# Probability and Statistics

Y. Lebbah

Université Oran1, Algeria
Email: ylebbah@gmail.com
Web: https://github.com/ylebbah/teaching

November 10, 2024

# Motivations of the course

1. Perform statistics, from basics to advanced computations with R.
2. Work on practical datasets

# Example

We are interested in the effect of some medication A on infections by a given disease. 16 mice were infected with the disease and then randomly distributed among two groups. The first group of 8 mice received medication A, at 10 mg per kilo, 60 hours after infection. The other 8 mice did not receive treatment. After one week, the following numbers of adult worms were found in the intestines:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Treated mice | 51 | 55 | 62 | 45 | 68 | 71 | 46 | 79 |
| Non treated mice | 45 | 53 | 52 | 51 | 57 | 51 | 68 | 88 |

What can be said about the possible effectiveness of the medication A, dosed at 10mg/kg for the treatment of infections of mice with the disease?

# Statistics

Three distinct domains in statistics:

1. Data collection
2. Descriptive statistics quantitatively describing or summarizing features from a collection of data
3. Statistical inference: drawing conclusions about an underlying population based on a sample or subset of the dataset
4. Statistical inference/Predictive inference: predicting future observations based on past observations (current dataset)

Thus, two main domains:

- **Descriptive statistics** [1]
- **Statistical inference** [2]

NB : Statistics is based on models and hypotheses coming from probability theory.

[1] Statistique descriptive
[2] Inférence statistique

# Installation of R

RStudio IDE:
https://www.rstudio.com/products/rstudio/download/
R Project for Statistical Computing:
https://www.r-project.org/

# case studies

https://www.kaggle.com/
https://r-dir.com/reference/datasets.html
https://quickstats.nass.usda.gov/
https://cloud.csiss.gmu.edu/Crop-CASMA/
https://registry.opendata.aws/tag/agriculture/

# Type of data formats

Inputs: ASCII, CSV, Excel, ...

# Descriptive statistics motivations

It quantitatively describes or summarizes features from a collection of data.
Types of representations:

- Tables
- Graphics
- Numerical indices

# Vocabulary

- **Population**: a collection of all the items under consideration
- **Sample** [3]: sub-set of the population
- **Variable/Characteristic/Attribute**: A characteristic about each individual element of a population or sample.
- **Parameter**: A numerical value summarizing all the data of an entire population.
- **Statistic**: A numerical value summarizing the sample data.

[3] Echantillon

# Kinds of variables

- **Quantitative, or Numerical, Variable :**
  - **discrete**
  - **continuous**
- **Qualitative, or Attribute, or Categorical, Variable :**
  - **ordinal**: data can be ordered
    Example Exam scores A+, A+, B+, B, C.
  - **nominal**: cannot be ordered
    Example Colors red, blue, yellow, ...

# Standard statistical indicators

They are used on quantitative variables:

Positioning indicators (Central Tendency):

- Mode
- Median [4]
- Mean [5]
- Quantiles

Dispersion indicators:

- Range [6]
- Interquartile
- Variance
- Standard deviation [7]

[4] Médiane
[5] Moyenne
[6] Largeur
[7] Ecart type

# Positioning indicators: Mode

- For a discrete variable, mode is the value of the variable having the highest frequency.
  Example: The mode in the set of numbers $\{21, 21, 21, 23, 24, 26, 26, 28, 29, 30, 31, 33\}$ is 21
- For a continuous variable, mode is the center of the interval/class having the highest frequency.

# Positioning indicators: Mean

The mean of a sample is the average value on the available observations:

$$\overline{x} = \frac{1}{n} \sum_{i=1..n} x_i$$

# Positioning indicators: Median

Let $x_1, ..., x_n$ be a set of observations and $x(1) \leq ... \leq x(n)$ are the ordered version.
The median is the quantile of order $1/2$:

- If $n$ is even: $\tilde{x} = x_{\frac{n+1}{2}}$
- If $n$ is odd: mean of all values between $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$

Example Let be the ordered values [1, 3, 3, 3, 5, 5, 6, 7, 7, 8, 8, 8, 9, 9, 10, 10, 10, 10, 11, 11, 12, 13, 13, 13, 14, 15, 16, 19]. $n = 28$, $n/2 = 14$, we have two values 9, 10. The median is $(9 + 10)/2 = 9.5$.

Example Let be the ordered values [3, 5, 5, 6, 7, 8, 8, 9, 9, 10, 10, 10, 10, 11, 11, 12, 13, 13, 13, 14, 15, 16, 19]. $n = 23$, $(n + 1)/2 = 12$, we have one single value 10. The median is 10.

# Positioning indicators: Quantiles

$q$-quantiles are values that partition a finite set of values $L$ into $q$ subsets of (nearly) equal sizes. Let $L\uparrow$ be $L$ with its increasing order.

There are $q-1$ of the $q$-quantiles, one for each integer $k$ satisfying $0 < k < q$.

The $q$-quantiles are the application of the quantile function to the values $\{1/q, 2/q, ..., (q-1)/q\}$:

- $(1/q)$ order quantile $Q_1 = L\uparrow[\lceil n \times (1/q)\rceil]$,
- $(2/q)$ order quantile $Q_2 = L\uparrow[\lceil n \times (2/q)\rceil]$
- ...
- $((q-1)/q)$ order quantile $Q_{q-1} = L\uparrow[\lceil n \times ((q-1)/q)\rceil]$.

# Positioning indicators: Quantiles with R

$$Q_i(p) = (1-\gamma)x_j + \gamma x_{j+1}$$

where $0 \leq i \leq 9$, $\frac{j-m}{n} \leq p \leq \frac{j-m+1}{n}$, $x_j$ is the jth order statistic, $n$ is the sample size, the value of $\gamma$ is a function of $j = \lfloor np + m \rfloor$ and $g = np + m - j$, and $m$ is a constant determined by the sample quantile type.
Example: Let be $i = 4, p = 0.25, m = 0$. $j = \lfloor 0.25 * n \rfloor$, which leads to the first quartile $Q_4(0.25)$.

- (1) If $g = 0$ then $\gamma = 0$, else $\gamma = 1$.
- (2) Similar to type 1 but with averaging at discontinuities. If $g = 0$ then $\gamma = 0.5$, else $\gamma = 1$.
- ...

Ref: R-doc-quantiles.

# Positioning indicators: Quantiles/Example

Example $n/q$ even Let $L = [1, 3, 3, 3, 5, 5, 6, 7, 7, 8, 8, 8, 9, 9, 10, 10, 10, 10, 11, 11, 12, 13, 13, 13, 14, 15, 16, 19]$ be a list of values sorted in increasing order.
There are $n = 28$ values, divided by 4, we got the 3 quantiles
$Q_1 = L[\lceil 28/4 \rceil] = L[7] = 6$, the seventh value, ..., the third
$Q_3 = L \uparrow [3(n/4)] = L \uparrow [21] = 13$, the 21th value. Thus the indices of the 4-quantiles partitions are: [1..7], [8..14], [15..21], [22..28].

# Positioning indicators: Quantiles/Example

Example $n/q$ odd Let $L = [3, 5, 5, 6, 7, 8, 8, 9, 9, 10, 10, 10, 10, 11, 11, 12, 13, 13, 13, 14, 15, 16, 19]$ be a list of values sorted in increasing order.
There are $n = 23$ values, divided by 4, we got the first quartile $Q_1 = L \uparrow [\lceil 23/4 \rceil] = L \uparrow [\lceil 5.75 \rceil] = L \uparrow [6] = 8$, the third quartile $Q_3 = L \uparrow [\lceil 3(n/4) \rceil] = L \uparrow [\lceil 3(23/4) \rceil] = L \uparrow [18] = 13$.
Thus the indices of 4-quantiles partitions are: [1..6], [7..12], [13..18], [19..23].

# Positioning indicators: Quantiles/Interpretation

How to interpret quartiles? Suppose that $Q_1$, $Q_2$, $Q_3$ are known.
What can we deduce?
At least a quarter (25%) of the values are less than or equal to $Q_1$.
At least three quarters (75%) of the values are less than or equal to $Q_3$.
About half of the values are in the interquartile range $[Q_1, Q_3]$.

# Positioning indicators: Quantiles

Particular quantiles:

- The only 2-quantile is called the median
- The 3-quantiles called tertiles or terciles $\rightarrow T$
- The 4-quantiles called quartiles $\rightarrow Q$
- The 5-quantiles called quintiles $\rightarrow QU$
- The 6-quantiles called sextiles $\rightarrow S$
- The 7-quantiles called septiles
- The 8-quantiles called octiles
- The 10-quantiles called deciles $\rightarrow D$
- The 12-quantiles called duo-deciles or dodeciles
- The 16-quantiles called hexadeciles $\rightarrow H$
- The 20-quantiles called ventiles, vigintiles (demi-deciles) $\rightarrow V$
- The 100-quantiles called percentiles $\rightarrow P$
- The 1000-quantiles have been called permilles or milliles

# Dispersion indicators: Range

The range measures the gap between the lowest and greatest values. It is defined as follows:

$$e_n = max(X_i) - min(X_i)$$

# Dispersion indicators: Interquartile distance

- **interquartile interval (range)** : interval of values comprised in $[Q_1, Q_3]$.
- **interquartile distance** : $\triangle_q = Q_3 - Q_1$.

# Dispersion indicators: Variance and standard deviation

Sample variance is defined as follows: $s^2 = \frac{\sum_{i=1..n}(x_i - \overline{x})^2}{n}$.

$s = \sqrt{\frac{\sum_{i=1..n}(x_i - \overline{x})^2}{n}}$ is the average of quadratic deviation from the mean of some sample.

$s$ is the sample standard-deviation.

Contrarily to variance, standard deviation is expressed in the same physical unity as the variable $X$.

Example: Standard deviation of the average temperatures recorded over a five-day period last winter: 18, 22, 19, 25, 12 (mean = 19.2).

$s^2 = 94.8/5 = 18.96$.

$s = \sqrt{23.7} = 4.35$.

# Probability

The sample space associated with an experiment is the set consisting of all possible outcomes [8] and is called the probability space $\Omega$. An event $A$ is a subset of outcomes in $\Omega$, that is, $A \subset \Omega$. Probability is defined as the real-valued function $\mathbb{P} : \Omega \to [0, 1]$ such that:

- for all $A \in \Omega$, we have $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\Omega) = 1$
- If $A_1$, $A_2$, $A_3$, ... is an infinite collection of disjoint events [9], then $\mathbb{P}(\cup_{i=1..\infty} A_i) = \sum_{i=1..\infty} \mathbb{P}(A_i)$.

[8] Résultat
[9] Evènement

# Probability/Example

Example 1: Examining a single fuse to see whether it is defective $D$ or $N$. $\Omega = \{D, N\}$.
Example 2: Examining the status of three fuses.
$\Omega = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$. In this case, we can explore the event where at least two fuses are defective, $A = \{DDN, DND, NDD, DDD\}$.

# Probability/Example

A die (dé) is loaded (not all outcomes are equally likely) such that the probability that the number $i$ shows up is $K \times i$, $i = 1, 2, ..., 6$, where $K$ is a constant. Find the value of $K$? Probability that a number greater than 3 shows up?

From axioms, we get $\mathbb{P}(1) + \mathbb{P}(2) + ... + \mathbb{P}(6) = 1$.

$(K)(1) + (K)(2) + ... + (K)(6) = 1$ or $(K)(1 + 2 + ... + 6) = (K)(21) = 1$. Hence $K = \frac{1}{21}$.

Let $A$ be the event that a number greater than 3 shows up. Then the outcomes in A are $\{4, 5, 6\}$ and they are mutually exclusive. Therefore,

$$P(A) = P(4) + P(5) + P(6) = \frac{4}{21} + \frac{5}{21} + \frac{6}{21} = \frac{15}{21}.$$

# Probability/properties

For two events $A$ and $B$ in $\Omega$, we have the following:

- Let $A^c$ be the complement of $A$ in $\Omega$, $P(A^c) = 1 - P(A)$.
- If $A \subset B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

# Probability: conditional probability

The conditional probability of an event $A$, given that an event $B$ has occurred, denoted by $P(A|B)$, is equal to

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

# Probability: conditional probability/Example

We toss two balanced dice, and let $A$ be the event that the sum of the face values of two dice is 8, and $B$ be the event that the face value of the first one is 3. Calculate $\mathbb{P}(A|B)$?
The elements of the events $A$ and $B$ are
$A = \{(2,6),(6,2),(3,5),(5,3),(4,4)\}$.
and $B = \{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$.
Now $A \cap B = \{(3,5)\}$.
$P(A) = 5/36$, $P(B) = 6/36$, and $P(A \cap B) = 1/36$.
$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/36}{6/36} = 1/6$.

# Probability: conditional probability / Total probability

Assume $S = A_1 \cup A_2 \cup ... \cup A_n$, where $P(A_i) > 0$, $i = 1, 2, ..., n$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. Then for any event $B$

$$\mathbb{P}(B) = \sum_{i=1..n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)$$

# Probability: Bays' rule

Assume $\mathbb{P}(A) \neq 0$ and $\mathbb{P}(B) \neq 0$ then

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

Assume $S = A_1 \cup A_2 \cup ... \cup A_n$, where $P(A_i) > 0$, $i = 1, 2, ..., n$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. Then for any event $B$, with $P(B) > 0$

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\sum_{i=1..n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}$$

# Probability: conditional probability / Bays' rule / Example

Suppose a statistics class contains 70% male and 30% female students. It is known that in a test, 5% of males and 10% of females got an "A" grade. If one student from this class is randomly selected and observed to have an "A" grade, what is the probability that this is a male student?

Let $A_1$ denote that the selected student is a male, and $A_2$ denote that the selected student is a female. Here the sample space $S = A_1 \cup A_2$. Let $D$ denote that the selected student has an "A" grade. We are given $\mathbb{P}(A_1) = 0.7$, $\mathbb{P}(A_2) = 0.3$, $\mathbb{P}(D|A_1) = 0.05$, and $\mathbb{P}(D|A_2) = 0.10$. Then by the total probability rule,

$$\mathbb{P}(D) = \mathbb{P}(A_1)\mathbb{P}(D|A_1) + \mathbb{P}(A_2)\mathbb{P}(D|A_2) = 0.035 + 0.030 = 0.065.$$

$$\mathbb{P}(A_1|D) = \frac{\mathbb{P}(A_1)\mathbb{P}(D|A_1)}{\mathbb{P}(A_1)\mathbb{P}(D|A_1)+\mathbb{P}(A_2)\mathbb{P}(D|A_2)} = 7/13 = 0.538.$$

# Random variables

An experiment may contain numerous characteristics that can be measured. However, in most cases, an experimenter will focus on some specific characteristics of the experiment. The concept of a random variable allows us to pass from the experimental outcomes to a numerical function of the outcomes, often simplifying the sample space.

- A random variable [10] (r.v.) $X$ is a function defined on a sample space, $\Omega$, that associates a real number $X(w) = x$, with each outcome $\omega \in \Omega$

$$X : \begin{array}{l} \Omega \to \mathbb{R} \\ \omega \mapsto x \end{array}$$

[10] Variable aléatoire

# Random variables

- Domain of variation of $X$ is the set $D_x \subseteq \mathbb{R}$ of values that can be taken by $X$.
- A discrete random variable can take a finite number of distinct values.
- A continuous random variable takes infinite (continuous) number of distinct values.

# Random variables/Example

Two balanced coins are tossed and face values are noted. Then the sample space $S = \{HH, HT, TH, TT\}$. Define the random variable $X(\omega) = n$, where $n$ is the number of heads and $\omega$ represents a simple event such as $HH$. Then

$$X(\omega) = \begin{cases} 0, & if\ \omega = (TT) \\ 1, & if\ \omega \in \{HT, TH\} \\ 2, & if\ \omega = (HH). \end{cases}$$

# Random variables/Example

In the tossing of three fair coins, let the random variable $X$ be defined as $X =$ number of tails. Then $X$ can assume values 0, 1, 2, and 3. We can associate these values with probabilities in the following way:

- $P(X = 0) = P(\{H, H, H\}) = 1/8.$
- $P(X = 1) = P(\{H, H, T\} \cup \{H, T, H\} \cup \{T, H, H\}) = 3/8.$
- $P(X = 2) = P(\{T, T, H\} \cup \{T, H, T\} \cup \{H, T, T\}) = 3/8.$
- $P(X = 3) = P(\{T, T, T\}) = 1/8.$

# Discrete probability mass function

- The discrete probability mass function (pmf) [11] of a discrete random variable $X$ is the function

$$p(x_i) = \mathbb{P}(X = x_i), i = 1, 2, 3, ...$$

- The cumulative distribution function (cdf) [12] $F$ of the discrete random variable $X$ is defined by

$$F(x) = \mathbb{P}(X \leq x) = \sum_{y \leq x} p(y)$$

A cumulative distribution function is also called a probability distribution function (PDF) or simply the distribution function.

[11] Fonction de densité de probabilité
[12] Fonction de répartition

# Discrete probability mass function/Example

Suppose that a fair coin is tossed twice so that the sample space is $S = \{HH, HT, TH, TT\}$. Let $X$ be number of heads. Find the probability function for $X$? Find the cumulative distribution function of $X$?

We have $1/4 = P(\{HH\}) = P(\{HT\}) = P(\{TH\}) = P(\{TT\})$.

Hence, the pmf is given by

$p(0) = P(X = 0) = 1/4, p(1) = 1/2, p(2) = 1/4$.

$F(1.5) = \mathbb{P}(X \leq 1.5) = \mathbb{P}(X = 0 \text{ or } 1) = P(X = 0) + P(X = 1) = 1/4 + 1/2 = 3/4$.

$$F(x) = \begin{cases} 0, & \text{if } -\infty < x < 0 \\ 1/4, & \text{if } 0 \leq x < 1 \\ 3/4, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } 2 \leq x < \infty \end{cases}$$

# Probability mass function

Let $X$ be a random variable. Suppose that there exists a non-negative real-valued function $f : \boldsymbol{R} \to [0, \infty)$ such that for any interval $[a, b]$

$$P(X \in [a, b]) = \int_a^b f(t)dt$$

Then $X$ is called a continuous random variable. The function $f$ is called the probability density function (pdf) of $X$. The cumulative distribution function (cdf) is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

For a given function $f$ to be a pdf, it needs to satisfy the following two conditions: $f(x) \geq 0$ for all values of $x$, and $\int_{-\infty}^{+\infty} f(t)dt = 1$.

