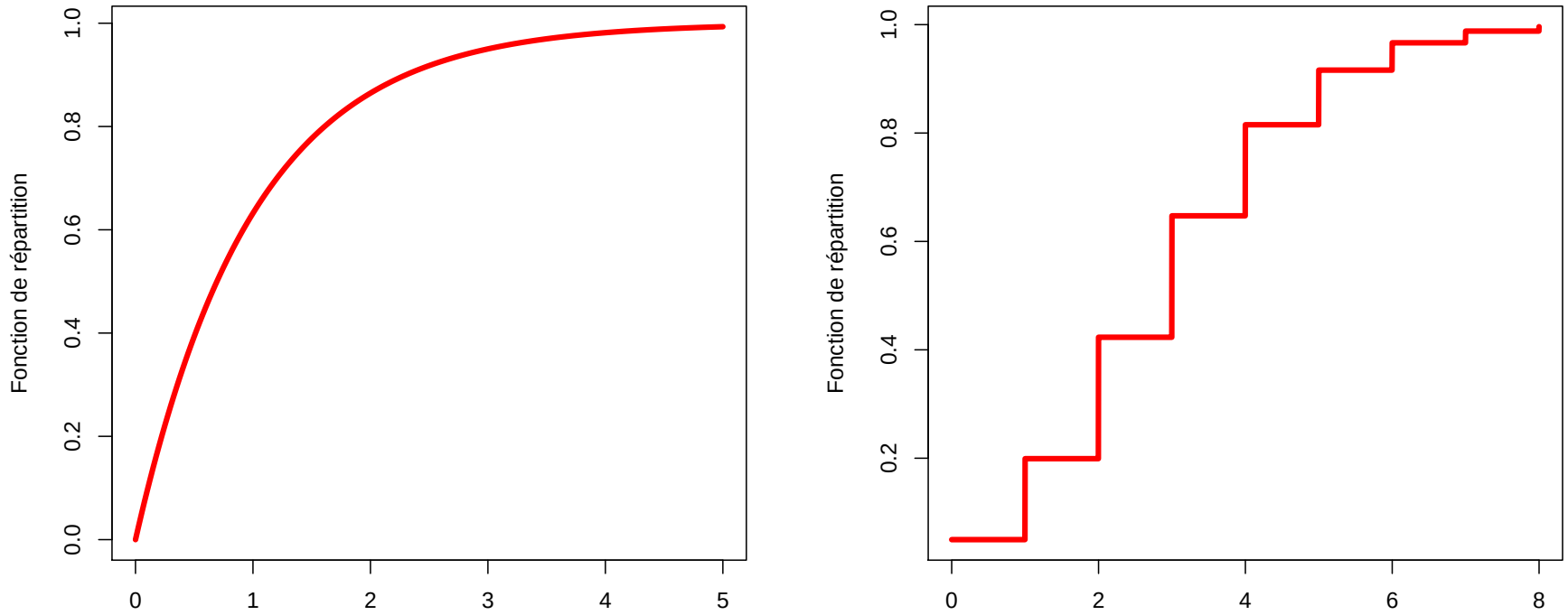


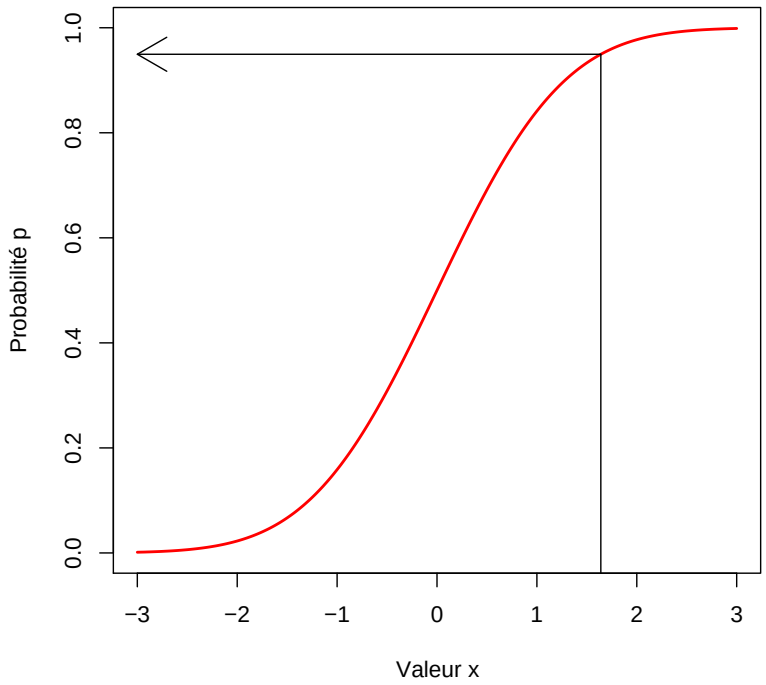

Figure 1: Cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$.

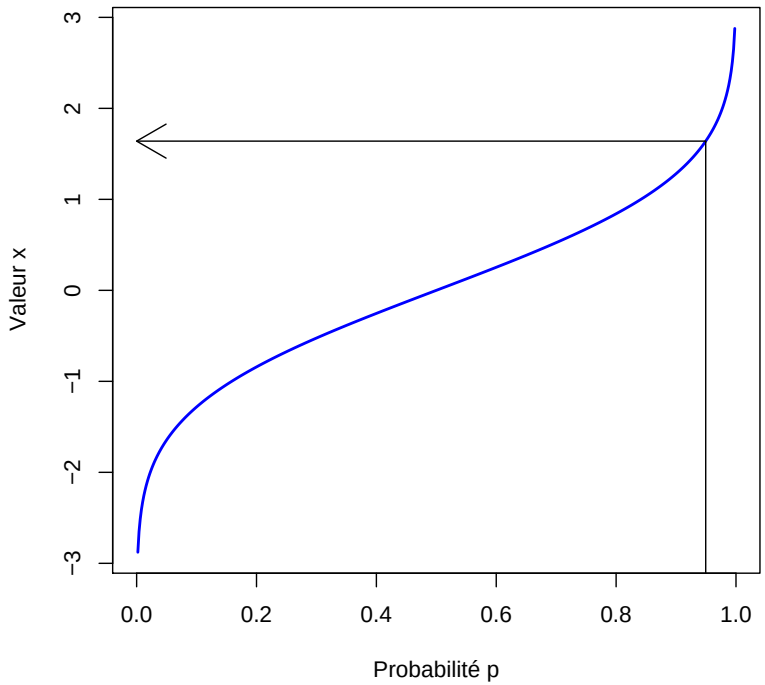
# Usual Functions

**Definition** Let $X$ denote a random variable, its quantile function is

$$Q(p) = F^{-1}(p) = \inf\{x \in \mathbb{R} \text{ tel que } F(x) > p\}, \text{ for all } p \in [0,1].$$

see quantile



With R, `qexp()` and `qpois()` are quantile functions of the exponential ($\mathcal{E}(1)$) and the Poisson distribution.

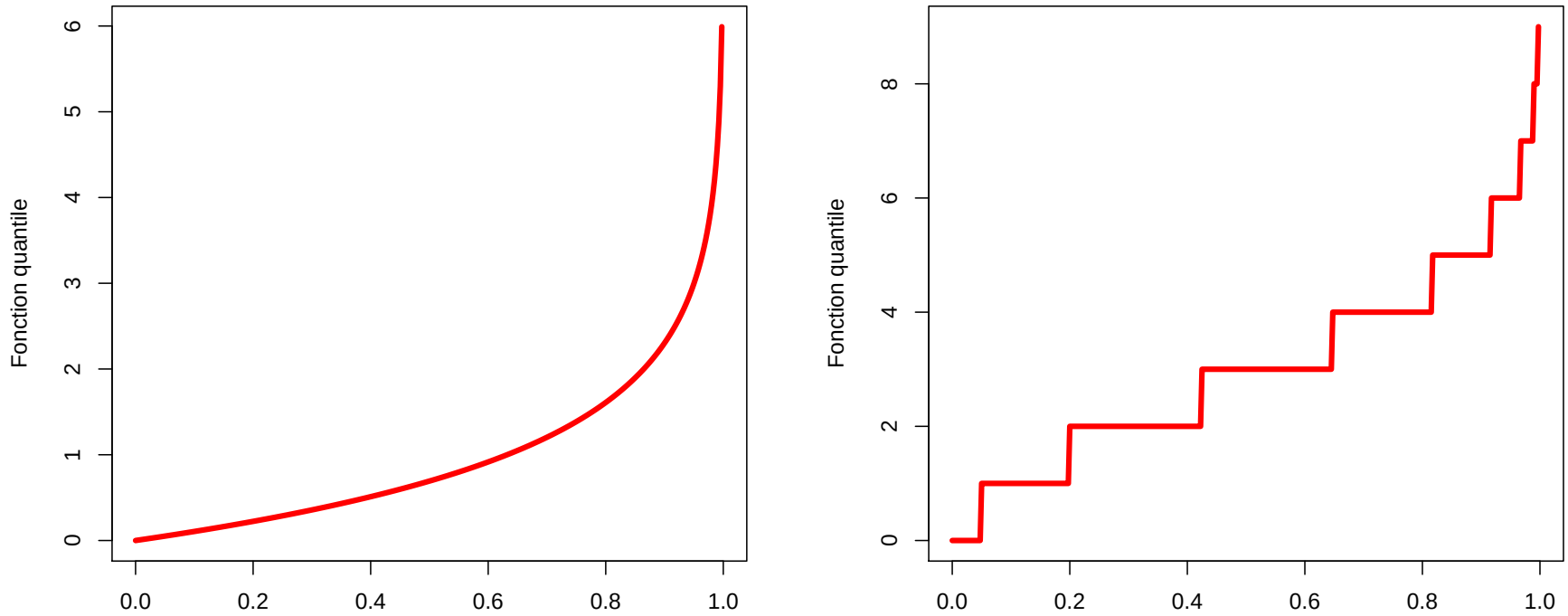



Figure 2: Quantile function $Q(p) = F^{-1}(p)$.

# Usual Functions

**Definition** Let $X$ be a random variable. The density or probablity function of $X$ is

$$f(x) = \begin{cases} \dfrac{dF(x)}{dx} = F'(x) \text{ in the (absolutely) continous case}, x \in \mathbb{R} \\ \mathbb{P}(X = x) \text{ in the discret case}, x \in \mathbb{N} \\ dF(x), \text{ in a more general context} \end{cases}$$

$F$ being an increasing function (if $A \subset B$, $\mathbb{P}[A] \leq \mathbb{P}[B]$), a density is always positive. For continuous distributions, we can have $f(x) > 1$.

Further, $F(x) = \displaystyle\int_{-\infty}^{x} f(s)ds$ for continuous distributions, $F(x) = \displaystyle\sum_{s=0}^{x} f(s)$ for discrete ones, see density or probability function

With R, `dexp()` and `dpois()` return density of the exponential ($\mathcal{E}(1)$) and the Poisson distributions .

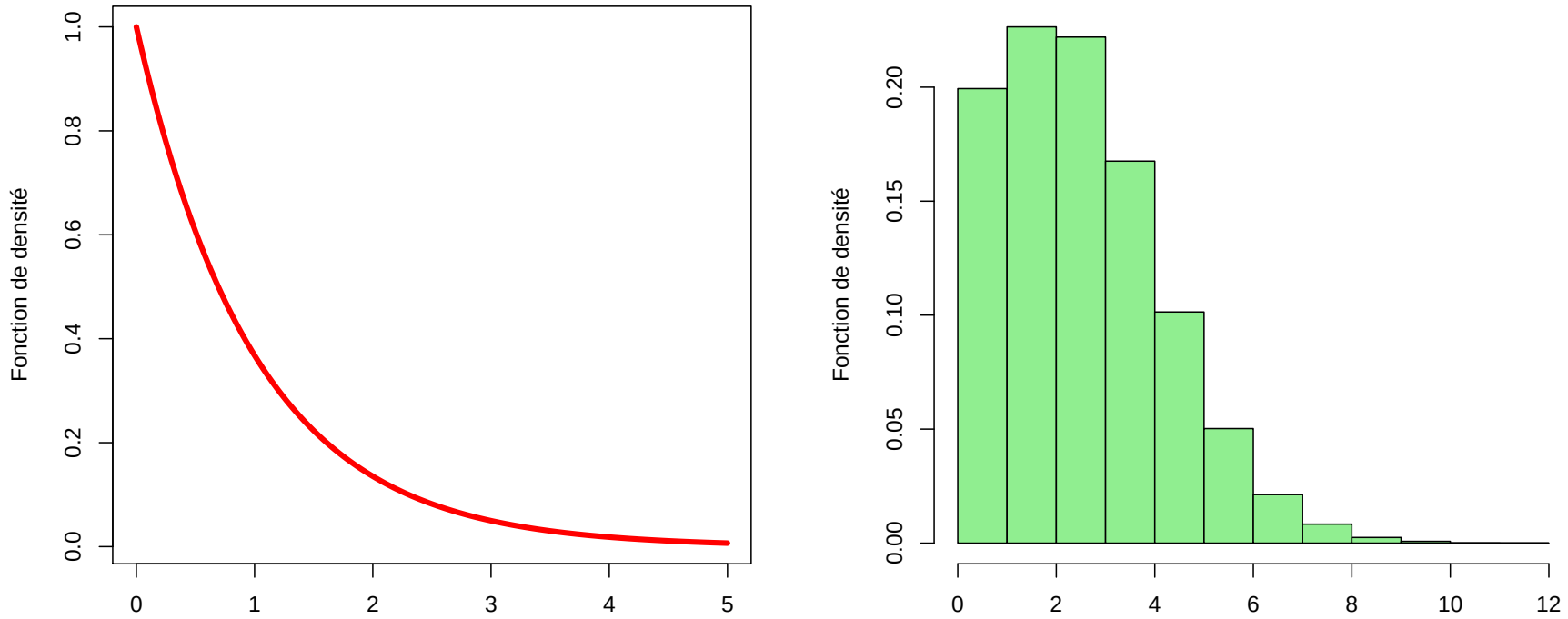



Figure 3: Densities $f(x) = F'(x)$ or $f(x) = \mathbb{P}(X = x)$.

$$\mathbb{P}(X \in [a,b]) = \int_a^b f(s)ds \text{ or } \sum_{s=a}^{b} f(s).$$

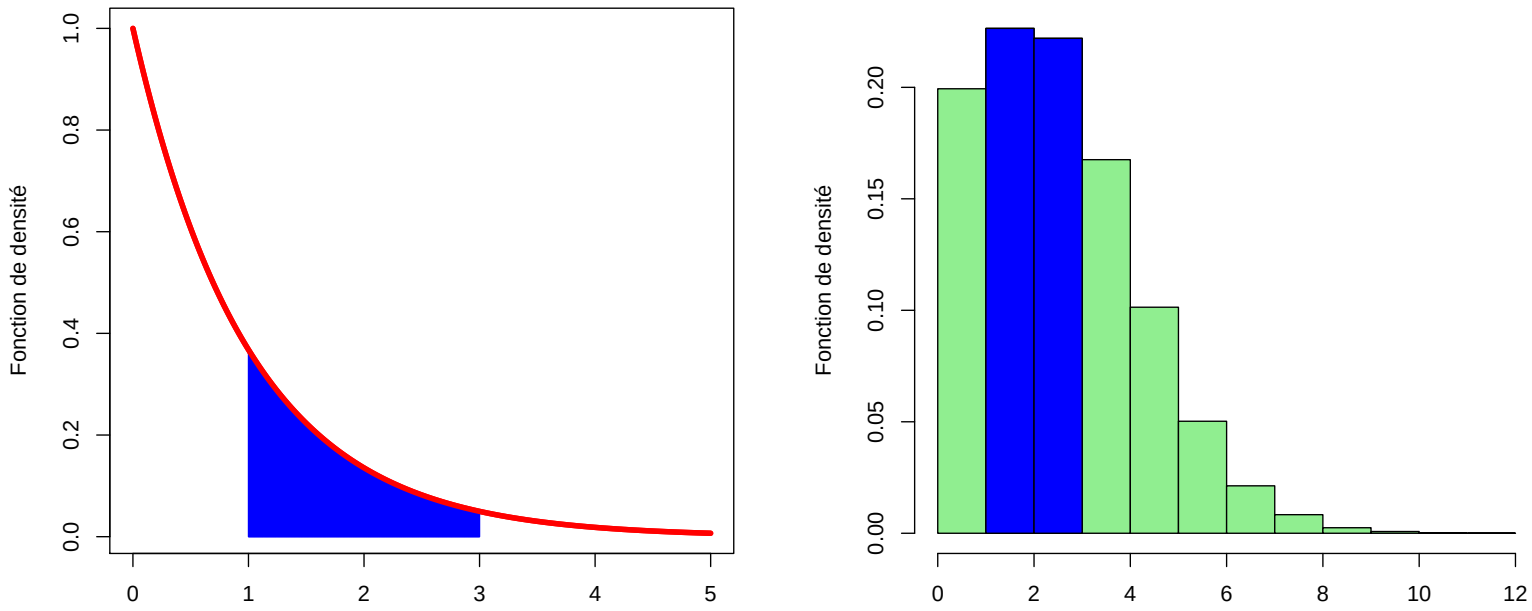



Figure 4: Probability $\mathbb{P}(X \in [1,3[)$.

# On Random Vectors

**Definition** Let $\boldsymbol{Z} = (X, Y)$ be a random vector. The cumulative distribution function of $\boldsymbol{Z}$ is

$$F(\boldsymbol{z}) = F(x, y) = \mathbb{P}(X \leq x, Y \leq y), \text{ for all } \boldsymbol{z} = (x, y) \in \mathbb{R} \times \mathbb{R}.$$

**Definition** Let $\boldsymbol{Z} = (X, Y)$ be a random vector. The density of $\boldsymbol{Z}$ is

$$f(\boldsymbol{z}) = f(x, y) = \begin{cases} \dfrac{\partial^2 F(x, y)}{\partial x \partial y} \text{ in the continuous case, } \boldsymbol{z} = (x, y) \in \mathbb{R} \times \mathbb{R} \\ \mathbb{P}(X = x, Y = y) \text{ in the discrete case, } \boldsymbol{z} = (x, y) \in \mathbb{N} \times \mathbb{N} \end{cases}$$

see random vectors

## On Random Vectors

Consider a random vector $\boldsymbol{Z} = (X, Y)$ with cdf $F$ and density $f$, one can extract marginal distributions of $X$ and $Y$ from

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y \leq +\infty) = \lim_{y \to \infty} F(x, y),$$

$$f_X(x) = \mathbb{P}(X = x) = \sum_{y=0}^{\infty} \mathbb{P}(X = x, Y = y) = \sum_{y=0}^{\infty} f(x, y), \text{ for a discrete distribution}$$

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \text{ for a continuous distribution}$$

# Conditional distribution $Y|X$

Define the conditionnal distribution of $Y$ given $X = x$, with density given by Bayes formula (see see conditional probabilities)

$$\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} \text{ in the discrete case,}$$

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_X(x)}, \text{ in the continuous case.}$$

One can also derive the conditional cdf

$$\mathbb{P}(Y \leq y|X = x) = \sum_{t=0}^{y} \mathbb{P}(Y = t|X = x) = \sum_{t=0}^{y} \frac{\mathbb{P}(X = x, Y = t)}{\mathbb{P}(X = x)} \text{ in the discrete case,}$$

$$F_{Y|X=x}(y) = \int_{-\infty}^{x} f_{Y|X=x}(t)dt = \frac{1}{f_X(x)} \int_{-\infty}^{x} f(x, t)dt, \text{ in the continuous case.}$$

# On Margins of Random Vectors

We have seen that

$$f_Y(y) = \sum_{x=0}^{\infty} f(x,y) \text{ or } \int_{-\infty}^{\infty} f(x,y)dx$$

Let us focus on the continuous case.

From Bayes formula,

$$f(x,y) = f_{Y|X=x}(y) \cdot f_X(x)$$

and we can write

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

known as the law of total probability.

## Independence

**Definition** Consider two random variables $X$ and $Y$. $X$ and $Y$ are independent if one of the following statements is valid

- $F(x,y) = F_X(x)F_Y(y)\ \forall x,y$, or $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \times \mathbb{P}(Y \leq y)$,
- $f(x,y) = f_X(x)f_Y(y)\ \forall x,y$, or $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \times \mathbb{P}(Y = y)$,
- $F_{Y|X=x}(y) = F_Y(y)\ \forall x,y$, or $f_{Y|X=x}(y) = f_Y(y)$,
- $F_{X|Y=y}(y) = F_X(x)\ \forall x,y$, or $f_{X|Y=y}(y) = f_X(x)$.

We will use notations $X \perp\!\!\!\perp Y$ when variables are independent, see see independence

# Independence

Consider the following (joint) probabilities for $X$ and $Y$, i.e. $\mathbb{P}(X = \cdot, Y = \cdot)$

| | $X=0$ | $X=1$ |
|---|---|---|
| $Y=0$ | 0.1 | 0.15 |
| $Y=1$ | 0.5 | 0.25 |

| | $X=0$ | $X=1$ |
|---|---|---|
| $Y=0$ | 0.15 | 0.1 |
| $Y=1$ | 0.45 | 0.3 |

In those two cases $\mathbb{P}(X = 1) = 0.4$, i.e. $X \sim \mathcal{B}(0.4)$ while $\mathbb{P}(Y = 1) = 0.75$, i.e. $Y \sim \mathcal{B}(0.75)$.

In the first case $X$ and $Y$ are not independent, but they are in the second case.

## Conditional Independence

Two variables $X$ and $Y$ are conditionnally independent given $Z$ if for all $z$ (such that $\mathbb{P}(Z = z) > 0$)

$$\mathbb{P}(X \leq x, Y \leq y \mid Z = z) = \mathbb{P}(X \leq x \mid Z = z) \cdot \mathbb{P}(Y \leq y \mid Z = z)$$

For instance, let $Z \in [0, 1]$, and consider $X|Z = z \sim \mathcal{B}(z)$ and $Y|Z = z \sim \mathcal{B}(z)$ independent (given $Z$). Variables are conditionally independent, but not independent.

# Moments of a distribution

**Definition** Let $X$ be a random variable. Its expected value is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx \text{ or } \sum_{x=0}^{\infty} x \cdot \mathbb{P}(X = x)$$

**Definition** Let $\boldsymbol{Z} = (X, Y)$ be a random vector. Its expected value is

$$\mathbb{E}(\boldsymbol{Z}) = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(Y) \end{pmatrix}$$

**Proposition.** *The expected value of* $Y = g(X)$*, where* $X$ *has density* $f$*, is*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) \cdot f(x)dx.$$

*If* $g$ *is nonlinear* $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$.

# On the expected value

**Proposition.** *Let $X$ and $Y$ two random variables with finite expected value*

- $\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$, $\forall \alpha, \beta$, *i.e. the expected vallue is linear*
- $\mathbb{E}(XY) \neq \mathbb{E}(X) \cdot \mathbb{E}(Y)$ *in general, but if* $X \perp\!\!\!\perp Y$, *equality holds.*

The expected value of any random variable is a number in $\mathbb{R}$.

Consider a uniform distribution on $[a, b]$, with density $f(x) = \dfrac{1}{b-a}\mathbf{1}(x \in [a, b])$,

$$\begin{aligned}\mathbb{E}(X) &= \int_{\mathbb{R}} xf(x)dx = \frac{1}{b-a}\int_a^b x dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\ &= \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{1}{b-a}\frac{(b-a)(a+b)}{2} = \frac{a+b}{2}.\end{aligned}$$

see expected value

If $\mathbb{E}[|X|] < \infty$, we note $X \in L^1$.

There are cases where expected value is infinite (does not exist)

Consider a repeated head/tail game, where gains are double when 'head' is obtained, and we can play again, until we get a 'tail'

$$\begin{aligned}\mathbb{E}(X) &= 1 \times \mathbb{P}(\text{'tail' at 1st draw}) \\ &\quad +1 \times 2 \times \mathbb{P}(\text{'tail' at 2nd draw}) \\ &\quad +2 \times 2 \times \mathbb{P}(\text{'tail' at 3rd draw}) \\ &\quad +4 \times 2 \times \mathbb{P}(\text{'tail' at 4th draw}) \\ &\quad +8 \times 2 \times \mathbb{P}(\text{'tail' at 5th draw}) + \cdots \\ &= \frac{1}{2} + \frac{2}{4} + \frac{4}{8} + \frac{8}{16} + \frac{16}{32} + \frac{32}{64} + \cdots = \infty.\end{aligned}$$

(so called St Petersburg paradox)

# Conditional Expectation

**Definition** Let $X$ and $Y$ be two random variables. The conditional expectation of $Y$ given $X = x$ is the expected value of the conditional distribution $Y|X = x$,

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X=x}(y)dy \text{ ou } \sum_{x=0}^{\infty} y \cdot \mathbb{P}(Y = y|X = x).$$

$\mathbb{E}(Y|X = x)$ is a function of $x$, $\mathbb{E}(Y|X = x) = \varphi(x)$. Random variable $\varphi(X)$ might be denoted $\mathbb{E}(Y|X)$.

**Proposition.** *$\mathbb{E}(Y|X)$ being a random variable, observe that*

$$\mathbb{E}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(Y)$$

see conditional expectation

*Proof.*

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}(X|Y)\right) &= \sum_y \mathbb{E}(X|Y=y)\cdot\mathbb{P}(Y=y)\\
&= \sum_y \left(\sum_x x\cdot\mathbb{P}(X=x|Y=y)\right)\cdot\mathbb{P}(Y=y)\\
&= \sum_y\sum_x x\cdot\mathbb{P}(X=x|Y=y)\cdot\mathbb{P}(Y=y)\\
&= \sum_x\sum_y x\cdot\mathbb{P}(Y=y|X=x)\cdot\mathbb{P}(X=x)\\
&= \sum_x x\cdot\mathbb{P}(X=x)\cdot\left(\sum_y \mathbb{P}(Y=y|X=x)\right)\\
&= \sum_x x\cdot\mathbb{P}(X=x) = \mathbb{E}(X).
\end{aligned}$$

□

# Higher Order Moments

Before introducting the order 2 moment, recall that

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

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y) \cdot f(x,y)dxdy.$$

**Definition** Let $X$ be a random variable. The variance of $X$ is

$$\text{Var}(X) = \mathbb{E}[(X-\mathbb{E}(X))^2] = \int_{-\infty}^{\infty}(x-\mathbb{E}(X))^2 \cdot f(x)dx \text{ or } \sum_{x=0}^{\infty}(x-\mathbb{E}(X))^2 \cdot \mathbb{P}(X=x).$$

Equivalently $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

The variance measures the dispersion of $X$ around $\mathbb{E}(X)$, and it is a positive number. $\sqrt{\text{Var}(X)}$ is called the standard deviation, see see variance and standard deviation

# Higher Order Moments

**Definition** Let $\boldsymbol{Z} = (X, Y)$ be a random vector. The variance-covariance matrix of $\boldsymbol{Z}$ is

$$\text{Var}(\boldsymbol{Z}) = \mathbb{E}\big((\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^\intercal\big) = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Var}(Y) \end{pmatrix}$$

where $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$ and

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

**Definition** Let $\boldsymbol{Z} = (X, Y)$ be a random vector. The (Pearson) correlation between $X$ and $Y$ is

$$\text{corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{\mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]}{\sqrt{\mathbb{E}[(X - \mathbb{E}(X))]^2 \cdot \mathbb{E}[(Y - \mathbb{E}(Y))]^2}}.$$

# On the Variance

See covariance and Pearson's correlation

**Proposition.** *The variance is always positive, and* $Var(X) = 0$ *if and only if* $X$ *is a constant.*

**Proposition.** *The variance is not linear, but*

$$\text{Var}(\alpha X + \beta Y) = \alpha^2 \text{Var}(X) + 2\alpha\beta \text{Cov}(X, Y) + \beta^2 \text{Var}(Y).$$

A consequence is that

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}\left(X_i\right) + \sum_{j \neq i} \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} \text{Var}\left(X_i\right) + 2 \sum_{j > i} \text{Cov}(X_i, X_j).$$

**Proposition.** *Variance is (usually) nonlinear, but* $\text{Var}(\alpha + \beta X) = \beta^2 \text{Var}(X)$.

If $\text{Var}[X] < \infty$ - or $\mathbb{E}[X^2] < \infty$ - we note $X \in L^2$.

# On covariance

**Proposition.** *Consider random variables $X$, $X_1$, $X_2$ and $Y$, then*

- $\text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$,
- $\text{Cov}(\alpha X_1 + \beta X_2, Y) = \alpha\text{Cov}(X_1,Y) + \beta\text{Cov}(X_2,Y)$.

$$\text{Cov}(X,Y) = \sum_{\omega\in\Omega}[X(\omega) - \mathbb{E}(X)]\cdot[Y(\omega) - \mathbb{E}(Y)]\cdot\mathbb{P}(\omega)$$

Heuristically, a positive covariance should mean that for a majority of events $\omega$, the following inequality should hold

$$[X(\omega) - \mathbb{E}(X)]\cdot[Y(\omega) - \mathbb{E}(Y)] \geq 0.$$

- $X(\omega) \geq \mathbb{E}(X)$ and $Y(\omega) \geq \mathbb{E}(Y)$, i.e. $X$ and $Y$ take together large values
- $X(\omega) \leq \mathbb{E}(X)$ and $Y(\omega) \leq \mathbb{E}(Y)$, i.e. $X$ and $Y$ take together small values

**Proposition.** *If $X$ and $Y$ are independent, $(X \perp\!\!\!\perp Y)$, then $\text{Cov}(X,Y) = 0$ (denote $X \perp Y$) but the converse is usually false.*

# Conditionnal Variance

**Definition** Let $X$ and $Y$ be two random variables. The conditional variance of $Y$ given $X = x$ is the variance of the conditional distribution $Y|X = x$,

$$\text{Var}(Y|X = x) = \int_{-\infty}^{\infty} [y - \mathbb{E}(Y|X = x)]^2 \cdot f_{Y|X=x}(y)dy.$$

$\text{Var}(Y|X = x)$ is a function of $x$, $\text{Var}(Y|X = x) = \psi(x)$. Random variable $\psi(X)$ will be denoted $\text{Var}(Y|X)$.

**Proposition.** $\text{Var}(Y|X)$ *being a random variable,*

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

*which is the variance decomposition formula.*

see law of total variance, and recall that $\text{Var}(Y) \geq \text{Var}[\mathbb{E}(Y|X)]$

# Conditionnal Variance

*Proof.* Use the following decomposition

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[(Y-\mathbb{E}(Y))^2] = \mathbb{E}[(Y-\mathbb{E}(Y|X)+\mathbb{E}(Y|X)-\mathbb{E}(Y))^2] \\
&= \mathbb{E}[([Y-\mathbb{E}(Y|X)]+[\mathbb{E}(Y|X)-\mathbb{E}(Y)])^2] \\
&= \mathbb{E}[([Y-\mathbb{E}(Y|X)])^2]+\mathbb{E}[([\mathbb{E}(Y|X)-\mathbb{E}(Y)])^2] \\
&\quad +2\mathbb{E}[[Y-\mathbb{E}(Y|X)]\cdot[\mathbb{E}(Y|X)-\mathbb{E}(Y)]]
\end{aligned}$$

Then observe that

$$\mathbb{E}[([Y-\mathbb{E}(Y|X)])^2] = \mathbb{E}\left(\mathbb{E}((Y-\mathbb{E}(Y|X))^2|X)\right) = \mathbb{E}[\ \mathrm{Var}(Y|X)],$$

$$\mathbb{E}[([\mathbb{E}(Y|X)-\mathbb{E}(Y)])^2] = \mathbb{E}[([\mathbb{E}(Y|X)-\mathbb{E}(\mathbb{E}(Y|X))])^2] = \ \mathrm{Var}[\mathbb{E}(Y|X)].$$

The expected value of the cross-product is null (given $X$). □

# Geometric Perspective

Recall that $L^2$ is the set of random variables with finite variance

- $< X, Y >= \mathbb{E}(XY)$ is a scalar product
- $\|X\| = \sqrt{\mathbb{E}(X^2)}$ is a norm (denoted $\|\cdot\|_2$).

$\mathbb{E}(X)$ is the orthogonal projection of $X$ on the set of constants

$$\mathbb{E}(X) = \underset{a\in\mathbb{R}}{\text{argmin}}\{\|X-a\|^2 = \mathbb{E}([X-a]^2)\}.$$

The correlation is the cosinus of the angle between $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$: if $\text{Corr}(X, Y) = 0$ variables are orthogonal, $X \perp Y$ (weaker than $X \perp\!\!\!\perp Y$).

If $L^2_X$ is the set of random variables generated from $X$ (that can be written $\varphi(X)$) with finite variance. $\mathbb{E}(Y|X)$ is the orthogonal projection of $Y$ on $L^2_X$

$$\mathbb{E}(Y|X) = \underset{\varphi}{\text{argmin}}\{\|Y-\varphi(X)\|^2 = \mathbb{E}([Y-\varphi(X)]^2)\}.$$

$\mathbb{E}(Y|X)$ is the best approximation of $Y$ by a function of $X$.

# Conditional Expectation

In an econometric model, we want to 'explain' $Y$ by $X$.

- linear econometrics, $\mathbb{E}(Y|X) \sim EL(Y|X) = \beta_0 + \beta_1 X$.
- nonlinear econometrics, $\mathbb{E}(Y|X) = \varphi(X)$.

or more generally, 'explain' $Y$ by $\boldsymbol{X}$.

- linear econometrics, $\mathbb{E}(Y|\boldsymbol{X}) \sim EL(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.
- nonlinear econometrics, $\mathbb{E}(Y|\boldsymbol{X}) = \varphi(\boldsymbol{X}) = \varphi(X_1, \cdots, X_k)$.

In a time series context, we want to 'explain' $X_t$ with $X_{t-1}, X_{t-2}, \cdots$.

- linear time series,
  $\mathbb{E}(X_t|X_{t-1}, X_{t-2}, \cdots) \sim EL(X_t|X_{t-1}, X_{t-2}, \cdots) = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_k X_{t-k}$
  (autoregressive).
- nonlinear time series, $\mathbb{E}(X_t|X_{t-1}, X_{t-2}, \cdots) = \varphi(X_{t-1}, X_{t-2}, \cdots)$.

# Sum of Random Variables

**Proposition.** *Let $X$ and $Y$ be two discrete random variables, then the distribution of $S = X + Y$ is*

$$\mathbb{P}(S = s) = \sum_{k=-\infty}^{\infty} \mathbb{P}(X = k) \times \mathbb{P}(Y = s - k).$$

*Let $X$ and $Y$ be two (abs) continuous random variables, then the distribution of $S = X + Y$ is*

$$f_S(s) = \int_{-\infty}^{\infty} f_X(x) \times f_Y(s - x)dx.$$

*Note $f_S = f_X \star f_Y$ where $\star$ is the convolution operator.*

see convolution including some particular distributions

# More on the Moments of a Distribution

$n$-th order moment of a random variable $X$ is $\mu_n = \mathbb{E}[X^n]$, if that value is finite. Let $\mu'_n$ denote centered moments.

Some of those moments :

- Order 1 moment $\mu = \mathbb{E}[X]$ is the expected value
- Centered order 2 moment: $\mu'_2 = \mathbb{E}\left[(X-\mu)^2\right]$ is the variance, $\sigma^2$.
- Centered and Reduced order 3 moment: $\mu'_3 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ is an assymmetric coefficient, called skewness.
- Centered and Reduced order 4 moment: $\mu'_4 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ is called kurtosis.

## Some Probabilistic Distributions: Bernoulli

The Bernoulli distribution $\mathcal{B}(p)$, $p \in (0,1)$

$$\mathbb{P}(X=0) = 1-p \text{ and } \mathbb{P}(X=1) = p.$$

Then $\mathbb{E}(X) = p$ and $\text{Var}(X) = p(1-p)$.

Let $Y$ denote some random variable on a set $\mathcal{A}$. Let $A \subset \mathcal{A}$, then $X = \mathbf{1}_A(Y) = \mathbf{1}(Y \in A)$ is $\mathcal{B}(p)$ distributed, with $p = \mathbb{P}[Y \in A]$.

see Bernoulli distribution

# Some Probabilistic Distributions: Binomial

The Binomial distribution $\mathcal{B}(n,p)$, $p\in(0,1)$ and $n\in\mathbb{N}^*$

$$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \text{ where } k=0,1,\cdots,n, \binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Then $\mathbb{E}(X)=np$ and $\text{Var}(X)=np(1-p)$. Hence $\text{Var}(X)<\mathbb{E}(X)$.

If $X_1,\cdots,X_n\sim\mathcal{B}(p)$ are independent, then $X=X_1+\cdots+X_n\sim\mathcal{B}(n,p)$.

