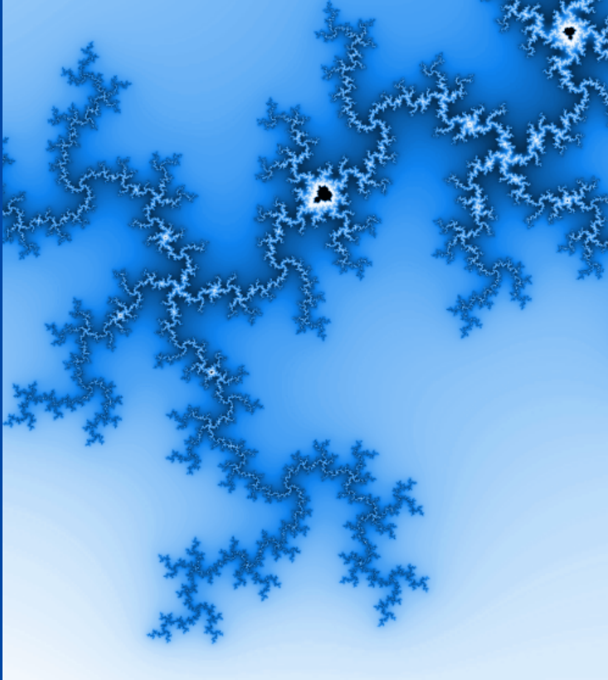
Statistics from a probability perspective : tricks and traps

Statistics Workshop for Zoologists

Laurent Loosveldt

11th December 2023

LIÈGE université
Mathématique



Why am I here?

Why am I here?

Doing Statistics : Expectation

Having a question concerning a population

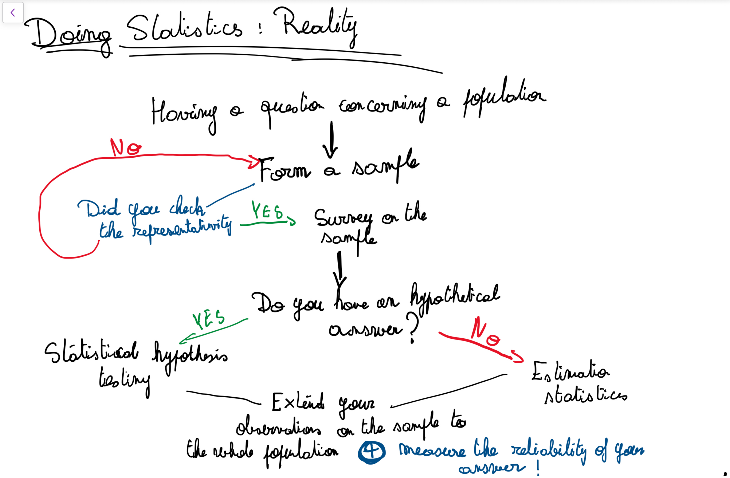
↓

Survey

↓

Answer

Why am I here?


Doing Statistics : Reality
Having a question concerning a population
No
Form a sample
Did you check the representativity
YES
Survey on the sample
Do you have an hypothetical answer ?
YES
No
Statistical hypothesis testing
Estimation statistics
Extend your observations on the sample to the whole population
+ measure the reliability of your answer !

Why am I here?

Doing Statistics : Reality

Having a question concerning a population

Form a sample

No

Did you check the representativity

YES

Survey on the sample

Do you have an hypothetical answer?

YES

Statistical hypothesis testing

No

Estimation statistics

Extend your observations on the sample to the whole population

+ measure the reliability of your answer !

The recipe to do probability

The recipe to do probability

- A random experiment

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.
- A σ-algebra $\mathscr{F}$ on Ω

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.
- A σ-algebra $\mathscr{F}$ on Ω i.e a collection of subsets of Ω

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.
- A σ-algebra $\mathscr{F}$ on Ω i.e a collection of subsets of Ω such that
  - $\Omega \in \mathscr{F}$;
  - if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$;
  - if $A_1, A_2, A_3, \dots \in \mathscr{F}$, then $\bigcup_j A_j \in \mathscr{F}$.

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.
- A σ-algebra $\mathscr{F}$ on Ω i.e a collection of subsets of Ω such that
  - $\Omega \in \mathscr{F}$;
  - if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$;
  - if $A_1, A_2, A_3, \dots \in \mathscr{F}$, then $\bigcup_j A_j \in \mathscr{F}$.

  The sets in $\mathscr{F}$ are called events.

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.
- A σ-algebra $\mathscr{F}$ on Ω i.e a collection of subsets of Ω such that
  - $\Omega \in \mathscr{F}$;
  - if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$;
  - if $A_1, A_2, A_3, \cdots \in \mathscr{F}$, then $\bigcup_j A_j \in \mathscr{F}$.