# Probability mass function/Example

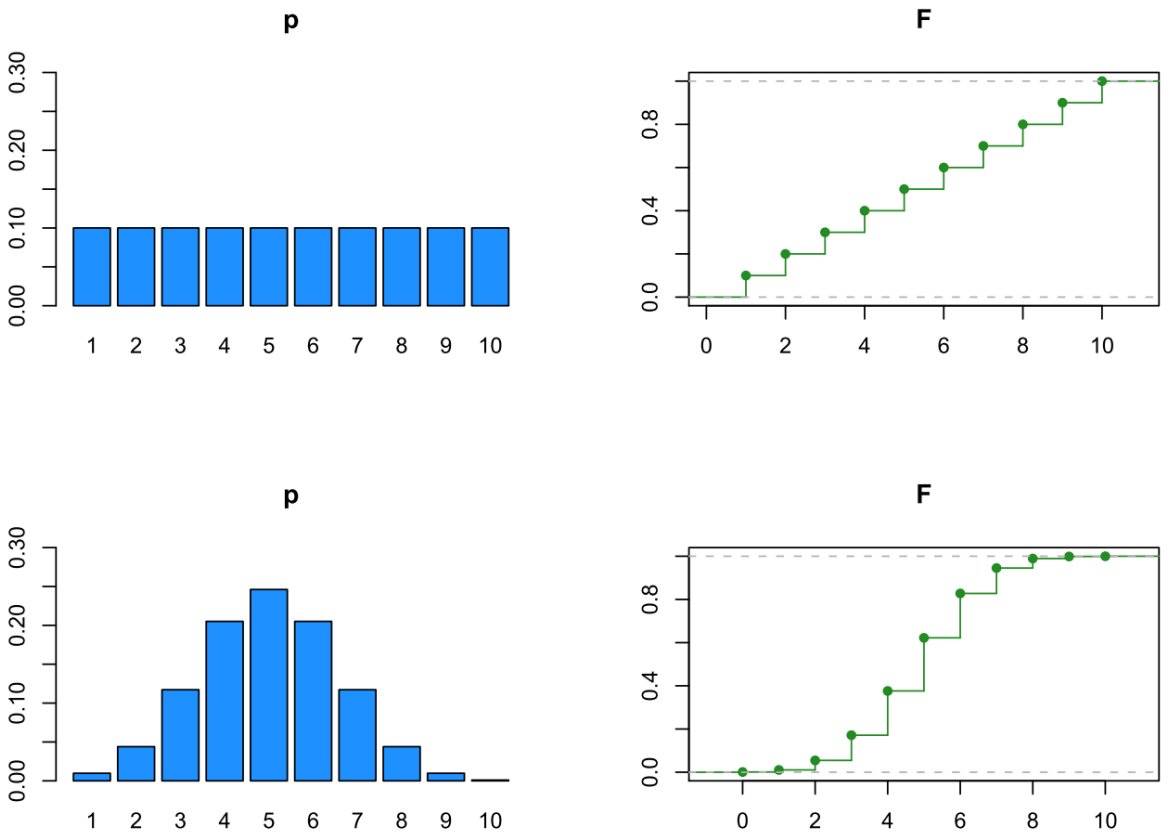



Figure: Discrete probability mass function/Example

# Probability mass function/Example

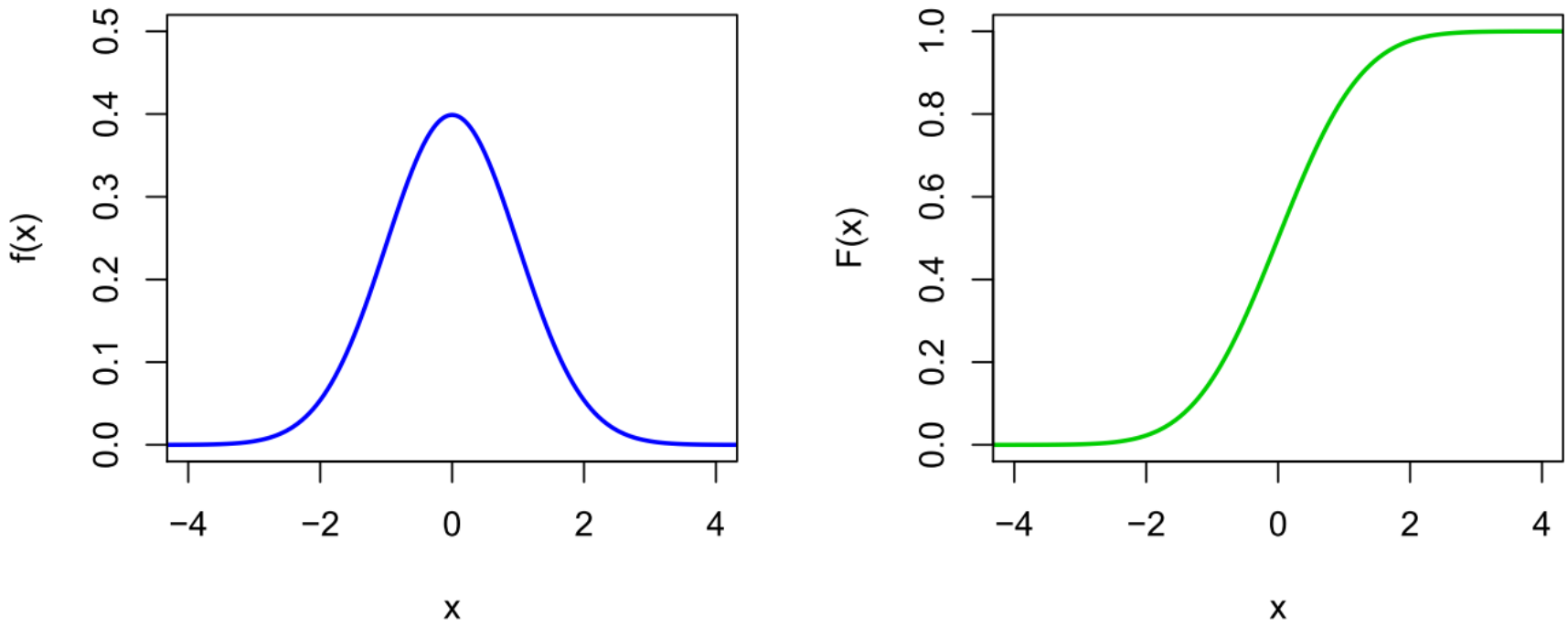



Figure: Continuous probability mass function/Example

# Probability mass function/Example

Let the function

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

For what value of $\lambda$ is $f$ a pdf?
First note that $f(x) \geq 0$. Now, for $f(x)$ to be a pdf, we need $\int_{-\infty}^{+\infty} f(x)dx = 1$.
Second, $1 = \int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} \lambda x e^{-x} dx == \lambda \int_{-\infty}^{+\infty} x e^{-x} dx =$ $\lambda[(-xe^{-x})|_0^{\infty} + \int_0^{+\infty} e^{-x} dx]$ (integration by parts).
$1 = \lambda[0 - e^{-x}|_0^{\infty} = \lambda$.

# Expected value, variance and standard deviation functions

- Let $X$ be a discrete random variable with pmf $p(x)$. Then the expected value of $X$, denoted by $E(X)$, is

$$\mu = E(X) = \sum_{all\ x} xp(x).$$

- The expected value [13] of a continuous random variable $X$ with pdf $f(x)$ is defined by

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

- The variance of a random variable $X$ is defined by

$$\sigma^2 = Var(X) = E[(X - \mu)^2]$$

$\sigma$ is called the standard deviation [14].

[13] Espérance mathématique
[14] Ecart type

# Expected value, variance and standard deviation functions / Examples

Let $X$ be a discrete random variable whose probability density function is given in the following table:

| X | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| P(X) | 1/7 | 1/7 | 1/14 | 2/7 | 1/14 | 1/7 | 1/7 |

$$E(X) = \sum xP(x) = -1 \times 1/7 + 0 \times 1/7 + 1 \times 1/14 + 2 \times 2/7 +$$

$$3 \times 1/14 + 4 \times 1/7 + 5 \times 1/7 = 2.$$

# Expected value, variance and standard deviation functions / Examples

Let $Y$ be a random variable with pdf:

$$f(y) = \begin{cases} 3/64(y^2(4-y)), & 0 \leq y \leq 4 \\ 0, \textit{elsewhere} \end{cases}$$

$$\begin{aligned} E(Y) &= \int_{-\infty}^{+\infty} yf(y)dy \\ &= 3/64 \int_0^4 yf(y)dy \\ &= 2.4 \end{aligned}$$

# Mode

The mode of a discrete variable $X$ is the value $x$, which is most likely to occur, with probability, $p(x)$.
The mode of a continuous variable $X$ is the value, $x$, at which the probability density function, $f(x)$, is at a maximum

# Mode

The mode of a discrete variable $X$ is the value $x$, which is most likely to occur, with probability, $p(x)$.
The mode of a continuous variable $X$ is the value, $x$, at which the probability density function, $f(x)$, is at a maximum
Example 1: rolling two dices simultaneously and getting the sum, the following probabilities are obtained:

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p(x) | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{2}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

# Mode

The mode of a discrete variable $X$ is the value $x$, which is most likely to occur, with probability, $p(x)$.
The mode of a continuous variable $X$ is the value, $x$, at which the probability density function, $f(x)$, is at a maximum
Example 1: rolling two dices simultaneously and getting the sum, the following probabilities are obtained:

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p(x) | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{2}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

The most likely value is 7 with its highest probability of 6/36, so the mode is 7.

# Mode

The mode of a discrete variable $X$ is the value $x$, which is most likely to occur, with probability, $p(x)$.
The mode of a continuous variable $X$ is the value, $x$, at which the probability density function, $f(x)$, is at a maximum
Example 1: rolling two dices simultaneously and getting the sum, the following probabilities are obtained:

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p(x) | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{2}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

The most likely value is 7 with its highest probability of 6/36, so the mode is 7.

$$\text{Example 2: } f(x) = \begin{cases} -x^2 + 2x - \frac{1}{6}, & 0 < x < 2 \\ 0, & \textit{otherwise} \end{cases}$$

# Mode

The mode of a discrete variable $X$ is the value $x$, which is most likely to occur, with probability, $p(x)$.
The mode of a continuous variable $X$ is the value, $x$, at which the probability density function, $f(x)$, is at a maximum
Example 1: rolling two dices simultaneously and getting the sum, the following probabilities are obtained:

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p(x) | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{2}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

The most likely value is 7 with its highest probability of 6/36, so the mode is 7.

$$\text{Example 2: } f(x) = \begin{cases} -x^2 + 2x - \frac{1}{6}, & 0 < x < 2 \\ 0, & \textit{otherwise} \end{cases}$$

$$f'(x) = -2x + 2 = 0 \Rightarrow x = 1$$

# Median

The Median of a random variable $X$ is the value $q_p$ such that:
$\mathbb{P}(X \leq q_p) = 0.5, \Leftrightarrow q_p = F^{-1}(0.5)$

# Median

The Median of a random variable $X$ is the value $q_p$ such that:
$\mathbb{P}(X \leq q_p) = 0.5, \Leftrightarrow q_p = F^{-1}(0.5)$
Example 1: For $x = 1, 4, p(x) = 0.2$ and $x = 2, 3, p(x) = 0.3$.

# Median

The Median of a random variable $X$ is the value $q_p$ such that:
$\mathbb{P}(X \leq q_p) = 0.5, \Leftrightarrow q_p = F^{-1}(0.5)$
Example 1: For $x = 1, 4, p(x) = 0.2$ and $x = 2, 3, p(x) = 0.3$.
$P(X \leq 2) = P(X = 1) + P(X = 2) = 0.2 + 0.3 = 0.5$

# Median

The Median of a random variable $X$ is the value $q_p$ such that:
$\mathbb{P}(X \leq q_p) = 0.5, \Leftrightarrow q_p = F^{-1}(0.5)$
Example 1: For $x = 1, 4, p(x) = 0.2$ and $x = 2, 3, p(x) = 0.3$.
$P(X \leq 2) = P(X = 1) + P(X = 2) = 0.2 + 0.3 = 0.5$
Example 2: $f(x) = \begin{cases} -x^2 + 2x - \frac{1}{6}, & 0 < x < 2 \\ 0, & \text{Otherwise} \end{cases}$

# Median

The Median of a random variable $X$ is the value $q_p$ such that:
$\mathbb{P}(X \leq q_p) = 0.5, \Leftrightarrow q_p = F^{-1}(0.5)$
Example 1: For $x = 1, 4, p(x) = 0.2$ and $x = 2, 3, p(x) = 0.3$.
$P(X \leq 2) = P(X = 1) + P(X = 2) = 0.2 + 0.3 = 0.5$
Example 2: $f(x) = \begin{cases} -x^2 + 2x - \frac{1}{6}, & 0 < x < 2 \\ 0, & \text{Otherwise} \end{cases}$

$$\begin{aligned} P(x < m) &= 0.5 \\ \Rightarrow \int_0^m (-x^2 + 2x - \tfrac{1}{6})dx &= 0.5 \\ = \left[-\tfrac{x^3}{3} + x^2 - \tfrac{1}{6}x\right]_{x=0}^{x=m} &= 0.5 \\ = -\tfrac{m^3}{3} + m^2 - \tfrac{1}{6}m &= 0.5 \\ \Rightarrow m &= 1 \end{aligned}$$

# Quantiles

Le $p$-quantile of a random variable $X$ is the value $q_p$ such that:

$$\mathbb{P}(X \leq q_p) = p, p \in [0, 1]$$

$$\Leftrightarrow q_p = F^{-1}(p)$$

- **Median** : quantile of order i/2
- **Quartiles** : quantile of order i/4
- **Deciles** : quantile of order i/10
- **Centiles** : quantile of order i/100
- ...

# Quantiles: Calculating the 75th Percentile percentile from a Discrete Random Variable

Given the following probability density function of a discrete random variable:

$$f(x) = \begin{cases} 0.2, & x = 1, 4 \\ 0.3, & x = 3, 4 \end{cases}$$

# Quantiles: Calculating the 75th Percentile percentile from a Discrete Random Variable

Given the following probability density function of a discrete random variable:

$$f(x) = \begin{cases} 0.2, & x = 1, 4 \\ 0.3, & x = 3, 4 \end{cases}$$

We need: $P(X < p_{0.75}) = 0.75$

# Quantiles: Calculating the 75th Percentile percentile from a Discrete Random Variable

Given the following probability density function of a discrete random variable:

$$f(x) = \begin{cases} 0.2, & x = 1, 4 \\ 0.3, & x = 3, 4 \end{cases}$$

We need: $P(X < p_{0.75}) = 0.75$

$$\begin{aligned} P(X < 3) &= P(X = 1) + P(X = 2) + P(X = 3) \\ &= 0.2 + 0.3 + 0.2 = 0.7 < 0.75 \\ P(X < 4) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= 0.2 + 0.3 + 0.2 + 0.3 = 1.0 > 0.75 \\ P(X \geq 4) &= P(X = 4) = 0.2 < 0.25 = 1?(\tfrac{75}{100}) \end{aligned}$$

So, the 75th percentile of the distribution is 4.

# Quantiles: Calculating the 25th Percentile percentile from a Continuous Random Variable

Given the following probability density function of a continuous random variable:

$$f(x) = \begin{cases} -x^2 + 2x - \frac{1}{6}, & 0 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

# Quantiles: Calculating the 25th Percentile percentile from a Continuous Random Variable

Given the following probability density function of a continuous random variable:

$$f(x) = \begin{cases} -x^2 + 2x - \frac{1}{6}, & 0 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

We need:

$$\begin{aligned} & P(x < c) && = 0.25 \\ \Rightarrow & \int_0^c \left(-x^2 + 2x - \tfrac{1}{6}\right) dx && = \tfrac{25}{100} = 0.25 \\ & \left[-\tfrac{x^3}{3} + 2x - \tfrac{1}{6}x\right]_{x=0}^{x=c} && = 0.25 \\ & -\tfrac{c^3}{3} + c^2 - \tfrac{1}{6}c && = 0.25 \\ & c && = 0.69 \end{aligned}$$

# Random couples

For a pair $(X, Y)$, the joint distribution function $F(x, y)$ is given by:

$$F(x, y) = \mathbb{P}(X \leq x \wedge Y \leq y)$$

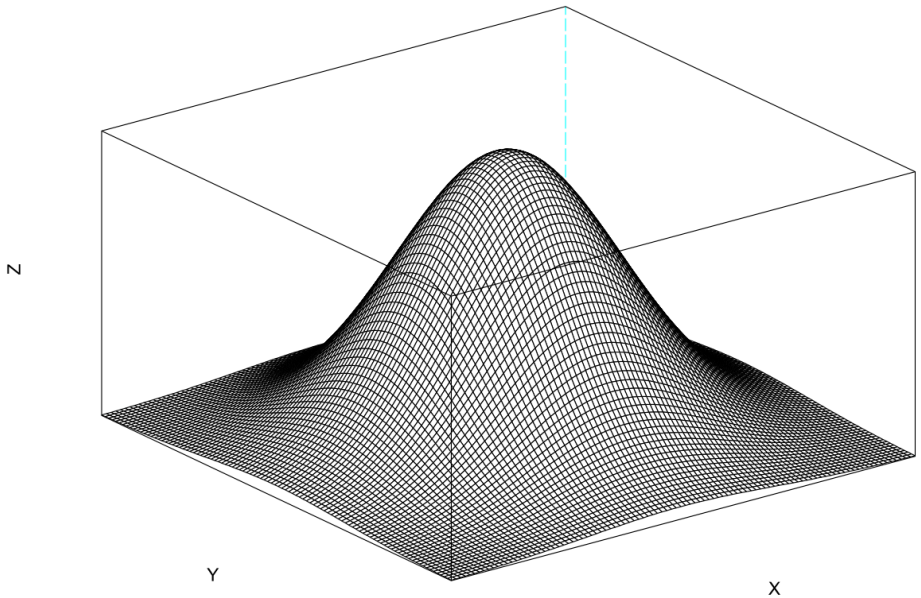



Figure: Joint distribution function/Example

# Random couples

For a couple of random variables $(X, Y)$, the covariance is defined by:

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

# Random couples: example

Daily Return for Two Stocks Using the Closing Prices:

| Day | ABC Returns | XYZ Returns |
|---|---|---|
| 1 | 1.1% | 3.0% |
| 2 | 1.7% | 4.2% |
| 3 | 2.1% | 4.9% |
| 4 | 1.4% | 4.1% |
| 5 | 0.2% | 2.5% |

# Random couples: example

Daily Return for Two Stocks Using the Closing Prices:

| Day | ABC Returns | XYZ Returns |
|---|---|---|
| 1 | 1.1% | 3.0% |
| 2 | 1.7% | 4.2% |
| 3 | 2.1% | 4.9% |
| 4 | 1.4% | 4.1% |
| 5 | 0.2% | 2.5% |

Next, calculate the average return for each stock:

- For ABC, it would be $\overline{ABC}$=(1.1 + 1.7 + 2.1 + 1.4 + 0.2) / 5 = 1.30.
- For XYZ, it would be $\overline{XYZ}$=(3 + 4.2 + 4.9 + 4.1 + 2.5) / 5 = 3.74.

# Random couples: example

Daily Return for Two Stocks Using the Closing Prices:

| Day | ABC Returns | XYZ Returns |
|---|---|---|
| 1 | 1.1% | 3.0% |
| 2 | 1.7% | 4.2% |
| 3 | 2.1% | 4.9% |
| 4 | 1.4% | 4.1% |
| 5 | 0.2% | 2.5% |

Next, calculate the average return for each stock:

- For ABC, it would be $\overline{ABC}$=(1.1 + 1.7 + 2.1 + 1.4 + 0.2) / 5 = 1.30.
- For XYZ, it would be $\overline{XYZ}$=(3 + 4.2 + 4.9 + 4.1 + 2.5) / 5 = 3.74.

$$Cov(ABC, XYZ) = 1/n \times \sum_{i=1..n} (ABC_i - \overline{ABC})(XYZ_i - \overline{XYZ})$$

$$... = [(1.1-1.30)x(3-3.74)]+[(1.7-1.30)x(4.2-3.74)]+[(2.1-1.30)x(4.9-3.74)]+?$$

$$... = 2.66/5 = 0.532$$

# Random couples: example

Daily Return for Two Stocks Using the Closing Prices:

| Day | ABC Returns | XYZ Returns |
|---|---|---|
| 1 | 1.1% | 3.0% |
| 2 | 1.7% | 4.2% |
| 3 | 2.1% | 4.9% |
| 4 | 1.4% | 4.1% |
| 5 | 0.2% | 2.5% |

Next, calculate the average return for each stock:

- For ABC, it would be $\overline{ABC}$=(1.1 + 1.7 + 2.1 + 1.4 + 0.2) / 5 = 1.30.
- For XYZ, it would be $\overline{XYZ}$=(3 + 4.2 + 4.9 + 4.1 + 2.5) / 5 = 3.74.

$$Cov(ABC, XYZ) = 1/n \times \sum_{i=1..n} (ABC_i - \overline{ABC})(XYZ_i - \overline{XYZ})$$

$$... = [(1.1-1.30)x(3-3.74)]+[(1.7-1.30)x(4.2-3.74)]+[(2.1-1.30)x(4.9-3.74)]+?$$

$$... = 2.66/5 = 0.532$$

The covariance between the two stock returns is 0.532. Because this number is positive, the stocks move in the same direction. When ABC had a high return, XYZ also had a high return.

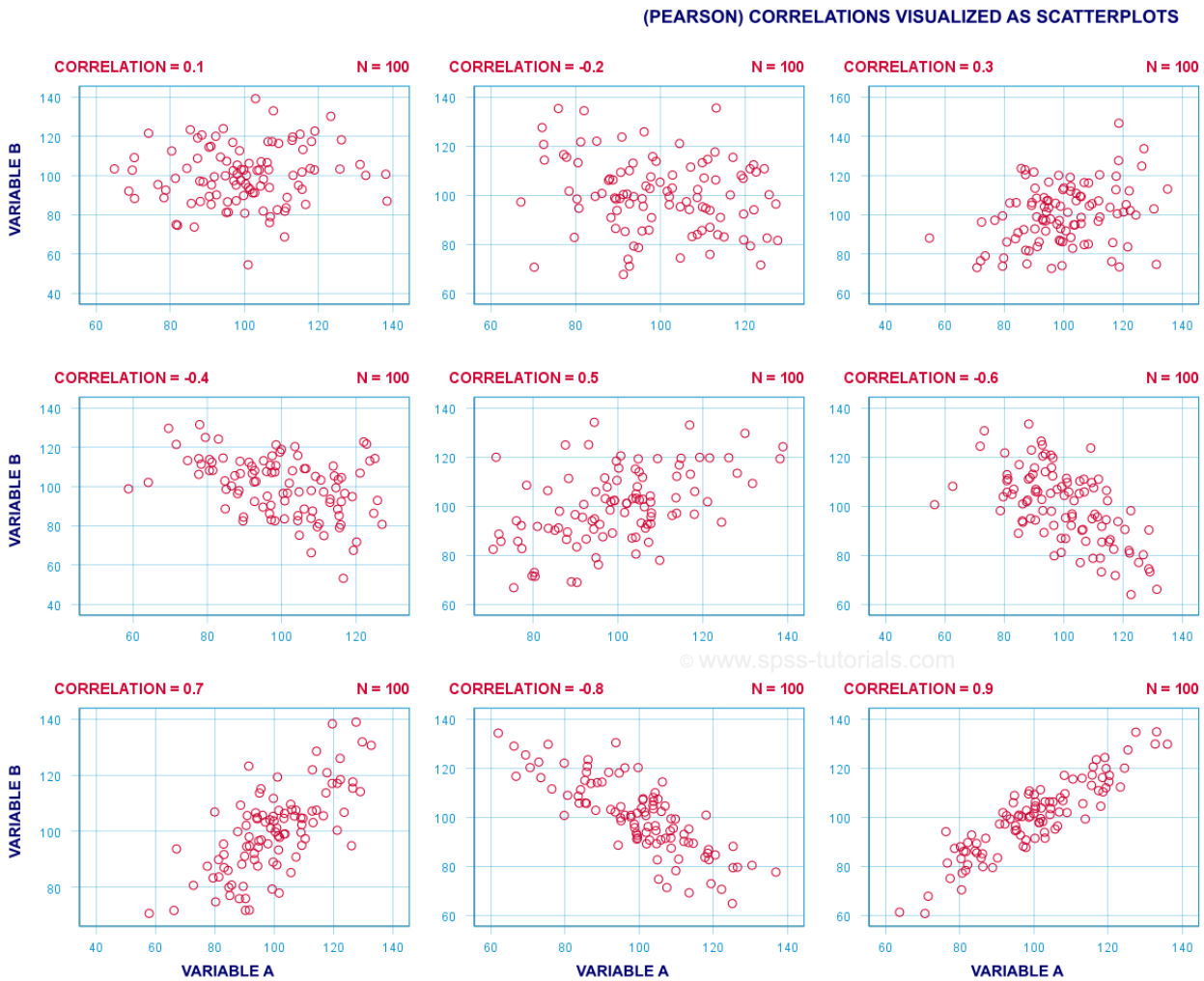
# Pearson correlation coefficient

For a pair of random variables (X, Y ), the coefficient of correlation is defined by:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$\rho_{X,Y} \in [-1,1]$
Correlations equal to +1 or -1 correspond to data points lying exactly on a line (in the case of the sample correlation). The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X). A value of +1 implies that all data points lie on a line for which $Y$ increases as $X$ increases, and vice versa for -1. A value of 0 implies that there is no linear dependency between the variables.