With R, `dbinom(x, size, prob)`, `qbinom()` and `pbinom()` are respectively the cdf, the quantile function and the probability function of $\mathcal{B}(n,p)$ where $n$ is the `size` and $p$ the `prob` parameter.

see binomial distribution

# Some Probabilistic Distributions: Binomial

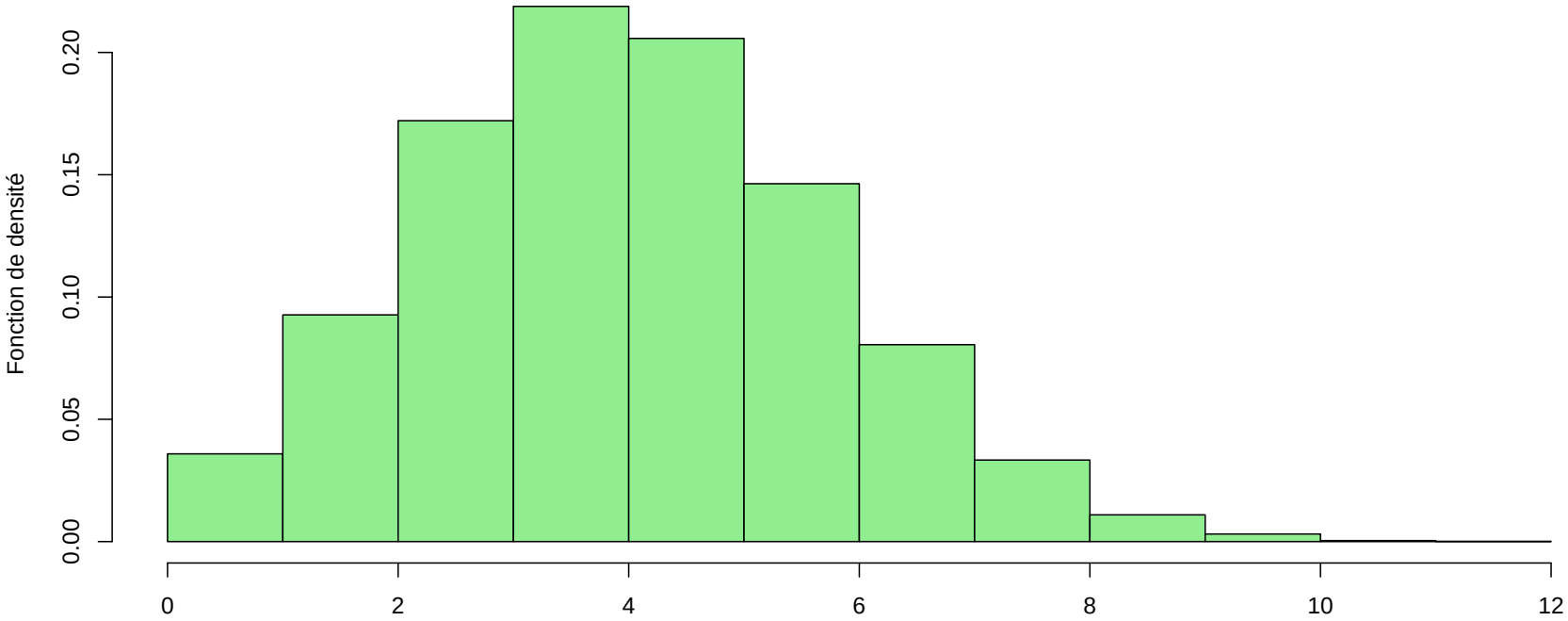



Figure 5: Binomial Distribution $\mathcal{B}(n, p)$.

# Some Probabilistic Distributions: Poisson

The Poisson distribution $\mathcal{P}(\lambda)$, $\lambda > 0$

$$\mathbb{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!} \text{ where } k = 0, 1, \cdots$$

Then $\mathbb{E}(X) = \lambda$ and $\text{Var}(X) = \lambda$ $(= \mathbb{E}(X))$.

Further, if $X_1 \sim \mathcal{P}(\lambda_1)$ and $X_2 \sim \mathcal{P}(\lambda_2)$ are independent, then

$$X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$$

Observe that a recursive equation can be obtained

$$\frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X = k)} = \frac{\lambda}{k+1} \text{ pour } k \geq 1$$

With R, `dpois(x, lambda)`, `qpois()` and `ppois()` are respectively the probability function, the quantile function and the cdf, see Poisson distribution

# Some Probabilistic Distributions: Poisson

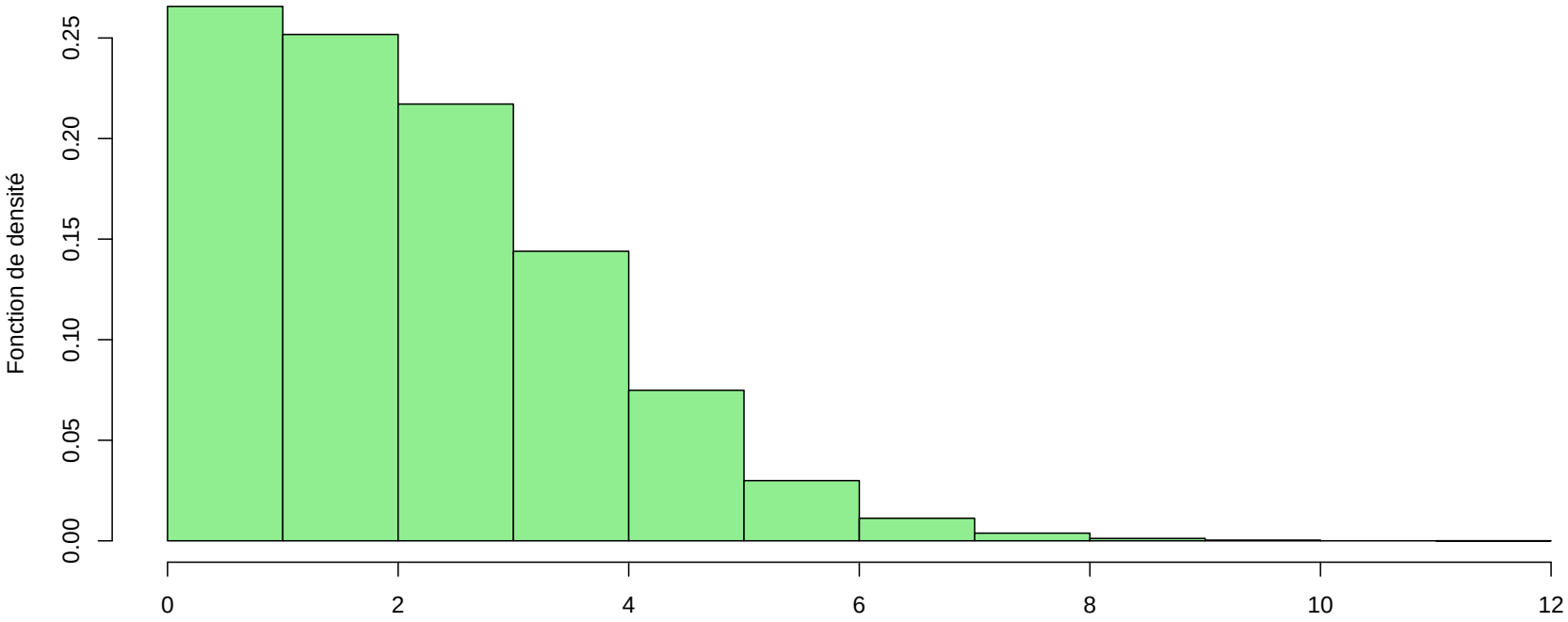



Figure 6: Poisson distribution, $\mathcal{P}(\lambda)$.

# Some Probabilistic Distributions: Geometric

The Geometric $\mathcal{G}(p)$, $p \in (0,1)$

$$\mathbb{P}(X = k) = p(1-p)^{k-1} \text{ for } k = 1, 2, \cdots$$

with cdf $\mathbb{P}(X \leq k) = 1 - p^k$.

Observe that this distribution satisfies the following relationship

$$\frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X = k)} = 1 - p \ (= \text{constant}) \text{ for } k \geq 1$$

First moments are here

$$\mathbb{E}(X) = \frac{1}{p} \text{ and } \text{Var}(X) = \frac{1-p}{p^2}.$$

see geometric distribution and the extention with the negative binomial distribution

## Some Probabilistic Distributions: Exponential

The exponential distribution $\mathcal{E}(\lambda)$, with $\lambda > 0$

$$F(x) = \mathbb{P}(X \leq x) = e^{-\lambda x} \text{ where } x \geq 0, f(x) = \lambda e^{-\lambda x}.$$

Then $\mathbb{E}(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$.

This is a memoryless distribution, since

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

In R, `dexp(x, rate)`, `qexp()` and `pexp()` are respectively the cdf, the quantile function and the density.

see exponential distribution

# Some Probabilistic Distributions: Exponential

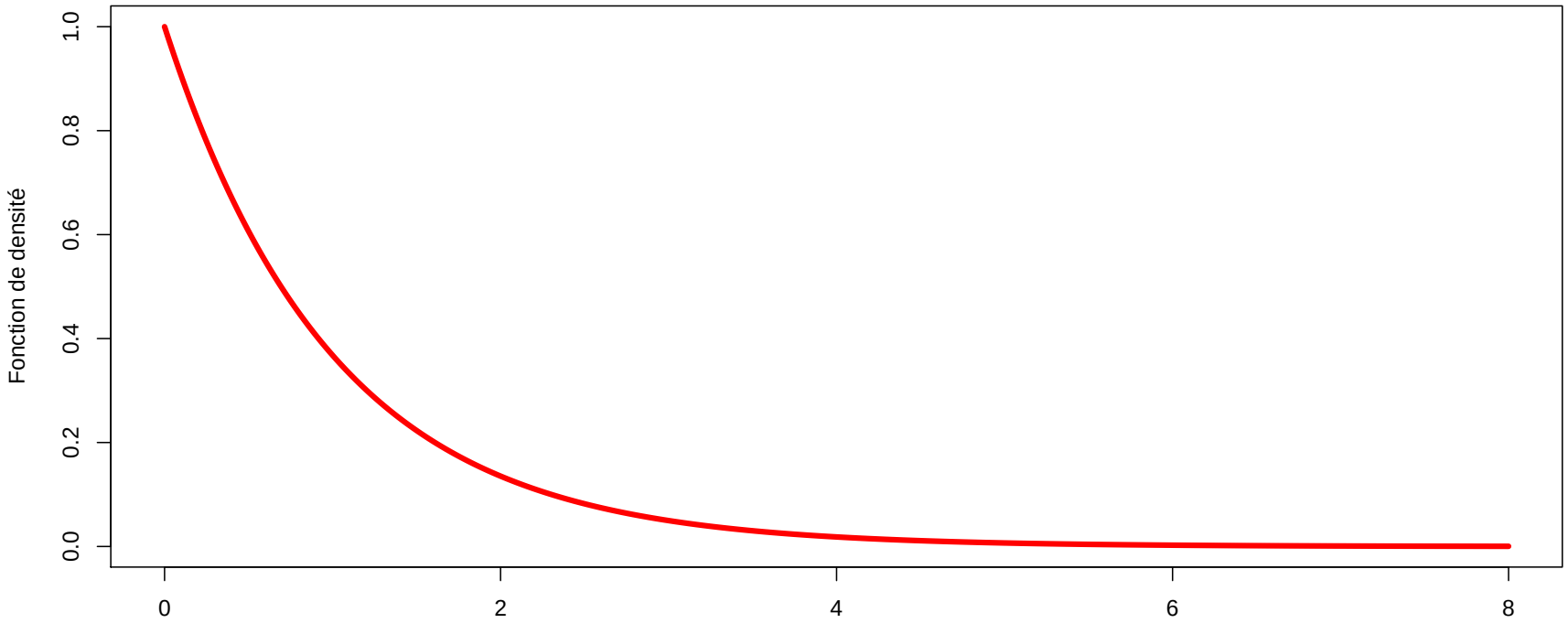



Figure 7: Exponential distribution, $\mathcal{E}(\lambda)$.

## Some Probabilistic Distributions: Gaussian

The Gaussian (or normal) distribution $\mathcal{N}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma > 0$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ for all } x \in \mathbb{R}.$$

Then $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

Observe that if $Z \sim \mathcal{N}(0,1)$, $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.

With `R`, `dnorm(x, mean, sd)`, `qnorm()` and `pnorm()` are respectively the cumulative distribution function, the quantile function and the density.

see Gaussian distribution

With `R`, `dnorm(x,mean=a,sd=b)` for the $\mathcal{N}(a, b)$ density.

# Some Probabilistic Distributions: Gaussian

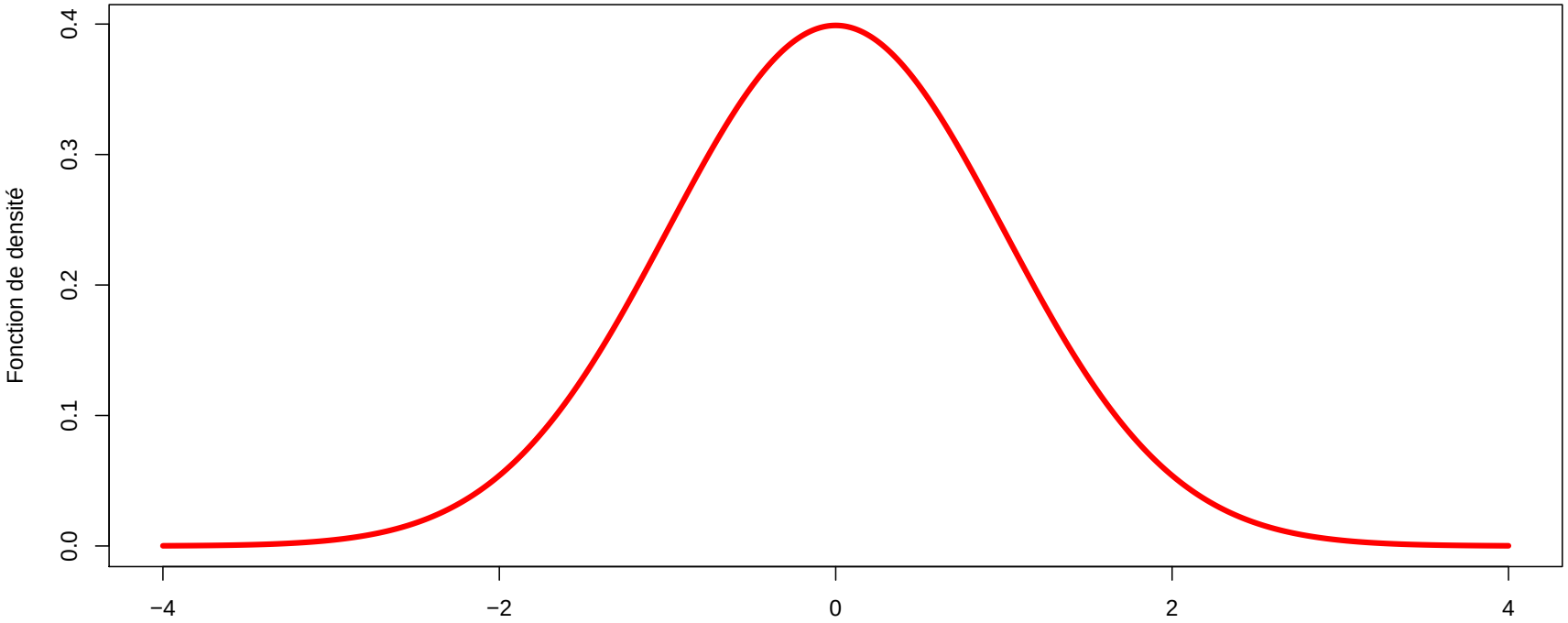



Figure 8: Normal distribution, $\mathcal{N}(0,1)$.

# Some Probabilistic Distributions: Gaussian

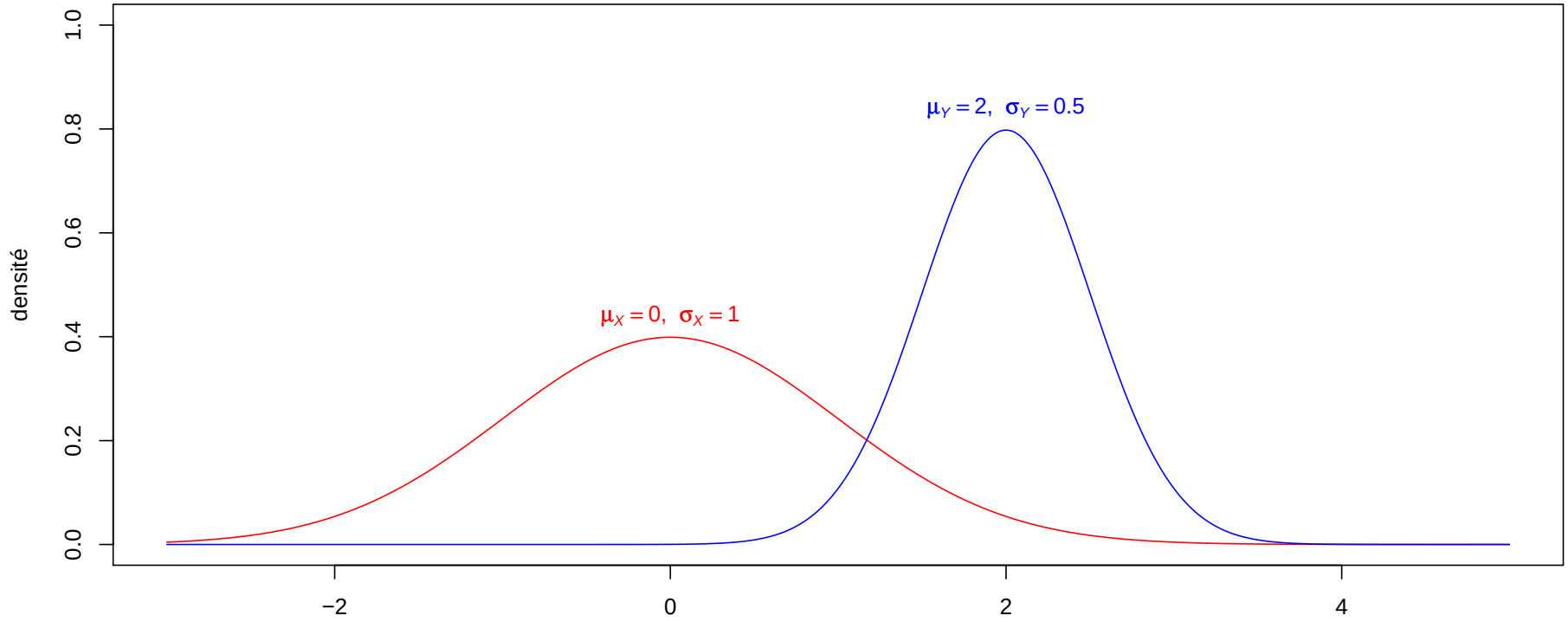



Figure 9: Densities of two Gaussian distributions, $X \sim \mathcal{N}(0, 1)$ and $X \sim \mathcal{N}(2, 0.5)$.

# Probability Distributions : The Gaussian Vector

The Gaussian **vector** $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ : $\boldsymbol{X} = (X_1, ..., X_n)$ is a Gaussian vector with mean $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}}\right)$ non-degenerated ($\Sigma$ is invertible) if its density is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^d,$$

**Proposition.** *Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a random vector with values in $\mathbb{R}^n$, then $\boldsymbol{X}$ is a Gaussian vector if and only if for any $\boldsymbol{a} = (a_1, ..., a_n) \in \mathbb{R}^d$, $\boldsymbol{a}^T\boldsymbol{X} = a_1X_1 + ... + a_nX_n$ has a (univariate) Gaussian distribution.*

see multivariate Gaussian distribution

## Probability Distributions : The Gaussian Vector

Hence, if $\boldsymbol{X}$ is a Gaussian vector, then for any $i$, $X_i$ has a (univariate) Gaussian distribution, but its converse it not necessarily true.

**Proposition.** *Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a random vector with mean $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and with covariance matrix $\boldsymbol{\Sigma}$, if $\boldsymbol{A}$ is a $k \times n$ matrix, and $\boldsymbol{b} \in \mathbb{R}^k$, then $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{b}$ is a Gaussian vector $\mathbb{R}^k$, with distribution $\mathcal{N}\left(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A}^T\right)$.*

Observe that if $(X_1, X_2)$ is a Gaussian vector $X_1$ and $X_2$ are independent if and only if

$$\text{Cov}(X_1, X_2) = \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) = 0.$$

# Probability Distributions : The Gaussian Vector

**Proposition.** *If* $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ *is a Gaussian vector with mean*

$$\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ and covariance matrix covariance } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \text{ then}$$

$$\boldsymbol{X}_2 | \boldsymbol{X}_1 = \boldsymbol{x}_1 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_1 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

Cf autoregressive time series : $X_t = \rho X_{t-1} + \varepsilon_t$, where $X_0 = 0$, $\varepsilon_1, \cdots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$, i.e. $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$. Then

$$\boldsymbol{X} = (X_1, \cdots, X_n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} = [\Sigma_{i,j}] = [\text{Cov}(X_i, X_j)] = [\rho^{|i-j|}].$$

# Probability Distributions : The Gaussian Vector

Let $\boldsymbol{Z} = (Y, X) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{Y,Y} & \Sigma_{Y,X} \\ \Sigma_{X,Y} & \Sigma_{X,X} \end{pmatrix}$$

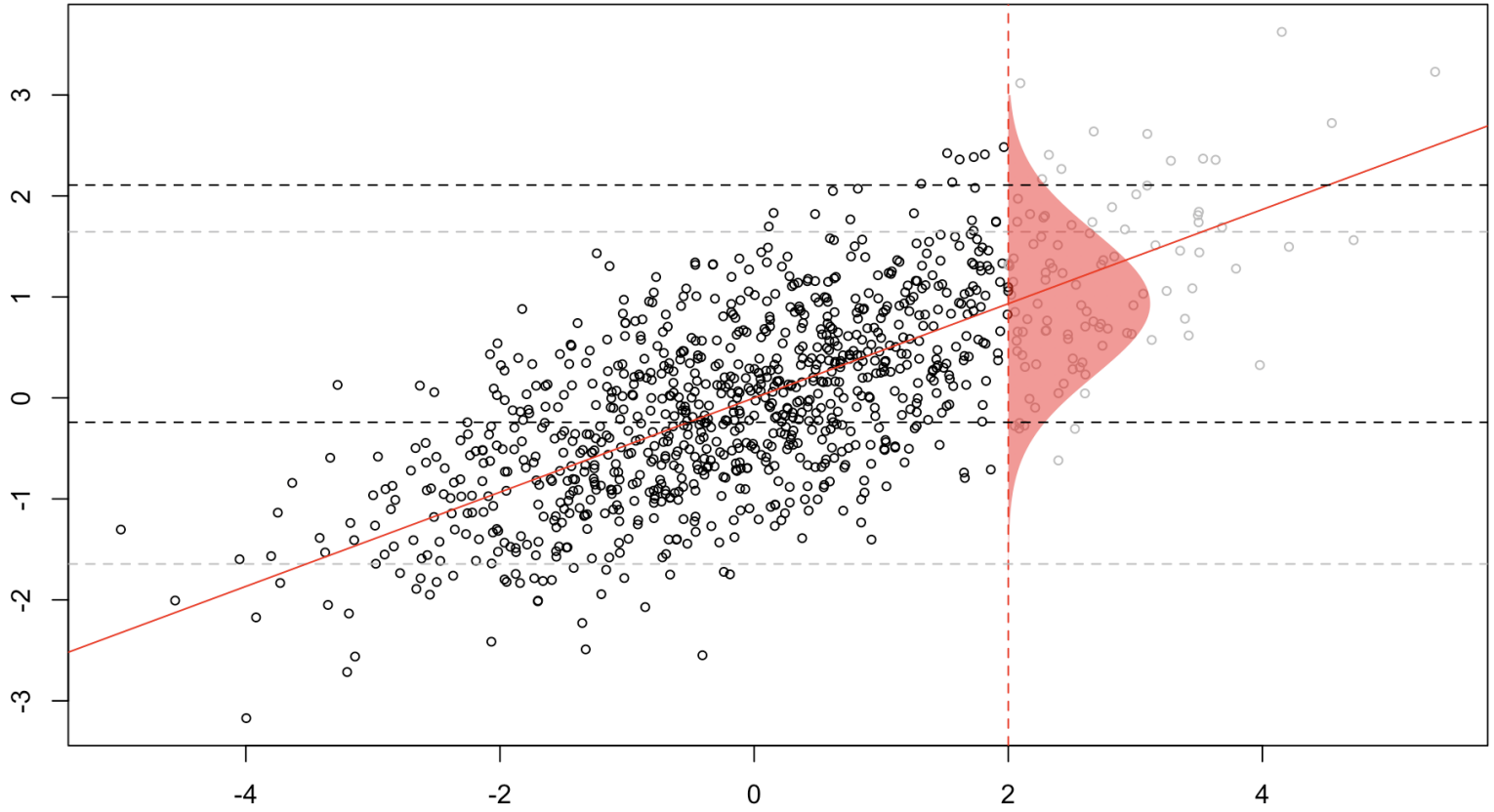
Then

$$Y|X = x \sim \mathcal{N}(\mu_{Y|x}, \sigma^2_{Y|x}) \text{ where } \begin{cases} \mu_{Y|x} = \mu_Y + \Sigma_{Y,X}\Sigma^{-1}_{X,X}(x - \mu_X) \\ \sigma^2_{Y|x} = \Sigma_{Y,Y} - \Sigma_{Y,X}\Sigma^{-1}_{X,X}\Sigma_{X,Y} \end{cases}$$

Hence, $\mathbb{E}[Y|X = x] = \mu_{Y|x}$ is linear in $x$, with slope $\text{Corr}(X, Y)\sqrt{\Sigma_{Y,Y}\Sigma^{-1}_{X,X}}$

and $\text{Var}[Y|X = x] = \sigma^2_{Y|x}$ is constant (furthermore $\text{Var}[Y|X = x] \leq \text{Var}[Y]$)

# Probability Distributions : The Gaussian Vector



## Probability Distribution

In dimension 2, a vector $(X, Y)$ centered (i.e. $\mu = \mathbf{0}$) is a Gaussian vector if its density is

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{(\sigma_x\sigma_y)}\right)\right)$$

with covariance matrix $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}.$$

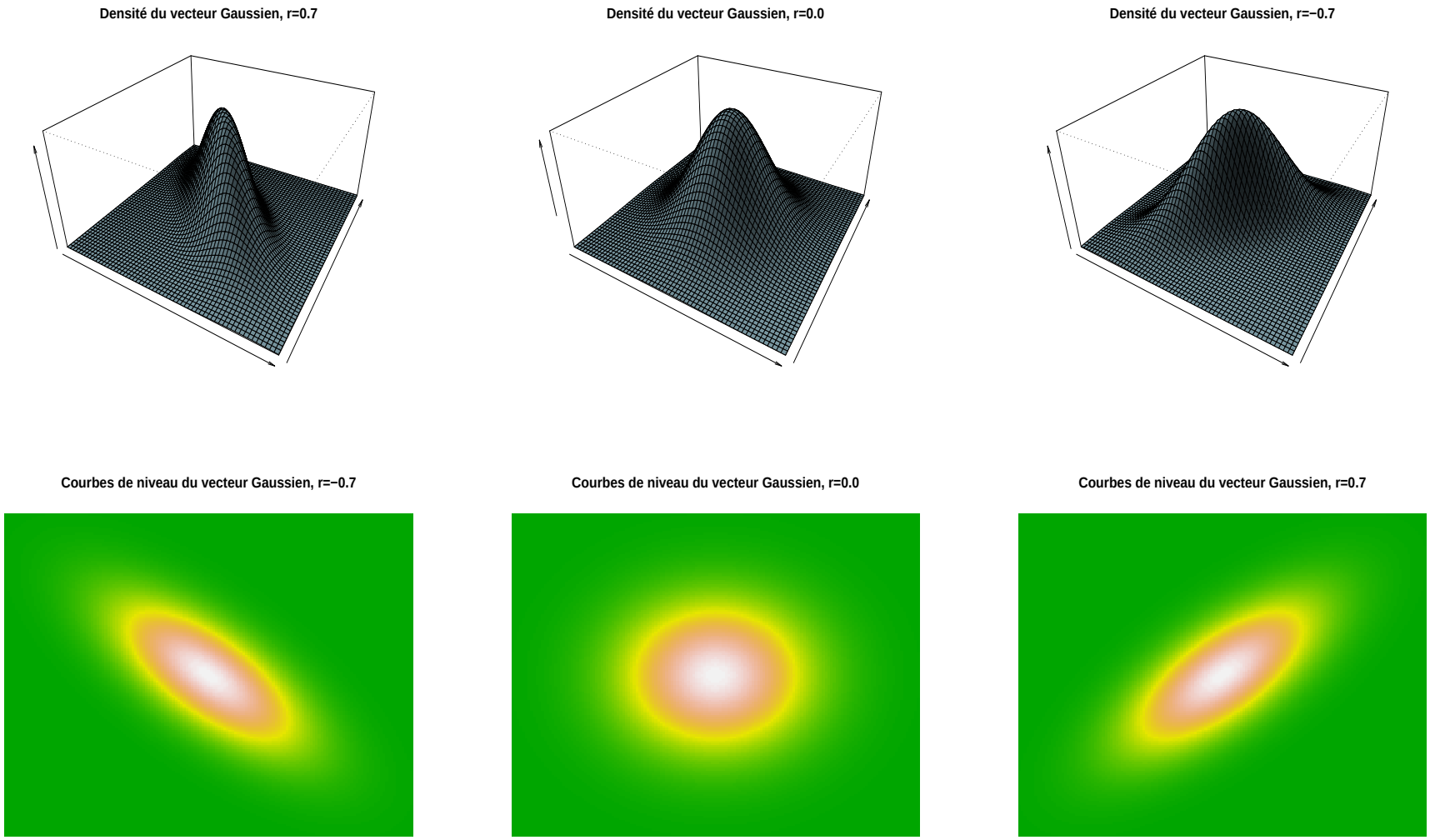



Figure 10: Bivariate Gaussien distribution.

# Probability Distributions

The chi-squared distribution $\chi^2(\nu)$, with $\nu \in \mathbb{N}^*$ has density

$$x \mapsto \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \text{ where } x \in [0; +\infty),$$

where $\Gamma$ denotes the Gamma function ($\Gamma(n+1) = n!$). Observe that $\mathbb{E}(X) = \nu$ et $\text{Var}(X) = 2\nu$. $\nu$ are the degrees of freedom, see chi-squared distribution

**Proposition.** *If $X_1, \cdots, X_\nu \sim \mathcal{N}(0,1)$ are independent variables, then*

$$Y = \sum_{i=1}^{\nu} X_i^2 \sim \chi^2(\nu), \text{ when } \nu \in \mathbb{N}.$$

With R, `dchisq(x, df)`, `qchisq()` and `pchisq()` are respectively the cdf, the quantile function and the density.

This is a particular case of the Gamma distribution, $X \sim \mathcal{G}\left(\frac{k}{2}, \frac{1}{2}\right)$, see see Gamma

# Probability Distributions

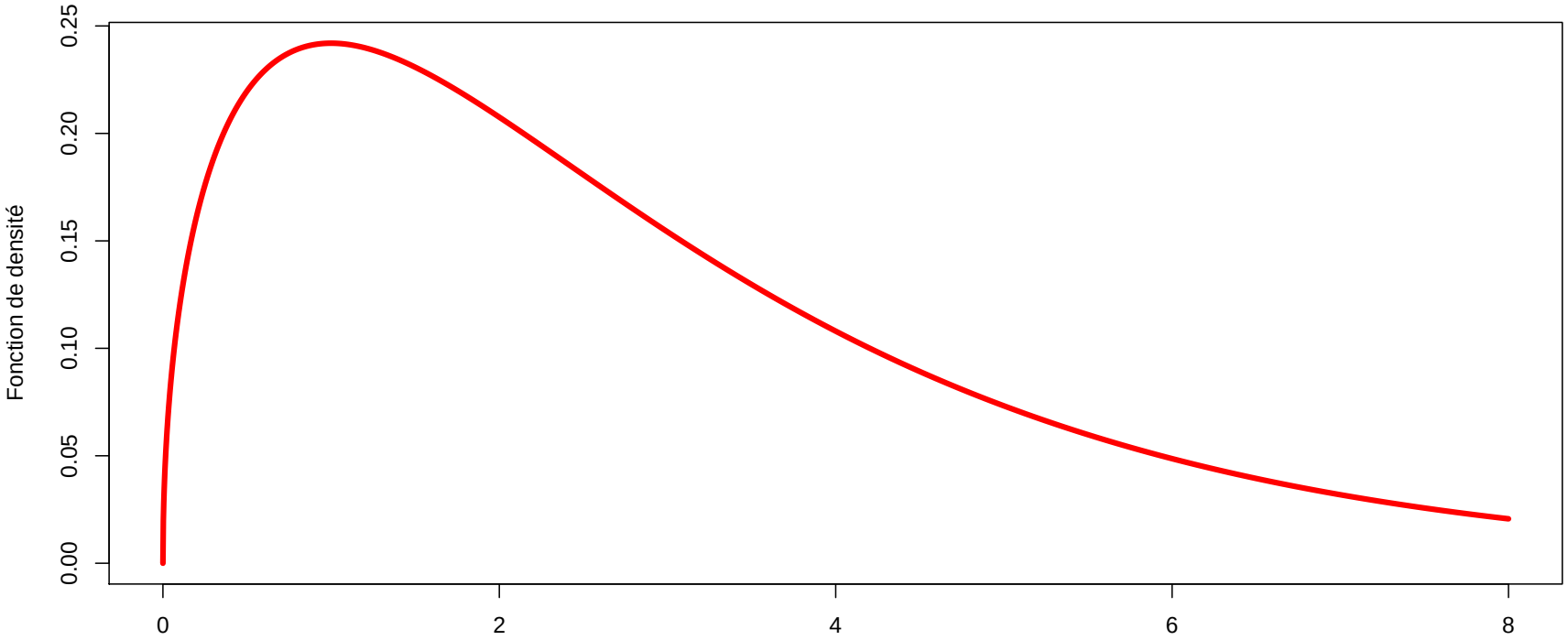



Figure 11: Chi-square distribution, $\chi^2(\nu)$.

# Probability Distributions

The Student's-$t$ distribution $\mathcal{S}t(\nu)$, has density

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-(\frac{\nu+1}{2})},$$

Observe that

$$\mathbb{E}(X) = 0 \text{ and } \text{Var}(X) = \frac{\nu}{\nu-2} \text{ when } \nu > 2.$$

**Proposition.** *If $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2(\nu)$ are independents, then*

$$T = \frac{X}{\sqrt{Y/\nu}} \sim \mathcal{S}t(\nu).$$

see Student's $t$

# Probability Distributions

Let $X_1, \cdots, X_n$ be $\mathcal{N}(\mu, \sigma^2)$ independent random variables. Let

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \text{ and } {S_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Then $\dfrac{(n-1)S_n^2}{\sigma^2}$ has a $\chi^2(n-1)$ distribution, and furthermore

$$T = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim \mathcal{S}t(n-1).$$

With R, `dt(x, df)`, `qt()` and `pt()` are respectively the cdf, the quantile and the density functions.