  The sets in $\mathscr{F}$ are called events.
- A *measure* $\mathbb{P} : \mathscr{F} \to [0,1] : A \mapsto \mathbb{P}(A)$

The recipe to do probability

- A random experiment i.e. an experiment with various outcomes that you can not predict in advance but you want to *measure* the chance that this experiment leads to certain outcomes.
- A set Ω collecting the outcomes from your experiment.
- A σ-algebra $\mathscr{F}$ on Ω i.e a collection of subsets of Ω such that
  - $\Omega \in \mathscr{F}$;
  - if $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$;
  - if $A_1, A_2, A_3, \dots \in \mathscr{F}$, then $\bigcup_j A_j \in \mathscr{F}$.

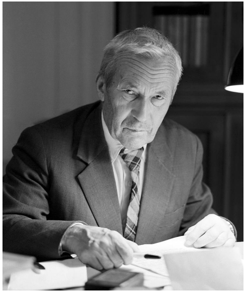
  The sets in $\mathscr{F}$ are called events.
- A *measure* $\mathbb{P} : \mathscr{F} \to [0,1] : A \mapsto \mathbb{P}(A)$ such that
  - $\mathbb{P}(\Omega) = 1$;
  - if $A_1, A_2, A_3, \dots \in \mathscr{F}$ are disjoint then $\mathbb{P}\left(\bigcup_j A_j\right) = \sum_j \mathbb{P}(A_j)$.

The recipe to do probability

We call $(\Omega, \mathscr{F}, \mathbb{P})$ a probability space.

The recipe to do probability

We call $(\Omega, \mathscr{F}, \mathbb{P})$ a probability space. This formalization is due to the Soviet Mathematician Andrey Kolmogorov (1903-1987)



Random variable

Definition

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A *random variable* is a function

$$X : \Omega \to \mathbb{R}$$

such that, for all $a < b$, the set

$$\{X \in [a,b]\} := \{\omega \in \Omega : X(\omega) \in [a,b]\}$$

belongs to $\mathscr{F}$.

Random variable

Definition

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A *random variable* is a function

$$X : \Omega \to \mathbb{R}$$

such that, for all $a < b$, the set

$$\{X \in [a,b]\} := \{\omega \in \Omega : X(\omega) \in [a,b]\}$$

belongs to $\mathscr{F}$.

It just means that we want to be able to compute, for all a, b the probability

$$\mathbb{P}(X \in [a,b]) = \mathbb{P}(a \leq X \leq b)$$

Random variable

Definition

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A *random variable* is a function

$$X : \Omega \to \mathbb{R}$$

such that, for all $a < b$, the set

$$\{X \in [a,b]\} := \{\omega \in \Omega : X(\omega) \in [a,b]\}$$

belongs to $\mathscr{F}$.

It just means that we want to be able to compute, for all a, b the probability

$$\mathbb{P}(X \in [a,b]) = \mathbb{P}(a \leq X \leq b)$$

thus we require that $\{X \in [a,b]\}$ is an event.

Random variable

It is mainly what is used in statistics by considering the variable that you are interested in as a random variable defined on your population.

Random variable

It is mainly what is used in statistics by considering the variable that you are interested in as a random variable defined on your population.

Example

- number of lion cubs per litter :

$$X : \text{lioness} \mapsto \{1, 2, 3, 4, 5, 6\}.$$

Random variable

It is mainly what is used in statistics by considering the variable that you are interested in as a random variable defined on your population.

Example

- number of lion cubs per litter :

$$X : \text{lioness} \mapsto \{1, 2, 3, 4, 5, 6\}.$$

- size of a gorilla :

$$X : \text{gorilla} \mapsto [0, 160]$$

Random variable

In practice:

Random variable

In practice:

- we have a “dictionary” of well-known laws for random variables;

Random variable

In practice:

- we have a "dictionary" of well-known laws for random variables;
- we assume that a variable of interest is distributed according to one of this law (and justify this assumption !).

Random variable

In practice:

- we have a “dictionary” of well-known laws for random variables;
- we assume that a variable of interest is distributed according to one of this law (and justify this assumption !).

Two important families of laws:

Definition

We say that a random variable X is *discrete* if it can only take countably many different values $(x_j)_j$.

Random variable

In practice:

- we have a “dictionary” of well-known laws for random variables;
- we assume that a variable of interest is distributed according to one of this law (and justify this assumption !).

Two important families of laws:

Definition

We say that a random variable X is *discrete* if it can only take countably many different values $(x_j)_j$.The law of this random variable is then characterized by the numbers

$$\mathbb{P}(X = x_j).$$

Random variable

In practice:

- we have a “dictionary” of well-known laws for random variables;
- we assume that a variable of interest is distributed according to one of this law (and justify this assumption !).