# Pearson correlation coefficient / Examples



# Spearman's correlation coefficient

For a sample of size $n$, the $n$ raw scores $X_i$, $Y_i$ are converted to ranks $R(X_i), R(Y_i)$. For a pair of random variables (X, Y ), the coefficient of Spearman correlation is defined by:

$$\rho_s = \frac{Cov(R(X), R(Y))}{\sqrt{Var(R(X))Var(R(Y))}}$$

$\rho_{X,Y} \in [-1, 1]$
Rewritten into the popular formula:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

where $d_i = R(X_i) - R(Y_i)$.

# Spearman's correlation coefficient

| IQ, $X_i$ | Hours of TV per week, $Y_i$ |
|---|---|
| 106 | 7 |
| 100 | 27 |
| 86 | 2 |
| 101 | 50 |
| 99 | 28 |
| 103 | 29 |
| 97 | 20 |
| 113 | 12 |
| 112 | 6 |
| 110 | 17 |

| IQ, $X_i$ | Hours of TV per week, $Y_i$ | rank $x_i$ | rank $y_i$ | $d_i$ | $d_i^2$ |
|---|---|---|---|---|---|
| 86 | 2 | 1 | 1 | 0 | 0 |
| 97 | 20 | 2 | 6 | −4 | 16 |
| 99 | 28 | 3 | 8 | −5 | 25 |
| 100 | 27 | 4 | 7 | −3 | 9 |
| 101 | 50 | 5 | 10 | −5 | 25 |
| 103 | 29 | 6 | 9 | −3 | 9 |
| 106 | 7 | 7 | 3 | 4 | 16 |
| 110 | 17 | 8 | 5 | 3 | 9 |
| 112 | 6 | 9 | 2 | 7 | 49 |
| 113 | 12 | 10 | 4 | 6 | 36 |

Figure: Examples

$$\rho = -29/165 = -0.175757575.$$

# Discrete uniform law

The discrete uniform distribution on $\{1, ..., n\}$ is the distribution of a random variable $X$ which can take the values $1, ..., n$ in an equiprobable way.

Notation: $X \sim U(\{1, ..., n\})$

Probability: $\mathbb{P}(X = k) = \frac{1}{n}; \forall k \in \{1, ..., n\}$.

Properties: $E(X) = \frac{n+1}{2}$, $Var(X) = \frac{n^2-1}{12}$.

Example: Roll a balanced die.

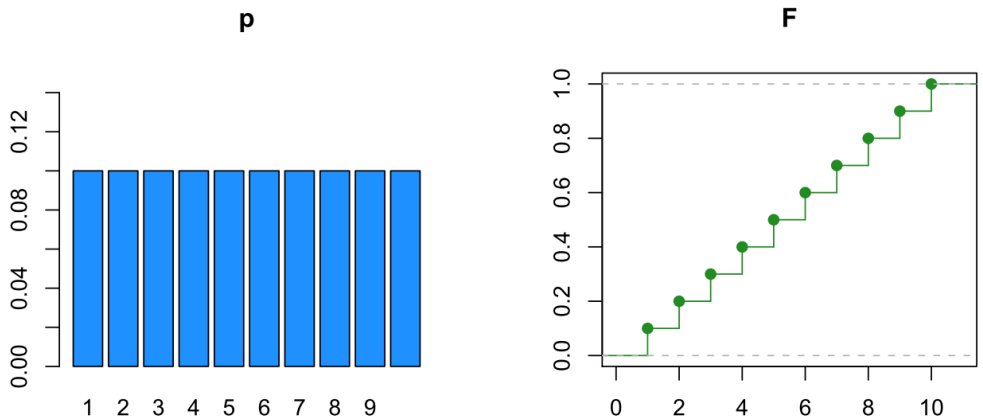


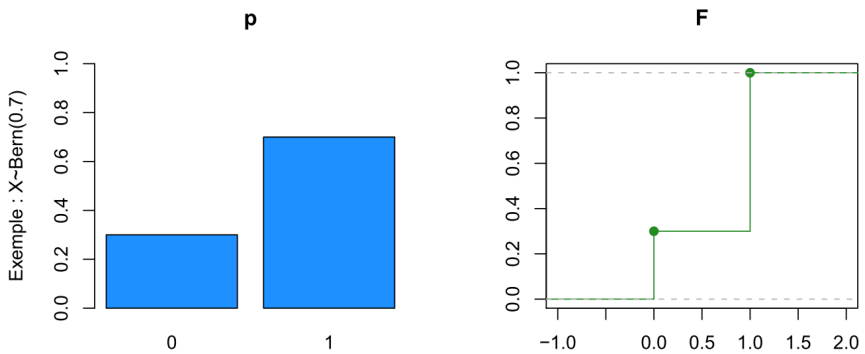

Figure: Illustration of the discrete uniform law

# Bernoulli's law

Bernoulli's law with parameter $p$ is the law of a discrete random variable $X$ which takes the value 1 with probability $p$ and the value 0 with probability $1-p$. The associated experiment is called a Bernoulli test. Notation: $X \sim \mathcal{B}(p)$.

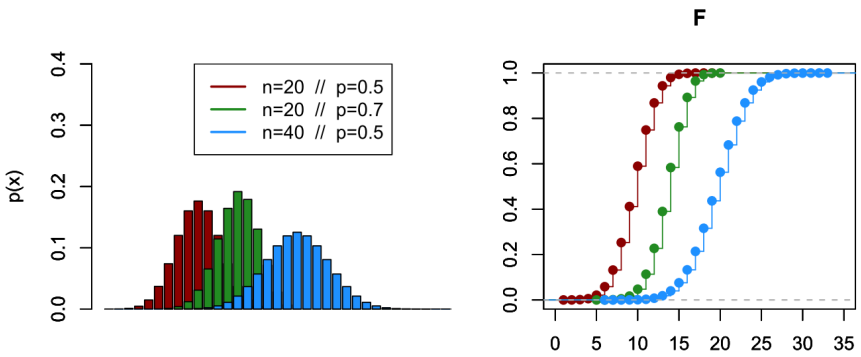
Probability: $\mathbb{P}(X = x) = \begin{cases} p & if\ x = 1 \\ 1-p & if\ x = 0 \\ 0 & else \end{cases}$

Properties: $E(X) = p$, $Var(X) = p(1-p)$. Example: heads or tails.



# Binomial law

The binomial law with parameters $n$ and $p$ is the law of the sum $X$ of $n$ independent random variables $Y_i$ such that $Y_i \sim \mathcal{B}(p)$.
Notation: $X = \sum_{i=1..n} Y_i \sim \mathcal{B}(n, p)$.
Probability: $\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}, k = 1..n$.
Properties: $E(X) = np$, $Var(X) = np(1-p)$.



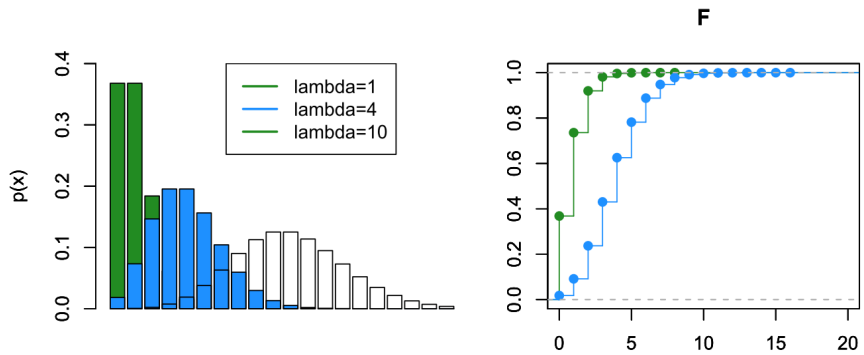

Example: We assume that a student has at most one coffee per day, that each day his probability of having a coffee is worth $p$, and that there is independence between his daily choices. The variable $X$ describing the number of coffees taken by the student in a week is a random variable with distribution $\mathcal{B}(5, p)$

# Geometric law

The geometric law with parameter $p \in [0,1]$ is the law of the random variable $Y$ which counts the number of independent repetitions of a test of Bernoulli (of parameter $p$) until the first success.
Notation: $X \sim \mathcal{G}(p)$.
Probability: $\mathbb{P}(X = k) = p(1-p)^k; k = 1..n$.
Properties: $E(X) = 1/p$, $Var(X) = (1-p)/p^2$.
Example: Counting the number of experiments needed to obtain a first success in repeating a Bernoulli test.

# Poisson's law

Poisson 's law (sometimes called law of number of events) of parameter $\lambda > 0$ is defined by the following probability function.
Notation: $X \sim \mathcal{P}(\lambda)$.
Probability: $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}; k = 1..n.$
Properties: $E(X) = \lambda$, $Var(X) = \lambda$.
Example: Count of the number of events during a time interval. If the average number of occurrences in a fixed time interval is $\lambda$, then the probability that there are exactly $k$ occurrences ($k$ being a natural number, $k = 0, 1, 2, ...$) is $\mathbb{P}(X = k)$.



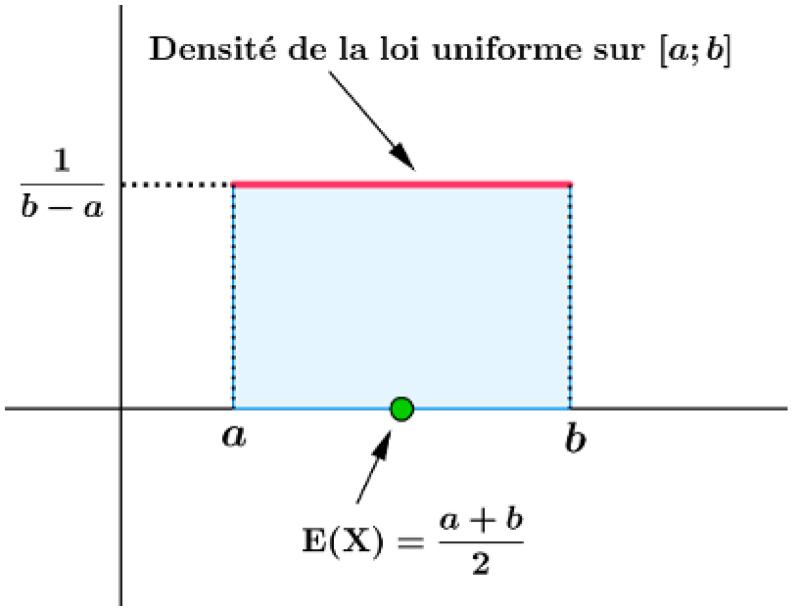
# Poisson's law

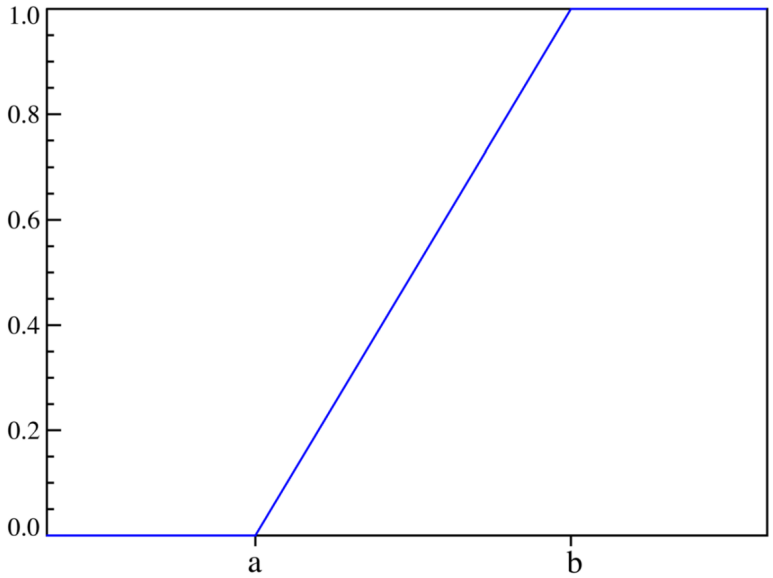
Theorem 1: When $n$ tends to infinity and, simultaneously, $p_n$ becomes small by so that $lim_{n\to\infty} np_n = \lambda > 0$, the binomial distribution with parameters $n$ and $np$ converges to the Poisson distribution with parameter $\lambda$. In practice, the approximation can be made when $n > 30$ and $np < 5$ or $n > 50$ and $p < 0.1$.
Theorem 2: If $X_1$ and $X_2$ are two independent random variables such that $X_1 \sim \mathcal{P}(\lambda_1)$ and $X_1 \sim \mathcal{P}(\lambda_1)$, then

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

# Uniform law

| Law | $f(x)$ | $E(X)$ | $Var(X)$ |
|---|---|---|---|
| $\mathcal{U}[a,b]$ | $f_X(x) = \frac{1}{b-a}, a \leq x \leq b$ | $\frac{a+b}{2}$ | $\frac{(a+b)^2}{12}$ |

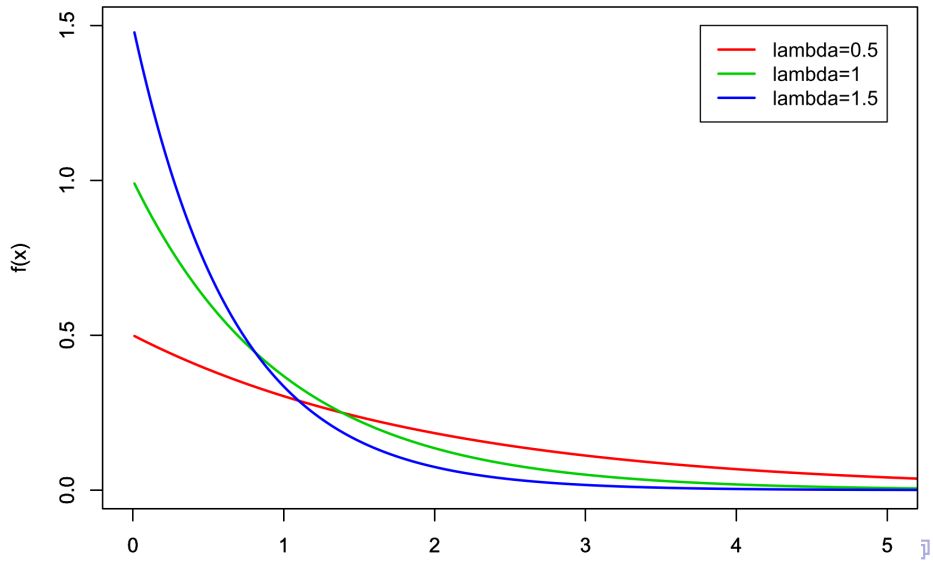


# Exponential law

$f(x) = \lambda e^{-\lambda x}$ if $x \geq 0$, $f(x) = 0$ if $x < 0$.
$E(X) = 1/\lambda$, $Var(X) = 1/\lambda^2$.

Example: The exponential law is mainly used in the problems of lifetime of some thing (e.g., equipment). The $\lambda$ parameter can represent the number of times an event occurs during a given period of time. For instance, the lifetime of electronic equipment follows an exponential law with parameter $\lambda = 1/10$ (the unit of time is the year). What is the probability that it will still work 5 years after it was made?



# Normal probability distribution

- A random variable $X$ is said to have a normal probability distribution with parameters $\mu$ and $\sigma^2$, if it has a pdf given by:

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

- We use the notation $X \backsim N(\mu, \sigma^2)$
- pdf

$$F(x) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

- there is no analytical form for $F$.
- $E(X) = \mu$
- $Var(X) = \sigma^2$
- If $\mu = 0$, and $\sigma = 1$, we call it a standard normal random variable.

# Normal probability distribution / Example

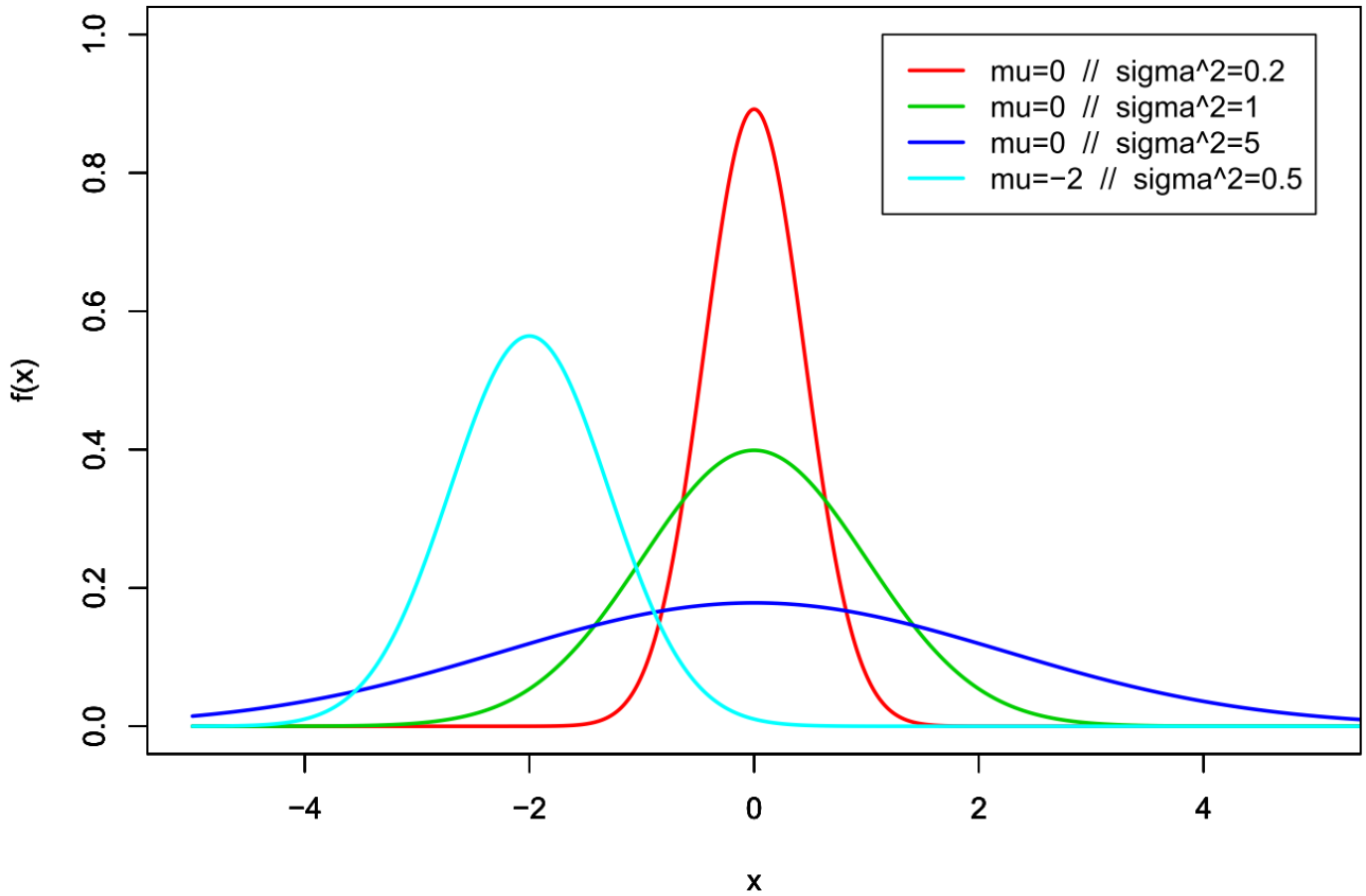


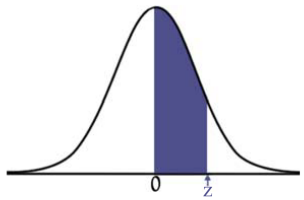

Figure: Examples

# Standard normal probability distribution

- If $\mu = 0$, and $\sigma = 1$, we call it a standard normal random variable. Its pdf is defined $\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.
- Every distribution $N(\mu, \sigma)$ is deductible from $N(0, 1)$ through the property: If $Y$ follows $N(\mu, \sigma)$, then $Z = \frac{Y-\mu}{\sigma}$ follows $N(0, 1)$.
- We denote $\Phi$ the pdf of standard normal probability distribution: $\Phi(x) = P(Z < x)$, where $Z$ is random variable following $N(0, 1)$.
- Examples: $\Phi(0) = 0.5, \Phi(1.645) \approx 0.95, \Phi(1.960) \approx 0.9750$.