# Probability Distributions

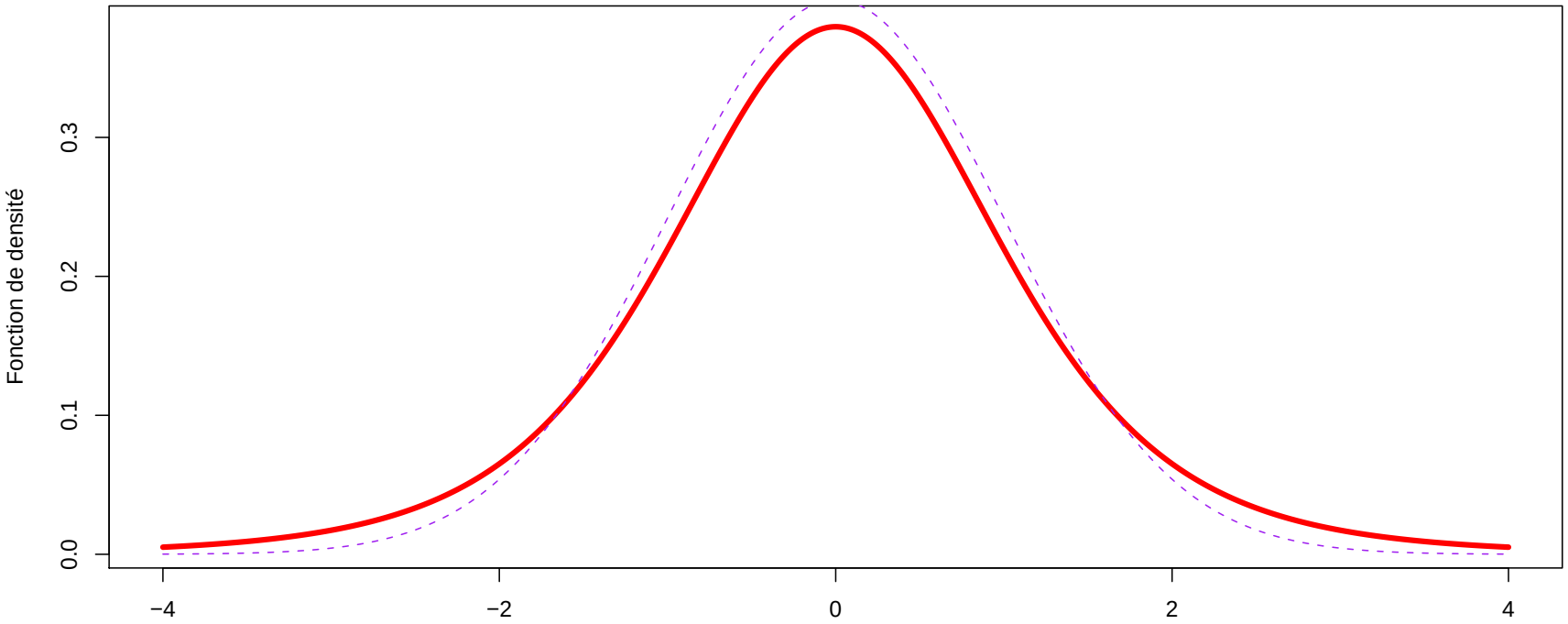



Figure 12: Student $t$ distributions, $\mathcal{S}t(\nu)$.

# Probability Distributions

The Fisher distribution $\mathcal{F}(d_1, d_2)$, has density

$$x \mapsto \frac{1}{x\,\mathrm{B}(d_1/2, d_2/2)} \left(\frac{d_1\,x}{d_1\,x + d_2}\right)^{d_1/2} \left(1 - \frac{d_1\,x}{d_1\,x + d_2}\right)^{d_2/2}$$

for $x \geq 0$ and $d_1, d_2 \in \mathbb{N}$, where $B$ denotes the Beta function.

$\mathbb{E}(X) = \dfrac{d_2}{d_2 - 2}$ when $d_2 > 2$ and $\mathrm{Var}(X) = \dfrac{2\,d_2^2\,(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}$ when $d_2 > 4$.

If $X \sim \mathcal{F}(\nu_1, \nu_2)$, then $\dfrac{1}{X} \sim \mathcal{F}(\nu_2, \nu_1)$.

If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$ are independent $Y = \dfrac{X_1/\nu_1}{X_2/\nu_2} \sim \mathcal{F}(\nu_1, \nu_2)$.

see Fisher's $\mathcal{F}$ on wikipedia...

# Probability Distributions

With R, `df(x, df1, df2)`, `qf()` and `pf()` denote the cdf, the quantile and the density functions.

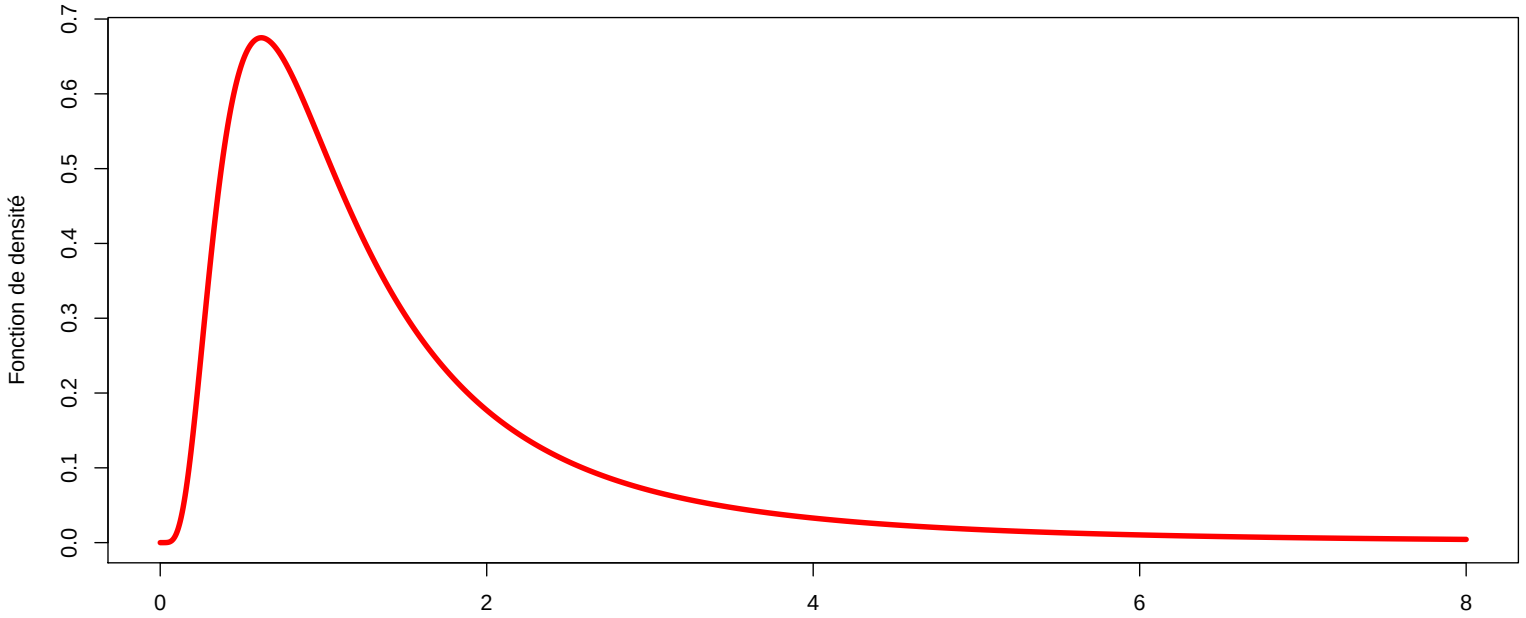



Figure 13: Fisher distribution, $\mathcal{F}(d_1, d_2)$.

# Gamma Distribution

We have seen the exponential distribution. We can extend it to get the Gamma distribution, see Gamma distribution

With a shape parameter $k$ and a scale parameter $\theta$,

$$f(x;k,\theta)=\frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)} \quad \text{for } x>0 \text{ and } k,\theta>0.$$

With a shape parameter $\alpha=k$ and an inverse scale parameter $\beta=1/\theta$, called a rate parameter,

$$f(x;\alpha,\beta)=\frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x>0 \text{ and } \alpha,\beta>0,$$

Observe that $\mathbb{E}[X]=k\theta=\dfrac{\alpha}{\beta}$ while $\text{Var}[X]=k\theta^2=\dfrac{\alpha}{\beta^2}$

## The Exponential Family

Considérons des lois de paramètres $\theta$ (et $\varphi$) dont la fonction de densité (par rapport à la mesure dominante adéquate (mesure de comptage sur $\mathbb{N}$ ou mesure de Lebesgue sur $\mathbb{R}$) s'écrit

$$f(y|\theta,\varphi)=\exp\left(\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi)\right),$$

où $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ sont des fonctions, et où $\theta$ est appelé paramètre naturel. Le paramètre $\theta$ est le paramètre d'intérêt tandis que $\varphi$ est considéré comme un paramètres de nuisance (et supposé connu, dans un premier temps).

cf exponential family

# La famille exponentielle

**Exemple** La loi Gaussienne de moyenne $\mu$ et de variance $\sigma^2$, $\mathcal{N}(\mu, \sigma^2)$ appartient à cette famille, avec $\theta = \mu$, $\varphi = \sigma^2$, $a(\varphi) = \varphi$, $b(\theta) = \theta^2/2$ et

$$c(y, \varphi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right), \quad y \in \mathbb{R},$$

**Exemple** La loi de Bernoulli de moyenne $\pi$, $\mathcal{B}(\pi)$ correspond au cas $\theta = \log\{p/(1-p)\}$, $a(\varphi) = 1$, $b(\theta) = \log(1+\exp(\theta))$, $\varphi = 1$ et $c(y, \varphi) = 0$.

**Exemple** La loi binomiale de moyenne $n\pi$, $\mathcal{B}(n, \pi)$ correspond au cas $\theta = \log\{p/(1-p)\}$, $a(\varphi) = 1$, $b(\theta) = n\log(1+\exp(\theta))$, $\varphi = 1$ et $c(y, \varphi) = \log\binom{n}{y}$.

**Exemple** La loi de Poisson de moyenne $\lambda$, $\mathcal{P}(\lambda)$ appartient à cette famille,

$$f(y|\lambda) = \exp(-\lambda)\frac{\lambda^y}{y!} = \exp\Big(y\log\lambda - \lambda - \log y!\Big), \quad y \in \mathbb{N},$$

avec $\theta = \log\lambda$, $\varphi = 1$, $a(\varphi) = 1$, $b(\theta) = \exp\theta = \lambda$ et $c(y, \varphi) = -\log y!$.

## La famille exponentielle

**Exemple** La loi Binomiale Négative, de paramètres $r$ et $p$,

$$f(k|r,p) = \binom{y+r-1}{y}(1-p)^r p^y, \quad y \in \mathbb{N}.$$

que l'on peut écrire

$$f(k|r,p) = \exp\left(y\log p + r\log(1-p) + \log\binom{y+r-1}{y}\right)$$

soit $\theta = \log p$, $b(\theta) = -r\log p$ et $a(\varphi) = 1$

# La famille exponentielle

**Exemple** La loi Gamma (incluant la loi exponentielle) de moyenne $\mu$ et de variance $\nu^{-1}$,

$$f(y|\mu,\nu)=\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu}y^{\nu-1}\exp\left(-\frac{\nu}{\mu}y\right),\quad y\in\mathbb{R}_+,$$

est également dans la famille exponentielle. Il faut choisir $\theta=-\frac{1}{\mu}$, $a(\varphi)=\varphi$, $b(\theta)=-\log(-\theta)$, $\varphi=\nu^{-1}$ et

$$c(y,\varphi)=\left(\frac{1}{\varphi}-1\right)\log(y)-\log\left(\Gamma\left(\frac{1}{\varphi}\right)\right)$$

On reviendra sur cette loi dans la section du cours sur la modélisation des coûts de sinistres.

# Espérance et variance

Pour une variable aléatoire $Y$ dont la densité est de la forme exponentielle, alors

$$\mathbb{E}(Y) = b'(\theta) \text{ et } \text{Var}(Y) = b''(\theta)\varphi,$$

i.e. la variance de $Y$ apparaît comme le produit de deux fonctions:

- la première, $b''(\theta)$ , qui dépend uniquement du paramètre $\theta$ est appelée *fonction variance*,
- la seconde est indépendante de $\theta$ et dépend uniquement de $\varphi$.

En notant $\mu = \mathbb{E}(Y)$, on voit que le paramètre $\theta$ est lié à la moyenne $\mu$. La fonction variance peut donc être définie en fonction de $\mu$ , nous la noterons dorénavant

$$\text{V}(\mu) = b''([b']^{-1}(\mu))\varphi.$$

**Exemple** Dans le cas de la loi normale, $\text{V}(\mu) = 1$, dans le cas de la loi de Poisson, $V(\mu) = \mu$ alors que dans le cas de la loi Gamma, $V(\mu) = \mu^2$.

## Espérance et fonction lien

Notons que la fonction variance caractérise complètement la loi de la famille exponentielle. Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, dite *fonction de lien canonique*, permettant de relier l'espérance $\mu$ au paramètre naturel (ou canonique) $\theta$. Le lien canonique est tel que $g_\star(\mu) = \theta$. Or, $\mu = b'(\theta)$ donc $g_\star(\cdot) = b'(\cdot)^{-1}$.

**Exemple** Pour la loi normale, $\theta = \mu$ (`link='identity'`),

**Exemple** Pour la loi de Poisson, $\theta = \log(\mu)$ (`link='log'`)

**Exemple** Pour la loi de Bernoulli, $\theta = logit(\mu) = \log \dfrac{\mu}{1-\mu}$, (`link='logit'`)

**Exemple** Pour la loi Gamma, $\theta = 1/\mu$ (`link='inverse'`)

# Espérance et variance

| Loi de probabilité | $V(\mu)$ |
|---|---|
| Normale | $1$ |
| Poisson | $\mu$ |
| Gamma | $\mu^2$ |
| Inverse gaussienne | $\mu^3$ |
| Binomiale | $\mu(1-\mu)$ |

# Espérance et variance, la famille Tweedie

Tweedie (1984) a suggéré la famille suivante

$$f(y|\mu,\varphi) = A(y,\varphi)\cdot\exp\left\{\frac{1}{\varphi}\Big[y\theta(\mu)-\kappa\big(\theta(\mu)\big)\Big]\right\},$$

où

$$\theta(\mu)=\begin{cases}\dfrac{\mu^{1-\gamma}}{1-\gamma} & \gamma\neq 1\\ \log\mu & \gamma=1\end{cases} \quad \text{et} \quad \kappa\big(\theta(\mu)\big)=\begin{cases}\dfrac{\mu^{2-\gamma}}{2-\gamma} & \gamma\neq 2\\ \log\mu & \gamma=2\end{cases}$$

La loi de $Y$ est alors une loi Poisson composée, avec des sauts Gamma,

$$Y\sim\mathcal{CP}oi\left(\mu^{2-\gamma}\varphi(2-\gamma),\mathcal{G}\left(-\frac{2-\gamma}{\varphi(1-\gamma)},\varphi(2-\gamma)\mu^{\gamma-1}\right)\right),$$

où $\gamma\in[1,2]$.

**Remarque** On a une mesure de Dirac en 0 avec distribution (continue) définie sur $\mathbb{R}^+$.

## Espérance et variance, la famille Tweedie

On obtient alors une fonction variance de la forme $\mathrm{V}(\mu)=\varphi\mu^\gamma$. On retrouve le modèle de Poisson quand $\gamma\to 1$ (ou $\alpha\to\infty$) et une loi Gamma quand $\gamma\to 2$ (ou $\alpha\to 0$). Il est en fait possible d'obtenir une classe beaucoup plus large, y compris dans le cas où $\gamma>2$ en considérant des lois stables.

## Paramètre naturel, et lien canonique

Le lien canonique est tel que $g(\mu_i) = \theta_i$. Or, $\mu_i = b'(\theta_i)$ d'où $g^{-1} = b'$.

| Loi de probabilité | Fonction de lien canonique |
|---|---|
| Normale | $\eta = \mu$ |
| Poisson | $\eta = \ln \mu$ |
| Gamma | $\eta = 1/\mu$ |
| Inverse gaussienne | $\eta = 1/\mu^2$ |
| Binomiale | $\eta = \ln \mu - \ln(1-\mu) = \text{logit}(\mu)$ |

# Conditional Distributions

- **Mixture of Bernoulli distribution $\mathcal{B}(\Theta)$**

Let $\Theta$ denote a random variable taking values $\theta_1, \theta_2 \in [0,1]$ with probabilities $p_1$ and $p_2$ (with $p_1 + p_2 = 1$). Assume that

$$X|\Theta = \theta_1 \sim \mathcal{B}(\theta_1) \text{ and } X|\Theta = \theta_2 \sim \mathcal{B}(\theta_2).$$

The non-conditionnal distribution of $X$ is

$$\mathbb{P}(X = x) = \mathbb{P}(X = x|\Theta = \theta_1) \cdot p_1 + \mathbb{P}(X = x|\Theta = \theta_2) \cdot p_2,$$

$$\mathbb{P}(X = 0) = \mathbb{P}(X = 0|\Theta = \theta_1) \cdot p_1 + \mathbb{P}(X = 0|\Theta = \theta_2) \cdot p_2 = 1 - \theta_1 p_1 - \theta_2 p_2$$

$$\mathbb{P}(X = 1) = \mathbb{P}(X = 1|\Theta = \theta_1) \cdot p_1 + \mathbb{P}(X = 1|\Theta = \theta_2) \cdot p_2 = \theta_1 p_1 + \theta_2 p_2$$

i.e. $X \sim \mathcal{B}(\theta_1 p_1 + \theta_2 p_2)$.

# Conditional Distributions

Observe that

$$\begin{aligned}\mathbb{E}(X) &= \theta_1 p_1 + \theta_2 p_2 \\ &= \mathbb{E}(X|\Theta=\theta_1)\mathbb{P}(\Theta=\theta_1) + \mathbb{E}(X|\Theta=\theta_2)\mathbb{P}(\Theta=\theta_2) = \mathbb{E}(\mathbb{E}(X|\Theta))\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) &= [\theta_1 p_1 + \theta_2 p_2][1-\theta_1 p_1 - \theta_2 p_2] \\ &= \theta_1^2 p_1 + \theta_2^2 p_2 - [\theta_1 p_1 + \theta_2 p_2]^2 \\ &+ [\theta_1(1-\theta_1)]p_1 + [\theta_2(1-\theta_2)]p_2 \\ &= \mathbb{E}(X|\Theta=\theta_1)^2\mathbb{P}(\Theta=\theta_1) + \mathbb{E}(X|\Theta=\theta_2)^2\mathbb{P}(\Theta=\theta_2) \\ &- [\mathbb{E}(X|\Theta=\theta_1)\mathbb{P}(\Theta=\theta_1) + \mathbb{E}(X|\Theta=\theta_2)\mathbb{P}(\Theta=\theta_2)]^2 \\ &+ \mathrm{Var}(X|\Theta=\theta_1)\mathbb{P}(\Theta=\theta_1) + \mathrm{Var}(X|\Theta=\theta_2)\mathbb{P}(\Theta=\theta_2) \\ &= \underbrace{\mathbb{E}([\mathbb{E}(X|\Theta)]^2) - [\mathbb{E}(\mathbb{E}(X|\Theta))]^2}_{\mathrm{Var}(\mathbb{E}(X|\Theta))} + \mathbb{E}(\mathrm{Var}(X|\Theta)\end{aligned}$$

# Conditional Distributions

- **Mixture of Poisson distributions** $\mathcal{P}(\Theta)$

Let $\Theta$ denote a random variable taking values $\theta_1, \theta_2 \in [0, 1]$ with probabilities $p_1$ and $p_2$ (with $p_1 + p_2 = 1$). Assume that

$$X|\Theta = \theta_1 \sim \mathcal{P}(\theta_1) \text{ and } X|\Theta = \theta_2 \sim \mathcal{P}(\theta_2).$$

Then

$$\mathbb{P}(X = x) = \frac{e^{-\theta_1}\theta_1^x}{x!} \cdot p_1 + \frac{e^{-\theta_2}\theta_2^x}{x!} \cdot p_2,$$

# Conditional Distributions

- **Continuous Mixture of Poisson $\mathcal{P}(\Theta)$ distributions**

Let $\Theta$ be a continuous random variable, taking values in $(0,\infty)$, with denisty $\pi(\cdot)$. Assume that

$$X|\Theta=\theta \sim \mathcal{P}(\theta) \text{ for all } \theta>0$$

Then

$$\mathbb{P}(X=x)=\int_0^\infty \mathbb{P}(X=x|\Theta=\theta)\pi(\theta)d\theta.$$

Further

$$\mathbb{E}(X)=\mathbb{E}(\mathbb{E}(X|\Theta))=\mathbb{E}(\Theta)$$

$$\text{Var}(X)=\text{Var}(\mathbb{E}(X|\Theta))+\mathbb{E}(\text{Var}(X|\Theta))=\text{Var}(\Theta)+\mathbb{E}(\Theta)>\mathbb{E}(\Theta).$$

# Conditional Distributions, Mixtures and Heterogeneity

$$f(x) = f(x|\Theta = \theta_1) \times \mathbb{P}(\Theta = \theta_1) + f(x|\Theta = \theta_2) \times \mathbb{P}(\Theta = \theta_2).$$

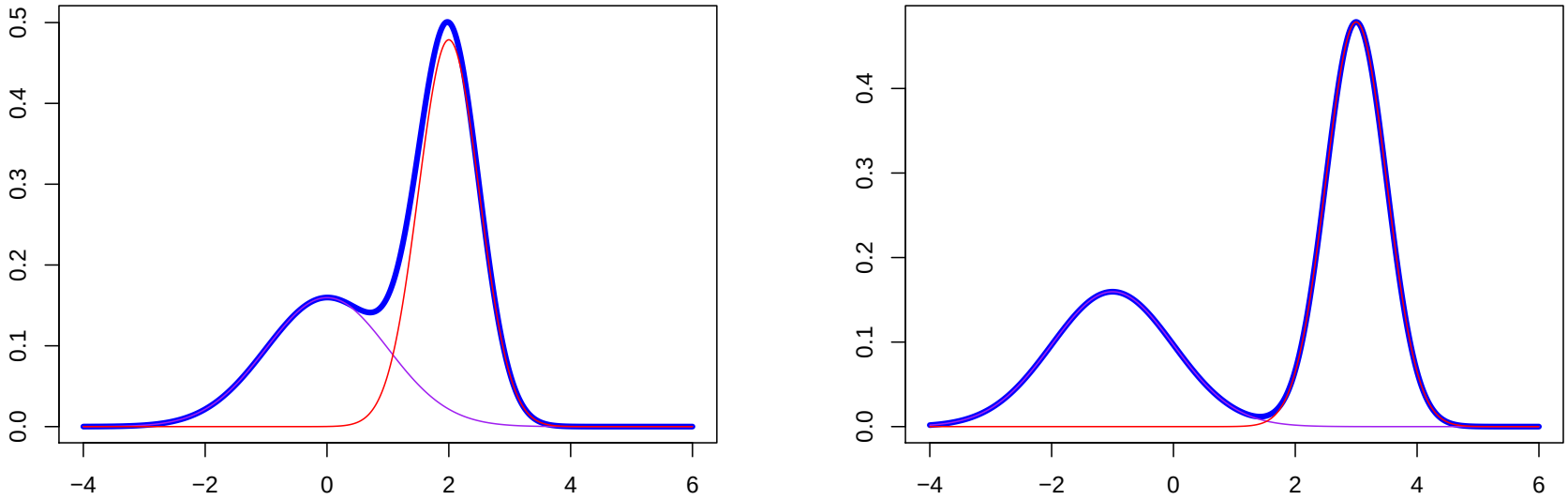



Figure 14: Mixture of Gaussian Distributions.

## Conditional Distributions, Mixtures and Heterogeneity

Mixtures are related to heterogeneity.

- In linear econometric models, $Y|\boldsymbol{X}=\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta},\sigma^2)$.
- In logit/probit models, $Y|\boldsymbol{X}=\boldsymbol{x}\sim\mathcal{B}(p[\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}])$ where $p[\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}]=\dfrac{e^{\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}}$.

E.g. $Y|X_1=\text{male}\sim\mathcal{B}(p_m)$ et $Y|X_1=\text{female}\sim\mathcal{B}(p_f)$ with only one categorical variable

E.g. $Y|(X_1=\text{male},X_2=x)\sim\mathcal{B}\left(\dfrac{e^{\beta_m+\beta_2 x}}{1+e^{\beta_m+\beta_2 x}}\right)$

## Some words on Convergence

Sequence of random variables $(X_n)$ converges almost surely towards $X$, denoted $X_n \overset{a.s.}{\to} X$, if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \text{ for all } \omega \in A,$$

where $A$ is a set such that $\mathbb{P}(A) = 1$. It is possible to say that $(X_n)$ converges towards $X$ with probability 1. Obserse that $X_n \overset{a.s.}{\to} X$ if and only if

$$\forall \varepsilon > 0, \mathbb{P}(\limsup\{|X_n - X| > \varepsilon\}) = 0.$$

It is also possible to control variation of the sequence $(X_n)$ : let $(\varepsilon_n)$ such that $\sum_{n\geq 0} \mathbb{P}(|X_n - X| > \varepsilon_n) < \infty$ where $\sum_{n\geq 0} \varepsilon_n < \infty$, then $(X_n)$ converges almost surely towards $X$.

cf convergence(s) of random variables

## Some words on Convergence

Sequence of random variables $(X_n)$ converges in $L^p$ towards $X$ - or on average of order $p$ - denoted $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} \mathbb{E}\left(|X_n - X|^p\right) = 0.$$

If $p = 1$ it is the convergence in mean and if $p = 2$, it is the quadratic convergence.

Suppose that $X_n \xrightarrow{a.s.} X$ and that there exists a random variable $Y$ such that for $n \geq 0$, $|X_n| \leq Y$ $\mathbb{P}$-almost surely with $Y \in L^p$, then $X_n \in L^p$ et $X_n \xrightarrow{L^p} X$.

## Some words on Convergence

The sequence $(X_n)$ converges in probability towards $X$, denoted $X_n \xrightarrow{\mathbb{P}} X$, if

$$\forall \varepsilon > 0, \lim_{n \to \infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function, if $X_n \xrightarrow{\mathbb{P}} X$ then $f(X_n) \xrightarrow{\mathbb{P}} f(X)$.

Furthermore, if either $X_n \xrightarrow{a.s.} X$ or $X_n \xrightarrow{L^1} X$ then $X_n \xrightarrow{\mathbb{P}} X$.

A sufficient condition to have $X_n \xrightarrow{\mathbb{P}} a$ is that

$$\lim_{n \to \infty} \mathbb{E}X_n = a \text{ and } \lim_{n \to \infty} \text{Var}(X_n) = 0$$

# Some words on Convergence

- (Strong) Law of Large Numbers

Suppose $X_i$'s are i.i.d. with finite expected value $\mu = \mathbb{E}(X_i)$, then $\overline{X}_n \overset{a.s.}{\to} \mu$ as $n \to \infty$.

- (Weak) Law of Large Numbers

Suppose $X_i$'s are i.i.d. with finite expected value $\mu = \mathbb{E}(X_i)$, then $\overline{X}_n \overset{\mathbb{P}}{\to} \mu$ as $n \to +\infty$.

see Law of Large Numbers

## Some words on Convergence

Sequence $(X_n)$ converges in distribution towards $X$, denoted $X_n \xrightarrow{\mathcal{L}} X$, if for any continuous function $h$

$$\lim_{n\to\infty} \mathbb{E}\left(h\left(X_n\right)\right) = \mathbb{E}\left(h\left(X\right)\right).$$

Convergence in distribution is the same as convergence of distribution function $X_n \xrightarrow{\mathcal{L}} X$if for any $t \in \mathbb{R}$ where $F_X$ is continuous

$$\lim_{n\to\infty} F_{X_n}\left(t\right) = F_X\left(t\right).$$

## Some words on Convergence

Let $h : \mathbb{R} \to \mathbb{R}$ denote a continuous function. If $X_n \xrightarrow{\mathcal{L}} X$ then $h(X_n) \xrightarrow{\mathcal{L}} h(X)$. Furthermore, if $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \xrightarrow{\mathcal{L}} X$ (the converse is valid if the limit is a constant).

- Central Limit Theorem

Let $X_1, X_2 \ldots$ denote i.i.d. random variables with mean $\mu$ and variance $\sigma^2$, then :

$$\frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{\text{Var}(\overline{X}_n)}} = \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{\mathcal{L}} X \text{ where } X \sim \mathcal{N}(0,1)$$

see Central Limit Theorem or the related blog post

# Visualization of Convergence

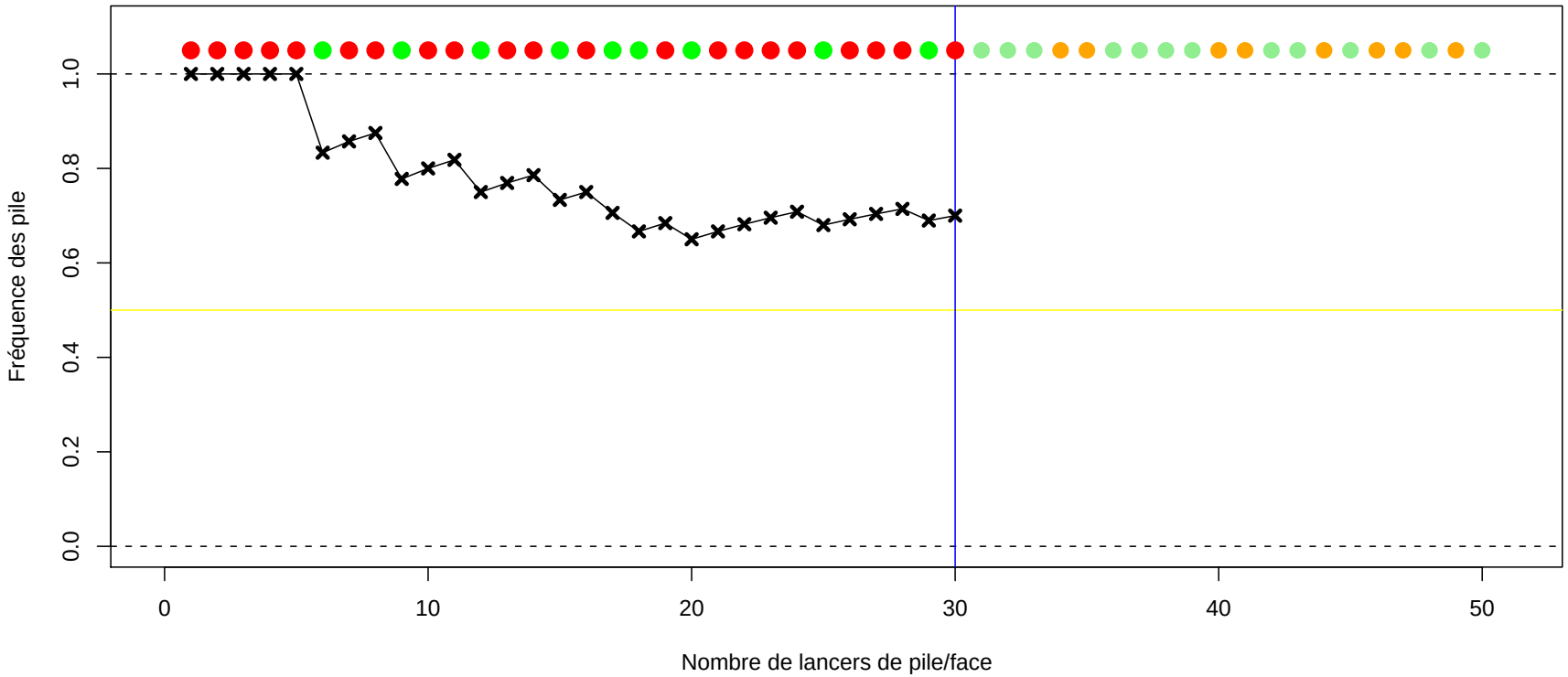



Figure 15: Convergence of the (empirical) mean $\overline{(x)}_n$.

# Visualization of Convergence

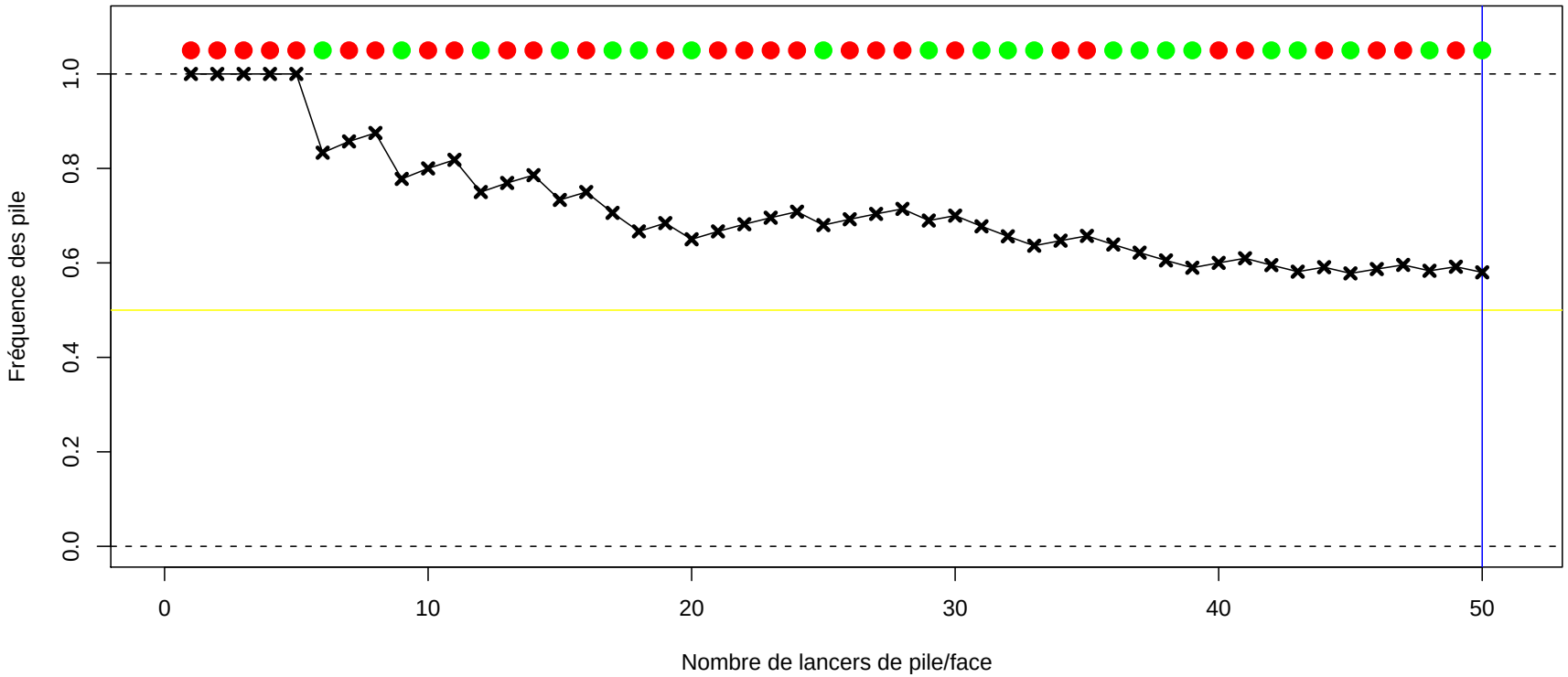



Figure 16: Convergence of the (empirical) mean $\overline{(x)}_n$.

# Visualization of Convergence

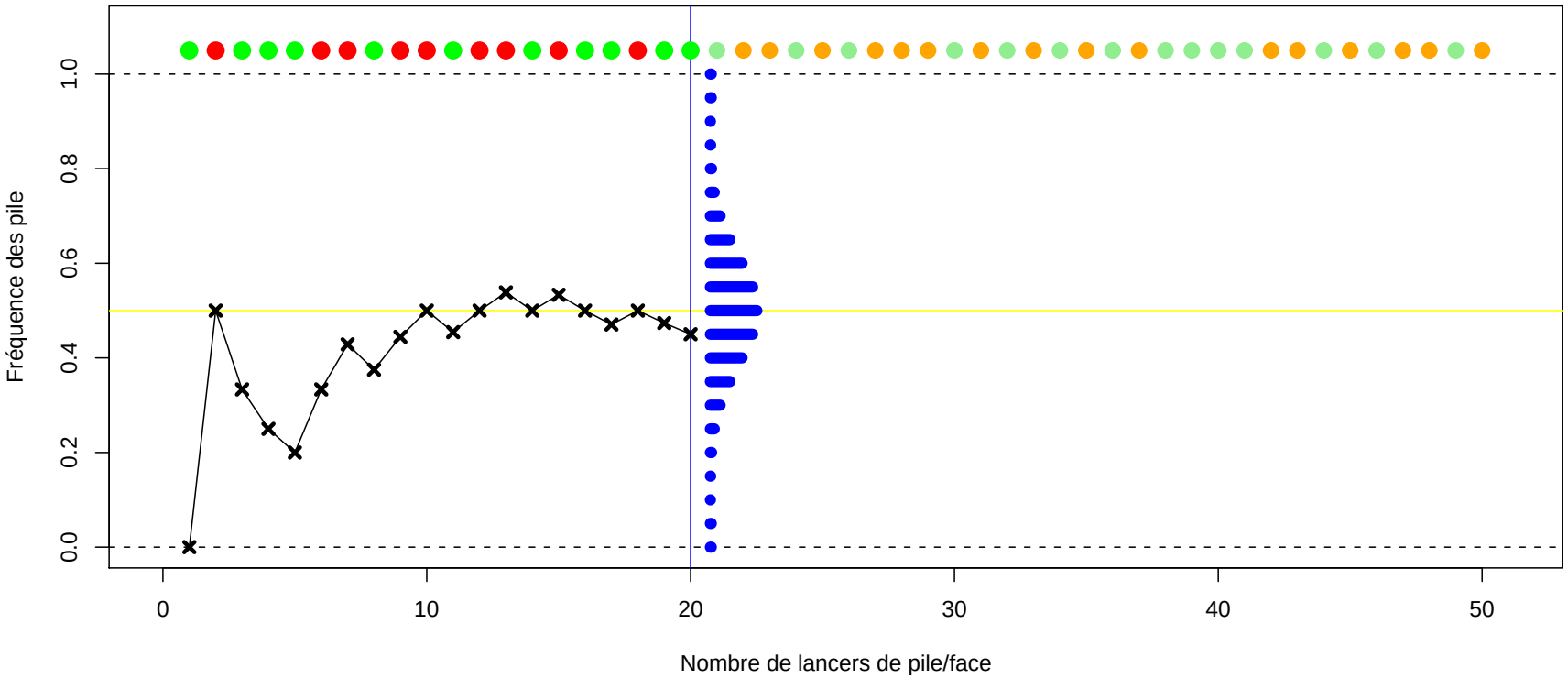



Figure 17: Convergence of the normalized (empirical) mean $\sqrt{n}(\overline{x}_n - \mu)\sigma^{-1}$.