Two important families of laws:

Definition

We say that a random variable X is *discrete* if it can only take countably many different values $(x_j)_j$.The law of this random variable is then characterized by the numbers

$$\mathbb{P}(X = x_j).$$

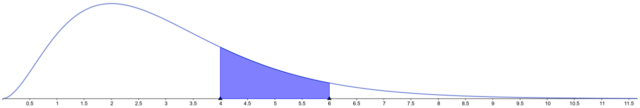
Remark: $\sum_j \mathbb{P}(X = x_j) = 1$.

Random variable

Definition

We say that a random variable X is *continuous* if there exists a positive integrable function f_X (called *density*) such that, for all $a < b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$



Random variable

Definition

We say that a random variable X is *continuous* if there exists a positive integrable function f_X (called *density*) such that, for all $a < b$,

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx.$$

Remark: $\int_{\mathbb{R}} f_X(x)\,dx = 1.$

Some parameters to characterize a random variable

Some parameters to characterize a random variable

Definition

The *expectation* of a discrete random values X with values $(x_j)_j$ is given by

$$\mathbb{E}[X] = \sum_j x_j \, \mathbb{P}(X = x_j)$$

if this sum is well-defined.

Some parameters to characterize a random variable

Definition

The *expectation* of a discrete random values X with values $(x_j)_j$ is given by

$$\mathbb{E}[X] = \sum_j x_j \, \mathbb{P}(X = x_j)$$

if this sum is well-defined. If the continuous random variable X satisfied $x \mapsto x f_X(x)$ is integrable, then X has an expectation which is given by

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) \, dx.$$

Some parameters to characterize a random variable

Definition

The *expectation* of a discrete random values X with values $(x_j)_j$ is given by

$$\mathbb{E}[X] = \sum_j x_j \, \mathbb{P}(X = x_j)$$

if this sum is well-defined. If the continuous random variable X satisfied $x \mapsto x f_X(x)$ is integrable, then X has an expectation which is given by

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) \, dx.$$

A center of gravity for the distribution, a central value.

Some parameters to characterize a random variable

Definition

The *variance* of the random variable X is given by

$$\mathbb{E}[(X-\mathbb{E}[X])^2]$$

if this quantity makes sense.

Some parameters to characterize a random variable

Definition

The *variance* of the random variable X is given by

$$\mathbb{E}[(X - \mathbb{E}[X])^2]$$

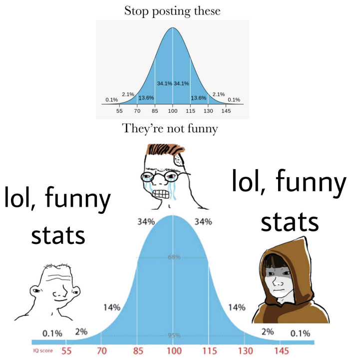
if this quantity makes sense.

The dispersion of the random variable around the expectation.

A superstar: the normal distribution

A superstar: the normal distribution

- The most popular law of probability;



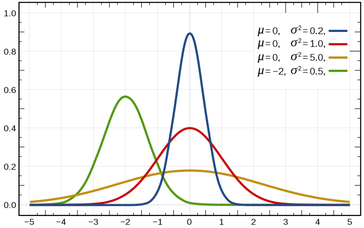
A superstar: the normal distribution

- The most popular law of probability;
- the most used both in theory and practice;

A superstar: the normal distribution

- The most popular law of probability;
- the most used both in theory and practice;
- recognisable by its famous “bell curve” which is the graph of the density

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



with, if $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.

A superstar: the normal distribution

- The most popular law of probability;
- the most used both in theory and practice;
- recognisable by its famous "bell curve" which is the graph of the density

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

  with, if $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.
- used to represent the distribution of:
  - many physiological data: size, weight, sleep time, the growth of hairs or nails, intelligence quotient,...;

A superstar: the normal distribution

- The most popular law of probability;
- the most used both in theory and practice;
- recognisable by its famous "bell curve" which is the graph of the density

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

  with, if $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.
- used to represent the distribution of:
  - many physiological data: size, weight, sleep time, the growth of hairs or nails, intelligence quotient,...;
  - many natural data: size and weight of seeds, animals,...