# Normal probability distribution / Table

Area between 0 and $z$




| $z$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| 0.7 | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |
| 3.1 | 0.4990 | 0.4991 | 0.4991 | 0.4991 | 0.4992 | 0.4992 | 0.4992 | 0.4992 | 0.4993 | 0.4993 |
| 3.2 | 0.4993 | 0.4993 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4995 | 0.4995 | 0.4995 |
| 3.3 | 0.4995 | 0.4995 | 0.4995 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4997 |
| 3.4 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4998 |

# Normal probability distribution / Example

**Example**

**(a)** For $X \sim N(0, 1)$, calculate $P(Z \geq 1.13)$.

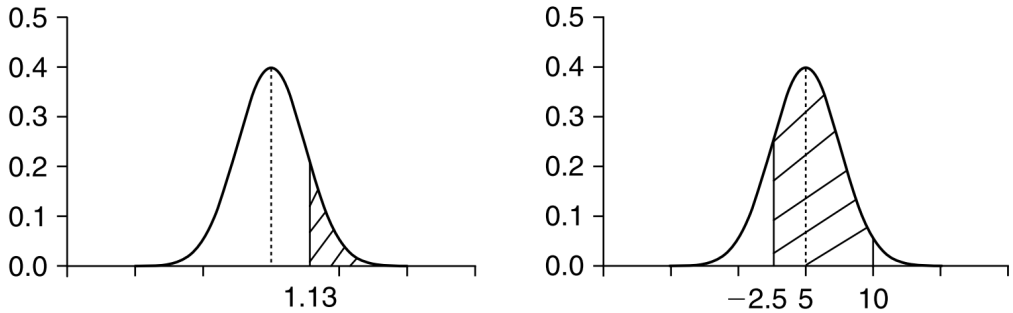
**(b)** For $X \sim N(5, 4)$, calculate $P(-2.5 < X < 10)$.

***Solution***

**(a)** *Using the normal table,*

$$P(Z \geq 1.13) = 1 - 0.8708 = 0.1292.$$

*The shaded part in the graph represents the* $P(Z \geq 1.13)$.




**(b)** *Using the z-transform, we have*

$$\begin{aligned} P(-2.5 < X < 10) &= P\left(\frac{-2.5-5}{2} < Z < \frac{10-5}{2}\right) \\ &= P(-3.75 < Z < 2.5) \\ &= P(-3.75 < Z < 0) + P(0 < Z < 2.5) \\ &= 0.9938. \end{aligned}$$

# Standard normal probability distribution

- For $|x| < 2$, we have a good approximation of $\Phi$: $\Phi(x) \approx 0.5 + \frac{1}{\sqrt{2\pi}}(x - \frac{x^3}{6} + \frac{x^5}{40})$.
- Conversely, from the probability value, we can look for the bound for which this probability is effective.
- We denote $z_{\alpha/2}$ the number for which $P(Z > z_{\alpha/2}) = \alpha/2$.
- Example:

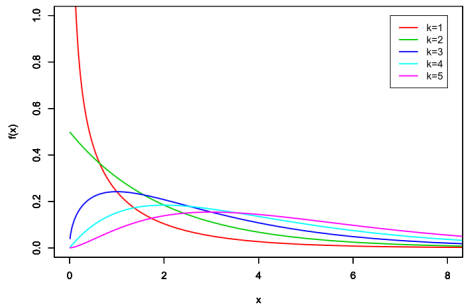
| Risk $\alpha/2$ | 0.01 | 0.02 | 0.05 | 0.10 |
|---|---|---|---|---|
| Critical value $z_{\alpha/2}$ | 2.58 | 2.33 | 1.96 | 1.645 |

# Loi du $\chi^2$

Let $X_1, ..., X_n$ be $n$ independent and identically random variables distributed with reduced centered normal law. The random variable $Y = X_1^2 + ... + X_n^2$ follows a continuous law called law of $\chi^2$ with $n$ degrees of freedom: $Y = \sum_{i=1..n} X_i^2 \sim \chi_n^2$

Density:

$$f(y) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} e^{-x/2}$$




Example: to test the quality of adjustment between an observed distribution and a theoretical distribution (even between two observed distributions).

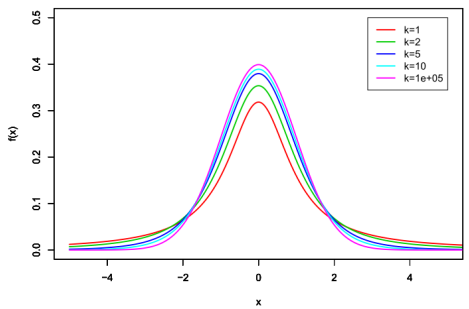
# Student 's law

Let $X$ and $Y$ be two independent random variables such that $X \sim N(0,1)$ and $Y \sim \chi^2_n$. The random variable $T = X/\sqrt{Y/n}$ follows a continuous law called Student's law with $n$ degrees of freedom.

Notation: $T = \dfrac{X}{\sqrt{Y/n}} \sim T$.

Properties: $E(T) = 0$, $Var(X) = \frac{n}{n-2}$ if $n > 2$. Example: Used for quality control.

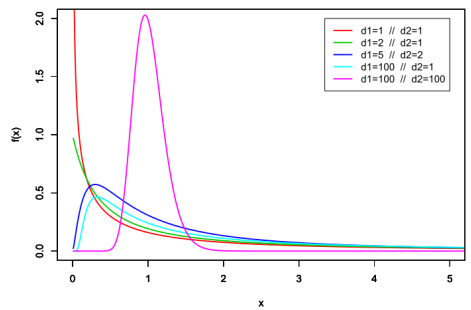
$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})(1+\frac{x^2}{n})^{\frac{n+1}{2}}}$$



# Fisher-Snedecor 's law

Let $Y_1$ and $Y_2$ be two independent random variables such that $Y_1 \sim \chi^2_{n_1}$ and $Y_2 \sim \chi^2_{n_2}$. The random variable $Z = (Y_1/n_1)/(Y_2/n_2)$ follows a continuous law called Fisher's law with $n_1$ and $n_2$ degrees of freedom: Notation: $Z \sim \mathcal{F}(n_1; n_2)$.
Probability: $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}; k = 1..n..$
Properties: $E(X) = \lambda$, $Var(X) = \lambda$. Example: It is used to test if variances are equal because the numerator and the denominator of $F$ can represent two variances.

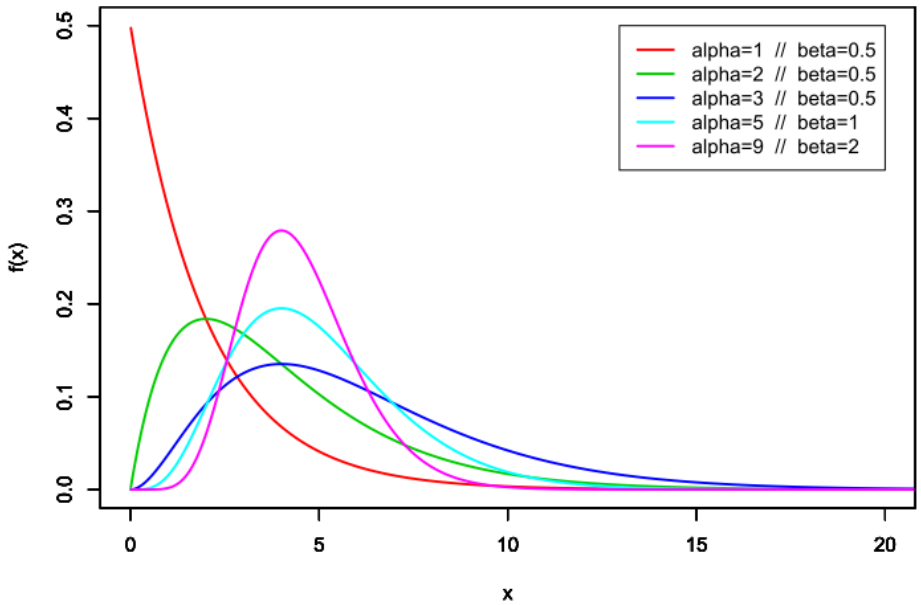
$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} n_1^{n_1/2} n_2^{n_2/2} \frac{x^{n_1/2-1}}{(n_2+n_1 x)}$$



# Gamma 's law

Example: It is used in particular for reliability (with usury).

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}$$



# Sum. Discrete laws

| Lois de probabilité discrètes | Fct de probabilité | $E(X)$ | $Var(X)$ | Genèse |
|---|---|---|---|---|
| Uniforme $X \sim U(\{1, 2, \ldots, n\})$ $n \in \mathbb{N}$ | $\frac{1}{n}$ | $\frac{n+1}{2}$ | $\frac{n^2-1}{12}$ | - |
| Bernoulli $X \sim B(p)$ $0 \leq p \leq 1$ | $p^x(1-p)^{1-x}$ si $x = 0, 1$ | $p$ | $p(1-p)$ | Lancer d'une pèce de monnaie avec $\mathbb{P}(pile) = p$ |
| Binomiale $X \sim B(n, p)$ $n$ entier $> 0$, $0 \leq p \leq 1$ | $C_n^x p^x (1-p)^{n-x}$ si x=0,1,...,n | $np$ | $np(1-p)$ | Loi de la somme de $n$ v.a. indépendantes de loi $B(p)$ |
| Poisson $X \sim P(\lambda)$ $\lambda > 0$ | $\frac{e^{-\lambda}\lambda^x}{x!}$ si $x = 0, 1, 2, \ldots$ | $\lambda$ | $\lambda$ | Limite de la loi binomiale lorsque $n \to \infty$, $np_n \to \lambda > 0$ et $p_n \to 0$ |
| Géométrique $X \sim G(p)$ $0 \leq p \leq 1$ | $p(1-p)^{x-1}$ si $x = 1, 2, \ldots$ | $\frac{1}{p}$ | $\frac{1-p}{p^2}$ | Nombre de lancers nécessaire pour l'obtention du premier $pile$ avec $\mathbb{P}(pile) = p$ |
| Binomiale négative $X \sim BN(n, p)$ $0 \leq p \leq 1$ | $C_{x-1}^{n-1} p^n (1-p)^{x-n}$ si $x = n, n+1, \ldots$ | $\frac{n}{p}$ | $\frac{n(1-p)}{p}$ | Loi de la somme de $n$ v.a. indépendantes de loi $G(p)$ |

# Sum. Continuous laws

| Lois de probabilité continues | Fct de densité | Fct de répartition | $E(X)$ | $Var(X)$ |
|---|---|---|---|---|
| Uniforme $X \sim U[a,\, b]$ $a < b$ | $\frac{1}{b-a} 1_{[a,b]}(x)$ | $\frac{x-a}{b-a} 1_{[a,b]}(x)$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Exponentielle $X \sim Exp(\lambda)$ $\lambda > 0$ | $\lambda e^{-\lambda x} 1_{x>0}$ | $1 - e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Normale $X \sim N(\mu, \sigma^2)$ $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$ | - | $\mu$ | $\sigma^2$ |
| Khi-deux $X \sim \chi^2_\nu$ $\nu > 0$ | $\frac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} e^{-x/2}$ | - | $\nu$ | $2\nu$ |
| Student $X \sim \chi^2_\nu$ $\nu > 0$ | $\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})(1+\frac{x^2}{n})^{\frac{n+1}{2}}}$ | - | $0$ si $\nu \geq 2$ | $\frac{\nu}{\nu-2}$ si $\nu \geq 3$ |
| Cauchy $X \sim C(\mu, \sigma)$ $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ | $\frac{\sigma}{\pi[(x-\mu)^2+\sigma^2]}$ | - | n'existe pas | n'existe pas |
| Gamma $X \sim \Gamma(\alpha, \beta)$ $\alpha > 0$, $\beta > 0$ | $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}$ si $x \geq 0$ | - | $\frac{\alpha}{\beta}$ | $\frac{\alpha}{\beta^2}$ |

# Central Limit Theorem (CLT)

Let $X_1, ..., X_n$ be a random sample from some population with mean $\mu$ and variance $\sigma^2$. Then for large $n$,

$$\overline{X} \simeq N(\mu, \frac{\sigma^2}{n})$$

even if the underlying distribution of individual observations in the population is not normal. The $\simeq$ symbol represents "approximately distributed", and the formula can be read as "the mean of $X$ is approximately normally distributed" with mean $\mu$ and variance $\frac{\sigma^2}{n}$. **A pillar of statistics, this theorem states also that the means of a large number of samples follow a normal law, even if they individually follow another law of probability.**

# Central Limit Theorem (CLT)

If $X_1, ..., X_n$ is a random sample from an infinite population with mean $\mu < \infty$, and variance $\sigma^2 < \infty$, then the limiting distribution of $Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is the standard normal probability distribution.

The CLT is extremely important in statistics because it says that we can approximate the distribution of certain statistics without much knowledge about the underlying probability distribution of that statistic for a relatively "large" sample size. How large the $n$ should be for this normal approximation to work depends on the distribution of the original distribution. **A rule of thumb is that the sample size n must be at least 30.**

# Central Limit Theorem (CLT) / Example

N.B./ 1 ounce $\sim$ 25 gram.

**Example**

drink vending machine is set so that the amount of drink dispensed is a random variable with a mean of 8 ounces and a standard deviation of 0.4 ounces. What is the approximate probability that the average of 36 randomly chosen fills exceed 8.1 ounces?

***Solution***

*From the CLT,* $((\overline{X} - 8)/(0.4/\sqrt{36})) \sim N(0, 1)$. *Hence, from the normal table,*

$$P\{\overline{X} > 8.1\} = P\left\{Z > \frac{8.1 - 8.0}{\frac{0.4}{\sqrt{36}}}\right\}$$
$$= p\{Z > 1.5\} = 0.0668.$$

When the sequence $(X_n)$ of random variables converges to rv $X$, it means that $X_n$ is getting closer when $n$ increases. Formally, sequence $(X_n)$ converges to $X$ if, for some $\epsilon$

$$P(|X_n - X| < \epsilon) \to 1 \text{ when } n \to \infty$$

We write $X_n \to_p X$.

**Property (Sufficient convergence condition)**

*If $(X_n)$ is a sequence of random variables such that:*

$$E(X_n) \to a$$
$$V(X_n) \to 0$$

*when $n \to \infty$, then $X_n \to_p a$.*

# Law of large numbers

**Theorem (Law of large numbers (weak))**

*The sequence of random variables $(X_n)$ mutually independent, such that $E(X_n) = m$ and $V(X_n) = \sigma^2$, then for $n \to \infty$: $\overline{X_n} \to m$.*

**Theorem (Law of large numbers (strong))**

*The sequence of random variables $(X_n)$ mutually independent, having the same law and the same expected value $m$, then for $n \to \infty$: $\overline{X_n} \to m$.*

# Law convergence

## Definition

*We state that the random variable $(X_n)$, with CDF (cumulative dist. funct.) $F_n$, converges in law to $X$, if the sequence $\{F_n(X)\}$ converges to $F(x)$.*

## Theorem (Probability convergence)

*The probability convergence of the sequence $(X_n)$ implies the convergence in law:*

$$X_n \to_p X \Rightarrow X_n \to_{law} X.$$

## Theorem (Normal Approximation to the Binomial Distribution)

**Normal Approximation to the Binomial Distribution**

If $X$ is a binomial random variable with parameters $n$ and $p$,

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable. To approximate a binomial probability with a normal distribution, a **continuity correction** is applied as follows:

$$P(X \leq x) = P(X \leq x + 0.5) \approx P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

and

$$P(x \leq X) = P(x - 0.5 \leq X) \approx P\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}} \leq Z\right)$$

The approximation is good for $np > 5$ and $n(1-p) > 5$.

## Theorem (Normal Approximation to the Poisson Distribution)

**Normal Approximation to the Poisson Distribution**

If $X$ is a Poisson random variable with $E(X) = \lambda$ and $V(X) = \lambda$,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

is approximately a standard normal random variable. The same continuity correction used for the binomial distribution can also be applied. The approximation is good for

$$\lambda > 5$$

# Main Approximations

$$X \sim B(n,p) \quad \Longrightarrow \quad X \sim P(\mu)$$

$$n \geq 30\,;\, p \leq 0.1$$

$$\mu = np$$

$$\Downarrow \mu \geq 20 \qquad \sigma^2 = \mu$$

$$X \sim B(n,p) \quad \Longrightarrow \quad N(\mu, \sigma^2)$$

$$n \geq 30\,;\, np \geq 5\,;\, n(1-p) \geq 5$$

$$\mu = np\,;\, \sigma^2 = np(1-p)$$

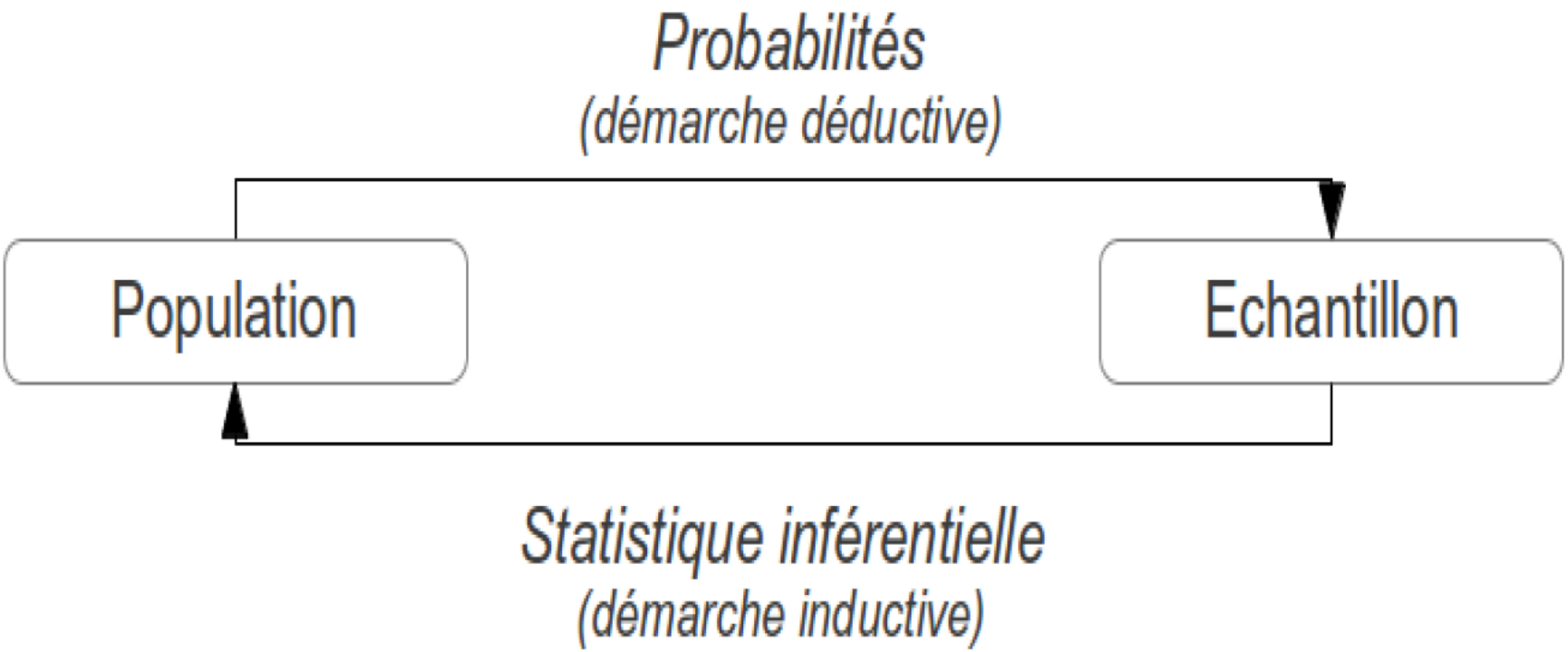
$$\Uparrow \; n \geq 30$$

$$X = Y_1 + ... + Y_n$$

$Y_1, ..., Y_n$ iid

# Motivations

Statistics and probabilities are two complementary aspects of the study random phenomena.