# Visualization of Convergence

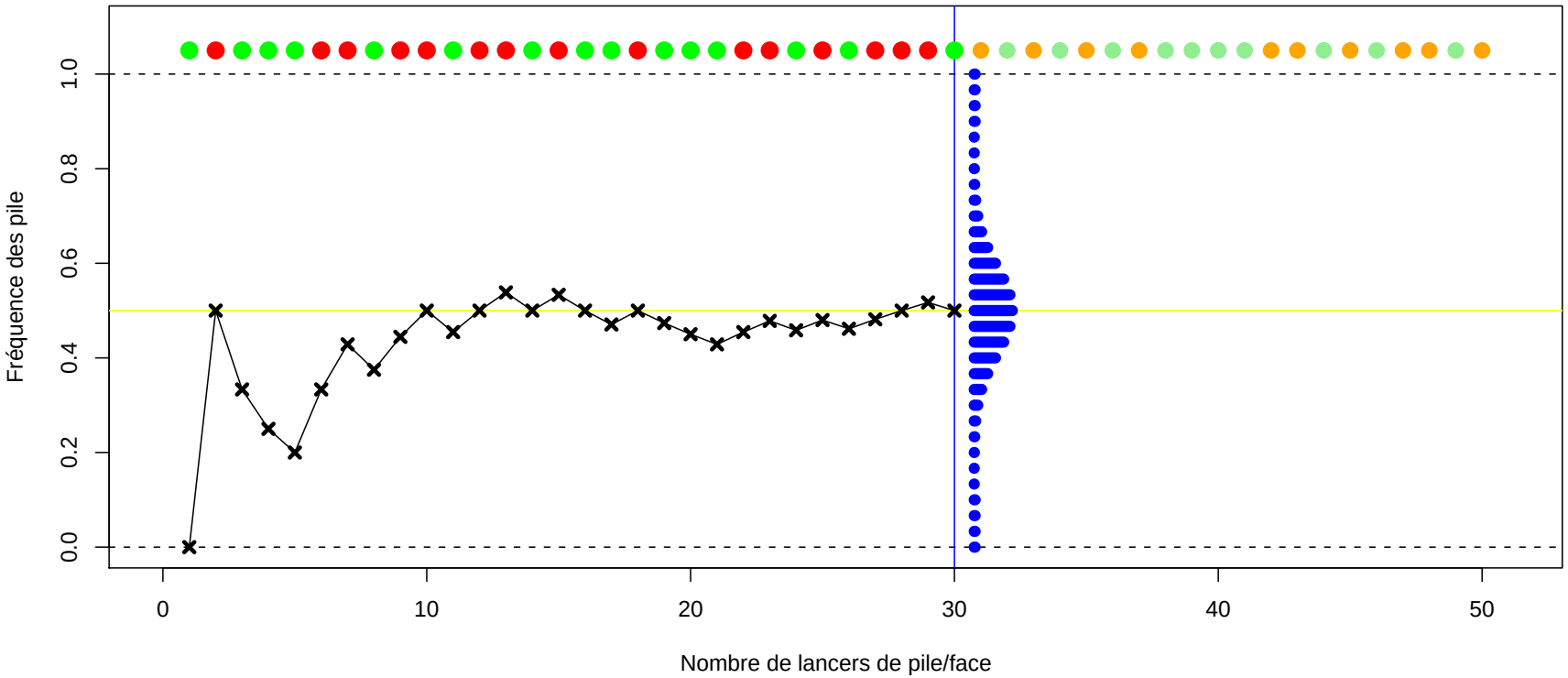



Figure 18: Convergence of the normalized (empirical) mean $\sqrt{n}(\overline{x}_n - \mu)\sigma^{-1}$.

# Visualization of Convergence

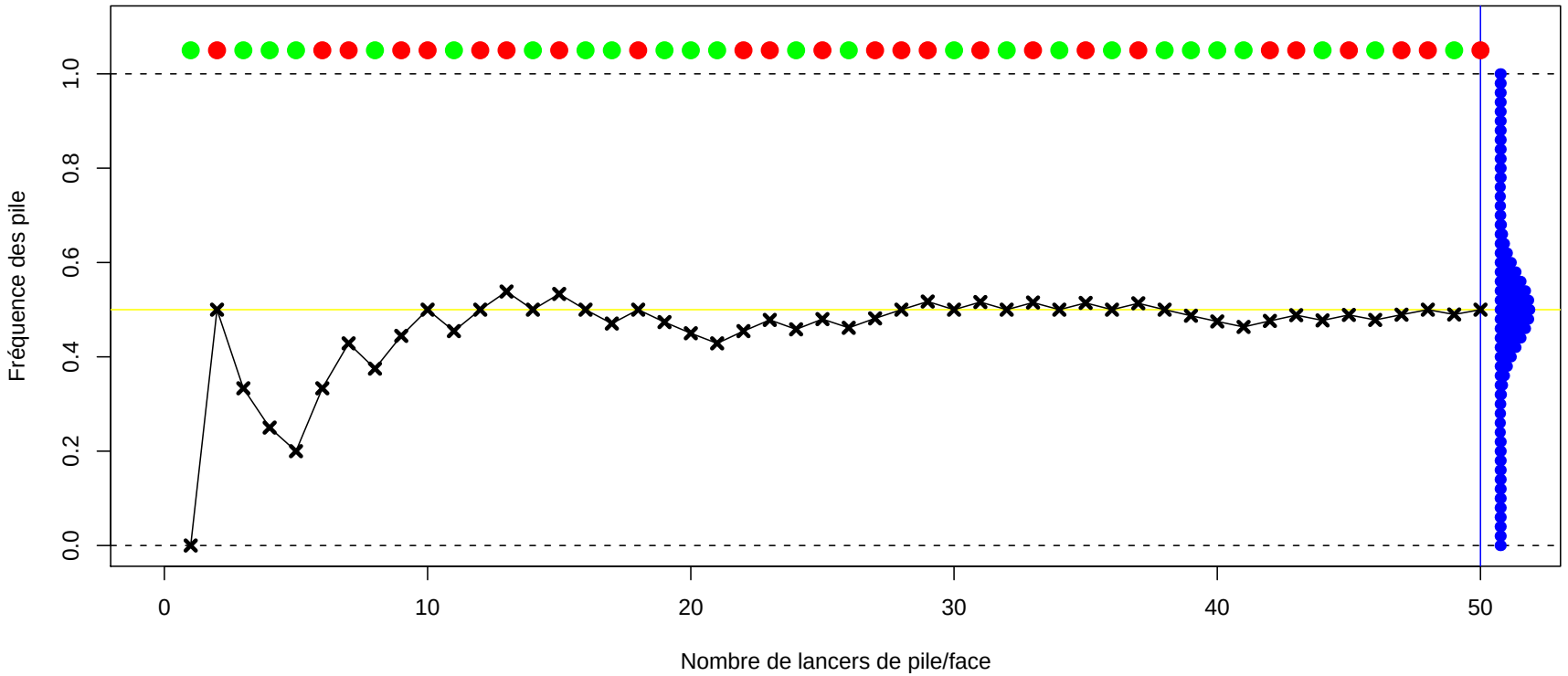



Figure 19: Convergence of the normalized (empirical) mean $\sqrt{n}(\overline{x}_n - \mu)\sigma^{-1}$.

## From Convergence to Approximations

**Proposition.** *Let $(X_n)$ denote a sequence of i.i.d. random variables $\mathcal{B}(n,p)$. If $n \to \infty$ and $p \to 0$ with $p \sim \lambda/n$, $X_n \xrightarrow{\mathcal{L}} X$ where $X \sim \mathcal{P}(\lambda)$.*

*Proof.* Based on

$$\binom{n}{k} p^k [1-p]^{n-k} \approx \exp[-np] \frac{[np]^k}{k!}$$

□

Poisson distribution $\mathcal{P}(np)$ is a good approximation of the Binomial $\mathcal{B}(n,p)$ when $n$ is large, as well as $np \to \infty$ (and thus $p$ small, with respect to $n$).

In practice, it can be used when $n > 30$ and $np < 5$.

## From convergence to approximations

**Proposition.** *Let $(X_n)$ be a sequence of i.i.d. $\mathcal{B}(n,p)$ variables. Then if $np \to \infty$,*

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} X \text{ with } X \sim \mathcal{N}(0,1)$$

In practice, the approximation is valid for $n > 30$ and $np > 5$, and $n(1-p) > 5$.

The Gaussian distribution $\mathcal{N}(np, np(1-p))$ is an approximation of the Binomial distribution $\mathcal{B}(n,p)$ for $n$ large enough, with $np, n(1-p) \to \infty$.

# From convergence to approximations

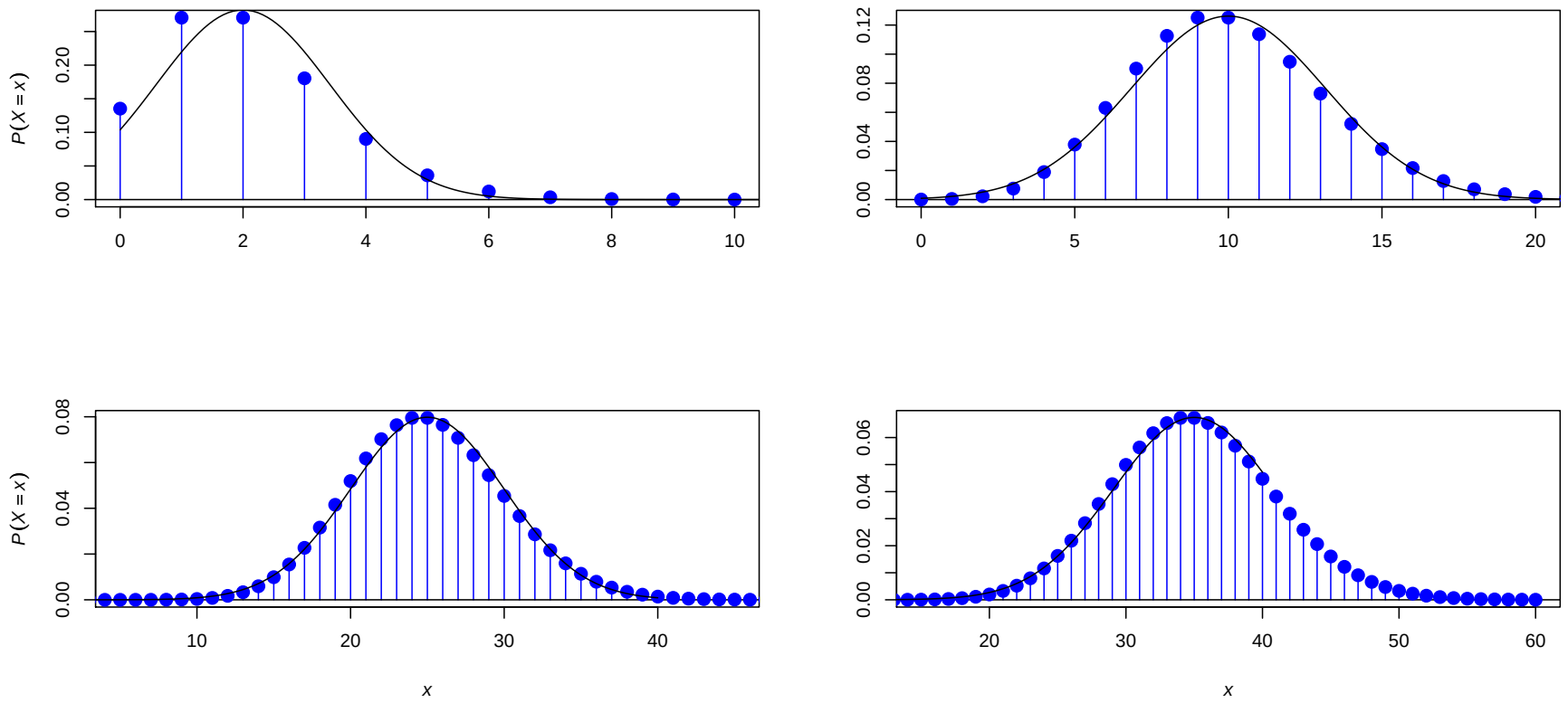



Figure 20: Gaussian Approximation of the Poisson distribution

## Transforming Random Variables

Let $X$ be an absolutely continuous random variable with density $f(x)$. We want to know the distribution of $Y = \phi(X)$.

**Proposition.** *If function $\phi$ is a differentiable one-to-one mapping, then variable $Y$ has a density $g$ satisfying*

$$g(y) = \frac{f(\phi^{-1}(y))}{\phi'(\phi^{-1}(y))}.$$

## Transforming Random Variables

**Proposition.** *Let $X$ be an absolutely continuous random variable with cdf $F$, i.e. $F(x) = \mathbb{P}(X \leq x)$. Then $Y = F(X)$ has a uniform distribution on $[0, 1]$.*

**Proposition.** *Let $Y$ be a uniform distribution on $[0, 1]$ and $F$ denote a cdf. Then $X = F^{-1}(Y)$ is a random variable with cdf $F$.*

This will be the startig point of Monte Carlo simulations.

## Transforming Random Variables

Let $(X, Y)$ be a random vector with absolutely continuous marginals, with joint density $f(x, y)$ . Let $(u, v) = \phi(x, y)$. If $J_\phi$ denotes the Jacobian associated with, i.e.

$$J_\phi = \left| \det \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial v}{\partial x} \\ \dfrac{\partial u}{\partial y} & \dfrac{\partial v}{\partial y} \end{pmatrix} \right|$$

then $(U, V) = \phi(X, Y)$ has the following joint density :

$$g(u, v) = \frac{1}{J_\phi} f\left(\phi^{-1}(u, v)\right)$$

# Transforming Random Variables

We have mentioned already that $\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$ unless $g$ is a linear function.

**Proposition.** *Let $g$ be a convex function, then $\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$.*

For instance, if $X$ takes values $\{1, 4\}$ with probability $1/2$.

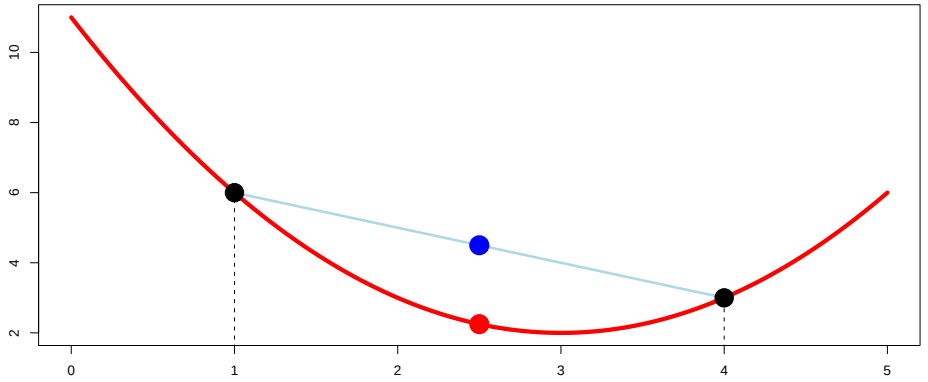



Figure 21: Jensen inequality: $g(\mathbb{E}(X))$ vs. $\mathbb{E}(g(X))$.

# Transforming Random Variables : the Δ-method

In the univariate case, if there is a sequence of random variables $X_n$ satisfying

$$\sqrt{n}[X_n - \theta] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

where $\theta$ and $\sigma^2$ are two constants, then

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 \cdot [g'(\theta)]^2)$$

for any function $g$ satisfying the property that $g'(\theta)$ exists and is non-zero valued.

In the multivariate case, if there is a sequence of random vectors $\boldsymbol{X}_n$ satisfying

$$\sqrt{n}[\boldsymbol{X}_n - \boldsymbol{\theta}] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma),$$

where $\boldsymbol{\theta}$ is a vector in $\mathbb{R}^d$ and $\Sigma$ is a symmetric positive $d \times d$ matrix, then

$$\sqrt{n}[g(\boldsymbol{X}_n) - g(\boldsymbol{\theta})] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \nabla g(\boldsymbol{\theta})^\mathsf{T} \Sigma \nabla g(\boldsymbol{\theta}))$$

for any function g satisfying the property that $\nabla g(\boldsymbol{\theta})$ exists and is non-zero valued.

## Transforming Random Variables : the Δ-method

**Example** Let $\theta \in (0,1)$, $Z_n \sim \mathcal{B}(n,\theta)$ and definie $X_n = Z_n/n$. Then

$$\sqrt{n}[X_n - \theta] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta)).$$

Consider transformation $g(\theta) = \log(\theta)$, then

$$\sqrt{n}\left[\log\left(X_n\right) - \log(\theta)\right] \xrightarrow{\mathcal{L}} N(0, \theta(1-\theta)[1/\theta]^2)$$

# Computer Based Randomness

Calculations of $\mathbb{E}[h(X)]$ can be complicated,

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

Sometimes, we simply want a numerical approximation of that integral. One can use numerical functions to compute those integrals. But one can also use Monte Carlo techniques. Assume that we can generate a sample $\{x_1, \cdots, x_n, \cdots\}$ i.i.d. from distribution $F$. From the law of large numbers we know that

$$\frac{1}{n}\sum_{i=1}^{n} h(x) \to \mathbb{E}[h(X)], \text{ as } n \to \infty.$$

or

$$\frac{1}{n}\sum_{i=1}^{n} h(F_X^{-1}(u_i)) \to \mathbb{E}[h(X)], \text{ as } n \to \infty$$

if $\{x_1, \cdots, x_n, \cdots\}$ i.i.d. from a uniform distribution on $[0, 1]$.

# Computer Based Randomness

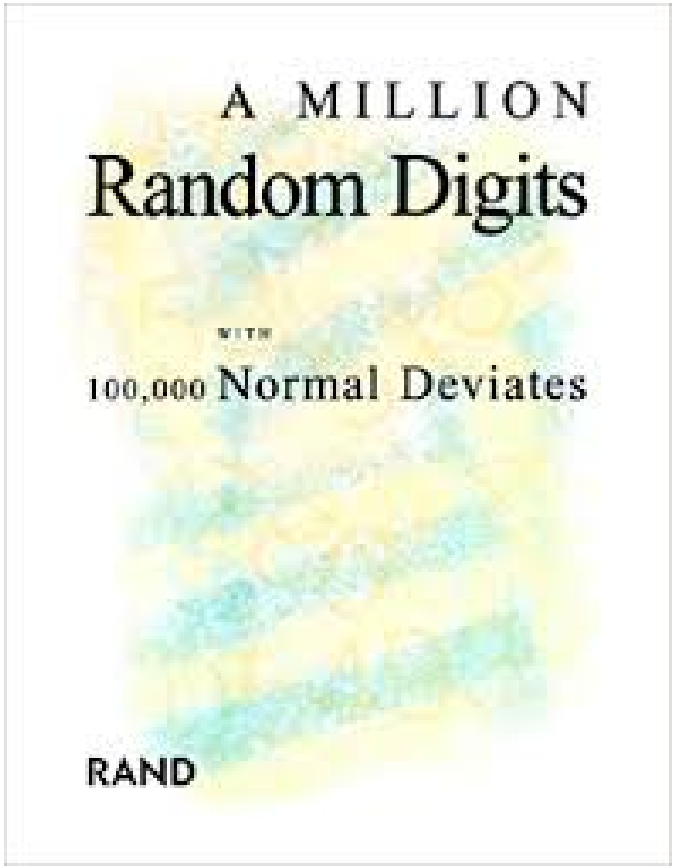



TABLE OF RANDOM DIGITS 1

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 00000 | 10097 | 32533 | 76520 | 13586 | 34673 | 54876 | 80959 | 09117 | 39292 | 74945 |
| 00001 | 37542 | 04805 | 64894 | 74296 | 24805 | 24037 | 20636 | 10402 | 00822 | 91665 |
| 00002 | 08422 | 68953 | 19645 | 09303 | 23209 | 02560 | 15953 | 34764 | 35080 | 33606 |
| 00003 | 99019 | 02529 | 09376 | 70715 | 38311 | 31165 | 88676 | 74397 | 04436 | 27659 |
| 00004 | 12807 | 99970 | 80157 | 36147 | 64032 | 36653 | 98951 | 16877 | 12171 | 76833 |
| 00005 | 66065 | 74717 | 34072 | 76850 | 36697 | 36170 | 65813 | 39885 | 11199 | 29170 |
| 00006 | 31060 | 10805 | 45571 | 82406 | 35303 | 42614 | 86799 | 07439 | 23403 | 09732 |
| 00007 | 85269 | 77602 | 02051 | 65692 | 68665 | 74818 | 73053 | 85247 | 18623 | 88579 |
| 00008 | 63573 | 32135 | 05325 | 47048 | 90553 | 57548 | 28468 | 28709 | 83491 | 25624 |
| 00009 | 73796 | 45753 | 03529 | 64778 | 35808 | 34282 | 60935 | 20344 | 35273 | 88435 |
| 00010 | 98520 | 17767 | 14905 | 68607 | 22109 | 40558 | 60970 | 93433 | 50500 | 73998 |
| 00011 | 11805 | 05431 | 39808 | 27732 | 50725 | 68248 | 29405 | 24201 | 52775 | 67851 |
| 00012 | 83452 | 99634 | 06288 | 98083 | 13746 | 70078 | 18475 | 40610 | 68711 | 77817 |
| 00013 | 88685 | 40200 | 86507 | 58401 | 36766 | 67951 | 90364 | 76493 | 29609 | 11062 |
| 00014 | 99594 | 67348 | 87517 | 64969 | 91826 | 08928 | 93785 | 61368 | 23478 | 34113 |
| 00015 | 65481 | 17674 | 17468 | 50950 | 58047 | 76974 | 73039 | 57186 | 40218 | 16544 |
| 00016 | 80124 | 35635 | 17727 | 08015 | 45318 | 22374 | 21115 | 78253 | 14385 | 53763 |
| 00017 | 74350 | 99817 | 77402 | 77214 | 43236 | 00210 | 45521 | 64237 | 96286 | 02655 |
| 00018 | 69916 | 26803 | 66252 | 29148 | 36936 | 87203 | 76621 | 13990 | 94400 | 56418 |
| 00019 | 09893 | 20505 | 14225 | 68514 | 46427 | 56788 | 96297 | 78822 | 54382 | 14598 |
| 00020 | 91499 | 14523 | 68479 | 27686 | 46162 | 83554 | 94750 | 89923 | 37089 | 20048 |
| 00021 | 80336 | 94598 | 26940 | 36858 | 70297 | 34135 | 53140 | 33340 | 42050 | 82341 |
| 00022 | 44104 | 81949 | 85157 | 47954 | 32979 | 26575 | 57600 | 40881 | 22222 | 06413 |
| 00023 | 12550 | 73742 | 11100 | 02040 | 12860 | 74697 | 96644 | 89439 | 28707 | 25815 |

# Monte Carlo Simulations

Let $X \sim \mathcal{C}auchy$ what is $\mathbb{P}[X > 2]$? Let

$$p = \mathbb{P}[X > 2] = \int_2^\infty \frac{dx}{\pi(1+x^2)} \qquad (\sim 0.15)$$

since $f(x) = \dfrac{1}{\pi(1+x^2)}$ and $Q(u) = F^{-1}(u) = \tan\left(\pi\left[u - \frac{1}{2}\right]\right)$.

Crude Monte Carlo: use the law of large numbers

$$\widehat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Q(u_i) > 2)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_1] \sim \frac{0.127}{n}$.

Crude Monte Carlo (with symmetry): $\mathbb{P}[X > 2] = \mathbb{P}[|X| > 2]/2$ and use the law

of large numbers

$$\widehat{p}_2 = \frac{1}{2n}\sum_{i=1}^{n}\mathbf{1}(|Q(u_i)|>2)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_2] \sim \frac{0.052}{n}$.

Using integral symmetries :

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{2} - \int_0^2 \frac{dx}{\pi(1+x^2)}$$

where the later integral is $\mathbb{E}[h(2U)]$ where $h(x) = \dfrac{2}{\pi(1+x^2)}$.

From the law of large numbers

$$\widehat{p}_3 = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n} h(2u_i)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_3] \sim \frac{0.0285}{n}$.

Using integral transformations :

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \int_0^{1/2} \frac{y^{-2}dy}{\pi(1-y^{-2})}$$

which is $\mathbb{E}[h(U/2)]$ where $h(x) = \dfrac{1}{2\pi(1+x^2)}$.

From the law of large numbers

$$\widehat{p}_4 = \frac{1}{4n}\sum_{i=1}^{n} h(u_i/2)$$

where $u_i$ are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_4] \sim \frac{0.0009}{n}$.

## The Estimator as a Random Variable

In descriptive statistics, estimators are functions of the observed sample, $\{x_1, \cdots, x_n\}$, e.g.

$$\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$$

In mathematical statistics, assume that $x_i = X_i(\omega)$, i.e. realizations of random variables,

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

$X_1$,..., $X_n$ being random variables, so that $\overline{X}_n$ is also a random variable.

For example, assume that we have a sample of size $n = 20$ from a uniform distribution on $[0, 1]$.

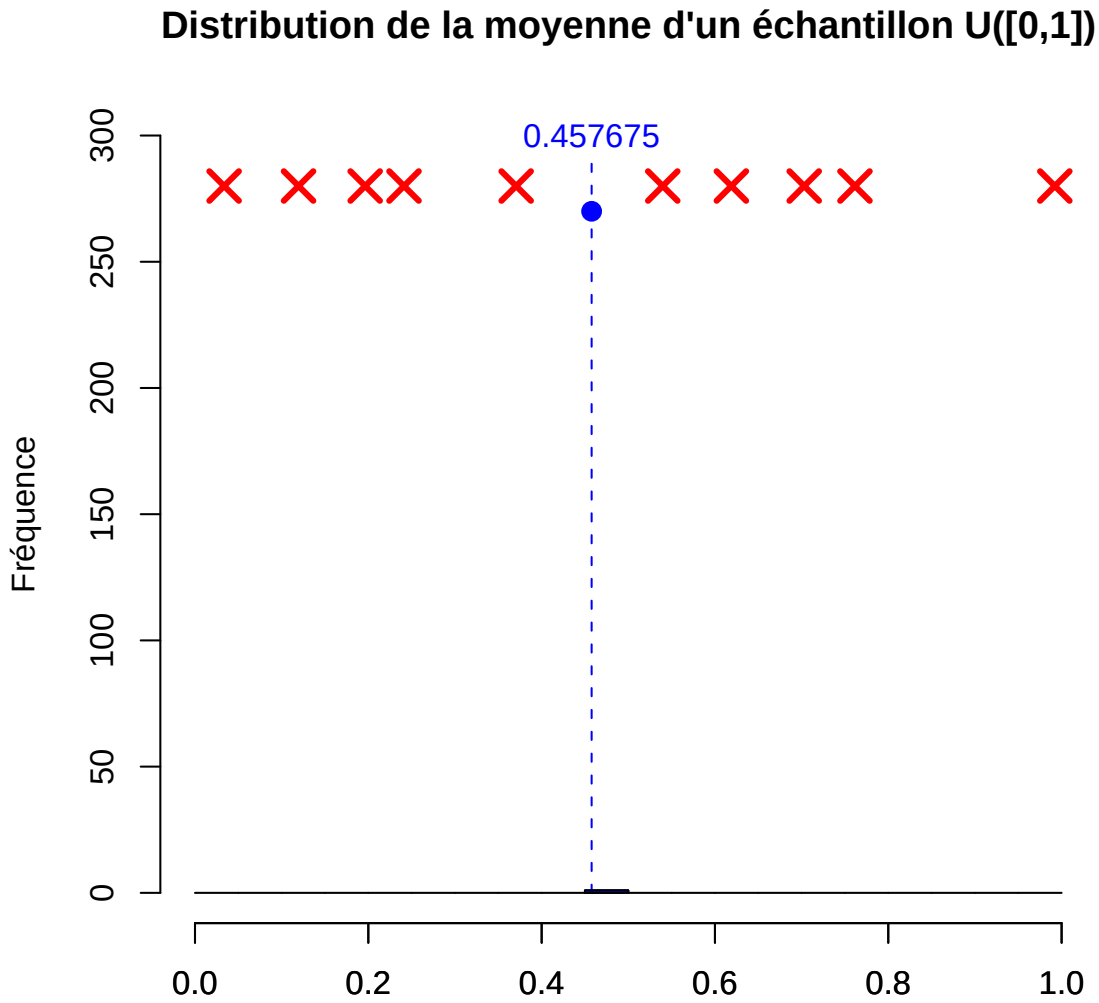



Figure 22: Distribution of the mean of $\{X_1, \cdots, X_{10}\}$, $X_i \sim \mathcal{U}([0,1])$.

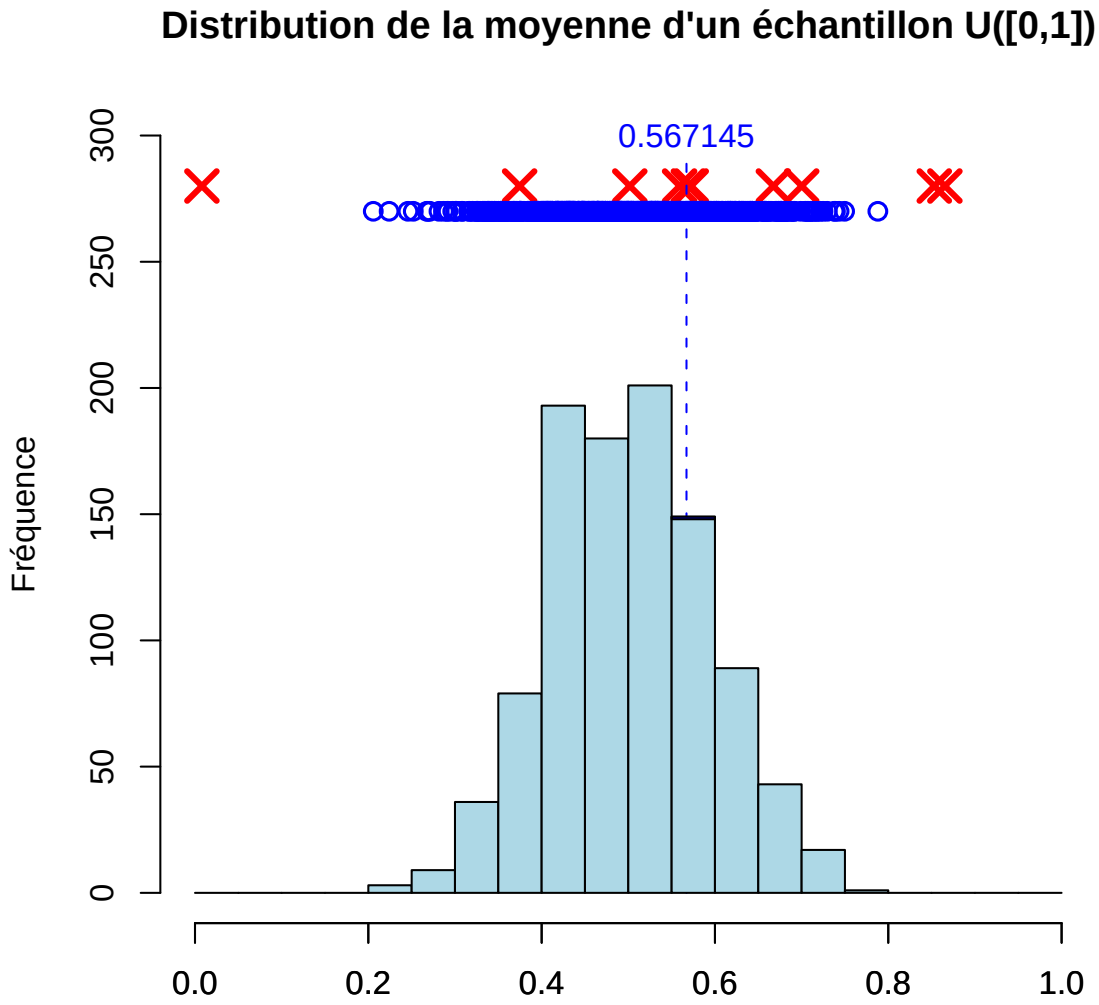



Figure 23: Distribution of the mean of $\{X_1, \cdots, X_{10}\}$, $X_i \sim \mathcal{U}([0,1])$.

## Some technical properties

Let $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ and set $\overline{x} = \dfrac{x_1 + \cdots + x_n}{n}$. then,

$$\min_{m \in \mathbb{R}} \left\{ \sum_{i=1}^{n} [x_i - m]^2 \right\} = \sum_{i=1}^{n} [x_i - \overline{x}]^2$$

while

$$\sum_{i=1}^{n} [x_i - \overline{x}]^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

# (Empirical) Mean

**Definition** Let $\{X_1, \cdots, X_n\}$ be i.i.d. random variables with cdf $F$. The (empirical) mean is

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Assume $X_i$'s i.i.d. with finite expected value (denoted $\mu$), then

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{*}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}n\mu = \mu$$

$*$ since the expected value is linear

**Proposition.** *Assume $X_i$'s i.i.d. with finite expected value (denoted $\mu$), then*

$$\mathbb{E}(\overline{X}_n) = \mu.$$

The average (or the mean) is an unbiased estimator of the expected value.

## (Empirical) Variance

Assume $X_i$'s i.i.d. with finite variance (denoted $\sigma^2$), then

$$\text{Var}(\overline{X}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{*}{=} \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$*$ because variables are independent, and variance is a quadratic function.

**Proposition.** *Assume $X_i$'s i.i.d. with finite variance (denoted $\sigma^2$),*

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

# (Empirical) Variance

**Definition** Let $\{X_1, \cdots, X_n\}$ be $n$ i.i.d. random variables with distribution $F$. The empirical variance is

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}[X_i - \overline{X}_n]^2.$$

Assume $X_i$'s i.i.d. with finite variance (denoted $\sigma^2$),

$$\mathbb{E}(S_n^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}[X_i - \overline{X}_n]^2\right) \overset{*}{=} \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}_n^2\right]\right)$$

$*$ from the same property as before

$$\mathbb{E}(S_n^2) = \frac{1}{n-1}[n\mathbb{E}(X_i^2) - n\mathbb{E}(\overline{X}^2)] \overset{*}{=} \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] = \sigma^2$$

$*$ since $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

## (Empirical) Variance

**Proposition.** *Asusme that $X_i$ independent, with finite variance (denoted $\sigma^2$),*

$$\mathbb{E}(S_n^2) = \sigma^2.$$

Empirical variance is an unbiased estimator of the variance.

Note that

$$\widetilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}[X_i - \overline{X}_n]^2$$

is also a popular estimator (but biased).