# Motivations

Statistical inference is the act of generalizing from the data ("sample") to a larger phenomenon ("population") with calculated degree of certainty.
The act of generalizing and deriving statistical judgements is the process of inference.
Statistical inference aims at learning characteristics of the population from a sample; the population characteristics are parameters and sample characteristics are statistics.
The two common forms of statistical inference are:

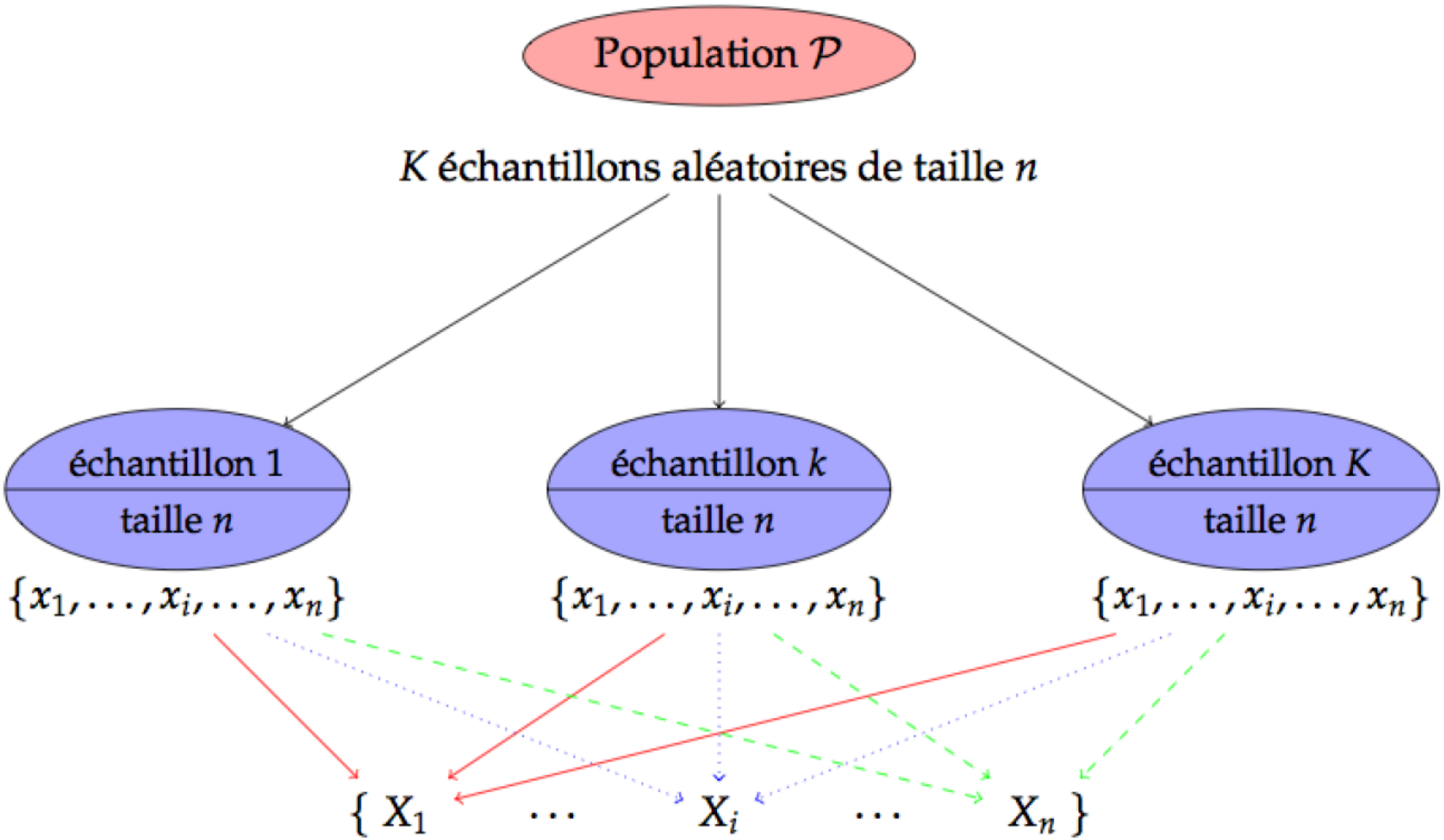
- Estimation
- Null hypothesis tests of significance (NHTS)
- Statistical model

# Motivations

A statistical model is a representation of a complex phenomena that generated the data :

- It has mathematical formulations that describe relationships between random variables and parameters.
- It makes assumptions about the random variables, and sometimes parameters.
- A general form: data = model + residuals
- Model should explain most of the variation in the data.
- Residuals are a representation of a lack-of-fit, that is of the portion of the data unexplained by the model.

# Sampling


Population $\mathcal{P}$
$K$ échantillons aléatoires de taille $n$
échantillon 1
taille $n$
échantillon $k$
taille $n$
échantillon $K$
taille $n$
$\{x_1, \ldots, x_i, \ldots, x_n\}$
$\{x_1, \ldots, x_i, \ldots, x_n\}$
$\{x_1, \ldots, x_i, \ldots, x_n\}$
$\{ X_1 \quad \cdots \quad X_i \quad \cdots \quad X_n \}$

# Sampling / Example

$K = 6$ different samples, each with $n = 10$, with the values of the sample mean, sample median, and sample standard deviation for each sample. Suppose that $\mu = 4.4311, \sigma = 2.316$.

| | Sample | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| *Observation* | | | | | | |
| 1 | 6.1171 | 5.07611 | 3.46710 | 1.55601 | 3.12372 | 8.93795 |
| 2 | 4.1600 | 6.79279 | 2.71938 | 4.56941 | 6.09685 | 3.92487 |
| 3 | 3.1950 | 4.43259 | 5.88129 | 4.79870 | 3.41181 | 8.76202 |
| 4 | 0.6694 | 8.55752 | 5.14915 | 2.49759 | 1.65409 | 7.05569 |
| 5 | 1.8552 | 6.82487 | 4.99635 | 2.33267 | 2.29512 | 2.30932 |
| 6 | 5.2316 | 7.39958 | 5.86887 | 4.01295 | 2.12583 | 5.94195 |
| 7 | 2.7609 | 2.14755 | 6.05918 | 9.08845 | 3.20938 | 6.74166 |
| 8 | 10.2185 | 8.50628 | 1.80119 | 3.25728 | 3.23209 | 1.75468 |
| 9 | 5.2438 | 5.49510 | 4.21994 | 3.70132 | 6.84426 | 4.91827 |
| 10 | 4.5590 | 4.04525 | 2.12934 | 5.50134 | 4.20694 | 7.26081 |
| Mean | 4.401 | 5.928 | 4.229 | 4.132 | 3.620 | 5.761 |
| Median | 4.360 | 6.144 | 4.608 | 3.857 | 3.221 | 6.342 |
| SD | 2.642 | 2.062 | 1.611 | 2.124 | 1.678 | 2.496 |

Figure: Material strengths for six different groups of ten specimens each

# Sampling

The rvs $X_1, X_2, ..., X_n$ are said to form a (simple) random sample of size $n$ if

1. The $X_i$'s are independent rvs.
2. Every $X_i$ has the same probability distribution.

Such a collection of random variables is also referred to as being independent and identically distributed (iid).

# Statistical model

- We study a variable $X$, for which we have observations. We assume that $X$ follows a known law $\mathcal{L}_\theta$, i.e. we choose from the existing models the most appropriate law to the observed phenomenon. Only the numerical value of the parameter $\theta$ involved in this probability law is unknown.
- We suppose that $X \sim \mathcal{L}_\theta$, and we look for how to set $\theta$.
- Example: Let $X$ be the size of the inhabitants of Oran. We assume that it follows a normal law, with unknown mean $\theta$ and known variance. We will therefore seek to estimate $\theta$ from a sample of data.

# Statistics

We call statistics any function of the n-sample $X_1, ..., X_n$:

$$T: \quad \begin{aligned} \boldsymbol{R}^n &\Rightarrow \boldsymbol{R} \\ (X_1, ..., X_n) &\Rightarrow T(X_1, ..., X_n) \end{aligned}$$

$t = T(X_1, ..., X_n)$ is a particular realization of the random variable $T$.

# Statistics

Examples:

- Sum: $S_n = \sum_{i=1..n} X_i$
- Empirical/experimental/observed mean: $\overline{X_n} = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1..n} X_i$.
- Empirical/experimental/observed variance: $S^2 = \frac{1}{n} \sum_{i=1..n} (X_i - \overline{X_n})^2 = \frac{1}{n} \sum_{i=1..n} X_i^2 - \overline{X_n}^2$.
- Corrected empirical variance: $S^{*2} = \frac{1}{n-1} \sum_{i=1..n} (X_i - \overline{X_n})^2 = \frac{n}{n-1} S^2$.
-

# Estimation

Let us consider the estimation problem:

- Let $X$ be a random variable with the probability law $\mathcal{L}_\theta, \theta \in \Theta$.
- Let $x_1, ..., x_n$ be an observation of the $n$-sample $X_1, ..., X_n$.
- How to estimate $\theta$ from $x_1, ..., x_n$?

There are two recipes to cope with this task:

- Point estimation.
- Estimation by confidence intervals.

# Point estimation

A point estimator is a statistic whose realization (for a given sample) constitutes an estimate of one of the parameters $\theta$ of the distribution (or one of the functions allowing it to be characterized). We denote by $\hat{\theta}$ the estimation of $\theta$.

# Quality of the estimation

- The bias $b$ of the estimation is the quantity: $b(\hat{\theta}) = E(\hat{\theta}) - \theta$.
- The estimator has asymptotically no bias iif $lim_{n \to +\infty} b(\hat{\theta}) = 0$.

# Usual estimators

Examples:

- Empirical mean: $\overline{X_n} = \frac{1}{n}\sum_{i=1..n} X_i$
- Empirical variance:
  $S^2 = \frac{1}{n}\sum_{i=1..n}(X_i - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1..n} X_i^2 - \overline{X_n}^2$.
- Corrected empirical variance: $S^{*2} = \frac{1}{n-1}\sum_{i=1..n}(X_i - \overline{X_n})^2$.

# Empirical mean

$$\overline{X_n} = \frac{1}{n} \sum_{i=1..n} X_i$$

Properties:

- Let $X_1, ..., X_n$ be $n$-sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$.
- $E(\overline{X_n}) = \mu$ and $V(\overline{X}_n) = \frac{\sigma^2}{n}$.
- $\overline{X}_n$ is an estimator with no bias, and converges to $\mu$.

# Empirical variance

$$S^2 = \frac{1}{n}\sum_{i=1..n}(X_i - \overline{X_n})^2 = \frac{1}{n}\sum_{i=1..n} X_i^2 - \overline{X_n}^2$$

$$S^{*2} = \frac{1}{n-1}\sum_{i=1..n}(X_i - \overline{X_n})^2$$

Properties:

- Let $X_1, ..., X_n$ be $n$-sample such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$.
- $E(S^2) = \frac{n-1}{n}\sigma^2$.
- $S^{*2} = \frac{n-1}{n}S^2$ is an estimator with no bias and converges to $\sigma^2$.

# Sampling / Example

An online florist offers three different sizes for Mother's Day bouquets: a small arrangement costing 80 DA (including shipping), a medium-sized one for 100 DA, and a large one with a price tag of 120 DA. If 20% of all purchasers choose the small arrangement, 30% choose medium, and 50% choose large, then the probability distribution of the cost of a single randomly selected flower arrangement is given by

| $x$ | 80 | 100 | 120 |
|---|---|---|---|
| $p(x)$ | 0.2 | 0.3 | 0.5 |

with $\mu = 106$ and $\sigma^2 = 244$.

# Sampling / Example

Suppose only two bouquets are sold today. Let $X_1$ = the cost of the first bouquet and $X_2$ = the cost of the second. Suppose that $X_1$ and $X_2$ are independent, each with the shown probability distribution, so that $X_1$ and $X_2$ constitute a random sample from the distribution. Following Table lists possible $(x_1, x_2)$ pairs, the probability of each pair computed and the assumption of independence, and the resulting $\bar{x}$ and $s^2 = (x_1 - \bar{x}) + (x_2 - \bar{x})$.

Outcomes, probabilities, and values of $\bar{x}$ and $s^2$ for Example

| $x_1$ | $x_2$ | $p(x_1, x_2)$ | $\bar{x}$ | $s^2$ |
|---|---|---|---|---|
| 80 | 80 | (.2)(.2) = .04 | 80 | 0 |
| 80 | 100 | (.2)(.3) = .06 | 90 | 200 |
| 80 | 120 | (.2)(.5) = .10 | 100 | 800 |
| 100 | 80 | (.3)(.2) = .06 | 90 | 200 |
| 100 | 100 | (.3)(.3) = .09 | 100 | 0 |
| 100 | 120 | (.3)(.5) = .15 | 110 | 200 |
| 120 | 80 | (.5)(.2) = .10 | 100 | 800 |
| 120 | 100 | (.5)(.3) = .15 | 110 | 200 |
| 120 | 120 | (.5)(.5) = .25 | 120 | 0 |

# Sampling / Example

For example, $\overline{x} = 100$ occurs three times in the table with probabilities .10, .09, and .10, so

$$P(\overline{X} = 100) = .10 + .09 + .10 = .29.$$

| $\bar{x}$ | 80 | 90 | 100 | 110 | 120 |
|---|---|---|---|---|---|
| $p_{\bar{X}}(\bar{x})$ | .04 | .12 | .29 | .30 | .25 |

| $s^2$ | 0 | 200 | 800 |
|---|---|---|---|
| $p_{S^2}(s^2)$ | .38 | .42 | .20 |

$$E(\overline{X} = \sum \overline{x} p_{\overline{X}}(\overline{x}) = 80(.40) + \ldots + 120(0.25) = 106 = \mu$$

$$V(\overline{X}) = \sum (\overline{x} - \mu_{\overline{X}})^2 p_{\overline{X}}(\overline{x}) = (80 - 106)^2(.40) + \ldots + (120 - 106)^2(0.25) = 122$$

$$= \sigma^2/2.$$

$$E(S^2) = \sum s^2 p_{s^2}(s^2) = 0(.38) + 200(0.42) + 800(0.20) = 244 = \sigma^2.$$

# Sampling / Example

Consider $n = 4$ in the current example:

| $\bar{x}$ | 80 | 85 | 90 | 95 | 100 | 105 | 110 | 115 | 120 |
|---|---|---|---|---|---|---|---|---|---|
| $p_{\overline{X}}(\bar{x})$ | .0016 | .0096 | .0376 | .0936 | .1761 | .2340 | .2350 | .1500 | .0625 |

# Sampling

The sample total $T_o = \sum_{i=1..n} X_i$, $\overline{X} = \frac{\sum_{i=1..n} X_i}{n} = T_o/n$.

**PROPOSITION** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value $\mu$ and standard deviation $\sigma$. Then

1. $E(T_o) = n\mu$
2. $V(T_o) = n\sigma^2$ and $\sigma_{T_o} = \sqrt{n}\sigma$
3. If the $X_i$'s are normally distributed, then $T_o$ is also normally distributed.

1. $E(\overline{X}) = \mu$
2. $V(\overline{X}) = \frac{\sigma^2}{n}$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$
3. If the $X_i$'s are normally distributed, then $\overline{X}$ is also normally distributed.

# Estimation by confidence intervals

- The estimation of the mean is not sufficient to draw a confident conclusion about the population mean. A confidence interval is an interval centred around the mean $m$ in which we have a confidence percentage of finding the average of the population. We want this percentage to be as high as possible.
- Let $\alpha$ be the risk and $1-\alpha$ the confidence. We know that

$$U = \frac{\tilde{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

follows the standard normal law.

# Estimation by confidence intervals

- We have to find a positive real $t$ such that $P(-t < U < t) = 1 - \alpha$. It is depicted in the next Figure.

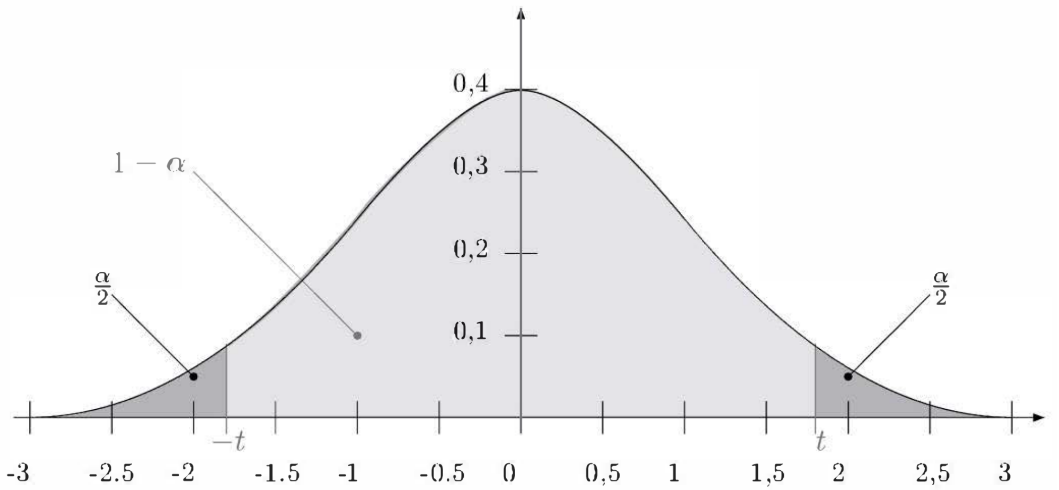


Figure 7-1 Intervalle de confiance

- $P(-t < U < t) = P(U < t) - P(U < -t) = P(U < t) - P(U > t)$
- $P(-t < U < t) = P(U < t) - (1 - P(U < t)) = 2P(U < t) - 1$
- Thus $P(U < t) = 1 - \alpha/2$.

# Estimation by confidence intervals

- The relation $P(U < t) = 1 - \alpha/2$ enables to compute $t$. In practice, the usual risks are $10\%, 5\%, 2\%, 1\%$.

| $\alpha$ | $1 - \frac{\alpha}{2}$ | $t$ |
|---|---|---|
| 10 % | 0,95 | 1,6449 |
| 5 % | 0,975 | 1,96 |
| 2 % | 0,99 | 2,3263 |
| 1 % | 0,995 | 2,5758 |

- $P(-t < U < t) = 1 - \alpha$, we have $P(-t < \frac{\tilde{X} - m}{\frac{\sigma}{\sqrt{n}}} < t) = 1 - \alpha$
- $P(-t\frac{\sigma}{\sqrt{n}} < \tilde{X} - m < t\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $P(m - t\frac{\sigma}{\sqrt{n}} < \tilde{X} < m + t\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- Thus the confidence interval is $[m - t\frac{\sigma}{\sqrt{n}}, m + t\frac{\sigma}{\sqrt{n}}) = 1 - \alpha]$.
- We can extend the framework of confidence intervals to other laws.

# Motivations

Estimators can be used in constructing hypothesis tests. We then take the problem inversely:

- We assume that we know the population according to a certain criterion.
- We have a new samples which seem to differ from the initial law followed by the population.
- We must conclude using a hypothesis test whether this change is the fact of random sampling or the fact of a real change in the parameters of the law in the population.

# Statistical Decision and Error Probabilities

$H_1$: hypothesis (claim/alternative)
$H_0$: nullification hypothesis

- Previous experiences. In this case, we want to know if conditions have changed.
- A theory or a model of the problem. In this case, we want to determine if the model is valid.
- External considerations, such as technical specifications. In this case we want to know if the object or the process is compliant.

Examples :

- $H_0 : \theta = \theta_0,\ H_1 : \theta \neq \theta_1.$
- $H_0 : \theta \geq \theta_0,\ H_1 : \theta = \theta_1.$ ...

# Example 1

The number of items sold per day during the year in a store follows a normal law with a mean of 300 and a standard deviation of 20. During sales in 15 days, the average number of items sold per day is 315. Do the sales lead to a significant improvement in sales? Let $m$ denote the mean of the normal law governing the number of articles sold during sales.
Two alternative hypotheses can be formulated:

- Null hypothesis $H_0$: "the sales bring no change: $m = 300$".
- Alternative hypothesis $H_1$: "Sales bring progress in sales: $m > 300$".
- We have to validate one of the two hypotheses. If the hypothesis $H_0$ is true, what is the probability that the observed mean is 315 or more?

# Example 1

- Let $X$ denotes the sample mean estimator; it follows a normal distribution with mean 300 and standard deviation $\frac{20}{\sqrt{15}} = 5.16398$.
- $P(\tilde{X} > 315) = P(U > \frac{315-300}{5.16398}) = P(U > 2.90)$
- $P(\tilde{X} > 315) = 1 - P(U < 2.90) = 1 - 0.9981 = 0.19\%$
- This probability is very low, then $H_0$ can be rejected.

# Example 1

- In general the probability is stated low with the values $1\%, 5\%$ and also $10\%$.
- If we adopt 5%, thus $P(\tilde{X} > X_c) = 0.05$
- $P(\tilde{X} < X_c) = 0.95$ and $P(U < \frac{X_c - 300}{5.16398}) = 0.95$
- $\frac{X_c - 300}{5.16398} = 1.6449$
- $X_c = 308.49$
- If the observed mean is less then 308.49, the null hypothesis is kept; otherwise it is rejected.