# Gaussian Sampling

**Proposition.** *Suppose $X_i$'s i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then*

- $\overline{X}_n$ *and* $S_n^2$ *are independent random variables*
- $\overline{X}_n$ *has distribution* $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$
- $(n-1)S_n^2/\sigma^2$ *has distribution* $\chi^2(n-1)$. *Assume that $X_i$'s are i.i.d. random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, then*
- $\sqrt{n}\dfrac{\overline{X}_n - \mu}{\sigma}$ *has a* $\mathcal{N}(0,1)$ *distribution*
- $\sqrt{n}\dfrac{\overline{X}_n - \mu}{S_n}$ *has a Student-t distribution with* $n-1$ *degrees of freedom*

# Gaussian Sampling

Indeed

$$\sqrt{n}\frac{\overline{X}_n-\mu}{S}=\underbrace{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}_{\mathcal{N}(0,1)}\Big/\underbrace{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}}}_{\chi^2(n-1)}\times\sqrt{n-1}$$

To get a better understanding of the $n-1$ degrees of freedom for a sum of $n$ terms,observe that

$$S_n^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right]=\frac{1}{n-1}\left[(X_1-\overline{X}_n)^2+\sum_{i=2}^{n}(X_i-\overline{X}_n)^2\right]$$

i.e. $S_n^2=\dfrac{1}{n-1}\left[\left(\sum_{i=2}^{n}(X_i-\overline{X}_n)\right)^2+\sum_{i=2}^{n}(X_i-\overline{X}_n)^2\right]$ because

$\sum_{i=1}^{n}(X_i-\overline{X}_n)=0$. Hence $S_n^2$ is a function of $n-1$ (centered) variables

$X_2-\overline{X}_n,\cdots,X_n-\overline{X}_n$

# Asymptotic Properties

**Proposition.** *Assume that $X_i$'s are i.i.d. random variables with cdf $F$, mean $\mu$ and variance $\sigma^2$ (finite). Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) = 0$$

*i.e. $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$ (convergence in probability).*

**Proposition.** *Assume that $X_i$'s are i.i.d. random variables with cdf $F$, mean $\mu$ and variance $\sigma^2$ (finite). Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|S_n^2 - \sigma^2| > \varepsilon) \leq \frac{\text{Var}(S_n^2)}{\varepsilon^2}$$

*i.e. a sufficient condition to get $S_n^2 \xrightarrow{\mathbb{P}} \sigma^2$ (convergence in probability) is that $\text{Var}(S_n^2) \to 0$ as $n \to \infty$.*

# Asymptotic Properties

**Proposition.** *Assume that $X_i$'s are i.i.d. random variables with cdf $F$, mean $\mu$ and variance $\sigma^2$ (finite). Then for any $z \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

*i.e.*

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

**Remark** If $X_i$'s have a $\mathcal{N}(\mu, \sigma^2)$ distribution, then

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

## Variance Estimation

Consider a Gaussian sample, then

$$\text{Var}\left(\frac{(n-1)S_n^2}{\sigma^2}\right) = \text{Var}(Z) \text{ with } Z \sim \chi^2_{n-1}$$

so that this quantity can be written

$$\frac{(n-1)^2}{\sigma^4}\text{Var}(S_n^2) = 2(n-1)$$

i.e.

$$\text{Var}(S_n^2) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}.$$

# Variance and Standard-Deviation Estimation

Assume that $X_i \sim \mathcal{N}(\mu, \sigma^2)$. A *natural* estimator of $\sigma$ is

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

One can prove that

$$\mathbb{E}(S_n) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma([n-1]/2)}\sigma \sim \left(1 - \frac{1}{4n} - \frac{7}{32n^2}\right)\sigma \neq \sigma$$

but

$$S_n \xrightarrow{\mathbb{P}} \sigma \text{ and } \sqrt{n}(S_n - \sigma) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma}{2}\right)$$

# Variance and Standard-Deviation Estimation

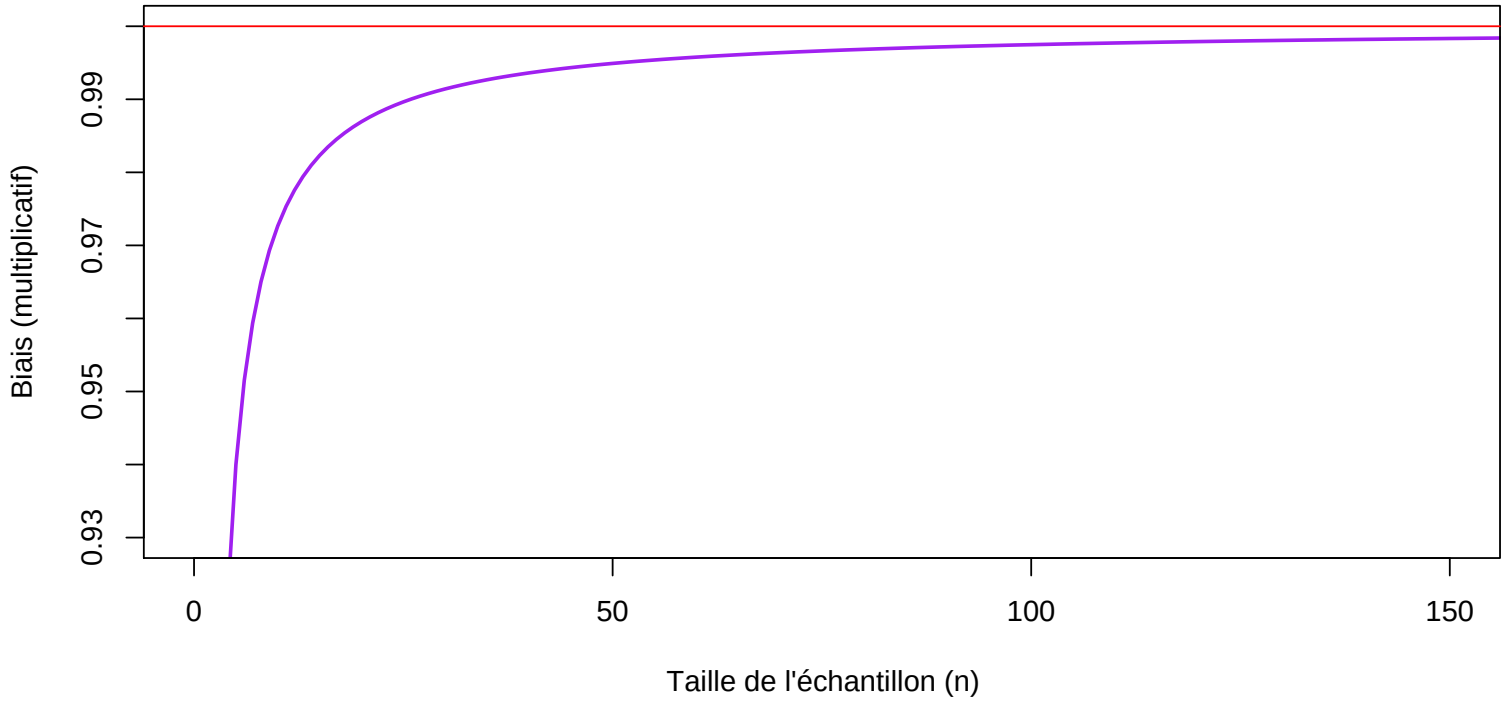



Figure 24: Bias when estimating Standard Deviation.

## Transformed Sample

Let $g : \mathbb{R} \to \mathbb{R}$ be sufficiently regular to write Taylor expansion

$$g(x) = g(x_0) + g'(x_0) \cdot [x - x_0] \quad + \text{some (small) additional term}$$

Let $Y_i = g(X_i)$. The, if $\mathbb{E}(X_i) = \mu$ with $g'(\mu) \neq 0$

$$Y_i = g(X_i) \approx g(\mu) + g'(\mu) \cdot [X_i - \mu]$$

so that

$$\mathbb{E}(Y_i) = \mathbb{E}(g(X_i)) \approx g(\mu)$$

and

$$\text{Var}(Y_i) = \text{Var}(g(X_i)) \approx [g'(\mu)]^2 \text{Var}(X_i)$$

Keep in mind that those are just approximations.

# Transformed Sample : the Δ-Method

The Delta-Method can be adapted in some cases...

**Proposition.** *Suppose $X_i$'s i.i.d. with distribution $F$, expected value $\mu$ and variance $\sigma^2$ (finite), then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

*And if $g'(\mu) \neq 0$, then*

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, [g'(\mu)]^2\sigma^2)$$

**Proposition.** *Suppose $X_i$'s i.i.d. with distribution $F$, expected value $\mu$ and variance $\sigma^2$ (finite), then if $g'(\mu) = 0$ but $g''(\mu) \neq 0$, we have*

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{\mathcal{L}} \frac{g''(\mu)}{2}\sigma^2\chi^2(1)$$

## Confidence Interval for $\mu$

The confidence interval for $\mu$ of order $1-\alpha$ (e.g. 95%) is the smallest interval $I$ such that

$$\mathbb{P}(\mu \in I) = 1-\alpha.$$

Let $u_\alpha$ denote the quantile of the $\mathcal{N}(0,1)$ of order $\alpha$, i.e.

$$u_{\alpha/2} = -u_{1-\alpha/2} = \Phi^{-1}(\alpha/2).$$

since $Z = \sqrt{n}\dfrac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$, we get $\mathbb{P}(Z \in [u_{\alpha/2}, u_{1-\alpha/2}]) = 1-\alpha$, and

$$\mathbb{P}\left(\mu \in \left[\overline{X} + \frac{u_{\alpha/2}}{\sqrt{n}}\sigma, \overline{X} + \frac{u_{1-\alpha/2}}{\sqrt{n}}\sigma\right]\right) = 1-\alpha.$$

## Confidence Interval, mean of a Gaussian Sample

- if $\alpha = 10\%$, $u_{1-\alpha/2} = 1.64$ and therefore, with probability 90%,

$$\overline{X} - \frac{1.64}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{1.64}{\sqrt{n}}\sigma,$$

- if $\alpha = 5\%$, $u_{1-\alpha/2} = 1.96$ and therefore, with probability 95%,

$$\overline{X} - \frac{1.96}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{1.96}{\sqrt{n}}\sigma,$$

# Confidence Interval, mean of a Gaussian Sample

If variance is unknown, plug-in $S_n^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^n X_i^2\right) - \overline{X}_n^2$.

We've seen that

$$\frac{(n-1)S_n^2}{\sigma^2} = \underbrace{\sum_{i=1}^n \left(\underbrace{\frac{X_i - \mathbb{E}(X)}{\sigma}}_{\mathcal{N}(0,1)}\right)^2}_{\chi^2(n) \text{ distribution}} - \underbrace{\left(\underbrace{\frac{\overline{X}_n - \mathbb{E}(X)}{\sigma/\sqrt{n}}}_{\mathcal{N}(0,1)}\right)^2}_{\chi^2(1) \text{ distribution}}$$

From Cochrane theorem $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$.

# Confidence Interval, mean of a Gaussian Sample

Since $\overline{X}_n$ and $S_n^2$ are independent,

$$T = \sqrt{n-1}\frac{\overline{X}_n - \mu}{S_n} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n-1}}}{\sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}}} \sim \mathcal{S}t(n-1).$$

If $t_{\alpha/2}^{(n-1)}$ denote the quantile of the $\mathcal{S}t(n-1)$ distribution with level $\alpha/2$, i.e.

$$t_{\alpha/2}^{(n)} = -t_{1-\alpha/2}^{(n-1)} \text{ satisfies } \mathbb{P}(T \leq t_{\alpha/2}^{(n-1)}) = \alpha/2$$

thus $\mathbb{P}(T \in [t_{\alpha/2}^{(n-1)}, t_{1-\alpha/2}^{(n-1)}]) = 1-\alpha$, and therefore

$$\mathbb{P}\left(\mu \in \left[\overline{X} + \frac{t_{\alpha/2}^{(n-1)}}{\sqrt{n-1}}\sigma, \overline{X} + \frac{t_{1-\alpha/2}^{(n-1)}}{\sqrt{n-1}}\sigma\right]\right) = 1-\alpha.$$

## Confidence Interval, mean of a Gaussian Sample

- if $n = 10$ and $\alpha = 10\%$, $u_{1-\alpha/2} = 1.833$ and with 90% chance,

$$\overline{X} - \frac{1.833}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{1.833}{\sqrt{n}}\sigma,$$

- if $n = 10$ and $\alpha = 5\%$, $u_{1-\alpha/2} = 2.262$ and with 95% chance,

$$\overline{X} - \frac{2.262}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{2.262}{\sqrt{n}}\sigma,$$

# Confidence Interval, mean of a Gaussian Sample

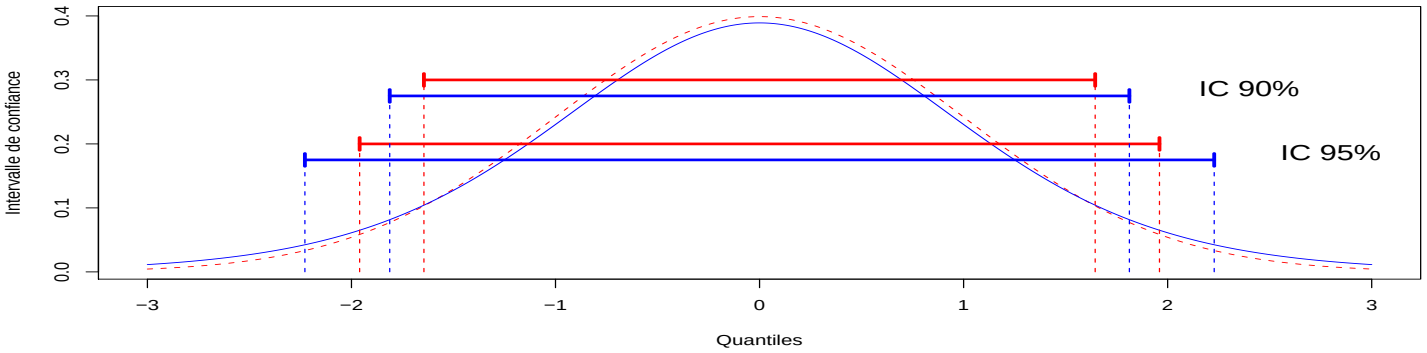



Figure 25: Quantiles for $n = 10$, $\sigma$ known or unknown.

## Confidence Interval, mean of a Gaussian Sample

- if $n = 20$ and $\alpha = 10\%$, $u_{1-\alpha/2} = 1.729$ and thus, with 90% chance

$$\overline{X} - \frac{1.729}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{1.729}{\sqrt{n}}\sigma,$$

- if $n = 20$ and $\alpha = 10\%$, $u_{1-\alpha/2} = 1.729$ and thus, with 95% chance

$$\overline{X} - \frac{2.093}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{2.093}{\sqrt{n}}\sigma,$$

# Confidence Interval, mean of a Gaussian Sample

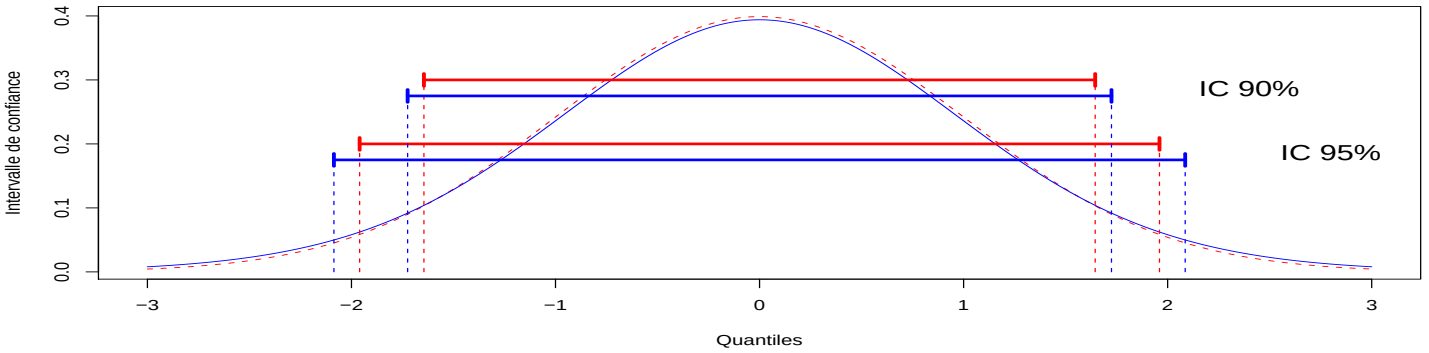



Figure 26: Quantiles for $n = 20$, $\sigma$ known or unknown.

## Confidence Interval, mean of a Gaussian Sample

- if $n = 100$ and $\alpha = 10\%$, $u_{1-\alpha/2} = 1.660$ and therefore, with 90% chance,

$$\overline{X} - \frac{1.660}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{1.660}{\sqrt{n}}\sigma,$$

- if $n = 100$ and $\alpha = 5\%$, $u_{1-\alpha/2} = 1.984$ and therefore, with 95% chance,

$$\overline{X} - \frac{1.984}{\sqrt{n}}\sigma \leq \mu \leq \overline{X} + \frac{1.984}{\sqrt{n}}\sigma,$$

# Confidence Interval, mean of a Gaussian Sample

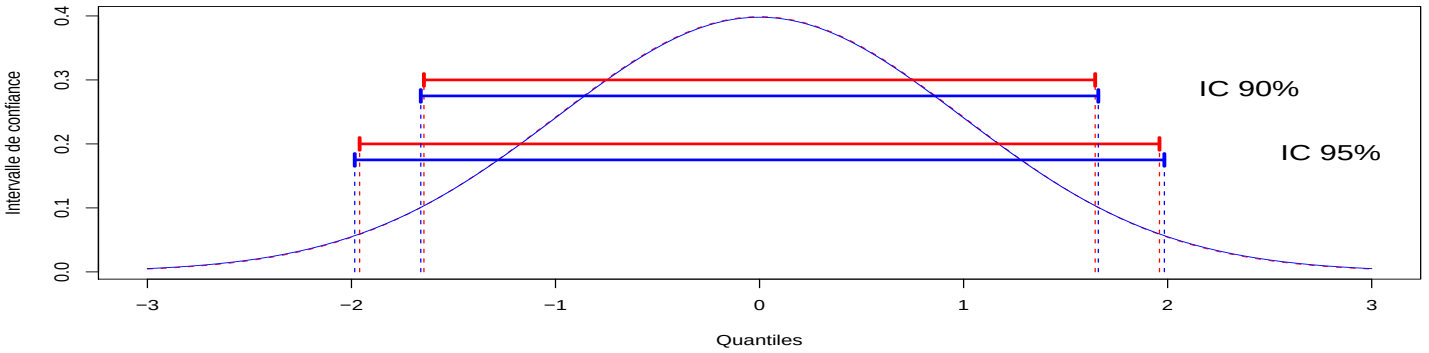



Figure 27: Quantiles for $n = 100$, $\sigma$ known or unknown.

# Using Statistical Tables

Cdf of $X \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(X \leq u) = \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

For example $\mathbb{P}(X \leq 1,96) = 0,975$.

| u | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | 0,5000 | 0,5040 | 0,5080 | 0,5120 | 0,5160 | 0,5199 | 0,5239 | 0,5279 | 0,5319 | 0,5359 |
| 0,1 | 0,5398 | 0,5438 | 0,5478 | 0,5517 | 0,5557 | 0,5596 | 0,5636 | 0,5675 | 0,5714 | 0,5753 |
| 0,2 | 0,5793 | 0,5832 | 0,5871 | 0,5910 | 0,5948 | 0,5987 | 0,6026 | 0,6064 | 0,6103 | 0,6141 |
| 0,3 | 0,6179 | 0,6217 | 0,6255 | 0,6293 | 0,6331 | 0,6368 | 0,6406 | 0,6443 | 0,6480 | 0,6517 |
| 0,4 | 0,6554 | 0,6591 | 0,6628 | 0,6664 | 0,6700 | 0,6736 | 0,6772 | 0,6808 | 0,6844 | 0,6879 |
| 0,5 | 0,6915 | 0,6950 | 0,6985 | 0,7019 | 0,7054 | 0,7088 | 0,7123 | 0,7157 | 0,7190 | 0,7224 |
| 0,6 | 0,7257 | 0,7290 | 0,7324 | 0,7357 | 0,7389 | 0,7422 | 0,7454 | 0,7486 | 0,7517 | 0,7549 |
| 0,7 | 0,7580 | 0,7611 | 0,7642 | 0,7673 | 0,7704 | 0,7734 | 0,7764 | 0,7794 | 0,7823 | 0,7852 |
| 0,8 | 0,7881 | 0,7910 | 0,7939 | 0,7967 | 0,7995 | 0,8023 | 0,8051 | 0,8078 | 0,8106 | 0,8133 |
| 0,9 | 0,8159 | 0,8186 | 0,8212 | 0,8238 | 0,8264 | 0,8289 | 0,8315 | 0,8340 | 0,8365 | 0,8389 |
| 1,0 | 0,8413 | 0,8438 | 0,8461 | 0,8485 | 0,8508 | 0,8531 | 0,8554 | 0,8577 | 0,8599 | 0,8621 |
| 1,1 | 0,8643 | 0,8665 | 0,8686 | 0,8708 | 0,8729 | 0,8749 | 0,8770 | 0,8790 | 0,8810 | 0,8830 |
| 1,2 | 0,8849 | 0,8869 | 0,8888 | 0,8907 | 0,8925 | 0,8944 | 0,8962 | 0,8980 | 0,8997 | 0,9015 |
| 1,3 | 0,9032 | 0,9049 | 0,9066 | 0,9082 | 0,9099 | 0,9115 | 0,9131 | 0,9147 | 0,9162 | 0,9177 |
| 1,4 | 0,9192 | 0,9207 | 0,9222 | 0,9236 | 0,9251 | 0,9265 | 0,9279 | 0,9292 | 0,9306 | 0,9319 |
| 1,5 | 0,9332 | 0,9345 | 0,9357 | 0,9370 | 0,9382 | 0,9394 | 0,9406 | 0,9418 | 0,9429 | 0,9441 |
| 1,6 | 0,9452 | 0,9463 | 0,9474 | 0,9484 | 0,9495 | 0,9505 | 0,9515 | 0,9525 | 0,9535 | 0,9545 |
| 1,7 | 0,9554 | 0,9564 | 0,9573 | 0,9582 | 0,9591 | 0,9599 | 0,9608 | 0,9616 | 0,9625 | 0,9633 |
| 1,8 | 0,9641 | 0,9649 | 0,9656 | 0,9664 | 0,9671 | 0,9678 | 0,9686 | 0,9693 | 0,9699 | 0,9706 |
| 1,9 | 0,9713 | 0,9719 | 0,9726 | 0,9732 | 0,9738 | 0,9744 | 0,9750 | 0,9756 | 0,9761 | 0,9767 |
| 2,0 | 0,9772 | 0,9779 | 0,9783 | 0,9788 | 0,9793 | 0,9798 | 0,9803 | 0,9808 | 0,9812 | 0,9817 |
| 2,1 | 0,9821 | 0,9826 | 0,9830 | 0,9834 | 0,9838 | 0,9842 | 0,9846 | 0,9850 | 0,9854 | 0,9857 |
| 2,2 | 0,9861 | 0,9864 | 0,9868 | 0,9871 | 0,9875 | 0,9878 | 0,9881 | 0,9884 | 0,9887 | 0,9890 |
| 2,3 | 0,9893 | 0,9896 | 0,9898 | 0,9901 | 0,9904 | 0,9906 | 0,9909 | 0,9911 | 0,9913 | 0,9916 |
| 2,4 | 0,9918 | 0,9920 | 0,9922 | 0,9925 | 0,9927 | 0,9929 | 0,9931 | 0,9932 | 0,9934 | 0,9936 |
| 2,5 | 0,9938 | 0,9940 | 0,9941 | 0,9943 | 0,9945 | 0,9946 | 0,9948 | 0,9949 | 0,9951 | 0,9952 |
| 2,6 | 0,9953 | 0,9955 | 0,9956 | 0,9957 | 0,9959 | 0,9960 | 0,9961 | 0,9962 | 0,9963 | 0,9964 |
| 2,7 | 0,9965 | 0,9966 | 0,9967 | 0,9968 | 0,9969 | 0,9970 | 0,9971 | 0,9972 | 0,9973 | 0,9974 |
| 2,8 | 0,9974 | 0,9975 | 0,9976 | 0,9977 | 0,9977 | 0,9978 | 0,9979 | 0,9979 | 0,9980 | 0,9981 |
| 2,9 | 0,9981 | 0,9982 | 0,9982 | 0,9983 | 0,9984 | 0,9984 | 0,9985 | 0,9985 | 0,9986 | 0,9986 |

# Interpretation of a confiance interval

Let us generate i.i.d. samples from a $\mathcal{N}(\mu, \sigma^2)$ distribution, with $\mu$ and $\sigma^2$ fixed, then there are 90% chances that $\mu$ belongs to

$$\left[\overline{X} + \frac{u_{\alpha/2}}{\sqrt{n}}\sigma, \overline{X} + \frac{u_{1-\alpha/2}}{\sqrt{n}}\sigma\right]$$

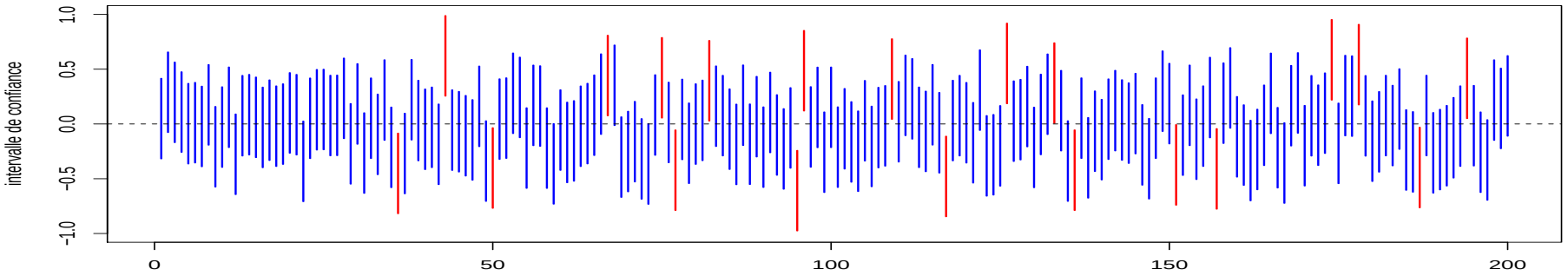



Figure 28: Confidence intervals for $\mu$ on 200 samples, with $\sigma^2$ known.

# Interpretation of a confiance interval

or, if $\sigma$ is unknown

$$\left[\overline{X} + \frac{t_{\alpha/2}^{(n-1)}}{\sqrt{n-1}}\sigma, \overline{X} + \frac{t_{1-\alpha/2}^{(n-1)}}{\sqrt{n-1}}\sigma\right]$$

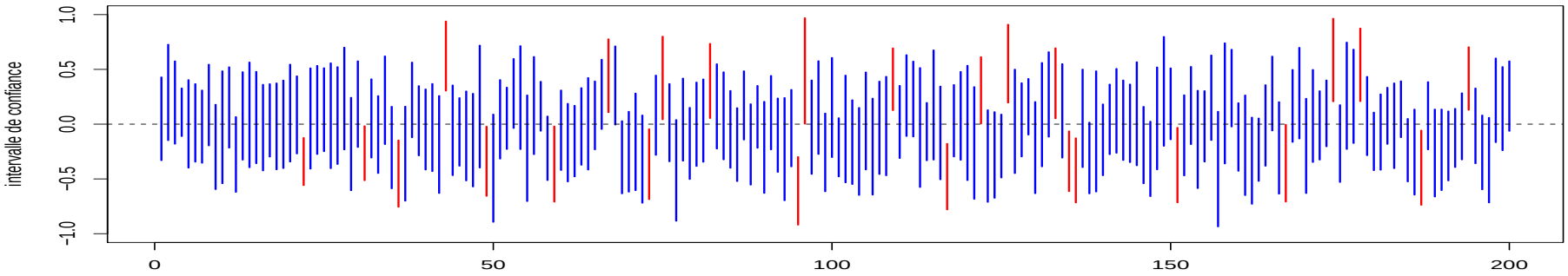



Figure 29: Confidence interval for $\mu$, with $\sigma^2$ unkown (estimated).

# Tests and Decision

A testing procedure yields a decision: either to reject or to accept $H_0$.

Decision $D_0$ is to accept $H_0$, decision $D_1$ is to reject $H_0$

| | $H_0$ true | $H_1$ true |
|---|---|---|
| Decision $d_0$ | Good decision | *error (type 2)* |
| Decision $d_1$ | *error (type 1)* | Good decision |

Type 1 error is the incorrect rejection of a true null hypothesis (a false positive)

Type 2 error is incorrectly retaining a false null hypothesis (a false negative)

The significance is

$$\alpha = \Pr\left(\text{reject } H_0 \mid H_0 \text{ is true}\right)$$

The power is

$$\text{power} = \Pr\left(\text{reject } H_0 \mid H_1 \text{ is true}\right) = 1 - \beta$$

# Usual Testing Procedures

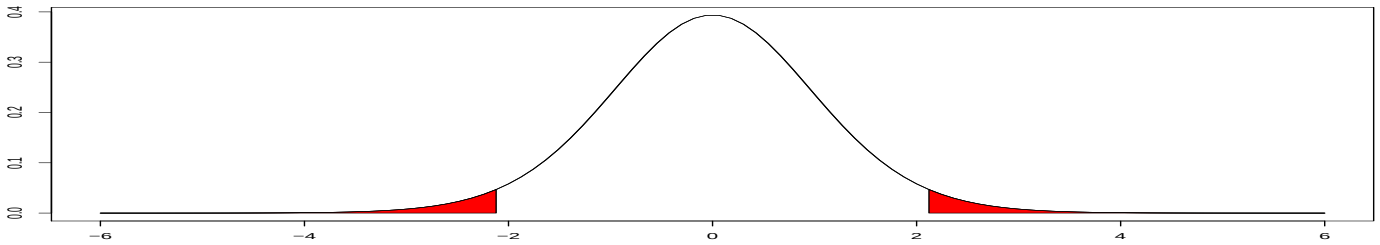
Consider the test on mean (equality) on a Gaussian sample

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_0 : \mu \neq \mu_0 \end{cases}$$

Test statistics is here

$$T = \sqrt{n}\frac{\overline{x} - \mu_0}{s} \text{ où } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which satisfies (under $H_0$) $T \sim \mathcal{S}t(n-1)$.



# Equal Means of Two (Independent) Samples

Consider a test of egality of means on two samples.

Consider two samples $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$. We wish to test

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_0 : \mu_X \neq \mu_Y \end{cases}$$

Assume furthermore that $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_j \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, i.e.

$$\overline{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right) \text{ and } \overline{Y} \sim \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$$

# Equal Means of Two (Independent) Samples

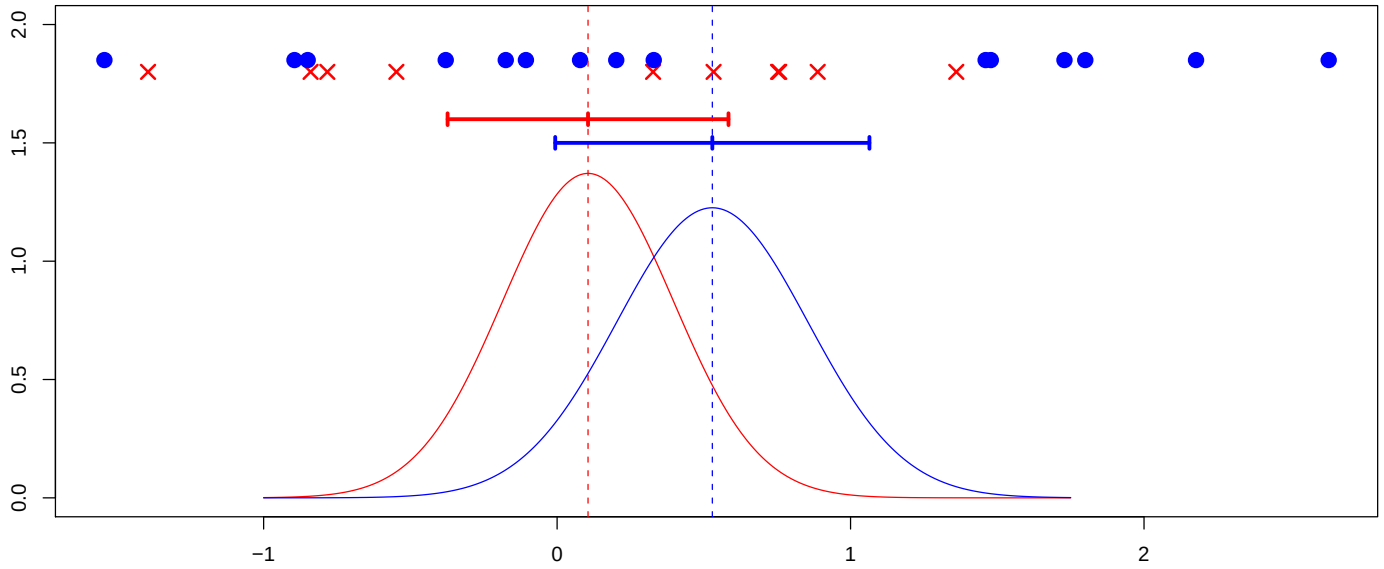



Figure 30: Distribution of $\overline{X}_n$ and $\overline{Y}_m$

## Equal Means of Two (Independent) Samples

Since $\overline{X}$ and $\overline{Y}$ are independent, $\Delta = \overline{X} - \overline{Y}$ has a Gaussian distribution,

$$\mathbb{E}(\Delta) = \mu_X - \mu_Y \text{ and } \text{Var}(\Delta) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$$

Thus, under $H_0$, $\mu_X - \mu_Y = 0$ and thus

$$D \sim \mathcal{N}\left(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right),$$

$$\text{i.e. } \Delta = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1).$$

## Equal Means of Two (Independent) Samples

If $\sigma_X^2$ and $\sigma_Y^2$ are unknown: we will substitute estimators $\widehat{\sigma}_X^2$ et $\widehat{\sigma}_Y^2$,

$$\text{i.e. } \Delta = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\widehat{\sigma}_X^2}{n} + \dfrac{\widehat{\sigma}_Y^2}{m}}} \sim \mathcal{S}t(\nu),$$

where $\nu$ is some complex (but known) function of $n_1$ and $n_2$.