# Example 2

Data collection carried out over many years in "Beauce" region have shown that the natural level of rainfall, in mm per year, follows a normal law N(600, 100). In the 1950s, rainmakers claimed to be able to increase the average level of rainfall by 50 mm by inseminating clouds with silver chloride. During the years 1951 to 1959, the process was tested and the following water heights were recorded:

| Year | 1951 | 1952 | 1953 | 1954 | 1955 | 1956 | 1957 | 1958 | 1959 |
|---|---|---|---|---|---|---|---|---|---|
| *mm* | 510 | 614 | 780 | 512 | 501 | 534 | 603 | 788 | 650 |

The question was whether insemination had an effect or not; so two hypotheses were competing: either the process had a definite effect on the average level of rainfall or it had none.

# Example 2

The problem can be formalized as follows: Let $X$ be the random variable equal to the annual rainfall level and let $m$ be its expectation; the two hypotheses in presence can be summarized as:

$H0$: $m = 600mm$

$H1$: $m = 650mm$

They decided that they would adopt the procedure if the result obtained by the measurements was part of an eventuality that had only a 5% chance of occurring. They then assumed the 5% risk of being wrong.

We want to test the average $m$ so we are interested in the average $X$ of the data collected.

# Example 2

Under the hypothesis $H0$ that farmers initially adopt, the sample mean follows a normal distribution:

$$L(\overline{X}) = N(600, \frac{100}{\sqrt{9}}) = N(600, 100/3)$$

The farmers' choice was therefore as follows: If $X$ is too large, that is to say if it is greater than a threshold which has only a 5% chance of being exceeded, we adopt the procedure with a 5% risk of being wrong.

$$P(\overline{X} > k) = 0.05 \Rightarrow P(\frac{\overline{X} - 600}{100/3} > \frac{k - 600}{100/3}) = 0.05$$

$$\Rightarrow P(Z > \frac{k - 600}{100/3}) = 0.05 \Rightarrow \frac{3(k - 600)}{100} = 1.64 \Rightarrow k = 655mm$$

# Example 2

The domain $\{\overline{X} > 655\}$ is called the critical region or $H0$ rejection region.
The complementary set $\{X < 655\}$ is called the acceptance region of $H0$ or more precisely the non-rejection region of $H0$.
The data collected indicated $X = 610.2mm$. The conclusion was therefore to retain $H0$ and reject the procedure.

# Example 2

The rainmakers were maybe right, but we didn't see it. If they were right, the law of $X$ was:

$$N(650, 100/3)$$

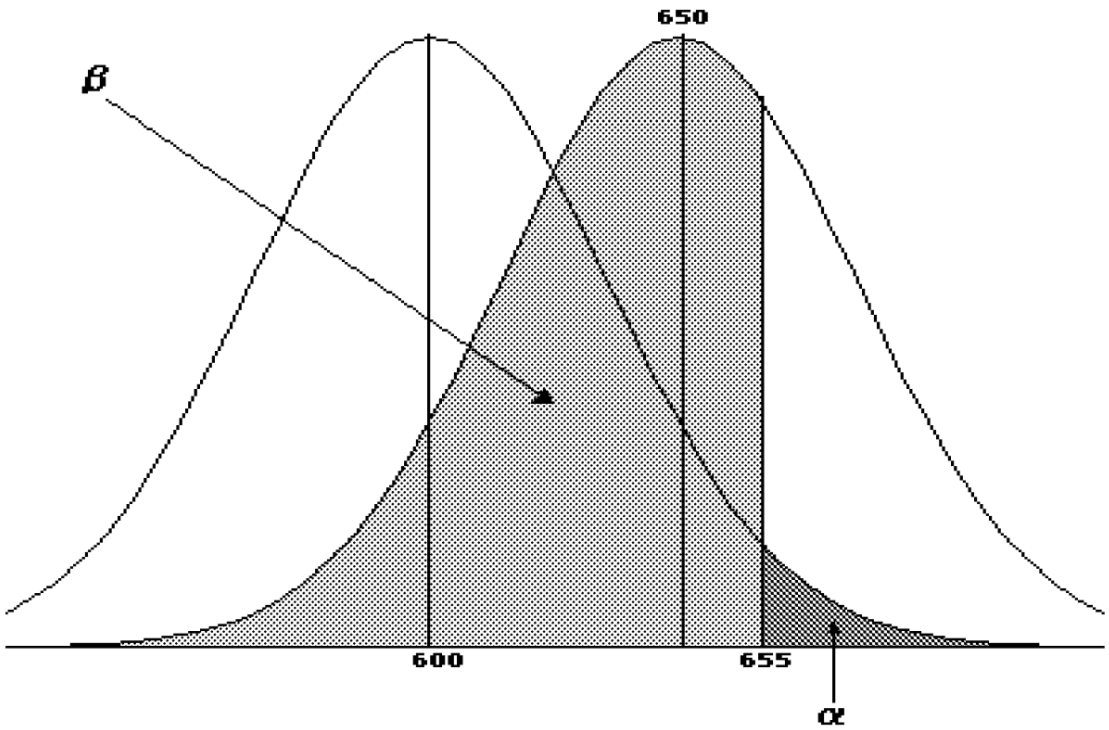
We made a mistake if $X$ was less than 655 mm because, then we chose to keep $H0$ while the makers were right.

$$\beta = P(\frac{\overline{X} - 650}{100/3} < \frac{655 - 650}{100/3}) \Rightarrow \beta = P(Z < 0.15) \Rightarrow \beta = 0.56$$

The probability of making a mistake is high... Usually $\beta \geq 0.20$

- $\alpha$ is called the risk of type I,
- $\beta$ is called the risk of type II.

# Example 2



# Example : propellant burning rate problem

- Suppose that an engineer is designing an air crew escape system that consists of an ejection seat and a rocket motor that powers the seat.
- For the ejection seat to function properly, the propellant should have a mean burning rate of $\mu = 50 cm/sec$ and a standard deviation $\sigma = 2.5 cm/s$.
- Suppose that a sample of $n = 10$ specimens is tested and that the sample mean burning rate $x$ is observed.
- Does the mean burning rate of the propellant equal 50 cm/sec, or is it some other value ?

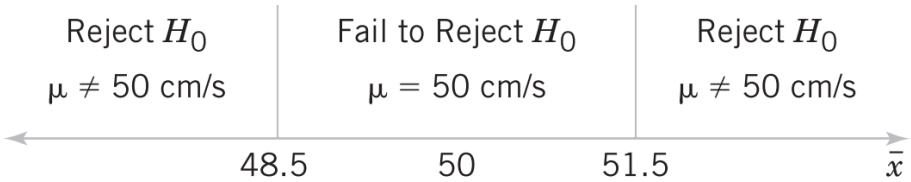
# Statistical Hypothesis

1. Statistical Hypothesis: A statistical hypothesis is a statement about the parameters of one or more populations.
2. Specifically, we are interested in deciding whether or not the mean burning rate is 50 cm per second. We may express this formally as:
   - null hypothesis $H0 : \mu = 50cm/s$,
   - alternative hypothesis $H1 : \mu \neq 50cm/s$.
3. It is called a two-sided alternative hypothesis.
4. In a one sided hypothesis, we could have $H1 : \mu \leq 50cm/s$ or also $H1 : \mu \geq 50cm/s$.

# Statistical Hypothesis: illustration, case 1

- The sample mean can take on many different values.
- Suppose that if $48.5 \leq x \leq 51.5$, we will not reject the null hypothesis $H0 : \mu = 50$, and if either $x < 48.5$ or $x > 51.5$, we will reject the null hypothesis in favor of the alternative hypothesis $H1$: $\mu \neq 50$.

**Decision criteria for testing $H_0$: $\mu = 50$ centimeters per second versus $H_1$: $\mu \neq 50$ centimeters per second.**



# Type I errors

1. Type I Error: Rejecting the null hypothesis $H0$ when it is true is defined as a type I error.
2. Probability of Type I Error: $\alpha = P(\text{type I error}) = P(\text{reject } H0 \text{ when } H0 \text{ is true })$
3. $\alpha$ is the probability of choosing $H1$ when $H0$ is true; it is called first-type risk.
4. The critical region is called the set of values $W$ of the decision variable leading to the rejection of $H0$ in favor of $H1$.

$$P(W|H0) = \alpha \; P(\overline{W}|H0) = 1 - \alpha$$

# Type I errors: illustration, case 1

- Let us compute $\alpha = P(\text{type I error}) = P(\text{reject } H0 \text{ when } H0 \text{ is true })$.

$$\alpha = P(\overline{X} < 48.5 \text{ or } \overline{X} > 51.5 \text{ when } \mu = 50)$$

$$\alpha = P(\overline{X} < 48.5 \text{ when } \mu = 50) + P(\overline{X} > 51.5 \text{ when } \mu = 50)$$

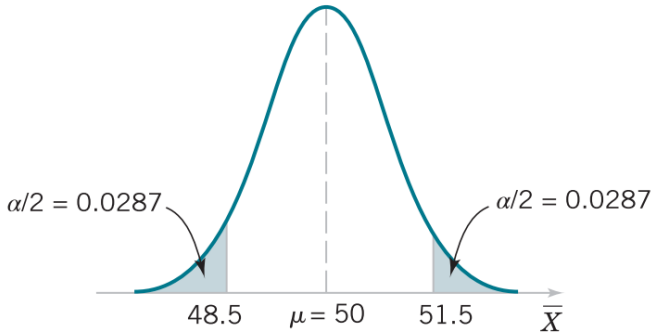
- The $z$-values for the critical values 48.5 and 51.5 are:

$$z_1 = \frac{48.5 - 50}{2.5\sqrt{10}} = -1.9 \text{ and } z_2 = \frac{51.5 - 50}{2.5\sqrt{10}} = +1.9$$

$$\alpha = P(z < -1.9) + P(z > 1.9) = 0.0287 + 0.0287 = 0.0574$$

# Type I errors: illustration, case 1

| Decision | $H_0$ Is True | $H_0$ Is False |
|---|---|---|
| Fail to reject $H_0$ | No error | Type II error |
| Reject $H_0$ | Type I error | No error |




**The critical region for $H_0$: $\mu = 50$ versus $H_1$: $\mu \neq 50$ and $n = 10$.**

This is the type I error probability. This implies that 5.74% of all random samples would lead to rejection of the hypothesis $H_0 : \mu = 50cm/s$ when the true mean burning rate is really $50cm/s$.

# Type I errors: illustration, case 2

- Notice that we can reduce $\alpha$ by widening the acceptance.
- For example, if we make the critical values 48 and 52, the value of $\alpha$ is:

$$\alpha = P(z < \frac{48-50}{2.5\sqrt{10}}) + P(z > \frac{52-50}{2.5\sqrt{10}})$$

$$\alpha = 0.0057 + 0.0057 = 0.0114$$

- If $n = 16$

$$\alpha = P(z < \frac{48-50}{2.5\sqrt{16}}) + P(z > \frac{52-50}{2.5\sqrt{15}})$$

$$\alpha = 0.0082 + 0.0082 = 0.0164$$

# Type II errors

1. Type II Error: Failing to reject the null hypothesis when it is false is defined as a type II error.
2. Probability of Type II Error: $\beta = P(\text{type II error}) = P(\text{fail to reject } H0 \text{ when } H0 \text{ is false})$
3. $\beta$ is the probability of retaining $H0$ while $H1$ is true; it is called risk of the second type.
4. The power of the test is defined by: $P(W|H1) = 1 - \beta$.
5. A power greater than 80% is desirable for the hypothesis test to be satisfactory.

# Type II error: illustration

1. Probability of Type II Error: $\beta = P(\text{type II error}) = P(\text{fail to reject } H0 \text{ when } H0 \text{ is false})$
2. To calculate $\beta$ ($\beta$-error), need a alternative hypothesis: need a particular value of $\mu$.
3. To reject the null hypothesis $H0 : \mu = 50$, $\mu > 52sm/s$ or $\mu < 48sm/s$: could calculate the probability of a type II error $\beta$ for the values $\mu = 52$ and $\mu = 48$ and use it to tell us how the test procedure would perform.
4. Reject $H0$, for a mean value of $\mu = 52$ or $\mu = 48$? (Because of symmetry, it is necessary to evaluate only one of the two cases?say).
5. Probability of failing to reject the null hypothesis $H0 : \mu = 50cm/s$ when the true mean is $\mu = 52cm/s$.

$$\beta = P(48.5 \leq X \leq 51.5 \text{ when } \mu = 52)$$

# Type II error: illustration

1. The $z$-values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$\beta = P(48.5 \leq X \leq 51.5 \text{ when } \mu = 52)$$

$$z_1 = \frac{48.5 - 52}{2.5\sqrt{10}} = -4.43 \text{ and } z_2 = \frac{51.5 - 52}{2.5\sqrt{10}} = -0.63$$

$$\beta = P(-4.43 \leq Z \leq ?0.63) = P(Z \leq ?0.63) - P(Z \leq -4.43)$$

$$\beta = 0.2643 - 0.0000 = 0.2643$$

2. The probability that we will fail to reject the false null hypothesis is 0.2643.

# Type II error: illustration

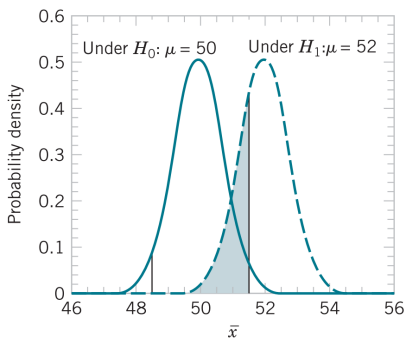
1. The probability of making a type II error $\beta$ increases rapidly as the true value of $\mu$ approaches the hypothesized value.

$$\beta = P(48.5 \leq X \leq 51.5 \text{ when } \mu = 50.5)$$

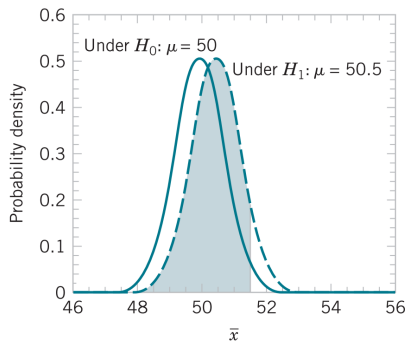
$$z_1 = \frac{48.5 - 50.5}{2.5\sqrt{10}} = -4.43 \text{ and } z_2 = \frac{51.5 - 50.5}{2.5\sqrt{10}} = -0.63$$

$$\beta = P(-2.53 \leq Z \leq 1.27) = P(Z \leq 1.27) - P(Z \leq -2.53)$$

$$\beta = 0.8980 - 0.0057 = 0.8923$$




**The probability of type II error when $\mu = 52$ and $n = 10$.**




**The probability of type II error when $\mu = 50.5$ and $n = 10$.**

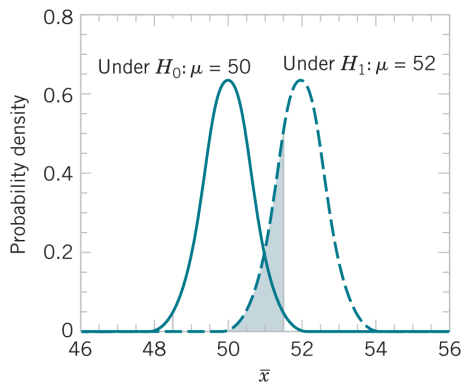
# Type II error: illustration

1. When $n = 16$, the standard deviation of $X$ is $\sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$, and the $z$-values corresponding to 48.5 and 51.5 when $\mu = 52$ are

$$z_1 = \frac{48.5 - 52}{2.5\sqrt{16}} = -5.60 \text{ and } z_2 = \frac{51.5 - 52}{2.5\sqrt{16}} = -0.80$$

$$\beta = P(-5.60 \leq Z \leq -0.80) = P(Z \leq -0.80) - P(Z \leq -5.60)$$

$$\beta = 0.2119 - 0.0000 = 0.2119$$




**The probability of type II error when $\mu = 52$ and $n = 16$.**

# Type II error: illustration

| Acceptance Region | Sample Size | $\alpha$ | $\beta$ at $\mu = 52$ | $\beta$ at $\mu = 50.5$ |
|---|---|---|---|---|
| $48.5 < \bar{x} < 51.5$ | 10 | 0.0576 | 0.2643 | 0.8923 |
| $48 < \bar{x} < 52$ | 10 | 0.0114 | 0.5000 | 0.9705 |
| $48.81 < \bar{x} < 51.19$ | 16 | 0.0576 | 0.0966 | 0.8606 |
| $48.42 < \bar{x} < 51.58$ | 16 | 0.0114 | 0.2515 | 0.9578 |

# Power

**Power**
The **power** of a statistical test is the probability of rejecting the null hypothesis $H_0$ when the alternative hypothesis is true.

The power is computed as $1 - \beta$, and power can be interpreted as *the probability of correctly rejecting a false null hypothesis*. We often compare statistical tests by comparing their power properties. For example, consider the propellant burning rate problem when we are testing $H_0$: $\mu = 50$ centimeters per second against $H_1$: $\mu \neq 50$ centimeters per second. Suppose that the true value of the mean is $\mu = 52$. When $n = 10$, we found that $\beta = 0.2643$, so the power of this test is $1 - \beta = 1 - 0.2643 = 0.7357$ when $\mu = 52$.

- That is, if the true mean is really 52 centimeters per second, this test will correctly reject $H0$: $\mu = 50$ and "detect" this difference 73.57% of the time.
- If this value of power is judged to be too low, the analyst can increase either $\alpha$ or the sample size $n$.

# Statistical Decision and Error Probabilities

| | *True state of* | *null hypothesis* |
|---|---|---|
| *Statistical decision* | $H_0$ true | $H_0$ false |
| Do not reject $H_0$ | Correct decision $(1-\alpha)$ | Type II error $(\beta)$ |
| Positive to $H_0$ | True Positive | False Positive |
| Reject $H_0$ | Type I error $(\alpha)$ | Correct decision $(1-\beta)$ |
| Negative to $H_0$ | False Negative | True Negative |

# P-Value, Statical significance

**P-Value**

The ***P*-value** is the smallest level of significance that would lead to rejection of the null hypothesis $H_0$ with the given data.

The Statical significance (p-value) is defined by:

$$p = \mathbb{P}(Hypothesis \mid H_0)$$

| | |
|---|---|
| One sided right test | $p = \mathbb{P}(T > t \vert H_0)$ |
| One sided left test | $p = \mathbb{P}(T < t \vert H_0)$ |
| Two sided test | $p = \mathbb{P}(\vert T \vert > \vert t \vert \vert H_0)$ |

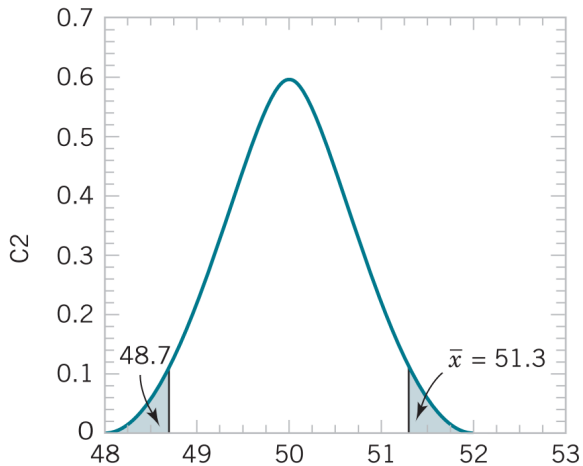
- The p-value is the probability to get the statistical test at least as extremal as the observed one when $H_0$ is true.
- In practice, we reject $H_0$ when $p < \alpha$.

# P-Value: illustration

- Consider the two-sided hypothesis test with $n = 16$ and $\sigma = 2.5$

$$H0 : \mu = 50 H1 : \mu \neq 50$$

- Suppose that the observed sample mean is $\overline{x} = 51.3 cm/s$ and the symmetric value 48.7:
- The P-value of the test is the probability above 51.3 plus the probability below 48.7




***P*-value is the area of the shaded region when $\bar{x} = 51.3$.**

# P-Value: illustration

$$\begin{aligned}\text{P-value} &= P(48.7 < \overline{X} < 51.3 \\ &= 1 - P(\tfrac{48.7-50}{2.5/\sqrt{16}} < Z < \tfrac{51.3-50}{2.5/\sqrt{16}}) \\ &= 1 - P(-2.08 < Z < 2.08) = 1 - 0.962 = 0.038\end{aligned}$$

- The P-value tells us that if the null hypothesis $H0 = 50$ is true, the probability of obtaining a random sample whose mean is at least as far from 50 as 51.3 (or 48.7) is 0.038.
- Therefore, an observed sample mean of 51.3 is a fairly rare event if the null hypothesis $H0 = 50$ is really true.
- Compared to the "standard" level of significance 0.05, our observed P-value is smaller, so if we were using a fixed significance level of 0.05, the null hypothesis would be rejected.
- P-value provides a measure of the credibility of the null hypothesis.
- It is the risk that we have made an incorrect decision if we reject the null hypothesis $H0$.

# T-test for means

We can formulate the null hypothesis and the alternative hypothesis for one sample as follows:

$$H_0 : \mu = m \quad H_1 : \mu \neq m$$

or

$$H_0 : \mu = m \quad H_1 : \mu < m$$

or

$$H_0 : \mu = m \quad H_1 : |\mu| > |m|$$

The two-sample independent t-test:

$$H_0 : \mu_A - \mu_B = 0 \quad H_1 : \mu_A - \mu_B \neq 0$$

...

We reject $H_0$ when $p - value < \alpha$, ($\alpha$ equal to 0.05 or 0.01).

# Principals

- Non parametric tests do not make any hypothesis on the probability distribution .
- When the data are quantitative, non parametric tests are performed by raking the data values.
- When the are qualitative, only non parametric tests are possible.

We study two popular tests for testing hypotheses about the population location, or median using the sign test and the Wilcoxon signed rank test.

# Non parametric tests: Sign test

Let $M$ be the median of a certain population. Then we know that

$$P(X \leq M) = 0.5 = P(X \geq M).$$

We consider the problem of testing the null hypothesis

$$H_0 : M = m_0 \text{ versus } H_a : M > m_0.$$

# Non parametric tests: Wilcoxon signed rank test

In the sign test, we have considered only whether each observation is greater than $m_0$ or less than $m_0$ without giving any importance to the magnitude of the difference from $m_0$.
An improved version of the sign test is the Wilcoxon signed rank test, in which one replaces the observations by their ranks of the ordered magnitudes of differences, $|x_i - m_0|$. The smallest observation is ranked as 1, the next smallest will be 2, and so on.

# Chi-squared test

A chi-squared test (also chi-square or $\chi^2$ test) is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis.
Chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.
Example: Rolling a die 600 times in a row gave the following results:

| number | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| cardinality | 88 | 109 | 107 | 94 | 105 | 97 |

We want to test the hypothesis that the die is not rigged, with a risk $\alpha = 0.05$.
$p - value = 0.6325$, no difference, thus the die is not rigged.

# ANOVA

It allows you to compare several averages at the same time. Using hypothesis testing terminology:

$$H_0 : \mu_1 = ... = \mu_L$$

$$H_a : \mu_1 \neq ... \neq \mu_L$$

# Linear Regression

We are looking for the best hypothesis function that will approximate the input data:

$$\text{Input } x \xrightarrow[\Longrightarrow]{\text{hypothèse } h} \text{Output } y.$$

We use several variables as input to the problem, which constitute as many degrees of freedom for the function $h$ to best approximate the input data.
With these new assumptions, $h$ takes a more general form for $n$ input variables:

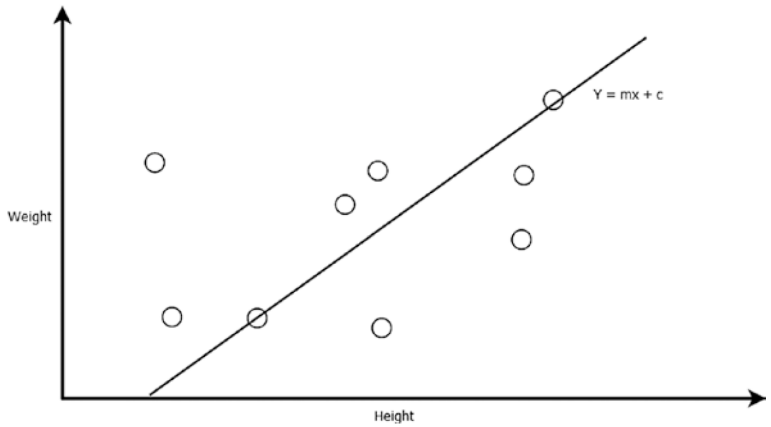
$$h(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n.$$

# Linear Regression

The input data is represented as a matrix of dimension $m \times n$:

$$X = \begin{pmatrix} X_{1,1}, \ldots, X_{1,n} \\ \ldots \\ X_{m,1}, \ldots, X_{m,n} \end{pmatrix}$$

$X_{i,j}$ corresponds to the value taken by the variable $j$ of the observation $i$.




*LInear Regressions*

# Linear Regression

The cost function is:

$$J(\theta) = \frac{1}{2m} \sum_{i=1..m} (h(x_i) - y_i)^2 \tag{1}$$

It is also necessary to normalize the data $X$ via the standard formula:

$$X_{std} = \frac{X - min(X)}{max(X) - min(X)}.$$

The aim is to find the min of the function $J$.

# Multiple Linear Regressions

A simple linear regression is for a single response variable, y, and a single independent variable, x. The equation for a simple linear regression is

$$y = b_0 + b_1 x$$

Multiple linear regression is built from a simple linear regression. Multiple linear regression is used when you have more than one independent variable. The equation of a multiple linear regression is

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \epsilon$$

When you have n observations or rows in the data set, you have the following model:

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \ldots + b_k x_{1k} + \epsilon_1$$

$$y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + \ldots + b_k x_{2k} + \epsilon_2$$

$$y_3 = b_0 + b_1 x_{31} + b_2 x_{32} + \ldots + b_k x_{3k} + \epsilon_3$$

$$\ldots$$

$$y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + \ldots + b_k x_{nk} + \epsilon_n$$

# Multiple Linear Regressions

Using a matrix, you can represent the equations as

$$y = Xb + \epsilon$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix} \quad and \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

To calculate the coefficients:

$$\hat{b} = \left(X'X\right)^{-1} X'y$$

# References


- http://www.jybaudot.fr/a_general/indexstats.html
- https://www.gerad.ca/Sebastien.Le.Digabel/MTH2302D/
- https://www.math.univ-toulouse.fr/~besse/Wikistat/

# Elements of a statistical hypothesis

- The null hypothesis, denoted by $H_0$, is usually the nullification of hypothesis/claim (or population property).
- The alternative hypothesis, denoted by $H_a$ (or sometimes denoted by $H_1$), is customarily the hypothesis/claim itself (also rejecting population property).
- The test statistic, denoted by $TS$, is a function of the sample measurements upon which the statistical decision, to reject or not to reject the null hypothesis, will be based.

# Statistical Decision and Error Probabilities

- **Type I error ($\alpha$)** : $H_0$ is rejected when in fact $H_0$ is true.

$$\mathbb{P}(\textit{rejecting } H_0 | H_0 \textit{ true})$$

- **Type II error ($\beta$)** : $H_0$ is accepted when in fact $H_a$ is true.

$$\mathbb{P}(\textit{not rejecting } H_0 | H_0 \textit{ false})$$

- **Statical significance** : risk of type I $\alpha$.
- Statical significance ($\alpha$) of the test usually fixed to 0.05 (or 0.01).

# Elements of a statistical hypothesis

- A rejection region $RR$ (or a critical region) $\Gamma$ is the region that specifies the values of the observed $TS$ for which the null hypothesis will be rejected. This is the range of values of the $TS$ that corresponds to the rejection of $H_0$ at some fixed level of significance, $\alpha$, which will be explained later.
- Conclusion: If the value of the observed $TS$ falls in the $RR$, the null hypothesis is rejected and we will conclude that there is enough evidence to decide that the alternative hypothesis is true. If the $TS$ does not fall in the $RR$, we conclude that we cannot reject the null hypothesis.

# Statical significance (p-value)

The Statical significance (p-value) is defined by:

$$p = min\{\alpha | T \in \Gamma_\alpha\}$$

| | |
|---|---|
| One sided right test | $p = \mathbb{P}(T > t \| H_0)$ |
| One sided left test | $p = \mathbb{P}(T < t \| H_0)$ |
| Two sided test | $p = \mathbb{P}(\|T\| > \|t\| \| H_0)$ |

- The p-value is the probability to get the statistical test at least as extremal as the observed one when $H_0$ is true.
- In practice, we reject $H_0$ when $p < \alpha$.

# Statical significance (p-value)

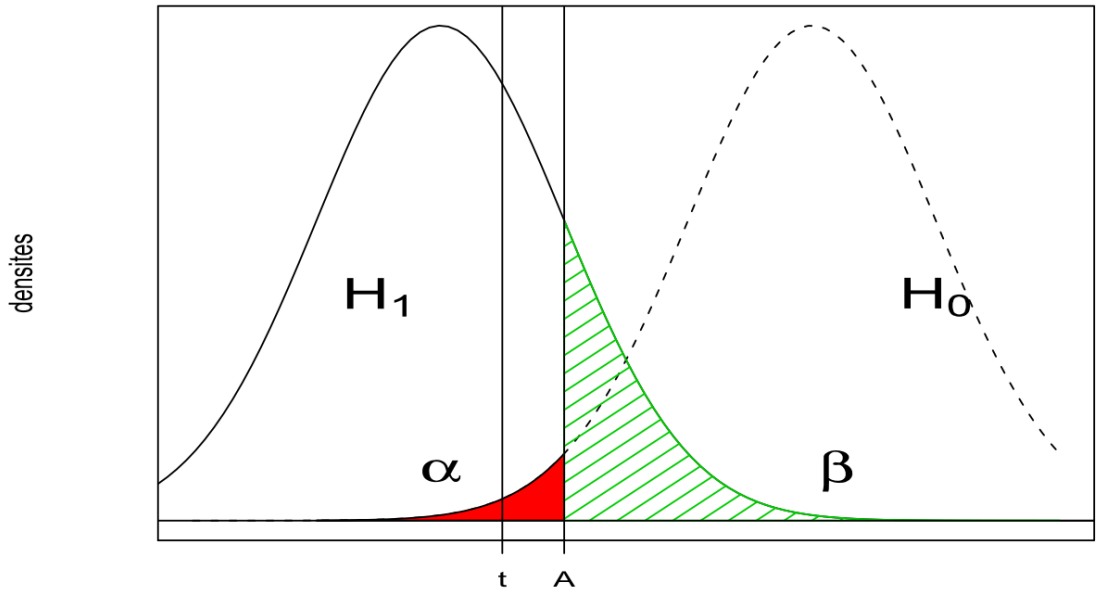



Figure: Type I and II error and critical region of shape, $\Gamma = ]-\infty, A]$

# Statical significance (p-value)

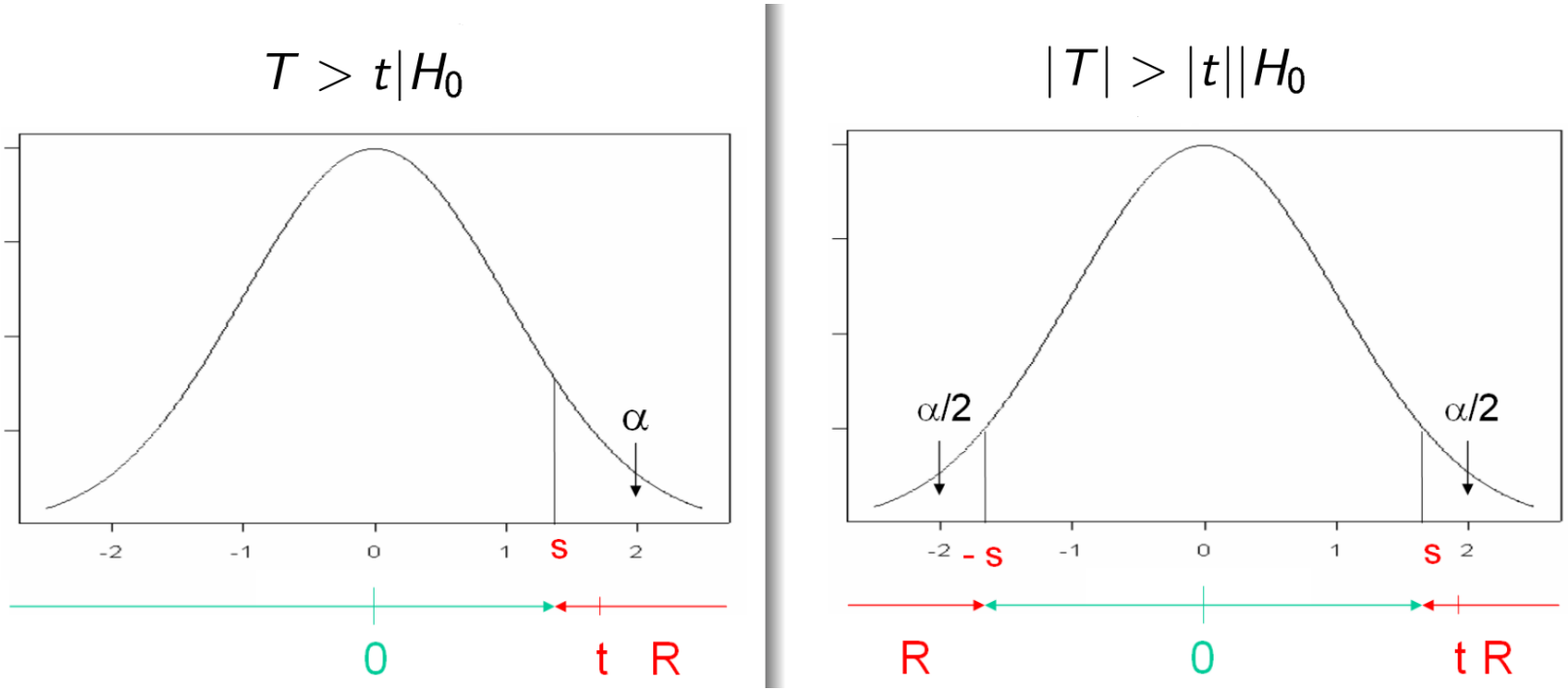



Figure: One sided right test vs Two sided test

# Statical significance (p-value)

In hypothesis testing, the choice of the value of $\alpha$ is somewhat arbitrary. For the same data, if the test is based on two different values of $\alpha$, the conclusions could be different. Many statisticians prefer to compute the so-called *p*-value, which is calculated based on the observed test statistic. For computing the *p*-value, it is not necessary to specify a value of $\alpha$. We can use the given data to obtain the *p*-value.

# Statical significance (p-value)

Corresponding to an observed value of a test statistic, the *p*-value (or attained significance level) is the lowest level of significance at which the null hypothesis would have been rejected.

**STEPS TO FIND THE *p*-VALUE**

1. Let TS be the test statistic.
2. Compute the value of TS using the sample $X_1, \ldots, X_n$. Say it is $a$.
3. The $p$-value is given by

$$p\text{-}value = \begin{cases} P(TS < a|H_0), & \text{if lower tail test} \\ P(TS > a|H_0), & \text{if upper tail test} \\ P(|TS| > |a||H_0), & \text{if two tail test.} \end{cases}$$

# Statical significance (p-value) / Example

The management of a local health club claims that its members lose on the average 15 kDA or more within the first 3 months after joining the club. To check this claim, a consumer agency took a random sample of 45 members of this health club and found that they lost an average of 13.8 kDA within the first 3 months of membership, with a sample standard deviation of 4.2 kDA.

- Find the $p$-value for this test.
- Based on the $p$-value in this case, would you reject the null hypothesis at $\alpha = 0.01$?

# Statical significance (p-value) / Example

a) Let $X$ be the weight loss. Let $\mu$ be the true mean weight loss in kDA within the first 3 months of membership in this club. Then we have to test the hypothesis:

$$H_0 : \mu = 15 \text{ versus } H_a : \mu < 15.$$

Here $n = 45, \overline{x} = 13.8$, and $s = 4.2$. Because $n = 45 > 30$, we can use normal approximation $X \sim N(\mu, \sigma/\sqrt{n})$.

# Statical significance (p-value) / Example

From $H_a : \mu < 15$:

$$p - value = P(X < 13.8) = 0.0274$$

b) No. Because the $p - value = 0.0274$ is greater than $\alpha = 0.01$, one cannot reject $H_0$. ($H_a$ is probable under $H_0$)
Suppose $\alpha = 0.05$: we reject $H_0$ ($H_a$ is not probable under $H_0$)

# Statical significance (p-value) / Example

To compute $P(X < 13.8)$ we should transform $X \sim N(\mu, \sigma/\sqrt{n})$ to the standard form $Z \sim N(0,1)$, where

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

For $X = 13.8$: $Z = \frac{X-\mu}{\sigma/\sqrt{n}} = \frac{13.8-15}{4.2/\sqrt{45}} = -1.9166$.

$$p-value = P(X < 13.8) = P(Z < -1.9166) \sim P(Z < -1.92) = 0.0274$$

# Statical significance (p-value) / Example 2

It is claimed that sports-car owners drive on average 18000 kms per year. A consumer firm believes that the average mileage is probably lower. To check, the consumer firm obtained information from 40 randomly selected sports-car owners that resulted in a sample mean of 17463 kms with a sample standard deviation of 1348 kms. What can we conclude about this claim? Use $\alpha = 0.01$. What is the $p$ value?

# Statical significance (p-value) / Example 2 / Solution

Let $\mu$ be the true population mean. We can formulate the hypotheses as $H0 : \mu = 18,000$ versus $Ha : m < 18000$. The observed $TS$ (for $n > 30$) is: For $X = 17463$:

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}} = \frac{17463 - 18000}{1348/\sqrt{40}} = -2.52$$

. $RR$ is $\{z \leq -Z_{0.01}\} = \{z < -2.33\}$.
Because $z = -2.52$ is less than $-2.33$, the null hypothesis is rejected at $\alpha = 0.01$. There is sufficient evidence to conclude that the mean mileage on sports cars is less than 18 000 miles per year. The $p$ value $= P(z < -2.52) = 0.0059$. This $p$ value is less than 0.01 and also supports rejection of the null hypothesis.

# Statical significance (p-value) / Example 3

In a salary equity study of faculty at a certain university, sample salaries of 50 male professors and 50 female professors yielded the following basic statistics.

| | Sample mean salary | Sample stand. dev. |
|---|---|---|
| Male professor | 46 400 | 360 |
| Female professor | 46 000 | 220 |

Test the hypothesis that the mean salary of male assistant professors is more than the mean salary of female assistant professors at this university. Use $\alpha = 0.05$.

# Statical significance (p-value) / Example 3

Let $\mu_1$ be the true mean salary for male assistant professors and $\mu_2$ be the true mean salary for female assistant professors at this university. To test: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 > 0$.

$$\text{the TS is: } z = \frac{\overline{x_1} - \overline{x_2} - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{46400 - 46000}{\sqrt{\frac{360^2}{50} + \frac{220^2}{50}}} = 6704$$

The *RR* for a $\alpha = 0.05$ is $\{z > 1645\}$. Because $z = 6704 > 1645$, we reject the null hypothesis at $\alpha = 0.05$. We conclude that the salary of male assistant professors at this university is higher than that of female assistant professors for $\alpha = 0.05$. Note that even though $\sigma_1^2$ and $\sigma_2^2$ are unknown, because $n_1 \geq 30$ and $n_2 \geq 30$, we could replace $\sigma_1$ and $\sigma_2$ with the respective sample variances. We are assuming that the salaries of male and female assistant professors are sampled independent of each other.

# Non parametric tests: Sign test

A test at a significance level $\alpha$ will reject $H_0$ if $n^+ \geq k$, where $k$ is chosen such that

$$P(N^+ \geq k \text{ when } M = m_0) = \alpha.$$

Similarly, if the alternative is of the form $H_a : M \neq m_0$, the critical region is of the form $N^+ \leq k$ or $N^+ \geq k_1$, where $P(N^+ \leq k) + P(N^+ \geq k_1) = \alpha$.

# Non parametric tests: Sign test

In order to determine such a $k$ and $k_1$, we need to determine the distribution of $N^+$. The test works on the principle that if the sample were to come from a population with a continuous distribution, then each of the observations falls above the median or below the median with probability $\frac{1}{2}$. Hence, the number of sample values falling below the median follows a binomial distribution with parameters $n$ and $p = 1/2$, $n$ being the sample size. If a sample value equals the hypothesized median $m_0$, that observation will be discarded and the sample size will be adjusted accordingly (we remark that such values should be very few). Thus, when $H_0$ is true, $N^+$ will have a binomial distribution with parameters $n$ and $p = 1/2$. For this reason, some authors call this test the binomial test. The following procedure summarizes the test statistic and the corresponding critical regions.

# Non parametric tests: Sign test

**SIGN TEST**

$$H_0 : M = m_0$$

| Alternative hypothesis | Critical region |
|---|---|
| $H_a : M > m_0$ | $N^+ \geq k$, where $\sum_{i=k}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n = \alpha$ |
| $H_a : M < m_0$ | $N^+ \leq k$, where $\sum_{i=0}^{k} \binom{n}{i} \left(\frac{1}{2}\right)^n = \alpha$ |
| and | |
| $H_a : M \neq m_0$ | $N^+ \leq k_1$, where $\sum_{i=k_1}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n = \frac{\alpha}{2}$ or $N^+ \leq k$, where $\sum_{i=0}^{k} \binom{n}{i} \left(\frac{1}{2}\right)^n = \frac{\alpha}{2}$. |

If $\alpha$ or $\alpha/2$ cannot be achieved exactly, choose $k$ (or $k$ and $k_1$) so that the probability comes as close to $\alpha$ (or $\alpha/2$) as possible.

# Non parametric tests: Sign test

**Hypothesis-testing procedure using the sign test**

We test

$$H_0 : M = m_0 \text{ vs. } H_1 : M > m_0.$$

1. Replace each value of the observation that is greater than $m_0$ by a plus sign and each sample value less than $m_0$ by a minus sign. If the sample value is equal to $m_0$, discard the observation and adjust the sample size $n$ accordingly.
2. Let $n^+$ be the number of $+$'s in the sample. For $n$ and $p = \frac{1}{2}$, from the binomial table, find

$$\gamma = P(N^+ \geq n^+).$$

3. **Decision:** If $\gamma$ is less than $\alpha$, $H_0$ must be rejected. Based on the sample, we will conclude that the median of the population is greater than $m_0$ at the significance level $\alpha$. Otherwise do not reject $H_0$.