With acceptation rate $\alpha \in [0, 1]$ (e.g. 10%),

$$\begin{cases} \text{accept } H_0 \text{ if } t_{\alpha/2} \leq \delta \leq t_{1-\alpha/2} \\ \text{reject } H_0 \text{ if } \delta < t_{\alpha/2} \text{ ou } \delta > t_{1-\alpha/2} \end{cases}$$

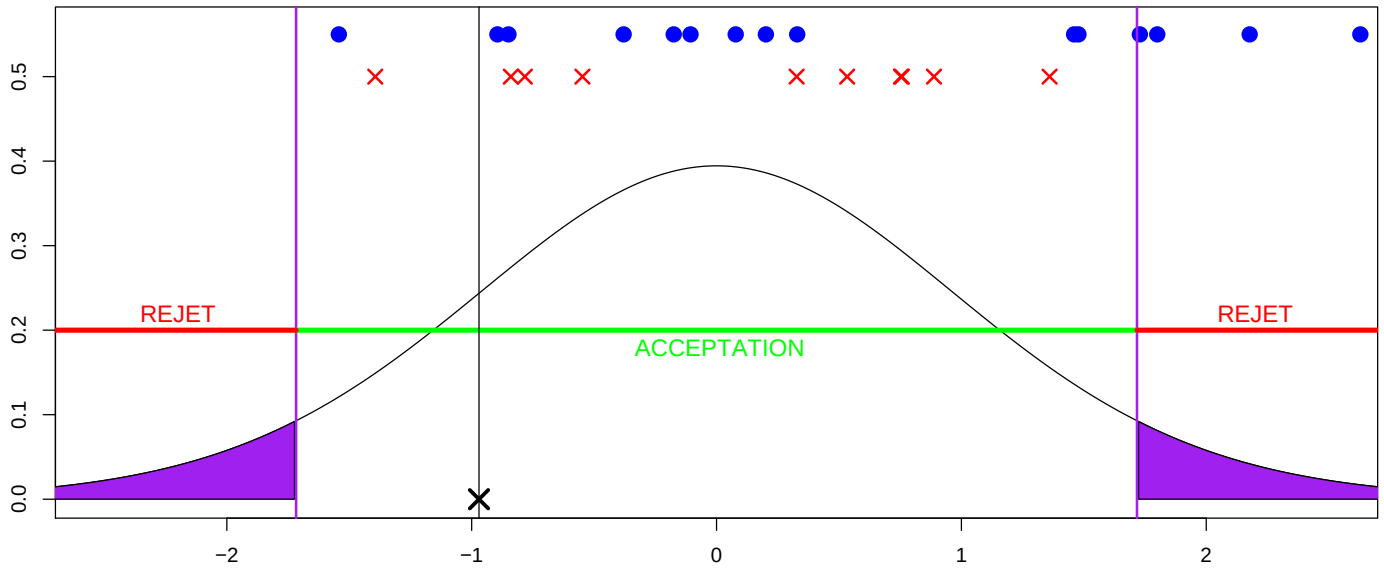



Figure 31: Acceptation and rejection regions

What is the probability $p$ to get a value at least as large as $\delta$ when $H_0$ is valid,

$$p = \mathbb{P}(|Z| > |\delta||H_0 \text{ vraie}) = \mathbb{P}(|Z| > |\delta||Z \sim \mathcal{S}t(\nu)).$$

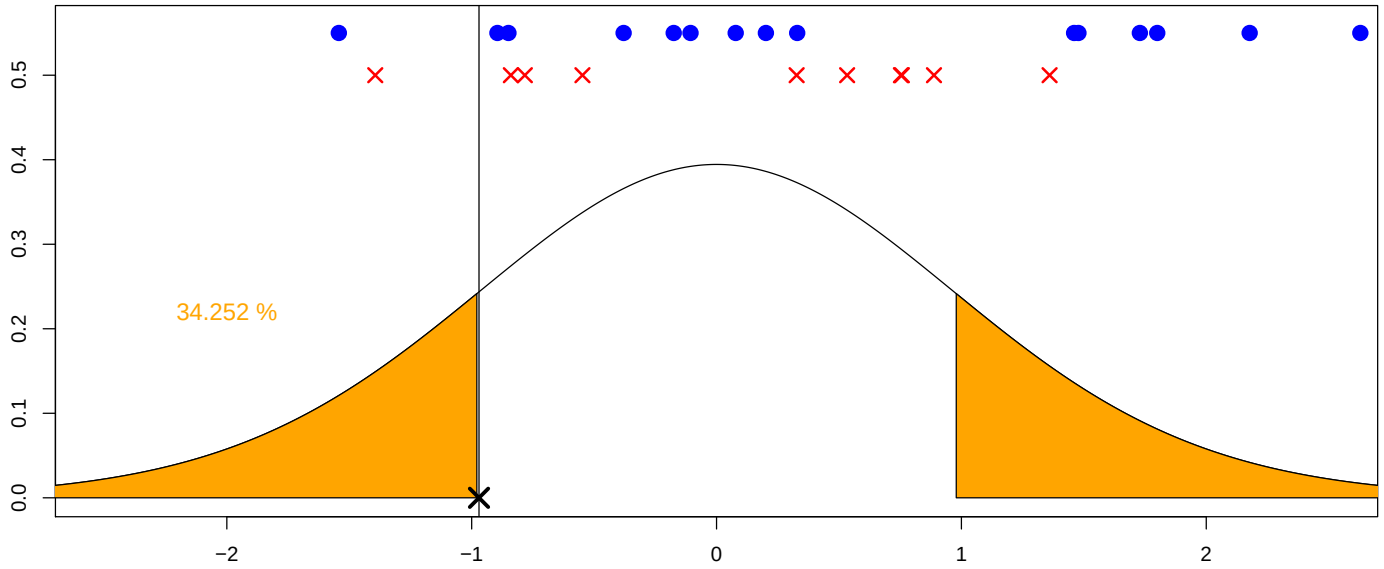
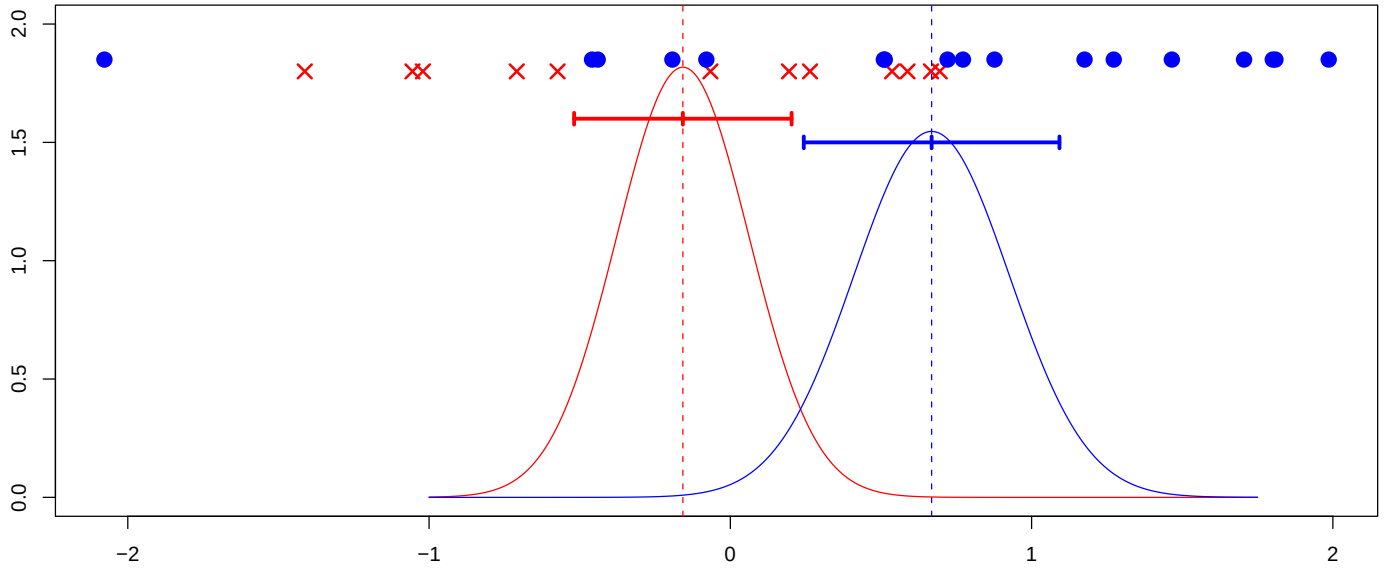



Figure 32: $p$-value of the test.

# Equal Means of Two (Independent) Samples

With R, use `t.test(x, y, alternative = c("two.sided", "less", "greater"), mu = 0, var.equal = FALSE, conf.level = 0.95)` to test if means of vectors `x` and `y` are equal (`mu=0`), against $H_1 : \mu_X \neq \mu_Y$ (`"two.sided"`).



# Equal Means of Two (Independent) Samples

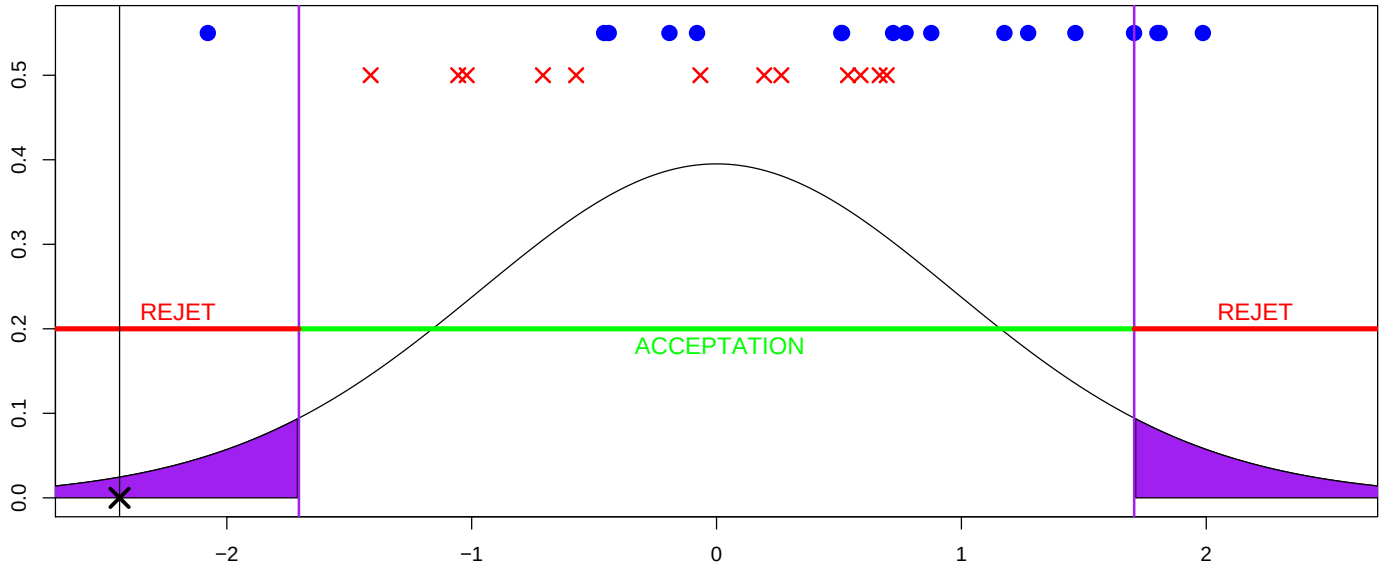



Figure 33: Comparing two means

# Equal Means of Two (Independent) Samples

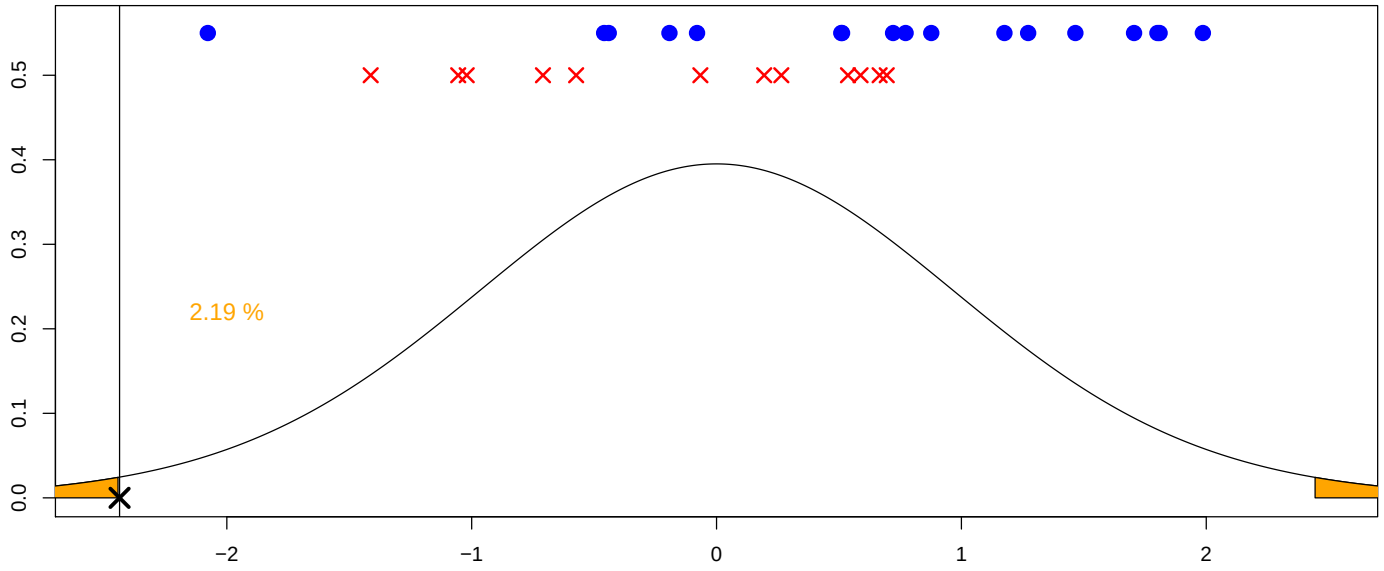



Figure 34: Comparing two means.

# Standard Usual Tests

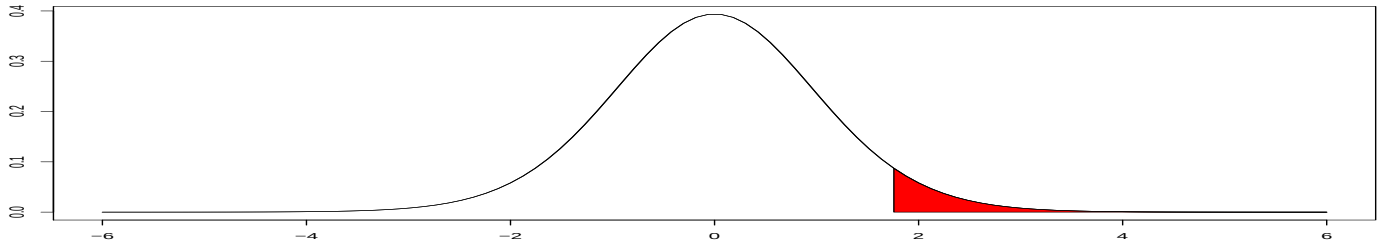
Consider the Mean Equality Test on One Sample

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_0 : \mu \geq \mu_0 \end{cases}$$

The testing statistics is

$$T = \sqrt{n}\frac{\overline{x} - \mu_0}{s} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which satisfies, under $H_0$, $T \sim \mathcal{S}t(n-1)$.



# Standard Usual Tests

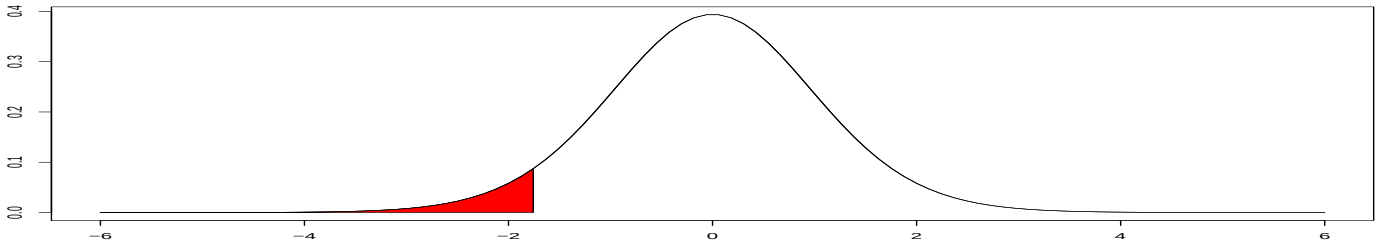
Consider an other alternative assumption (ordering instead of inequality)

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_0 : \mu \leq \mu_0 \end{cases}$$

The testing statistics is the same

$$T = \sqrt{n}\frac{\overline{x} - \mu_0}{s} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which satistifes, uner $H_0$, $T \sim \mathcal{S}t(n-1)$.



# Standard Usual Tests

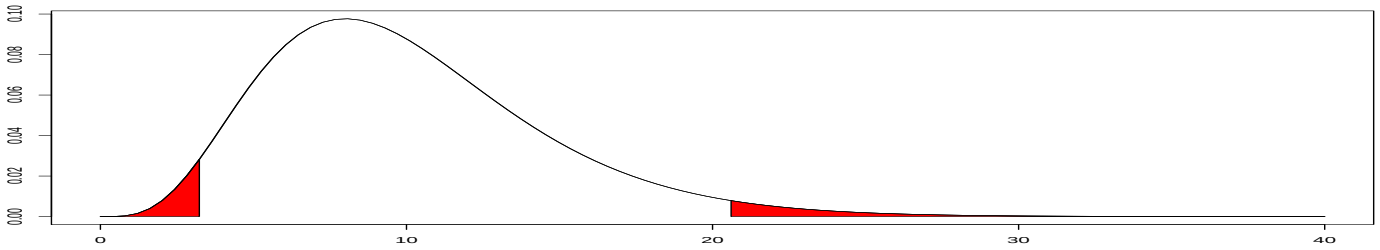
Consider a Test on the Variance (Equality)

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_0 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

The test statistics is here

$$T = \frac{(n-1)s^2}{\sigma_0^2} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which satisfies under $H_0$, $T \sim \chi^2(n-1)$.



# Standard Usual Tests

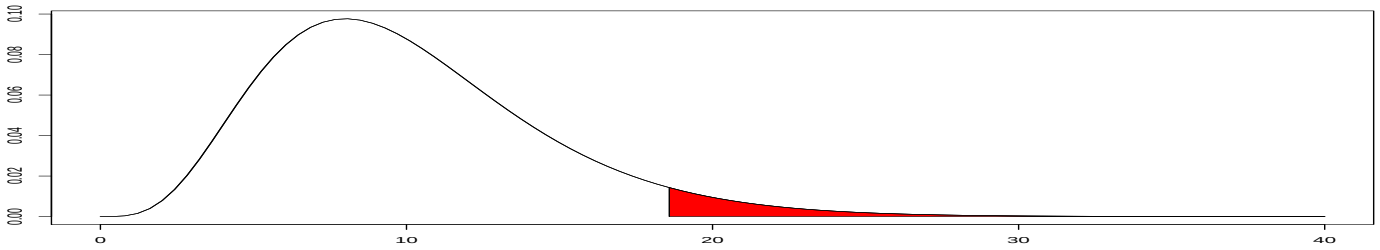
Consider a Test on the Variance (Inequality)

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_0 : \sigma^2 \geq \sigma_0^2 \end{cases}$$

The test statistics is here

$$T = \frac{(n-1)s^2}{\sigma_0^2} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which satisfies under $H_0$, $T \sim \chi^2(n-1)$.



# Standard Usual Tests

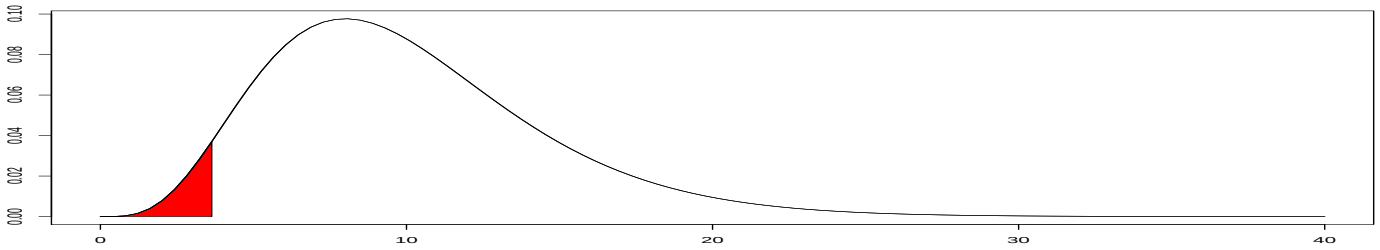
Consider a Test on the Variance (Inequality)

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_0 : \sigma^2 \leq \sigma_0^2 \end{cases}$$

The test statistics is here

$$T = \frac{(n-1)s^2}{\sigma_0^2} \text{ where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

which satisfies under $H_0$, $T \sim \chi^2(n-1)$.



# Standard Usual Tests

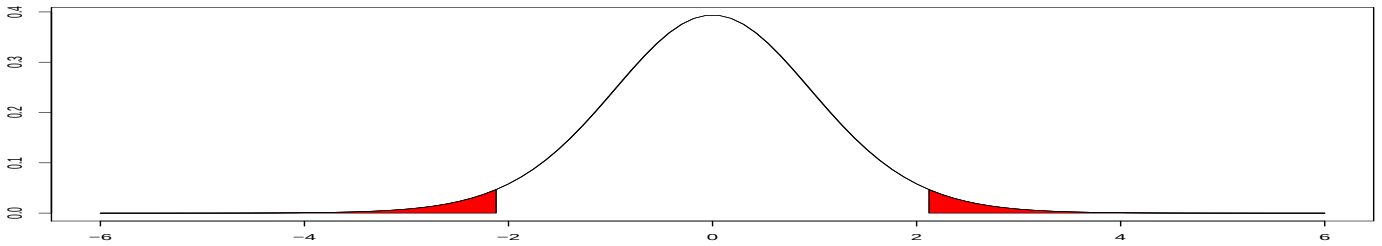
Testing Equality on two Means on two Samples

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_0 : \mu_1 \neq \mu_2 \end{cases}$$

The statistics test is here

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{[\overline{x}_1 - \overline{x}_2] - [\mu_1 - \mu_2]}{s} \text{ where } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

which satisfies under $H_0$, $T \sim \mathcal{S}t(n_1 + n_2 - 2)$.



# Standard Usual Tests

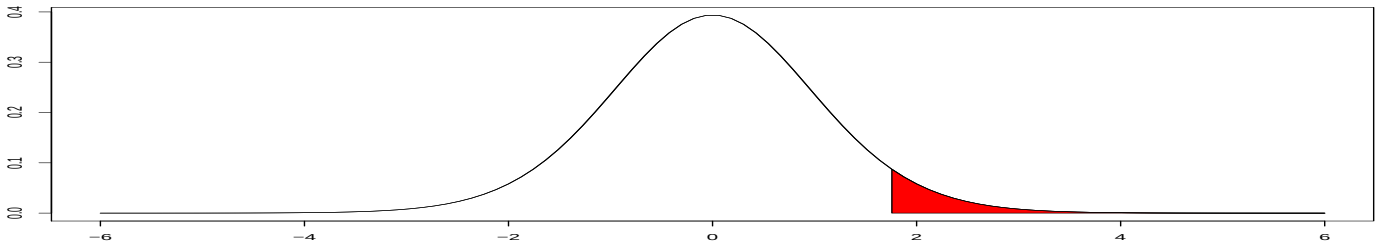
Testing Equality on two Means on two Samples

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_0 : \mu_1 \geq \mu_2 \end{cases}$$

The statistics test is here

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{[\overline{x}_1 - \overline{x}_2] - [\mu_1 - \mu_2]}{s} \text{ where } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

which satisfies under $H_0$, $T \sim \mathcal{S}t(n_1 + n_2 - 2)$.



# Standard Usual Tests

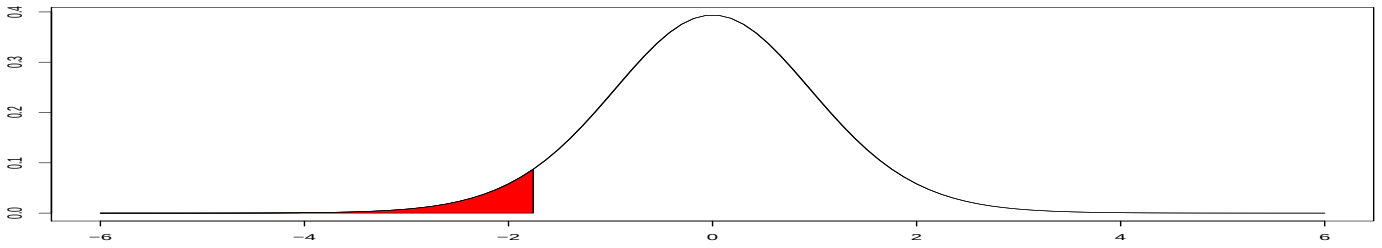
Testing Equality on two Means on two Samples

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_0 : \mu_1 \leq \mu_2 \end{cases}$$

The statistics test is here

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{[\overline{x}_1 - \overline{x}_2] - [\mu_1 - \mu_2]}{s} \text{ where } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

which satisfies under $H_0$, $T \sim \mathcal{S}t(n_1 + n_2 - 2)$.



# Standard Usual Tests

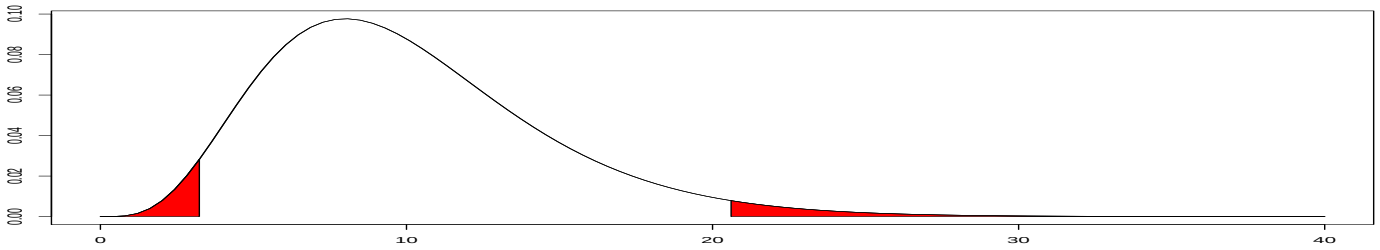
Consider a test of variance equality on two samples

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_0 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

The test statistics is

$$T = \frac{s_1^2}{s_2^2}, \text{ if } s_1^2 > s_2^2,$$

which should follow (with Gaussian samples) under $H_0$, $T \sim \mathcal{F}(n_1 - 1, n_2 - 1)$.



# Standard Usual Tests

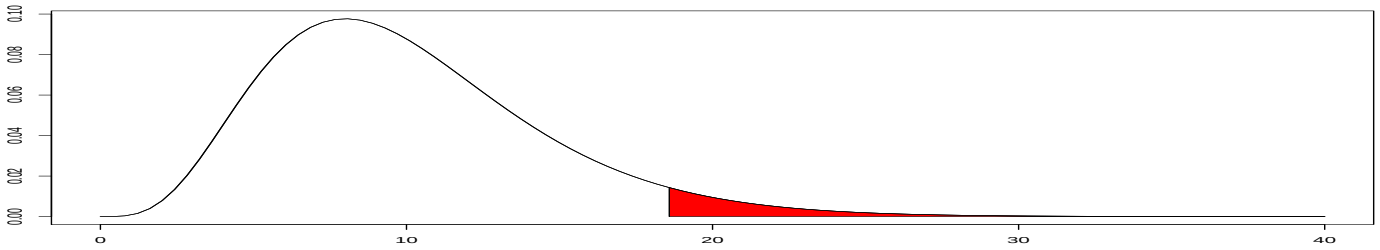
Consider a test of variance equality on two samples

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_0 : \sigma_1^2 \geq \sigma_2^2 \end{cases}$$

The test statistics is here

$$T = \frac{s_1^2}{s_2^2}, \text{ if } s_1^2 > s_2^2,$$

which satisfies, under $H_0$, $T \sim \mathcal{F}(n_1 - 1, n_2 - 1)$.



# Standard Usual Tests

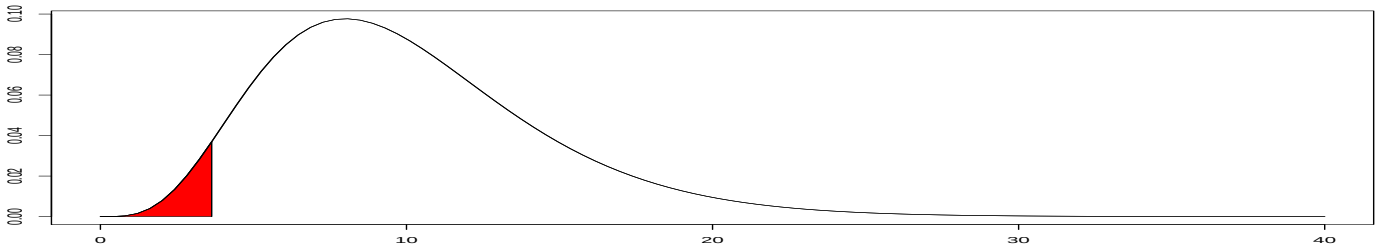
Consider a test of variance equality on two samples

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_0 : \sigma_1^2 \leq \sigma_2^2 \end{cases}$$

The test statistics is here

$$T = \frac{s_1^2}{s_2^2}, \text{ if } s_1^2 > s_2^2,$$

which satisfies under $H_0$, $T \sim \mathcal{F}(n_1 - 1, n_2 - 1)$.



# Multinomial Test

A multinomial distribution is the natural extension of the binomial distribution, from 2 classes $\{0,1\}$ to $k$ classes, say $\{1,2,\cdots,k\}$.

Let $\boldsymbol{p}=(p_1,\cdots,p_k)$ denote a probability distribution on $\{1,2,\cdots,k\}$.

For a multinomial distribution, let $\boldsymbol{n}$ denote a vector in $\mathbb{N}^k$ such that $n_1+\cdots+n_k=n$,

$$\mathbb{P}[\boldsymbol{N}=\boldsymbol{n}]=n!\prod_{i=1}^{n}\frac{p_i^{n_i}}{n_i!}$$

Pearson's chi-squared test has been introduced to test $H_0:\boldsymbol{p}=\boldsymbol{\pi}$ against $H_1:\boldsymbol{p}\neq\boldsymbol{\pi}$

$$X^2=\sum_{i=1}^{k}\frac{(n_i-n\pi_i)^2}{n\pi_i}$$

and under $H_0$, $X^2\sim\chi^2(k-1)$.

# Independence Test (Discrete)

This test is based on Pearson's chi-squared test on the contingency table.

Consider two variables $X \in \{1, 2, \cdots, I\}$ and $Y \in \{1, 2, \cdots, J\}$ and let $\boldsymbol{n} = [n_{i,j}]$ denote the contingency table

$$n_{i,j} = \sum_{k=1}^{n} \mathbf{1}(x_k = i, y_k = j)$$

Let $n_{i,\cdot} = \sum_{j=1}^{J} n_{i,j}$ and $n_{\cdot,j} = \sum_{i=1}^{I} n_{i,j}$.

If variables are independent, $\forall i, j$

$$\underbrace{\mathbb{P}[x = i, y = j]}_{\sim \frac{n_{i,j}}{n}} = \underbrace{\mathbb{P}[x = i]}_{\sim \frac{n_{i,\cdot}}{n}} \cdot \underbrace{\mathbb{P}[y = j]}_{\sim \frac{n_{\cdot,j}}{n}}$$

## Independence Test (Discrete)

Hence, $n_{i,j}^{\perp} = \dfrac{n_{i,\cdot} n_{\cdot,j}}{n}$ would be the value of the contingency table if variables were independent.

Here the statistics used to test $H_0 : X \perp\!\!\!\perp Y$ is

$$X^2 = \sum_{i=1}^{k} \frac{\left(n_{i,j} - n_{i,j}^{\perp}\right)^2}{n_{i,j}^{\perp}}$$

and under $H_0$, $X^2 \sim \chi^2([I-1][J-1])$.

With R, use `chisq.test()`.

# Independence Test (Discrete)

Consider the application with the color of the hair and the color of the eye.

```
> N = margin.table(HairEyeColor, c(1,2))
        Eye
Hair     Brown Blue Hazel Green
  Black     68   20    15     5
  Brown    119   84    54    29
  Red       26   17    14    14
  Blond      7   94    10    16
```

Joint probabilites are

```
> N/n
Hair     Brown  Blue Hazel Green
  Black 0.115 0.034 0.025 0.008
  Brown 0.201 0.142 0.091 0.049
  Red   0.044 0.029 0.024 0.024
  Blond 0.012 0.159 0.017 0.027
```

# Independence Test (Discrete)

Under the independence assumption,

$$p_{i,j} = \frac{N_{i,j}}{n} = \frac{N_{i,\cdot}}{n}\frac{N_{\cdot,j}}{n} = p_{i,j}^{\perp}$$

```
> n=sum(N)
> NHair=apply(N,1,sum)
> NEye =apply(N,2,sum)
> Nind=NHair%*%t(NEye)/n
> Nind/n
              Brown       Blue       Hazel       Green
Black  0.06779584 0.06625502 0.02865915 0.01972243
Brown  0.17953342 0.17545311 0.07589367 0.05222790
Red    0.04456949 0.04355654 0.01884074 0.01296567
Blond  0.07972288 0.07791100 0.03370104 0.02319211
```

# Independence Test (Discrete)

The test statistic is

$$\sum_{i,j} \underbrace{\frac{(n_{i,j}-n_{i,j}^{\perp})^2}{n_{i,j}^{\perp}}}_{\text{contribution}}$$

```
> X=(N-Nind)^2/Nind
> Qobs=sum(X)
> Qobs
[1] 138.2898
> 1-pchisq(Qobs,df=(ncol(N)-1)*(nrow(N)-1))
> chisq.test(N)

        Pearson's Chi-squared test

data:  N
X-squared = 138.2898, df = 9, p-value < 2.2e-16
```

# Independence Test (Continuous)

Pearson's Correlation,

$$r(X,Y)=\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}=\frac{\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{[\mathbb{E}(X^2)-\mathbb{E}(X)^2]\cdot[\mathbb{E}(Y^2)-\mathbb{E}(Y)^2]}}$$

Spearman's (Rank) Correlation

$$\rho(X,Y)=\frac{\text{Cov}(F_X(X),F_Y(Y))}{\sqrt{\text{Var}(F_X(X))\text{Var}(F_Y(Y))}}=12\ \text{Cov}(F_X(X),F_Y(Y))$$

Let $d_i=R_i-S_i=n(\widehat{F}_X(x_i)-\widehat{F}_Y(y_i))$ and define $R=\sum R_i^2$

Test on Correlation Coefficient

$$Z=\frac{6R-n(n^2-1)}{n(n+1)\sqrt{n-1}}$$

## Parametric Modeling

Consider a sample $\{x_1, \cdots, x_n\}$, with $n$ independent observations.

Assume that $x_i$'s are obtained from random variables with identical (unknown) distribution $F$.

In parametric statistics, $F$ belongs to some family $\mathcal{F} = \{F_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$.

- $X$ has a Bernoulli distribution, $X \sim \mathcal{B}(p)$, $\theta = p \in (0, 1)$,
- $X$ has a Poisson distribution, $X \sim \mathcal{P}(\lambda)$, $\theta = \lambda \in \mathbb{R}^+$,
- $X$ has a Gaussian distribution, $X \sim \mathcal{N}(\mu, \sigma)$, $\boldsymbol{\theta} = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,

We want to find the best choice for $\boldsymbol{\theta}$, the true unknown value of the parameter, so that $X \sim F_{\boldsymbol{\theta}}$.

# Heads and Tails

Consider the following sample

$$\{\text{head}, \text{head}, \text{tail}, \text{head}, \text{tail}, \text{head}, \text{tail}, \text{tail}, \text{head}, \text{tail}, \text{head}, \text{tail}\}$$

that we will convert using

$$X = \begin{cases} 1 \text{ if head} \\ 0 \text{ if tail.} \end{cases}$$

Our sample is now

$$\{1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0\}$$

Here $X$ has a Bernoulli distribution $X \sim \mathcal{B}(p)$, where parameter $p$ is unknown.

# Statistical Inference

What is the true unknown value of $p$ ?

- What is the value for $p$ that could be the most likely?

Over $n$ draws, the probability to get exactly our sample $\{x_1, \cdots, x_n\}$ is

$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n),$$

where $X_1, \cdots, X_n$ are $n$ independent verions of $X$, with distribution $\mathcal{B}(p)$. Hence,

$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} p^{x_i} \times (1-p)^{1-x_i},$$

because $p^{x_i} \times (1-p)^{1-x_i} = \begin{cases} p \text{ if } x_i \text{ equals } 1 \\ 1-p \text{ if } x_i \text{ equals } 0 \end{cases}$

## Statistical Inference

Thus,

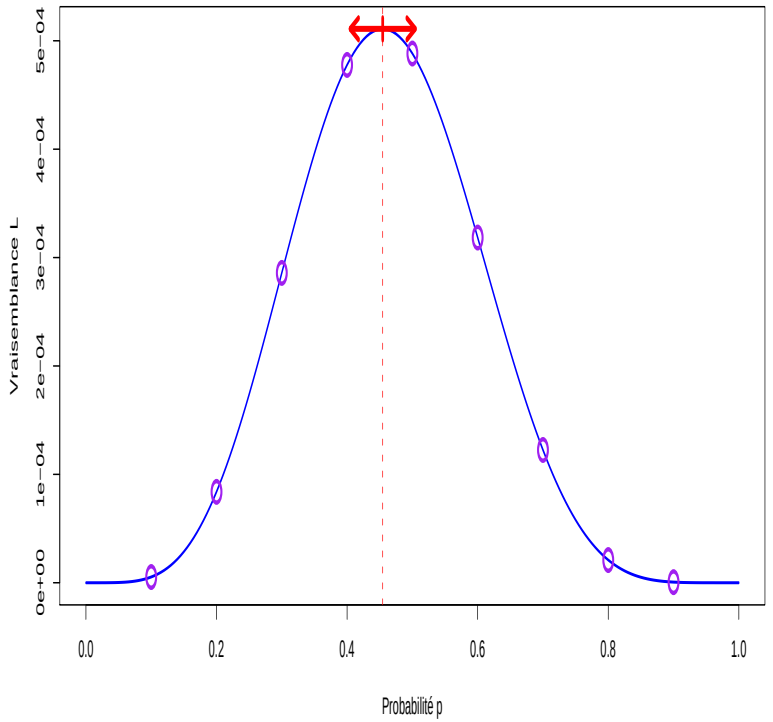
$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = p^{\sum_{i=1}^n x_i} \times (1-p)^{\sum_{i=1}^n 1-x_i}.$$

This function which depends on $p$ (but also $\{x_1, \cdots, x_n\}$) is called likelihood of the sample, and is denoted $\mathcal{L}$,

$$\mathcal{L}(p; x_1, \cdots, x_n) = p^{\sum_{i=1}^n x_i} \times (1-p)^{\sum_{i=1}^n 1-x_i}.$$

Here we have obtained 5 times 1's and 6 times 0's. As a function of $p$ we get the difference likelihoods,

| Value of $p$ | $\mathcal{L}(p; x_1, \cdots, x_n)$ |
|---|---|
| 0.1 | 5.314410e-06 |
| 0.2 | 8.388608e-05 |
| 0.3 | 2.858871e-04 |
| 0.4 | 4.777574e-04 |
| 0.5 | 4.882812e-04 |
| 0.6 | 3.185050e-04 |
| 0.7 | 1.225230e-04 |
| 0.8 | 2.097152e-05 |
| 0.9 | 5.904900e-07 |




The value with the highest likelihood $p$ is here 0.4545.

## Statistical Inference

- Why not use the (empirical) mean?

We have obtained the following sample

$$\{1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0\}$$

For a Bernoulli distribution, $\mathbb{E}(X) = p$. Thus, it can be seen as natural to use a estimator of $p$ an estimator of $\mathbb{E}(X)$, the average of 1's is our sample, $\overline{x}$.

A natural estimator for $p$ would be $\overline{x}$ $5/11 = 0.4545$.

# Maximum Likelihood

In a more general setting, let $f_{\boldsymbol{\theta}}$ denote the true (unknown) distribution of $X$,

- if $X$ is continuous, $f_{\boldsymbol{\theta}}$ denotes the density i.e. $f_{\boldsymbol{\theta}}(x) = \dfrac{dF(x)}{dx} = F'(x)$,
- if $X$ is discrete, $f_{\boldsymbol{\theta}}$ denotes the probability $f_{\boldsymbol{\theta}}(x) = \mathbb{P}(X = x)$,

Since $X_i$'s are i.i.d., the likelihood of the sample is

$$\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n) = \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)$$

A natural estimator for $\boldsymbol{\theta}$ is obtained as the maximum of the likelihood

$$\widehat{\boldsymbol{\theta}} \in \text{argmax}\{\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

One should keep in mind that for any increasing function $h$,

$$\widehat{\boldsymbol{\theta}} \in \text{argmax}\{h\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

# Maximum Likelihood

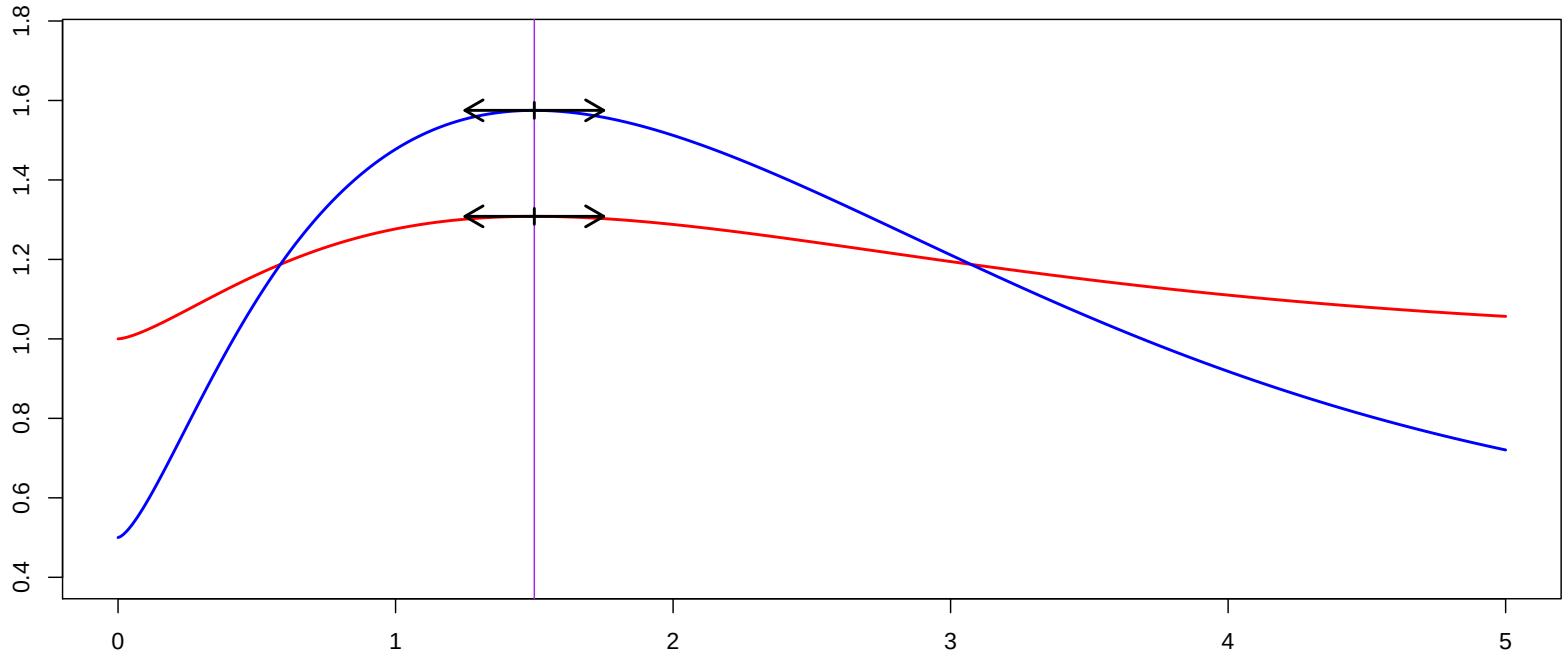



Figure 35: Invariance of the maximum's location.

## Maximum Likelihood

Consider the case here where $h = \log$

$$\widehat{\boldsymbol{\theta}} \in \text{argmax}\{\log\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

i.e. equivalently, we can look for the maximum of the log-likelihood, which can be written

$$\log \mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(x_i)$$

From a practical perspective, the first order condition will ask us to compute derivatives, and the derivative of a sum is easier to derive than the derivative of a product, assuming that $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})$ is differentiable.

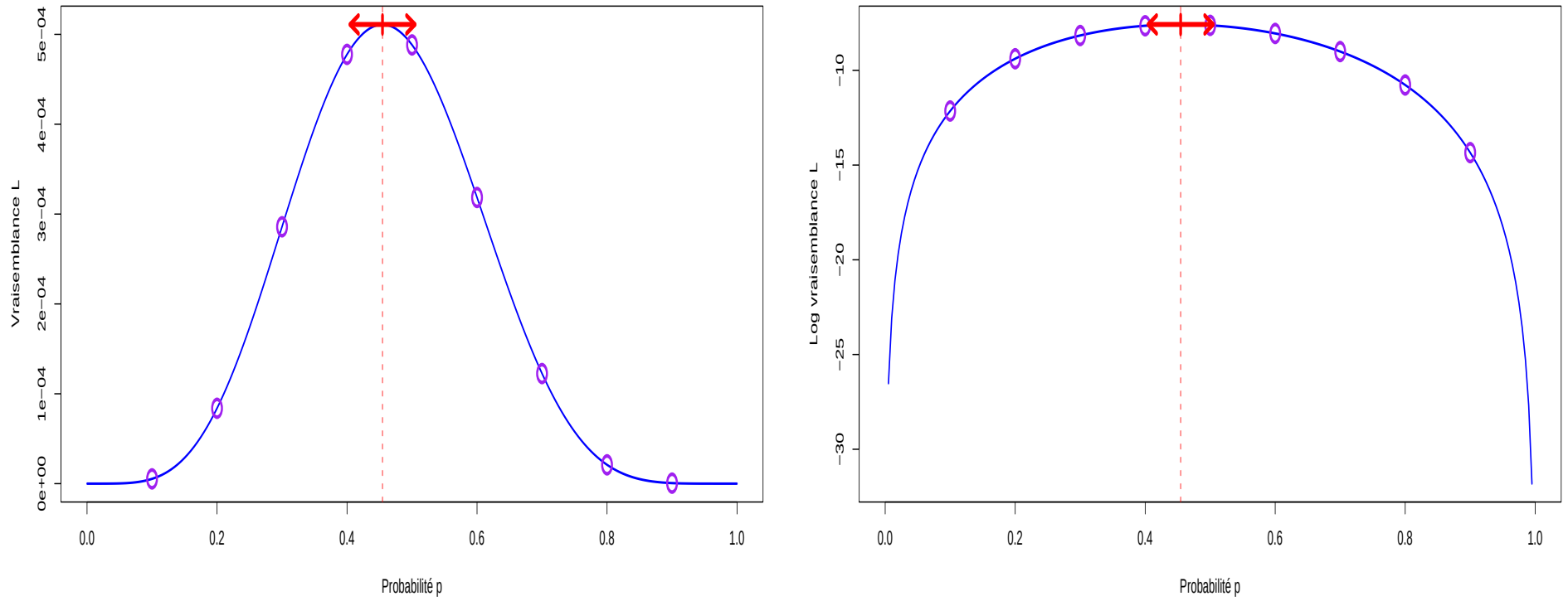



Figure 36: Likelihood and log-likelihood.

# Maximum Likelihood

Likelihood equations are

- First order condition

$$\text{if } \boldsymbol{\theta} \in \mathbb{R}^k, \left.\frac{\partial \log\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}$$

$$\text{if } \theta \in \mathbb{R}, \left.\frac{\partial \log\left(\mathcal{L}(\theta; x_1, \cdots, x_n)\right)}{\partial \theta}\right|_{\theta=\widehat{\theta}} = 0$$

- Second order condition

$$\text{if } \boldsymbol{\theta} \in \mathbb{R}^k, \left.\frac{\partial^2 \log\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \text{ is definite negative}$$

$$\text{if } \theta \in \mathbb{R}, \left.\frac{\partial^2 \log\left(\mathcal{L}(\theta; x_1, \cdots, x_n)\right)}{\partial \theta}\right|_{\theta=\widehat{\theta}} < 0$$

Function $\dfrac{\partial \log\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right)}{\partial \boldsymbol{\theta}}$ is the fonction score: at the maximum, the score is null.

# Fisher Information

An estimator $\widehat{\theta}$ of $\theta$ is said to be sufficient if it contains as much information about $\theta$ as the whole sample $\{x_1, \cdots, x_n\}$.

Fisher information associated with a density $f_\theta$, with $\theta\mathbb{R}$ is

$$I(\theta) = \mathbb{E}\left(\frac{d}{d\theta}\log f_\theta(X)\right)^2 \text{ where } X \text{ has distribution } f_\theta,$$

$$I(\theta) = \text{Var}\left(\frac{d}{d\theta}\log f_\theta(X)\right) = -\mathbb{E}\left(\frac{d^2}{d\theta^2}\log f_\theta(X)\right).$$

Fisher information is the variance of the score function (applied to some random variables).

This is information related to $X$, and in the case of a sample $X_1, \cdots, X_n$ i.id. with density $f_\theta$, the information is $I_n(\theta) = n \cdot I(\theta)$.

## Efficiency and Optimality

If $\widehat{\theta}$ is an unbiased estimator of $\theta$, then $\text{Var}(\widehat{\theta}) \geq \dfrac{1}{nI(\theta)}$. If that bound is attained, the estimator is said to beefficient.

Note that this lower bound is not necessarily reached.

An unbiased estimator $\widehat{\theta}$ is said to be optimal if it has the lowest variance among all unbiased estimators.

see bias, minimum variance unbiased estimator

## Fisher information in higher dimension

If $\boldsymbol{\theta} \in \mathbb{R}^k$, then Fisher information is the $k \times k$ matrix $I = [I_{i,j}]$ with

$$I_{i,j} = \mathbb{E}\left( \frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(X) \frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(X) \right).$$

see Fisher information

# Fisher Information & Computations

Assume that $X$ has a Poisson distribution $\mathcal{P}(\theta)$,

$$\log f_\theta(x) = -\theta + x\log\theta - \log(x!) \text{ and } \frac{d^2}{d\theta^2}\log f_\theta(x) = -\frac{x}{\theta^2}$$

$$I(\theta) = -\mathbb{E}\left(\frac{d^2}{d\theta^2}\log f_\theta(X)\right) = -\mathbb{E}\left(-\frac{X}{\theta^2}\right) = \frac{1}{\theta}$$

For a binomial distribution $\mathcal{B}(n,\theta)$, $I(\theta) = \dfrac{n}{\theta(1-\theta)}$

For a Gaussian distribution $\mathcal{N}(\theta,\sigma^2)$, $I(\theta) = \dfrac{1}{\sigma^2}$

For a Gaussian distribution $\mathcal{N}(\mu,\theta)$, $I(\theta) = \dfrac{1}{2\theta^2}$

## Heuristic Interpretations

Let $S_{\boldsymbol{\theta}}(\boldsymbol{x}) = \dfrac{\partial \log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta}}$ denote the score function.

Consider a sample $\boldsymbol{X}$ i.id. from $F_{\boldsymbol{\theta}}$, then $S_{\boldsymbol{\theta}}(\boldsymbol{X})$ is a random variable. Then

$$\mathbb{E}[S_{\boldsymbol{\theta}}(\boldsymbol{X})] = \mathbf{0}$$

while

$$\text{Var}[S_{\boldsymbol{\theta}}(\boldsymbol{X})] = I_n(\boldsymbol{\theta}).$$

# Maximum Likelihood

**Definition** Let $\{x_1, \cdots, x_n\}$ be a sample with distribution $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_n \in \text{argmax}\big\{\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n), \boldsymbol{\theta} \in \boldsymbol{\Theta}\big\}.$$

**Proposition.** *Under some technical assumptions $\widehat{\boldsymbol{\theta}}_n$ converges almost surely towards $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}$, as $n \to \infty$.*

**Proposition.** *Under some technical assumptions $\widehat{\boldsymbol{\theta}}_n$ is asymptotically efficient,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\boldsymbol{\theta})).$$

Results are only asymptotic, there is no reason, e.g., to have an unbiased estimator, see maximum likelihood estimation

## Gaussian case, $\mathcal{N}(\mu, \sigma^2)$

Let $\{x_1, \cdots, x_n\}$ be a sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, with density

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

The likelihood is here

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right),$$

i.e.

$$\mathcal{L}(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right).$$

## Gaussian case, $\mathcal{N}(\mu, \sigma^2)$

The maximum likelihood estimator of $\mu$ is obtained from the first order equations

$$\begin{aligned}
& \frac{\partial}{\partial \mu} \log \mathcal{L} \\
= & \frac{\partial}{\partial \mu} \log \left( \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left( - \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right) \right) \\
= & \frac{\partial}{\partial \mu} \left( \log \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} - \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right) \\
= & 0 - \frac{-2n(\bar{x} - \mu)}{2\sigma^2} = 0.
\end{aligned}$$

i.e. $\widehat{\mu} = \bar{x} = \dfrac{1}{n} \sum_{i=1}^n x_i.$

The second part of the first order condition is here

$$\begin{aligned}
& \frac{\partial}{\partial\sigma}\log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^n(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right)\\
= & \frac{\partial}{\partial\sigma}\left(\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right)-\frac{\sum_{i=1}^n(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right)\\
= & -\frac{n}{\sigma}+\frac{\sum_{i=1}^n(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{\sigma^3}=0.
\end{aligned}$$

The first order condition yields

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^n(x_i-\widehat{\mu})^2=\frac{1}{n}\sum_{i=1}^n(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^n x_i^2-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n x_ix_j.$$

Observe that here $\mathbb{E}\left[\widehat{\mu}\right]=\mu$, while $\mathbb{E}\left[\widehat{\sigma}^2\right]\neq\sigma^2$.

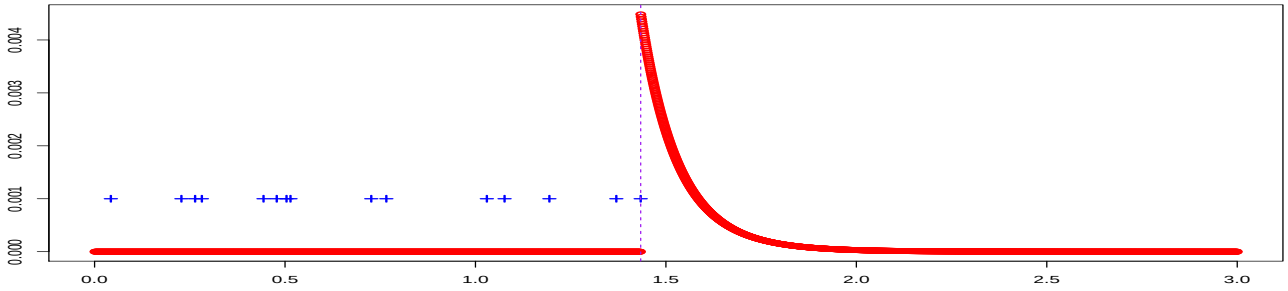
# Uniform Distribution on $[0, \theta]$

The density of the $X_i$'s is $f_\theta(x) = \dfrac{1}{\theta}\mathbf{1}(0 \leq x \leq \theta)$.

The likelihood function is here

$$\mathcal{L}(\theta; x_1, \cdots, x_n) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}(0 \leq x_i \leq \theta) = \frac{1}{\theta^n}\mathbf{1}(0 \leq \inf\{x_i\} \leq \sup\{x_i\} \leq \theta).$$

Unfortunately, that function is not differentiable in $\theta$, it we can see that $\mathcal{L}$ is maximal when $\theta$ is as small as possible, i.e. $\widehat{\theta} = \sup\{x_i\}$.



## Uniform Distribution on $[\theta, \theta+1]$

In some case, the maximum likelihood is not unique.

Assume that $\{x_1, \cdots, x_n\}$ are uniformly distributed on $[\theta, \theta+1]$. If

$$\widehat{\theta}^- = \sup\{x_i\} - 1 < \inf\{x_i\} = \widehat{\theta}^+$$

then any estimator $\widehat{\theta} \in [\widehat{\theta}^-, \widehat{\theta}^+]$ is a maximum likelihood estimator of $\theta$.

And as mentioned already, the maximum likelihood estimator is not necessarily unbiased. For the exponential distribution, $\widehat{\theta} = 1/\overline{x}$. One can prove that in that case

$$\mathbb{E}(\widehat{\theta}) = \frac{n}{n-1}\theta > \theta.$$

# Gamma distribution

The log-likelihood function is

$$\log \mathcal{L}(k,\theta) = = (k-1)\sum_{i=1}^{n}\ln(x_i) - \sum_{i=1}^{n}\frac{x_i}{\theta} - nk\ln(\theta) - n\log(\Gamma(k))$$

The first order condition for $\theta$ yields

$$\hat{\theta} = \frac{1}{kn}\sum_{i=1}^{n} x_i$$

Substituting this into the log-likelihood function gives

$$\log \mathcal{L} = (k-1)\sum_{i=1}^{n}\ln(x_i) - nk - nk\log\left(\frac{\sum x_i}{kn}\right) - n\ln(\Gamma(k))$$

which cannot be solve explicitely...

# Numerical Aspects

For standard distribution, in R, use `library(MASS)` to get the maximum likelihood estimator, e.g. `fitdistr(x.norm,"normal")` for a normal distribution and a sample `x`.

One can also use numerical algorithm, in R. It is necessary to define the log-likelihood `LV <- function(theta){-sum(log(dexp(x,theta)))}` and the use `optim(2,LV)` to get the minimum of that function (since it computes a minimum, use the opposite of the log-likelihood).

Numerically, those function are based on Newton-Rahpson also called Fisher's score to approximate the maximum of that function.

Let $S(x,\theta)=\dfrac{\partial}{\partial\theta}\log f(x,\theta)$ the score function. Set

$$S_n(\theta)=\sum_{i=1}^{n}S(X_i,\theta).$$

## Numerical Aspects

Then use Taylor approximation of $S_n$ in the neighbourhood of $\theta_0$,

$$S_n(x) = S_n(\theta_0) + (x - \theta_0)S_n'(y) \text{ for some } y \in [x, \theta_0]$$

Set $x = \widehat{\theta}_n$, then

$$S_n(\widehat{\theta}_n) = 0 = +(\widehat{\theta}_n - \theta_0)S_n'(y) \text{ for some } y \in [\theta_0, \widehat{\theta}_n]$$

Hence, $\widehat{\theta}_n = \theta_0 - \dfrac{S_n(\theta_0)}{S_n'(y)}$ for $y \in [\theta_0, \widehat{\theta}_n]$

## Numerical Aspects

Let us now construct the following sequence (Newton-Raphson)

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} - \frac{S_n(\widehat{\theta}_n^{(i)})}{S_n'(\widehat{\theta}_n^{(i)})},$$

from some starting value $\widehat{\theta}_n^{(0)}$ (hopefully well chosen).

This can be seen as the Score technique

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} - \frac{S_n(\widehat{\theta}_n^{(i)})}{nI(\widehat{\theta}_n^{(i)})},$$

again from some starting value.

# Numerical Aspects : Non-Identifiable Model

**Definition** Consider a family of distributions $\mathcal{F} = \{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ is identifiable if if the mapping $\boldsymbol{\theta} \mapsto F_{\boldsymbol{\theta}}$ is one-to-one:

$$F_{\boldsymbol{\theta}_1} = F_{\boldsymbol{\theta}_2} \text{ implies } \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

**Example** The Gaussian distribution, $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, $f_{\boldsymbol{\theta}}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$. If $f_{\boldsymbol{\theta}_1} = f_{\boldsymbol{\theta}_2}$ then

$$\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right)$$

$$\frac{1}{\sigma_1^2}(x-\mu_1)^2 + \ln\sigma_1 = \frac{1}{\sigma_2^2}(x-\mu_2)^2 + \ln\sigma_2$$

$$x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} + \ln\sigma_1 - \ln\sigma_2\right) = 0 \ \forall x \in \mathbb{R}$$

hence $\sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$.

## Numerical Aspects : Non-Identifiable Model

**Example** Mixture of two distributions : $\boldsymbol{\theta} = (p, \lambda, \mu) \in (0,1) \times \mathbb{R}_+ \times \mathbb{R}_+$,

$$f_{\boldsymbol{\theta}}(x) = p \cdot \left(\lambda e^{-\lambda x}\right) + (1-p) \cdot \left(\mu e^{-\mu x}\right)$$

Observe that $\boldsymbol{\theta}_1 = (p, \lambda, \mu)$ and $\boldsymbol{\theta}_2 = (1-p, \mu, \lambda)$ yield the same distributions, since $f_{\boldsymbol{\theta}_1}(x) = f_{\boldsymbol{\theta}_2}(x)$ $\forall x \in \mathbb{R}_+$.

It is necessary to add a (linear) constraint : either $p > 1-p$ or $\lambda > \mu$.

With `R`, one can solve numerically $\min\left\{\log \mathcal{L}(\boldsymbol{\theta})\right\}$ for $\boldsymbol{\theta} \in \mathbb{R}^p$,

but also more generally $\min\left\{\log \mathcal{L}(\boldsymbol{\theta})\right\}$ for $\boldsymbol{\theta} \in \mathbb{R}^p$ subject to $\boldsymbol{U}\boldsymbol{\theta} - \boldsymbol{c} \geq \boldsymbol{0}$ for some $k \times p$ matrix $\boldsymbol{U}$ and $k$ dimensional vector $\boldsymbol{c}$.

# Numerical Aspects : Non-Identifiable Model

In a general context, consider an optimisation problem

$$\min\big\{f(\boldsymbol{\theta})\big\},\ \text{subject to } g_i(\boldsymbol{\theta})\leq 0,\ \forall i=1,\cdots,k$$

Define the Lagrangian of the problem,

$$\ell(\boldsymbol{\theta},\boldsymbol{\lambda})=f(\boldsymbol{\theta})+\sum_{j=1}^{k}\lambda_i g_i(\boldsymbol{\theta})$$

First order conditions are the following $p+k$ equations

$$\left.\frac{\partial \ell(\boldsymbol{\theta},\boldsymbol{\lambda})}{\theta_i}\right|_{(\boldsymbol{\theta},\boldsymbol{\lambda})=(\boldsymbol{\theta}^\star,\boldsymbol{\lambda}^\star)}=0,\ \forall i=1,\cdots,p,$$

$$\left.\frac{\partial \ell(\boldsymbol{\theta},\boldsymbol{\lambda})}{\lambda_j}\right|_{(\boldsymbol{\theta},\boldsymbol{\lambda})=(\boldsymbol{\theta}^\star,\boldsymbol{\lambda}^\star)}=0,\ \forall j=1,\cdots,k,$$

(see Microeconomics course).

# Numerical Aspects : Non-Identifiable Model

The log-likelood for the mixture of Exponential distributions is

```
logL = function(param){
-sum(log(param[1]*dexp(X,param[2])+(1-param[1])*dexp(X,param[3])))
}
Amat = matrix(c(1,0,1,0,0,-1), 2, 3)
bvec = c(0,0)
constrOptim(c(.25,2,.5), logL, NULL, ui = Amat, ci = bvec)$par
```

# Testing Procedures Based on Maximum Likelihood

Consider the heads/tails problem.

We can derive an asyptotic confidence interval from properties of the maximum likelihood

$$\sqrt{n}(\pi - \widehat{\pi}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\pi))$$

where $I(\pi)$ denotes Fisher's information, i.e.

$$I(\pi) = \frac{1}{\pi[1-\pi]}$$

which yields the following (95%) confidence interval for $\pi$

$$\left[\widehat{\pi} \pm \frac{1.96}{\sqrt{n}} \sqrt{\widehat{\pi}[1-\widehat{\pi}]}\right].$$

# Testing Procedures Based on Maximum Likelihood

Consider the following (simulated) sample $\{y_1, \cdots, y_n\}$

```
> set.seed(1)
> n=20
> (Y=sample(0:1,size=n,replace=TRUE))
 [1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1
```

Here $Y_i \sim \mathcal{B}(\pi)$, with $\pi = \mathbb{E}(Y)$. Set $\widehat{\pi} = \overline{y}$, i.e.

```
> mean(Y)
[1] 0.55
```

Consider some test $H_0 : \pi = \pi_\star$ against $H_1 : \pi \neq \pi_\star$ (with e.g. $\pi^\star = 50\%$)

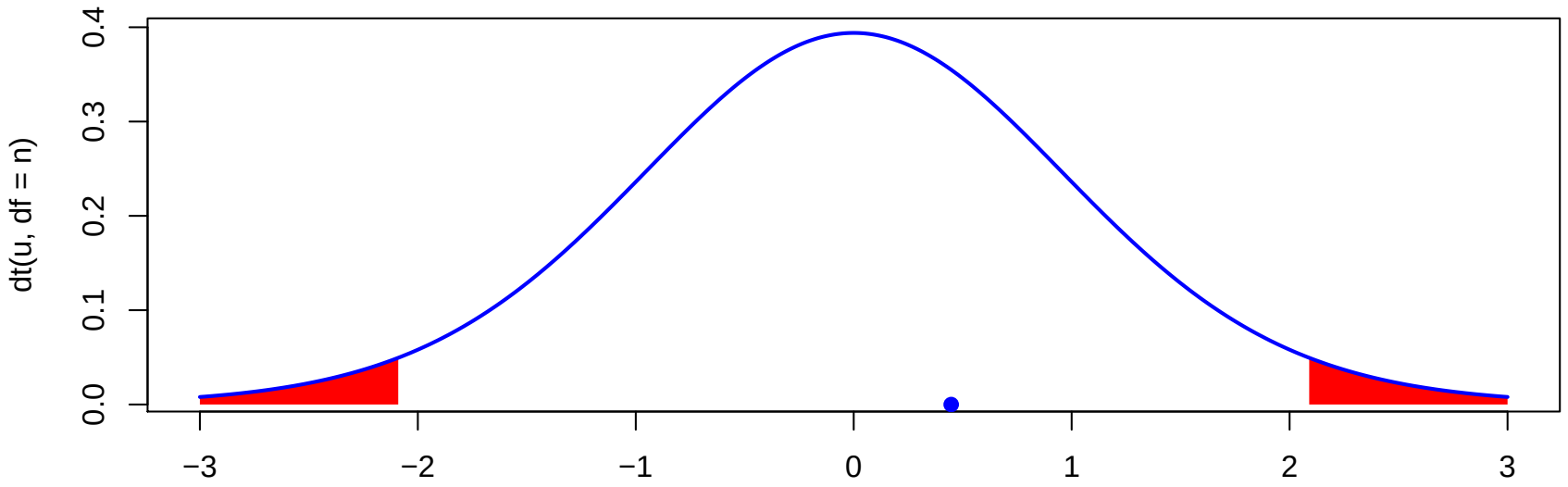
One can use Student $t$-test

$$T = \sqrt{n}\frac{\widehat{\pi} - \pi_\star}{\sqrt{\pi_\star(1-\pi_\star)}}$$

which has, under $H_0$, a Student $t$ distribution with $n$ degrees of freedom.

# Testing Procedures Based on Maximum Likelihood

```
> (T=sqrt(n)*(pn-p0)/(sqrt(p0*(1-p0))))
[1] 0.4472136
> abs(T)<qt(1-alpha/2,df=n)
[1] TRUE
```

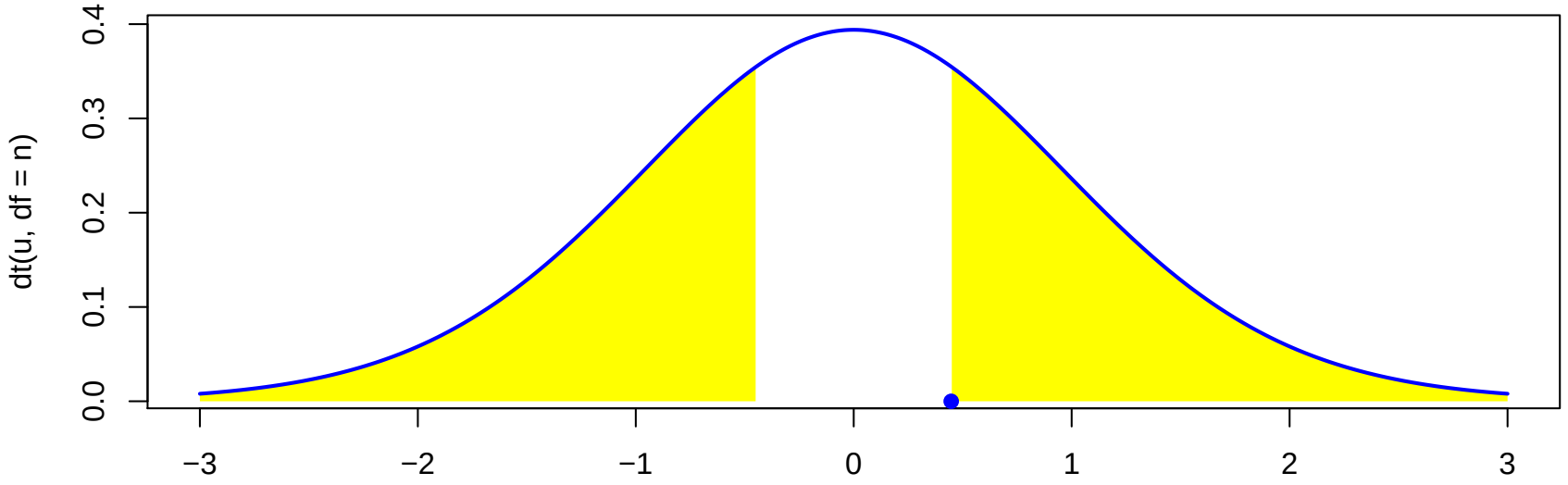


# Testing Procedures Based on Maximum Likelihood

We are here in the acceptance region of the test.

One can also compute the $p$-value, $\mathbb{P}(|T| > |t_{obs}|)$, see $p$-value

```
>   2*(1-pt(abs(T),df=n))
[1] 0.6595265
```



# Testing Procedures Based on Maximum Likelihood

The idea of Wald test is to look at the difference between $\widehat{\pi}$ and $\pi_\star$. Under $H_0$,

$$T = n\frac{(\widehat{\pi}-\pi_\star)^2}{I^{-1}(\pi_\star)} \xrightarrow{\mathcal{L}} \chi^2(1)$$

The idea of the likelihood ratio test is to look at the difference between $\log\mathcal{L}(\widehat{\theta})$ and $\log\mathcal{L}(\theta_\star)$ (i.e. the logarithm of the ratio). Under $H_0$,

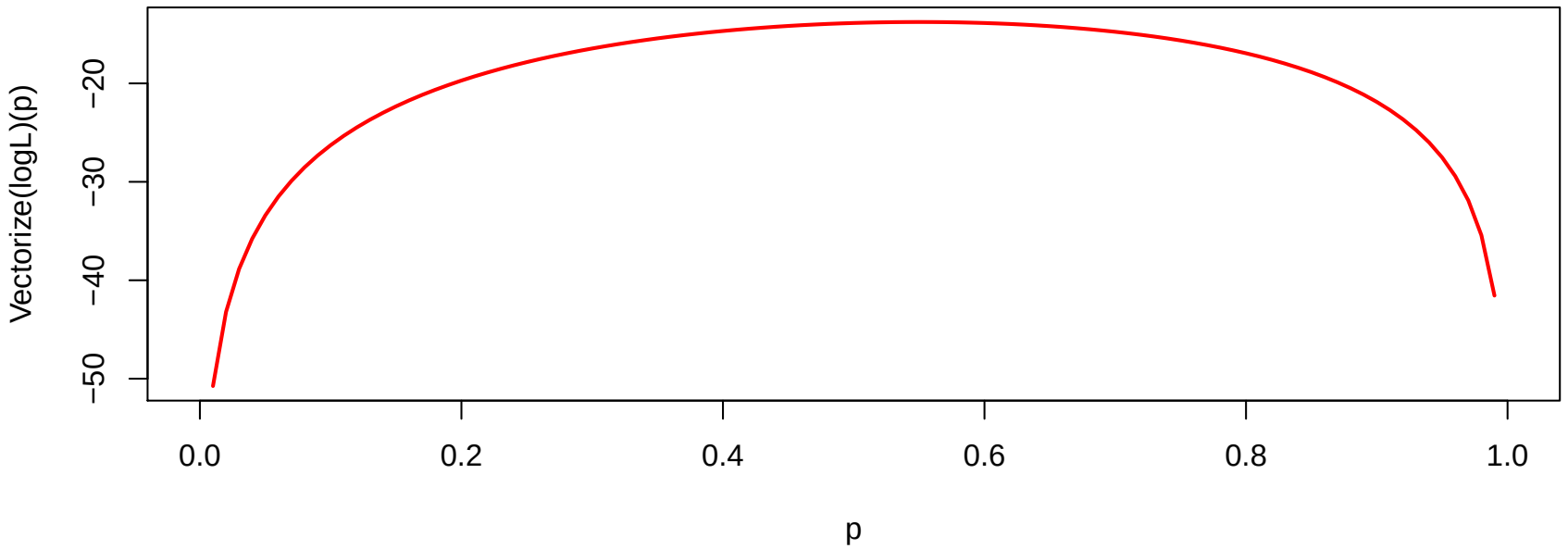
$$T = 2\log\left(\frac{\log\mathcal{L}(\theta_\star)}{\log\mathcal{L}(\widehat{\theta})}\right) \xrightarrow{\mathcal{L}} \chi^2(1)$$

The idea of the Score test is to look at the difference between $\dfrac{\partial\log\mathcal{L}(\pi_\star)}{\partial\pi}$ and 0. Under $H_0$,

$$T = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log f_{\pi_\star}(x_i)}{\partial\pi}\right)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$$

# Testing Procedures Based on Maximum Likelihood

```
> p=seq(0,1,by=.01)
> logL=function(p){sum(log(dbinom(X,size=1,prob=p)))}
> plot(p,Vectorize(logL)(p),type="l",col="red",lwd=2)
```




see Wald test, likelihood-ratio test and also score test,

# Testing Procedures Based on Maximum Likelihood

Numerically, we get the maximum of $\log L$ using

```
> neglogL=function(p){-sum(log(dbinom(X,size=1,prob=p)))}
> pml=optim(fn=neglogL,par=p0,method="BFGS")
> pml
$par
[1] 0.5499996

$value
[1] 13.76278
```

i.e. we obtain (numerically) $\widehat{\pi} = \overline{y}$.