**Assumptions:** The population distribution is continuous. The number of ties is small (less than 10% of the sample).

The $p$ value is computed from its definition given by the formula:

$$p \text{ value} = P(N^+ \geq n^+) = \sum_{i=k..n} \binom{n}{i} (1/2)^n = \gamma.$$

# Non parametric tests: Sign test - Example

For the given data from an experiment

$$1.51 \quad 1.35 \quad 1.69 \quad 1.48 \quad 1.29 \quad 1.27 \quad 1.54 \quad 1.39 \quad 1.45$$

test the hypothesis that $H_0$: $M = 1.4$ versus $H_a$: $M > 1.4$ at $\alpha = 0.05$.

**Solution**
*We test*

$$H_0 : M = 1.4 \ \textit{versus} \ H_a : M > 1.4.$$

*Replacing each value greater than 1.4 with a plus sign and each value less than 1.4 with a minus sign, we have*

$$+ - + + - - + - +.$$

*Thus,* $\mathrm{n}^+ = 5$. *From the binomial table with* $\mathrm{n} = 9$ *and* $\mathrm{p} = \frac{1}{2}$, *we have*

$$P(N^+ \geq 5) = 0.50.$$

*Hence, the* p *value is 0.5. Because* $\alpha = 0.05 < 0.50$, *the null hypothesis is not rejected. We conclude that the median does not exceed* 1.4.

# Non parametric tests: Sign test

When the sample size $n$ is large, we can apply the normal approximation to the binomial distribution. That is, the test statistic $N^+$ is approximately normally distributed. Thus, under $H_0$, $N^+$ will have approximate normal distribution with mean $np = n/2$ and variance of $np(1-p) = n/4$. By the $z$-transform, we have

$$Z = \frac{N^+ - n/2}{\sqrt{n/4}} = \frac{2N^+ - n}{\sqrt{n}} \sim N(0,1).$$

We could utilize this test if $n$ is large, that is, if $np \geq 5$ and $n(1-p) \geq 5$. Hence, under $H_0$, because $p = 1/2$, if $n \geq 10$, we could use the large sample test. The following table summarizes the method for a large sample sign test.

# Non parametric tests: Wilcoxon signed rank test

**Hypothesis testing procedure using Wilcoxon signed rank test**

We wish to test

$$H_0 : M = m_0 \text{ versus } H_1 : M \neq m_0.$$

1. Compute the absolute differences $z_i = |x_i - m_0|$ for each observation. Replace each value of the observation that is greater than $m_0$ by a plus sign and each sample value that is less than $m_0$ by a minus sign. If the sample value is equal to $m_0$, discard the observation and adjust the sample size $n$ accordingly.
2. Assign each $z_i$ a value equal to its rank. If two values of $z_i$ are equal, assign each $z_i$ a rank equal to the average of the ranks each should receive if there were not a tie.
3. Let $W^+$ be the sum of the ranks associated with plus signs and $W^-$ be the sums of ranks with negative signs.
4. **Decision:** If $m_0$ is the true median, then the observations should be evenly distributed about $m_0$. For a given $\alpha$ critical region, reject $H_0$ if

$$W^+ \leq c_1, \text{ where } P(W^+ \leq c_1) = \frac{\alpha}{2},$$

or

$$W^+ \geq c_2, \text{ where } P(W^+ \geq c_2) = \frac{\alpha}{2}.$$

**Assumptions:** The population distribution is continuous and symmetrical. The number of ties is small, less than 10% of the sample size.

# Non parametric tests: Wilcoxon signed rank test - Example

For the given data that resulted from an experiment

$$1.51 \quad 1.35 \quad 1.69 \quad 1.48 \quad 1.29 \quad 1.27 \quad 1.54 \quad 1.39 \quad 1.45$$

test the hypothesis that $H_0$: $M = 1.4$ versus $H_a$: $M \neq 1.4$. Use $\alpha = 0.05$.

**Solution**
*We wish to test*

$$H_0 : M = 1.4 \ \textit{versus} \ H_a : M \neq 1.4.$$

*Here,* $\alpha = 0.05$, *and* $m_0 = 1.4$. *The results of steps 1 to 3 are given in* Table 12.1.

*Thus, we have* $W^+ = 29$ *and* $n = 9$. *From the Wilcoxon signed-rank test table in the appendix, we should reject* $H_0$ *if* $W^+ \leq 6$ *or* $W^+ \geq 38$ *with actual level of* $\alpha = 0.054$. *Because* $W^+ = 29$ *does not fall in the rejection region, we do not reject the null hypothesis that* $M = 1.4$.

**TABLE 12.1** Data Summary for Wilcoxon Signed Rank Test.

| $x_i$ | $z_i = \lvert x_i - 1.4 \rvert$ | Sign | Rank |
|---|---|---|---|
| 1.51 | 0.11 | + | 5.5 |
| 1.35 | 0.05 | − | 3 |
| 1.69 | 0.29 | + | 9 |
| 1.48 | 0.08 | + | 4 |
| 1.29 | 0.11 | − | 5.5 |
| 1.27 | 0.13 | − | 7 |
| 1.54 | 0.14 | + | 8 |
| 1.39 | 0.01 | − | 1.5 |
| 1.45 | 0.01 | + | 1.5 |

# References

https://analystprep.com/study-notes/actuarial-exams/soa/p-probability/univariate-random-variables/explain-and-calculate-expected-value-mode-median-percentile-and-higher-moments/

https://math.unm.edu/~james/hw6_sol.pdf

https://www.investopedia.com/articles/financial-theory/11/calculating-covariance.asp

# Introduction

Ref:
https://www.geeksforgeeks.org/naive-bayes-classifiers/
Naive Bayes classifiers are a collection of classification algorithms based on Bayes? Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
To start with, let us consider a dataset.
One of the most simple and effective classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities.

# Motivating example

Consider a fictional dataset that describes the weather conditions for running (sport).

| | Outlook | Temperature | Humidity | Windy | Running |
|---|---|---|---|---|---|
| 0 | Rainy | Hot | High | False | No |
| 1 | Rainy | Hot | High | True | No |
| 2 | Overcast | Hot | High | False | Yes |
| 3 | Sunny | Mild | High | False | Yes |
| 4 | Sunny | Cool | Normal | False | Yes |
| 5 | Sunny | Cool | Normal | True | No |
| 6 | Overcast | Cool | Normal | True | Yes |
| 7 | Rainy | Mild | High | False | No |
| 8 | Rainy | Cool | Normal | False | Yes |
| 9 | Sunny | Mild | Normal | False | Yes |
| 10 | Rainy | Mild | Normal | True | Yes |
| 11 | Overcast | Mild | High | True | Yes |
| 12 | Overcast | Hot | Normal | False | Yes |
| 13 | Sunny | Mild | High | True | No |

# Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

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

# Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

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

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$\mathbb{P}(y|X) = \frac{\mathbb{P}(y)\mathbb{P}(X|y)}{\mathbb{P}(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:

$$X = (x_1, x_2, x_3, ....., x_n)$$

# Example

Just to clear, an example of a feature vector and corresponding class variable can be: (refer 1st row of dataset)

- X = (Rainy, Hot, High, False)
- y = ?

$P(y|X)$ means the probability of "Not playing golf" given that the weather conditions are "Rainy outlook", "Temperature is hot", "high humidity" and "no wind".

# Naive Bayes approach

Now, if any two events A and B are independent, then,

$$P(A, B) = P(A)P(B)$$

# Naive Bayes approach

Now, if any two events A and B are independent, then,

$$P(A, B) = P(A)P(B)$$

Hence, we reach to the result:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

which can be expressed as:

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

# Naive Bayes approach

Now, if any two events A and B are independent, then,

$$P(A, B) = P(A)P(B)$$

Hence, we reach to the result:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

which can be expressed as:

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, as the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

# Naive Bayes approach

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

# Naive Bayes approach

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

So, finally, we are left with the task of calculating $P(y)$ and $P(x_i|y)$.

# Naive Bayes approach

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

So, finally, we are left with the task of calculating $P(y)$ and $P(x_i|y)$.
Please note that P(y) is also called class probability and $P(x_i|y)$ is called conditional probability.

# Naive Bayes approach

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

So, finally, we are left with the task of calculating $P(y)$ and $P(x_i|y)$.
Please note that P(y) is also called class probability and $P(x_i|y)$ is called conditional probability.
The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i|y)$.

# Naive Bayes approach illustration

**Outlook**

| | Yes | No | P(yes) | P(no) |
|---|---|---|---|---|
| Sunny | 3 | 2 | 2/9 | 3/5 |
| Overcast | 4 | 0 | 4/9 | 0/5 |
| Rainy | 3 | 2 | 3/9 | 2/5 |
| **Total** | 9 | 5 | 100% | 100% |

**Temperature**

| | Yes | No | P(yes) | P(no) |
|---|---|---|---|---|
| Hot | 2 | 2 | 2/9 | 2/5 |
| Mild | 4 | 2 | 4/9 | 2/5 |
| Cool | 3 | 1 | 3/9 | 1/5 |
| **Total** | 9 | 5 | 100% | 100% |

**Humidity**

| | Yes | No | P(yes) | P(no) |
|---|---|---|---|---|
| High | 3 | 4 | 3/9 | 4/5 |
| Normal | 6 | 1 | 6/9 | 1/5 |
| **Total** | 9 | 5 | 100% | 100% |

**Wind**

| | Yes | No | P(yes) | P(no) |
|---|---|---|---|---|
| False | 6 | 2 | 6/9 | 2/5 |
| True | 3 | 3 | 3/9 | 3/5 |
| **Total** | 9 | 5 | 100% | 100% |

| Play | | P(Yes)/P(No) |
|---|---|---|
| Yes | 9 | 9/14 |
| No | 5 | 5/14 |
| **Total** | 14 | 100% |

Figure: $P(x_i|y_j)$ for each $x_i$ in $X$ and $y_j$ in $y$.

# Naive Bayes approach illustration

Let us test it on a new set of features (let us call it today):

- today = (Sunny, Hot, Normal, False)

# Naive Bayes approach illustration

Let us test it on a new set of features (let us call it today):

- today = (Sunny, Hot, Normal, False)

$$P(Yes|today) = \frac{\begin{pmatrix} P(SunnyOutlook|Yes)P(HotTemperature|Yes) \\ P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes) \end{pmatrix}}{P(today)}$$

# Naive Bayes approach illustration

Let us test it on a new set of features (let us call it today):

- today = (Sunny, Hot, Normal, False)

$$P(Yes|today) = \frac{\left( \begin{array}{c} P(SunnyOutlook|Yes)P(HotTemperature|Yes) \\ P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes) \end{array} \right)}{P(today)}$$

and probability to not play golf is given by:

$$P(No|today) = \frac{\left( \begin{array}{c} P(SunnyOutlook|No)P(HotTemperature|No) \\ P(NormalHumidity|No)P(NoWind|No)P(No) \end{array} \right)}{P(today)}$$

# Naive Bayes approach illustration

Since, $P(today)$ is common in both probabilities, we can ignore $P(today)$ and find proportional probabilities as:

$$P(Yes|today) \propto \frac{3}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.02116$$

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

$$P(Yes|today) + P(No|today) = 1$$

# Naive Bayes approach illustration

Since, $P(today)$ is common in both probabilities, we can ignore $P(today)$ and find proportional probabilities as:

$$P(Yes|today) \propto \frac{3}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.02116$$

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

$$P(Yes|today) + P(No|today) = 1$$

These numbers can be converted into a probability by making the sum equal to 1 (normalization):

$$P(Yes|today) = \frac{0.02116}{0.02116 + 0.0068} \approx 0.0237$$

$$P(No|today) = \frac{0.0068}{0.0141 + 0.0068} \approx 0.33$$

$$P(Yes|today) > P(No|today)$$

# Naive Bayes approach illustration

Since, $P(today)$ is common in both probabilities, we can ignore $P(today)$ and find proportional probabilities as:

$$P(Yes|today) \propto \frac{3}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.02116$$

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

$$P(Yes|today) + P(No|today) = 1$$

These numbers can be converted into a probability by making the sum equal to 1 (normalization):

$$P(Yes|today) = \frac{0.02116}{0.02116 + 0.0068} \approx 0.0237$$

$$P(No|today) = \frac{0.0068}{0.0141 + 0.0068} \approx 0.33$$

$$P(Yes|today) > P(No|today)$$

So, prediction that running would be played is 'Yes'.

# Annex: Combinatorics / Multiplication

The goal combinatorics analysis (enumeration techniques) is to learn how to count the number of elements of a finite set.

# Annex: Combinatorics / Multiplication

The goal combinatorics analysis (enumeration techniques) is to learn how to count the number of elements of a finite set.
We suppose that an experiment is the succession of $m$ sub-experiments. If the $i$th experiment has $n_i$ possible results for $i = 1, ..., n$, then the number total possible outcomes of the overall experiment is

$$n = \prod_{i=1..m} n_i = n_1 n_2 ... n_m.$$

# Annex: Combinatorics / Multiplication

The goal combinatorics analysis (enumeration techniques) is to learn how to count the number of elements of a finite set.
We suppose that an experiment is the succession of $m$ sub-experiments. If the $i$th experiment has $n_i$ possible results for $i = 1, ..., n$, then the number total possible outcomes of the overall experiment is

$$n = \prod_{i=1..m} n_i = n_1 n_2 ... n_m.$$

Example : You are purchasing a suitcase with a 4-digit code. How many possibilities do you have to choose a code?

# Annex: Combinatorics / Multiplication

The goal combinatorics analysis (enumeration techniques) is to learn how to count the number of elements of a finite set.
We suppose that an experiment is the succession of $m$ sub-experiments.
If the $i$th experiment has $n_i$ possible results for $i = 1, ..., n$, then the number total possible outcomes of the overall experiment is

$$n = \prod_{i=1..m} n_i = n_1 n_2 ... n_m.$$

Example : You are purchasing a suitcase with a 4-digit code. How many possibilities do you have to choose a code?
Answer: $m = 4$,

$$n_1 = 10, n_2 = 10, n_3 = 10, n_4 = 10$$

# Annex: Combinatorics / Multiplication

The goal combinatorics analysis (enumeration techniques) is to learn how to count the number of elements of a finite set.
We suppose that an experiment is the succession of $m$ sub-experiments.
If the $i$th experiment has $n_i$ possible results for $i = 1, ..., n$, then the number total possible outcomes of the overall experiment is

$$n = \prod_{i=1..m} n_i = n_1 n_2 ... n_m.$$

Example : You are purchasing a suitcase with a 4-digit code. How many possibilities do you have to choose a code?
Answer: $m = 4$,

$$n_1 = 10, n_2 = 10, n_3 = 10, n_4 = 10$$

so the total number of possible codes is

$$10 \cdot 10 \cdot 10 \cdot 10 = 10^4.$$

# Annex: Combinatorics / Permutations

A permutation of $n$ distinct elements $e_1, ..., e_n$ is an ordered rearrangement, without repetition of these $n$ elements.

# Annex: Combinatorics / Permutations

A permutation of $n$ distinct elements $e_1, ..., e_n$ is an ordered rearrangement, without repetition of these $n$ elements.
Example: "a", "b" and "c" are three elements.

# Annex: Combinatorics / Permutations

A permutation of $n$ distinct elements $e_1, ..., e_n$ is an ordered rearrangement, without repetition of these $n$ elements.
Example: "a", "b" and "c" are three elements.
Possible arrangements are : abc, acb, bac, bca, cab, cba.

# Annex: Combinatorics / Permutations

A permutation of $n$ distinct elements $e_1, ..., e_n$ is an ordered rearrangement, without repetition of these $n$ elements.
Example: "a", "b" and "c" are three elements.
Possible arrangements are : abc, acb, bac, bca, cab, cba.
The factor function is the domain function $\mathcal{N} = \{0, 1, 2, ...\}$, which associates to each

$$n \in \mathcal{N}, n! = n \cdot (n-1) \cdot (n-2) \cdots (2) \cdot (1).$$

# Annex: Combinatorics / Arrangements

An arrangement is a permutation of $k$ elements taken from $n$ distinct elements ($k \leq n$). The elements are taken without repetition and are ordered.

# Annex: Combinatorics / Arrangements

An arrangement is a permutation of $k$ elements taken from $n$ distinct elements ($k \leq n$). The elements are taken without repetition and are ordered.
The number of permutations of $k$ among $n$ is denoted $A_{n,k}$.

# Annex: Combinatorics / Arrangements

An arrangement is a permutation of $k$ elements taken from $n$ distinct elements ($k \leq n$). The elements are taken without repetition and are ordered.
The number of permutations of $k$ among $n$ is denoted $A_{n,k}$.
The arrangements of 2 elements taken in $\{1, 2, 3, 4\}$ are

# Annex: Combinatorics / Arrangements

An arrangement is a permutation of $k$ elements taken from $n$ distinct elements ($k \leq n$). The elements are taken without repetition and are ordered.
The number of permutations of $k$ among $n$ is denoted $A_{n,k}$.
The arrangements of 2 elements taken in $\{1, 2, 3, 4\}$ are

$$[1,2], [1,3], [1,4], [2,1], [2,3], [2,4], [3,1], [3,2], [3,4], [4,1], [4,2], [4,3]$$

# Annex: Combinatorics / Arrangements

An arrangement is a permutation of $k$ elements taken from $n$ distinct elements ($k \leq n$). The elements are taken without repetition and are ordered.
The number of permutations of $k$ among $n$ is denoted $A_{n,k}$.
The arrangements of 2 elements taken in $\{1, 2, 3, 4\}$ are

$$[1,2], [1,3], [1,4], [2,1], [2,3], [2,4], [3,1], [3,2], [3,4], [4,1], [4,2], [4,3]$$

Thus 12.

# Annex: Combinatorics / Arrangements

An arrangement is a permutation of $k$ elements taken from $n$ distinct elements ($k \leq n$). The elements are taken without repetition and are ordered.
The number of permutations of $k$ among $n$ is denoted $A_{n,k}$.
The arrangements of 2 elements taken in $\{1, 2, 3, 4\}$ are

$$[1,2], [1,3], [1,4], [2,1], [2,3], [2,4], [3,1], [3,2], [3,4], [4,1], [4,2], [4,3]$$

Thus 12.

$$\begin{aligned} A_{n,k} &= n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

How many words of 3 distinct letters can be formed in an alphabet of 26 letters?

# Annex: Combinatorics / Combinations

A combination of $k$ elements taken from an $n$-element set distinct is a $k$-element subset of this set. The elements are taken without repetition and are not ordered.

# Annex: Combinatorics / Combinations

A combination of $k$ elements taken from an $n$-element set distinct is a $k$-element subset of this set. The elements are taken without repetition and are not ordered.

The number of combinations of $k$ among $n$ is denoted $C_{n,k}$ or $\binom{n}{k}$, called binomial coefficient.

# Annex: Combinatorics / Combinations

A combination of $k$ elements taken from an $n$-element set distinct is a $k$-element subset of this set. The elements are taken without repetition and are not ordered.

The number of combinations of $k$ among $n$ is denoted $C_{n,k}$ or $\binom{n}{k}$, called binomial coefficient.

The combinations of 2 elements taken from $\{1, 2, 3, 4\}$ are $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}$. There are 6.

# Annex: Combinatorics / Combinations

A combination of $k$ elements taken from an $n$-element set distinct is a $k$-element subset of this set. The elements are taken without repetition and are not ordered.

The number of combinations of $k$ among $n$ is denoted $C_{n,k}$ or $\binom{n}{k}$, called binomial coefficient.

The combinations of 2 elements taken from $\{1,2,3,4\}$ are $\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}$. There are 6.

$$\begin{aligned} C_{n,k} &= \frac{A_{n,k}}{k!} \\ &= \frac{n!}{k!(n-k)!} \end{aligned}$$

# Annex: Combinatorics / Combinations

A combination of $k$ elements taken from an $n$-element set distinct is a $k$-element subset of this set. The elements are taken without repetition and are not ordered.

The number of combinations of $k$ among $n$ is denoted $C_{n,k}$ or $\binom{n}{k}$, called binomial coefficient.

The combinations of 2 elements taken from $\{1,2,3,4\}$ are $\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}$. There are 6.

$$\begin{aligned} C_{n,k} &= \frac{A_{n,k}}{k!} \\ &= \frac{n!}{k!(n-k)!} \end{aligned}$$

Example: we have 15 medications and we want to test their compatibility in a group of 4. How many possible groups are there?

# Annex: Combinatorics / Multinomial theorem

For any positive integer $k$ and any non-negative integer $p$, the multinomial formula describes how a sum with $k$ terms expands when raised to an arbitrary power $p$:

$$(z_1 + \cdots + z_k)^p$$

$$= \sum_{n_1+n_2+\cdots+n_k=p, n_1\geq 0, n_2\geq 0, \ldots, n_k\geq 0} \binom{p}{n_1, n_2, \ldots, n_k} z_1^{n_1} \cdots z_k^{n_k}$$

where

$$\binom{p}{n_1, n_2, \ldots, n_k} = \frac{p!}{n_1! \cdots n_k!}$$ is a multinomial coefficient.