# Testing Procedures Based on Maximum Likelihood

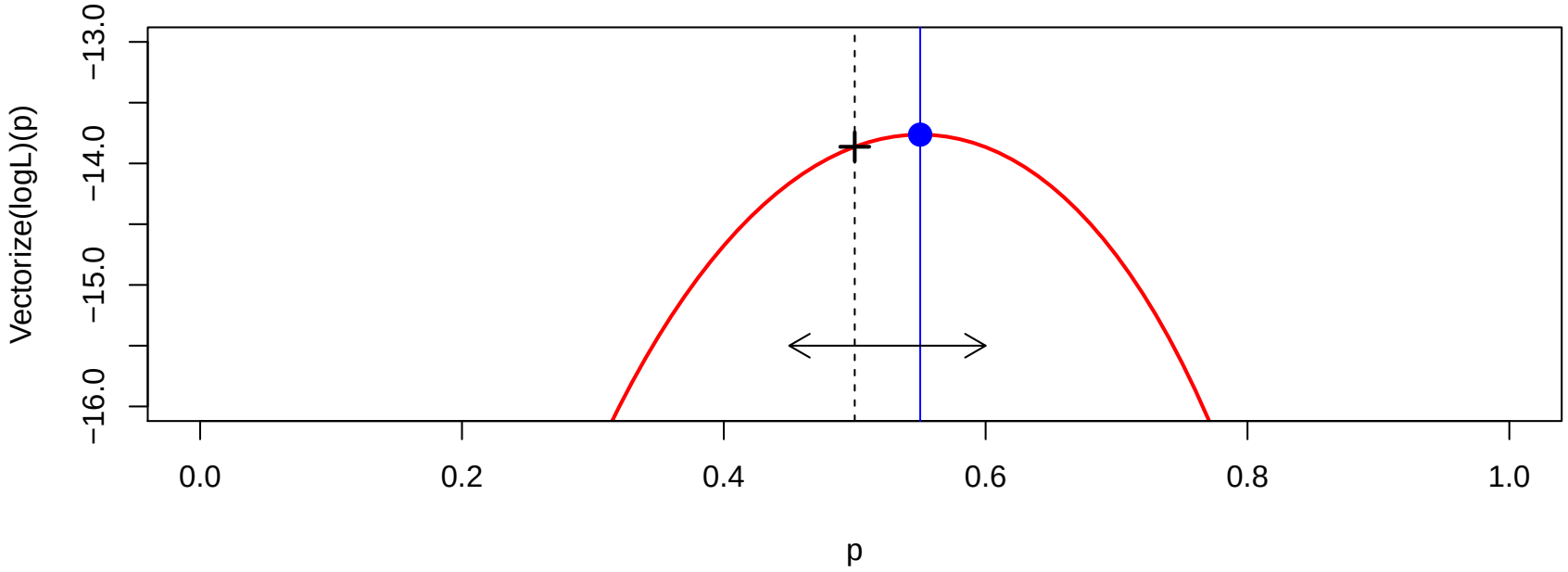
Let us test $H_0 : \pi = \pi_\star = 50\%$ against $H_1 : \pi \neq 50\%$. For Wald test, we need to compute $nI(\theta_\star)$, i.e.

```
> nx=sum(X==1)
> f = expression(nx*log(p)+(n-nx)*log(1-p))
> Df = D(f, "p")
> Df2 = D(Df, "p")
> p=p0=0.5
> (IF=-eval(Df2))
[1] 80
```

# Testing Procedures Based on Maximum Likelihood

Here we can compare it with the theoretical value, since we can derive it $I(\pi)^{-1} = \pi(1-\pi)$

```
> 1/(p0*(1-p0)/n)
[1] 80
```



# Testing Procedures Based on Maximum Likelihood

Wald statistics is here

```
> pml=optim(fn=neglogL,par=p0,method="BFGS")$par
> (T=(pml-p0)^2*IF)
[1] 0.199997
```

that should be compared with a $\chi^2$ quantile,

```
> T<qchisq(1-alpha,df=1)
[1] TRUE
```

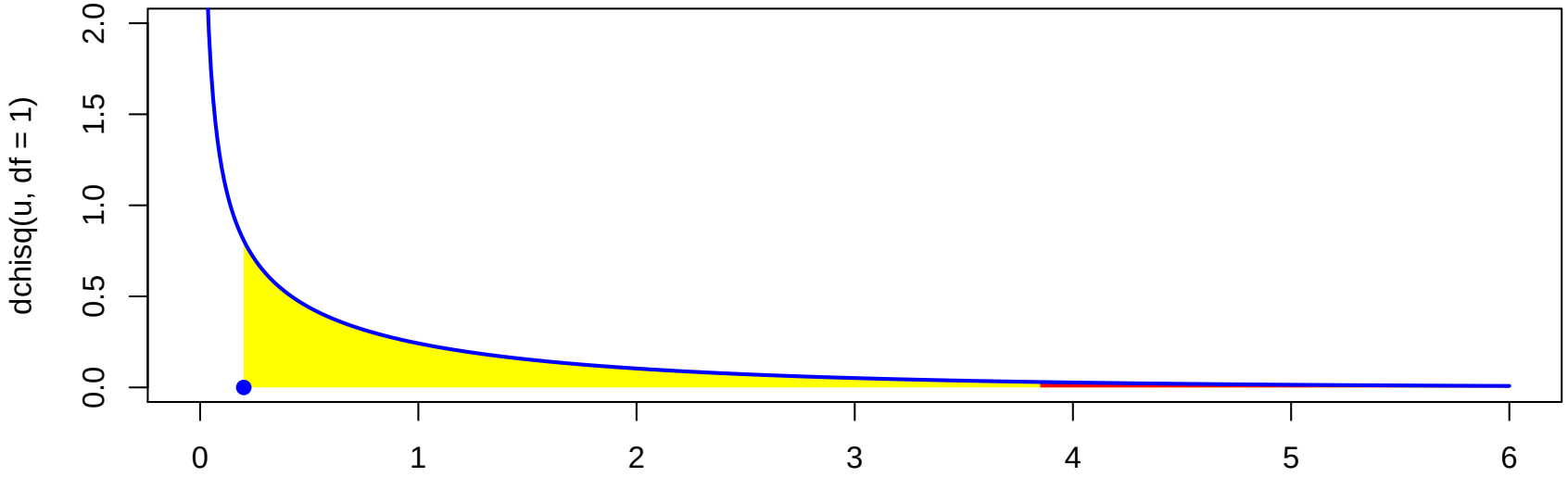
i.e. we are in the acceptance region.

# Testing Procedures Based on Maximum Likelihood

One can also compute the $p$-value of the test

```
> 1-pchisq(T,df=1)
[1] 0.6547233
```

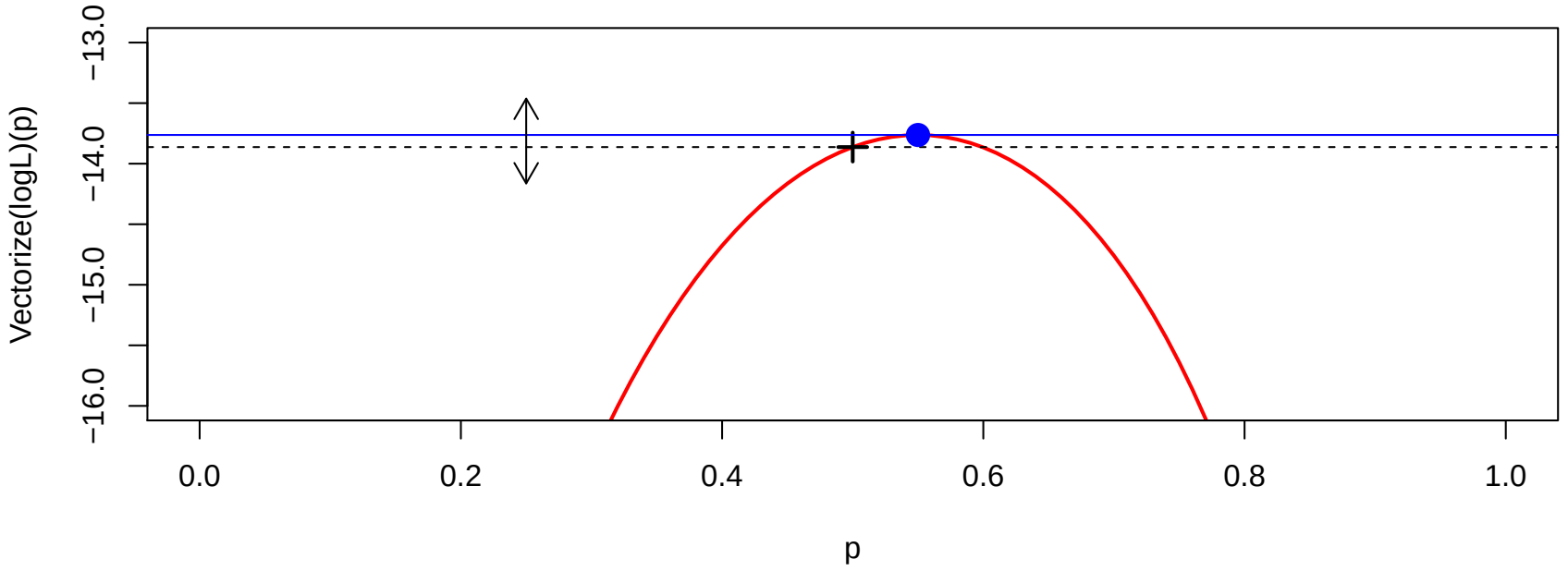
i.e. we should not reject $H_0$.



# Testing Procedures Based on Maximum Likelihood

For the likelihood ratio test, $T$ is here

```
> (T=2*(logL(pml)-logL(p0)))
[1] 0.2003347
```




Why is there a 2 in the likelihood ratio test ?

# Testing Procedures Based on Maximum Likelihood

Again, we are in the acceptance region

```
> T<qchisq(1-alpha,df=1)
[1] TRUE
```

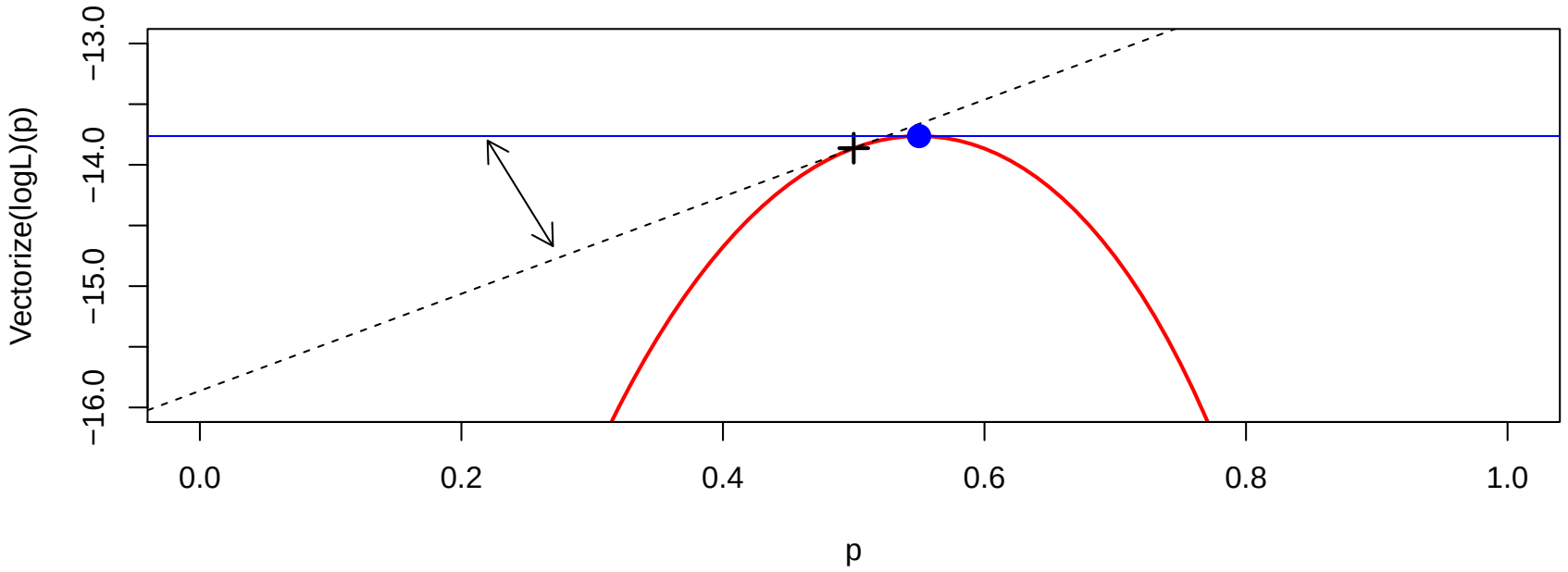
Last be not least, the score test

```
> nx=sum(X==1)
> f = expression(nx*log(p)+(n-nx)*log(1-p))
> Df = D(f, "p")
> p=p0
> score=eval(Df)
```

Here the statistics is

```
> (T=score^2/IF)
[1] 0.2
```

# Testing Procedures Based on Maximum Likelihood




which is also in the acceptance region

```
> T<qchisq(1-alpha,df=1)
[1] TRUE
```

# Profile Likelihood

Consider some model with a multivariate parameter $\boldsymbol{\theta}$, so that (classical) maximum likelihood yields

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \{\log \mathcal{L}(\boldsymbol{\theta})\}$$

Assume that $\boldsymbol{\theta} = (\alpha, \beta)$ where $\alpha$ is the parameter of interest. Set

$$\widehat{\alpha} = \underset{\alpha}{\text{argmax}} \left\{ \max_{\beta} \{\log \mathcal{L}(\alpha, \beta)\} \right\} = \underset{\alpha}{\text{argmax}} \{\log \mathcal{L}_p(\alpha)\}$$

where $\mathcal{L}_p$ is the profile-likelihood of $\alpha$.

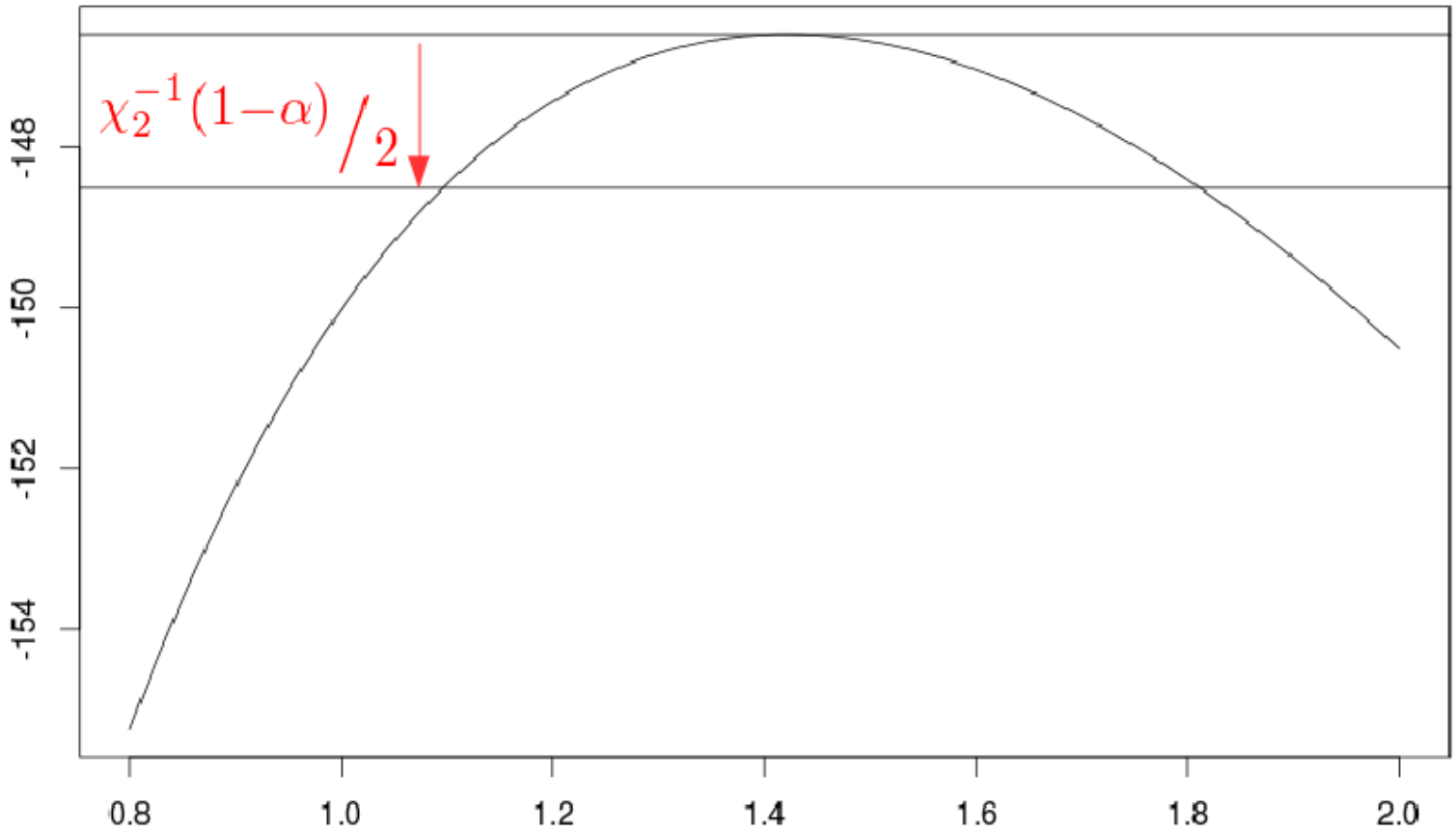
From the likelihood ratio test, one gets

$$2[\log \mathcal{L}(\widehat{\boldsymbol{\theta}}) - \log \mathcal{L}(\boldsymbol{\theta})] \sim [\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}]^{\mathsf{T}} I(\boldsymbol{\theta})[\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \sim \chi^2(k)$$

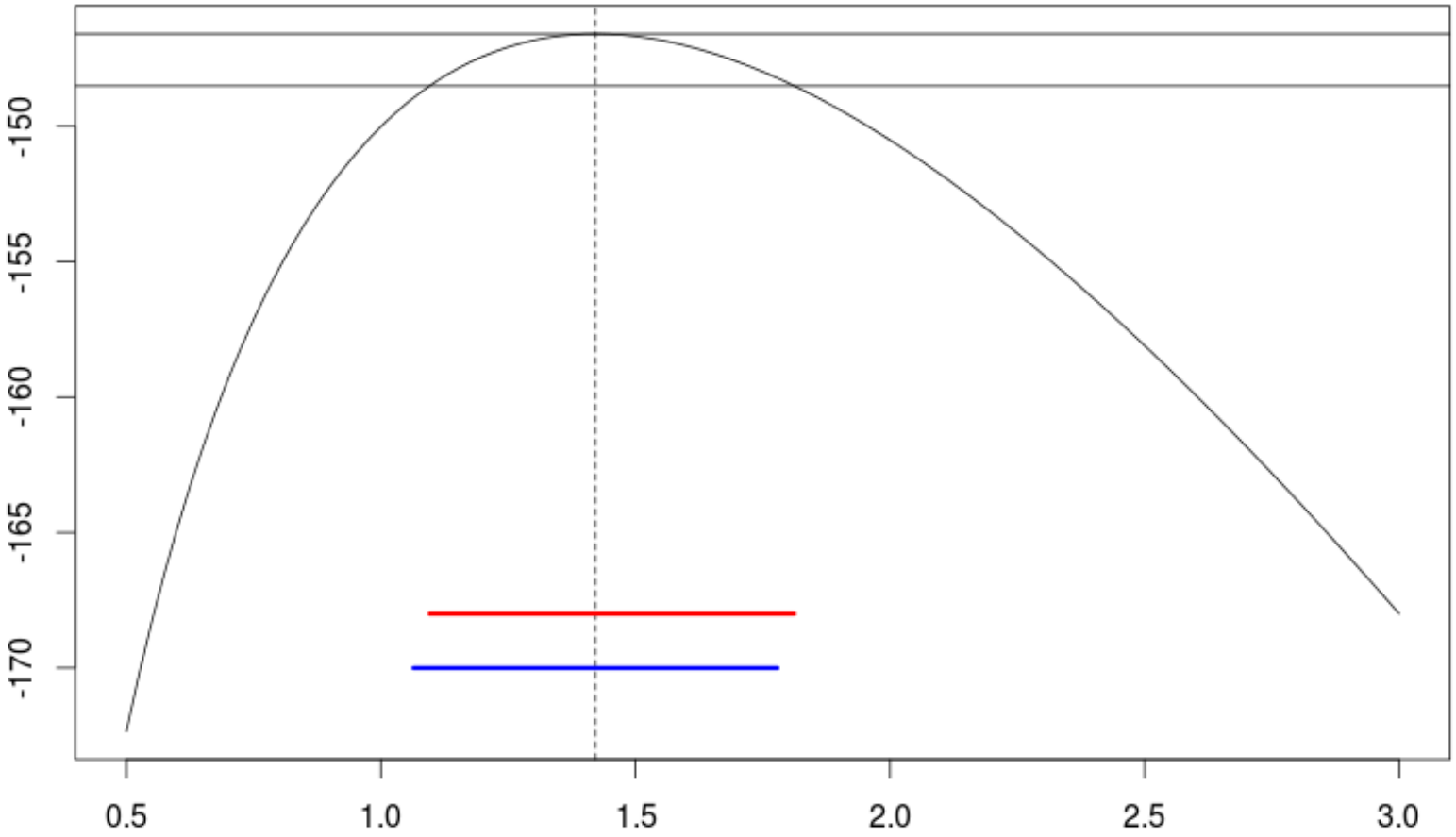
and here, one can write $2[\log \mathcal{L}_p(\widehat{\alpha}) - \log \mathcal{L}_p(\alpha)] \sim \chi^2(1)$.

Consider the case where $\boldsymbol{\theta} = (\alpha, \beta)$ denote the paramter of a Gamma distribution

# Profile Likelihood



# Profile Likelihood



# Multinomial Distribution and Factors / Categorical Variable

Consider variable $X$ taking values in $\mathcal{X} = \{\chi_1, \cdots, \chi_k\}$, with probabilities $\boldsymbol{p} = (p_1, \cdots, p_k)$. Here $p_j > 0$ and $p_1 + \cdots, p_k = 1$, see multinomial distribution. Consider now a sample $\{x_1, \cdots, x_n\}$. Let $\boldsymbol{n} = (n_1, \cdots, n_k)$ denote counts in each category. Then

$$\mathbb{P}[\boldsymbol{N} = \boldsymbol{n}] = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

The maximumum likelihood estimator of parameter $\boldsymbol{p}$ is

$$\widehat{\boldsymbol{p}} = \text{argmax}\{\log \mathcal{L}(\boldsymbol{p})\} \text{ subject to } \boldsymbol{p}^\mathsf{T}\mathbf{1} = 1.$$

The solution $\widehat{\boldsymbol{p}}$ is such that $\widehat{p}_j n_j/n$. One can prove that

$$\mathbb{E}[\widehat{\boldsymbol{p}}] = \boldsymbol{p} \text{ and } \text{Var}[\widehat{\boldsymbol{p}}] = \Sigma = [\Sigma_{i,j}]$$

where

$$\Sigma_{i,i} = \frac{p_i(1-p_i)}{n} \text{ while } \Sigma_{i,j} = -\frac{p_i p_j}{n}.$$

# Method of Moments

The method of moments is probably the most simple and intuitive technique to derive an estimator of $\theta$. If $\mathbb{E}(X) = g(\theta)$, we should consider $\widehat{\theta}$ such that $\overline{x} = g(\widehat{\theta})$, i.e.

$$\mathbb{E}(X) = g(\theta) \ \leftrightarrow \widehat{\theta} = g^{-1}(\overline{x}).$$

For an exponential distribution $\mathcal{E}(\theta)$, $\mathbb{P}(X \leq x) = 1 - e^{-\theta x}$, $\mathbb{E}(X) = 1/\theta$, and $\widehat{\theta} = 1/\overline{x}$.

For a uniform distribution on $[0, \theta]$, $\mathbb{E}(X) = \theta/2$, so $\widehat{\theta} = 2\overline{x}$.

If $\boldsymbol{\theta} \in \mathbb{R}^2$, we should use two moments, i.e. either $\text{Var}(X)$ or $\mathbb{E}(X^2)$.

# Comparing Estimators

Standard propoerties of statistical estimators are

- unbiasedness, $\mathbb{E}(\widehat{\theta}_n) = \theta$,
- convergence, $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$, as $n \to \infty$
- asymptotic normality, $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ as $n \to \infty$,
- efficiency
- optimality

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ denote two unbiased estimators, $\widehat{\theta}_1$ is said to be more efficient than $\widehat{\theta}_2$ if its variance is smaller.

see unbiased estimator, efficiency, and asymptotic properties.

# Comparing Estimators

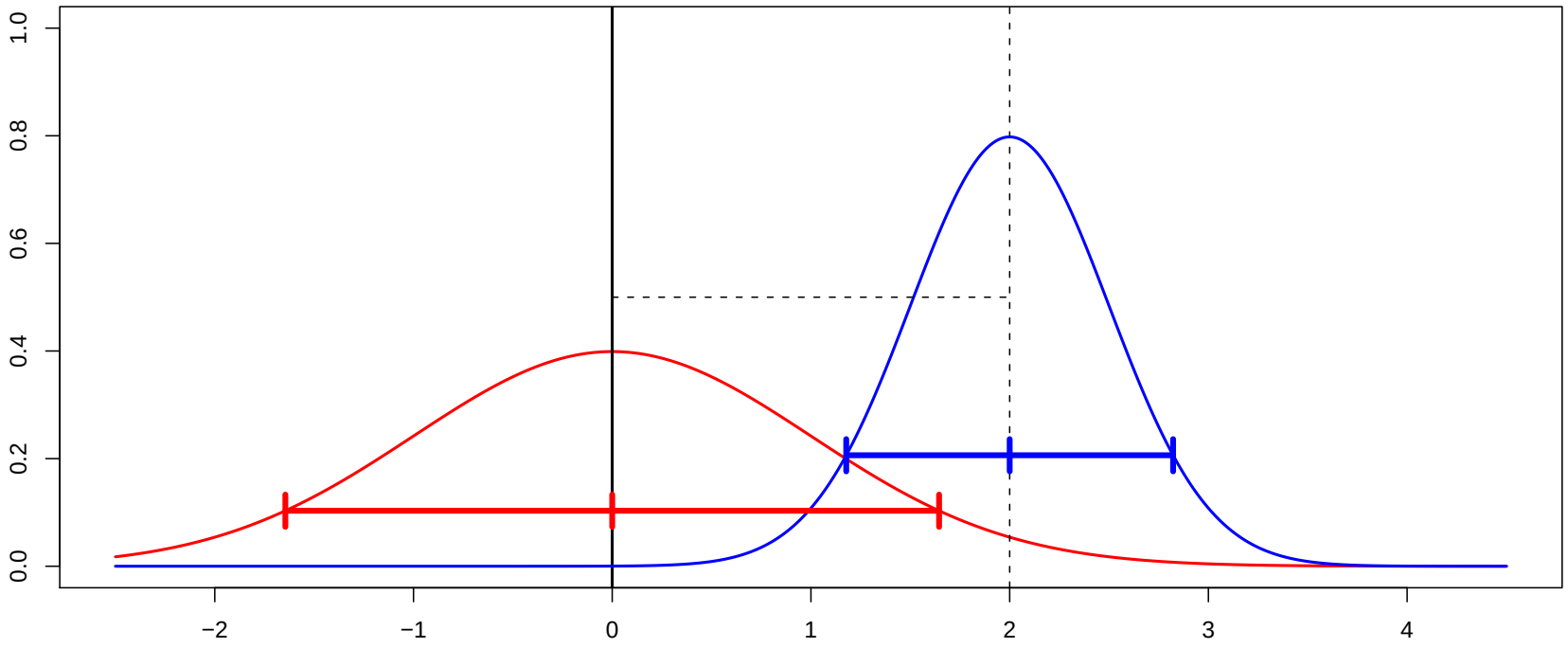



Figure 37: Chosing an estimator, $\widehat{\theta}_1$ versus $\widehat{\theta}_2$.

## Comparing Estimators

- $\widehat{\theta}_1$ is a biased estimator of $\theta$ ($\mathbb{E}(\widehat{\theta}_1) \neq \mathbb{E}(\theta)$),
- $\widehat{\theta}_2$ is an unbiased estimator of $\theta$ ($\mathbb{E}(\widehat{\theta}_2) = \mathbb{E}(\theta)$),
- $\text{Var}(\widehat{\theta}_1) \leq \text{Var}(\widehat{\theta}_2)$.

Estimator $\widehat{\theta}_1$ can be interesting if it bias can be estimated, but usually

- bias is a function of $\theta$ (which is unkown),
- bias is a complicated function of $\theta$.

# A short introduction to non-parametric statistics

For $x \in \mathbb{R}$, the cdf is defined as $F(x) = \mathbb{P}[X \leq x]$, so, given a sample $\{x_1, x_2, \cdots, x_n\}$, a natural estimator for $F(x)$ is

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x),$$

called the empirical cumulative distribution function, see e.c.d.f

Consider now a probabilist sample, $\{X_1, X_2, \cdots, X_n\}$, so that

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{1}(X_i \leq x)}_{Y_i},$$

is now a random variable. $Y_i$'s are i.i.d. random variables, with a Bernoulli distribution, $Y_i \sim \mathcal{B}(p)$, with $p = F(x)$.

# A short introduction to non-parametric statistics

Thus, $n\widehat{F}(x) \sim \mathcal{B}(n, F(x))$, and therefore

$$\mathbb{E}[\widehat{F}(x)] = F(x) \text{ and } \text{Var}(\widehat{F}(x)) = \frac{F(x) \cdot [1 - F(x)]}{n}.$$

By the strong law of large numbers, the estimator $\hat{F}_n(t)$ converges towards $F(t)$ almost surely, for every $t$

$$\widehat{F}_n(t) \xrightarrow{a.s.} F(t)$$

thus the estimator $\widehat{F}_n(t)$ is consistent.

This is a pointwise convergence of the empirical distribution function

There is a stronger result, called the Glivenko-Cantelli theorem, which states that the convergence in fact happens uniformly over $t$

$$\|\widehat{F}_n - F\|_\infty \equiv \sup_{t\in\mathbb{R}} \left|\widehat{F}_n(t) - F(t)\right| \xrightarrow{a.s.} 0.$$

## A short introduction to non-parametric statistics

The left part is called the Kolmogorov-Smirnov statistic, used for testing the goodness-of-fit between the empirical distribution $\widehat{F}_n$ and the assumed true cumulative distribution function $F$.

One can also prove that

$$\sqrt{n}\big(\widehat{F}_n(t) - F(t)\big) \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}\Big(0, F(t)\big(1 - F(t)\big)\Big).$$

(which can be used to get a pointwise confidence interval).

See Kolmogorov-Sminrnov test

# A short introduction to non-parametric statistics

We've seen how to estimate (non-parametrically) $F$. What about the density?

Its kernel density estimator is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where $K$ is the kernel — a non-negative function that integrates to one

$h > 0$ is a smoothing parameter called the bandwidth

A rule-of-thumb bandwidth estimator $h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}$,

See kernel density estimation

# Goodness of Fit Test (Discrete)

Consider $k$ possible values, $\{1, \cdots, k\}$ for convenience. Let $n_j$ the number of observations that took value $j$. We've seen Pearson's Chi-square test, that can be used here.

The test statistics of $H_0 : \boldsymbol{p} = \boldsymbol{p}^\star$ is $Q = \sum_{j=1}^{k} \frac{(n_j - np_j^\star)^2}{np_j} \sim \chi^2(k-1)$.

If is possible to use that test for a Poisson distribution. Consider a sample $\{y_1, \cdots, y_n\}$ and classes $\{0, 1, \cdots, k^+\}$ where $k^+$ means $k$ or more. Set

$$n_j = \sum_{i=1}^{n} \mathbf{1}(y_i = j) \text{ and } n_{k^+} = \sum_{i=1}^{n} \mathbf{1}(y_i \geq k)$$

In that case $p_j = \mathbb{P}[Y = j]$ when $Y \sim \mathcal{P}(\lambda)$ and $p_{k^+} = \mathbb{P}[Y \geq k]$, and the test stastitics is

$$Q = \sum_{j=0}^{k^+} \frac{(n_j - np_j^\star)^2}{np_j} \sim \chi^2(k^+)$$

In a general context, recall that

$$\widetilde{X}_j = \frac{N_j - np_j}{\sqrt{np_j(1-p_j)}} \to \mathcal{N}(0,1)$$

and more generally, $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \cdots, \widetilde{X}_k) \to \mathcal{N}(0, \star)$... but $\star \neq \mathbb{I}$.

Set

$$X_j = \frac{N_j - np_j}{\sqrt{np_j}} \to \mathcal{N}(0, \star)$$

and observe also that $\boldsymbol{X} = (X_1, \cdots, X_k) \to \mathcal{N}(0, \star)$. And one can prove that

$$\sum_{j=1}^{k} X_j^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \to \chi^2(k-1).$$

The appropriate test-statistic is

$$Q = \sum_{j=1}^{k} X_j^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \text{ and not } \sum_{j=1}^{k} X_j^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j(1-p_j)}$$

# Goodness of Fit Test (Continuous)

Recall that the empirical distribution function $\widehat{F}_n$ for $n$ i.i.d. observations $x_i$ is

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x])$$

The Kolmogorov-Smirnov statistic for a given cdf $F$ is

$$D_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)|$$

One can also use it to test if two samples have the same distribution

$$D_{n,m} = \sup_{x} |\widehat{F}_{1,n}(x) - \widehat{F}_{2,m}(x)|,$$

where $\widehat{F}_{1,n}$ and $\widehat{F}_{2,m}$ are the empirical distribution functions of the first and the second sample respectively.

# Goodness of Fit Test (Continuous)

The null hypothesis (that $F_1 = F_2$) is rejected at level $\alpha$ if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}.$$

Where the value of $c(\alpha)$ is given in the Kolmogorov-Smirnov table

| $\alpha$ | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
|---|---|---|---|---|---|---|
| $c(\alpha)$ | 1.22 | 1.36 | 1.48 | 1.63 | 1.73 | 1.95 |

and more generally, use

$$c(\alpha) = \sqrt{-\frac{1}{2}\log\left(\frac{\alpha}{2}\right)}.$$

# Appendix : Results and Notations on Matrices

Products : $c_{ij} = \boldsymbol{a}_{i\cdot}^{\mathsf{T}}\boldsymbol{b}_{\cdot j} = a_{i1}b_{1j} + \cdots + a_{im}b_{mj} = \sum_{k=1}^{m} a_{ik}b_{kj}$,

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{np} \end{pmatrix}$$

A sum can be expressed as a product of two vectors : $\sum_{i=1}^{n} a_i b_i = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{b}$

Example : $\sum_{i=1}^{n} a_i = \boldsymbol{a}^{\mathsf{T}}\mathbf{1} = \mathbf{1}^{\mathsf{T}}\boldsymbol{a}$ or $\sum_{i=1}^{n}(x_i - \overline{x})^2 = (\boldsymbol{x} - \overline{x})^{\mathsf{T}}(\boldsymbol{x} - \overline{x})$.

# Appendix : Results and Notations on Matrices

A $n \times n$ real matrix $A$ is called invertible if there exists an $n \times n$ square matrix $B$ such that $AB = BA = \mathbb{I}$. Then $B$ is its inverse, $B = A^{-1}$.

a symmetric $n \times n$ real matrix $M$ is said to be positive definite if the scalar $\boldsymbol{z}^\mathsf{T} M \boldsymbol{z}$ is strictly positive for every non-zero column vector $\boldsymbol{z}$ of $n$ real numbers.

For any real invertible matrix $A$, the product $A^\mathsf{T} A$ is a positive definite matrix

# Appendix : Results and Notations on Matrices

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} = \nabla f \text{ where } y = f(\mathbf{x})$$

is the gradient of function $f : \mathbb{R}^n \to \mathbb{R}$, while

$$\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}^\intercal} = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_1} \\ \frac{\partial^2 y}{\partial x_1 \partial x_2} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_1 \partial x_n} & \frac{\partial^2 y}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{bmatrix} = H$$

is the Hessian matrix

# Appendix : Results and Notations on Matrices

$$\frac{\partial \mathbf{a}^\mathsf{T}\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\mathsf{T}\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\mathsf{T})\mathbf{x}$$

see matrix calculus for more formulas.

# Appendix : Results on Optimisation

Suppose that $f$ is continuous at $a$ and differentiable on some open interval containing $a$.

If there exists a positive number $r$ such that for every $x$ in $(a-r, a)$ we have $f'(x) \geq 0$, and for every $x$ in $(a, a+r)$ we have $f'(x) \leq 0$, then $f$ has a local maximum at $a$.

If there exists a positive number $r$ such that for every $x$ in $(a-r, a)$ we have $f'(x) \leq 0$, and for every $x$ in $(a, a+r)$ we have $f'(x) \geq 0$, then $f$ has a local minimum at $a$.

From this first order condition, we search solutions $x$ to $f'(x) = 0$.

In the multivariate contexte, we search solutions $\boldsymbol{x}$ to $\nabla f(\boldsymbol{x}) = \mathbf{0}$.

# Appendix : Results on Optimisation

More generally, see Karush–Kuhn–Tucker conditions : we want to solve

$$\min\{f(\boldsymbol{x})\} \text{ subject to } \begin{cases} g_i(\boldsymbol{x}) \leq 0, \forall i \\ h_j(\boldsymbol{x}) = 0, \forall j \end{cases}$$

$f$ is the objective function, $g_i$'s are the inequality constraint functions, and $h_j$'s are the equality constraint functions.

We have the following first order (necessary) condition.

# Appendix : Results on Optimisation

Suppose that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable at a point $\boldsymbol{x}^*$. If x$\boldsymbol{x}^*$ is a local minimum and the optimization problem satisfies some regularity conditions, then there exist constants $\mu_i$'s and $\lambda_j$'s called KKT (or Lagrange) multipliers, such that at the minimum, $\boldsymbol{x}^*$,

$$-\nabla f(\boldsymbol{x}^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{\ell} \lambda_j \nabla h_j(\boldsymbol{x}^*),$$

or $\nabla \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0$ where

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) + \sum_{j=1}^{\ell} \lambda_j h_j(\boldsymbol{x}),$$

see Boyd & Vandenberghe (2009) convex optimization