A superstar: the normal distribution

- The most popular law of probability;
- the most used both in theory and practice;
- recognisable by its famous "bell curve" which is the graph of the density

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

  with, if $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.
- used to represent the distribution of:
  - many physiological data: size, weight, sleep time, the growth of hairs or nails, intelligence quotient,...;
  - many natural data: size and weight of seeds, animals,...
  - industrial production;

A superstar: the normal distribution

- The most popular law of probability;
- the most used both in theory and practice;
- recognisable by its famous "bell curve" which is the graph of the density

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

  with, if $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.
- used to represent the distribution of:
  - many physiological data: size, weight, sleep time, the growth of hairs or nails, intelligence quotient,...;
  - many natural data: size and weight of seeds, animals,...
  - industrial production;
  - ...

So... is everything normal?

So... is everything normal?

Of course not!

So... is everything normal?

Of course not!

Exponential
- Waiting time
- Life time of electonic components

Poisson
Number of times that a rare event occurs

Bernouilli
Any phenomena with two possible outcomes (sucess/fail)

Binomial
Counting the number of sucess while repeating independant Bernouilli trials

Geometric
Time needed for the first sucess while repeating independant Bernouilli trials

Discrete Uniform
Any phenomena with finitely many outcomes, all with the same probability

Continuous Uniform
Any phenomena with a continuum of outcomes, with the same probability all along

But...

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x \right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

But...

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x \right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

- if $X \sim \mathcal{N}(0,1)$, $\mathbb{P}(X \leq x) = \Phi(x)$;

But...

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x\in\mathbb{R}$, we have

$$\lim_{n\to+\infty}\mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)-\mu}{\sqrt{\frac{\sigma^2}{n}}}\leq x\right)=\Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

- if $X\sim\mathcal{N}(0,1)$, $\mathbb{P}(X\leq x)=\Phi(x)$;
- $\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)$ is the mean of the random variables X_1, $X_2,\ldots,X_n$

But...

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^n X_j\right)-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x\right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

- if $X \sim \mathcal{N}(0,1)$, $\mathbb{P}(X \leq x) = \Phi(x)$;
- $\left(\frac{1}{n}\sum_{j=1}^n X_j\right)$ is the mean of the random variables X_1, $X_2,\ldots,X_n$

$\Rightarrow$ in average, any phenomena repeated sufficiently many times has a distribution which can be approximated by the normal distribution!

Central Limit Theorem: consequences

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x\in\mathbb{R}$, we have

$$\lim_{n\to+\infty}\mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)-\mu}{\sqrt{\frac{\sigma^2}{n}}}\leq x\right)=\Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

Central Limit Theorem: consequences

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x\right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

- it explains the omnipresence of the normal distribution in the nature: many phenomenon is the sum of small independent phenomena.

Central Limit Theorem: consequences

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x\right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

- it explains the omnipresence of the normal distribution in the nature: many phenomenon is the sum of small independent phenomena.
- it allows to base a lot of statistical methods on the normal distribution.

In practice

In practice

Many questions in statistics can be translated into determining the value of a/some parameters of a probability distribution.

In practice

Many questions in statistics can be translated into determining the value of a/some parameters of a probability distribution.

What is the average size of the adults living in Belgium?

In practice

Many questions in statistics can be translated into determining the value of a/some parameters of a probability distribution.

What is the average size of the adults living in Belgium?

Size is normally distributed and we try to determine the expectation of the distribution.

In practice

Many questions in statistics can be translated into determining the value of a/some parameters of a probability distribution.

What is the average size of the adults living in Belgium?

Size is normally distributed and we try to determine the expectation of the distribution.

What is the proportion of the population living below the poverty threshold?

In practice

Many questions in statistics can be translated into determining the value of a/some parameters of a probability distribution.

What is the average size of the adults living in Belgium?

Size is normally distributed and we try to determine the expectation of the distribution.

What is the proportion of the population living below the poverty threshold?

Each person has two options: living below the poverty threshold (score 1) or not (score 0).

In practice

Many questions in statistics can be translated into determining the value of a/some parameters of a probability distribution.

What is the average size of the adults living in Belgium?

Size is normally distributed and we try to determine the expectation of the distribution.

What is the proportion of the population living below the poverty threshold?

Each person has two options: living below the poverty threshold (score 1) or not (score 0). Therefore the population is distributed according to a Bernoulli random variable of parameter p =the proportion of the population living below the poverty threshold.

In practice: random sample

To answer your question, a sample $\{x_1, \ldots, x_n\}$ of n observations is available to you.

In practice: random sample

To answer your question, a sample $\{x_1, \ldots, x_n\}$ of n observations is available to you.

Each of this observation comes from your population which is distributed according to a probability distribution F_θ, where θ is/are (a) parameter(s) which characterize(s) the distribution ((μ, σ^2) for a normal distribution, p for a Bernoulli,...).

In practice: random sample

To answer your question, a sample $\{x_1,\ldots,x_n\}$ of n observations is available to you.

Each of this observation comes from your population which is distributed according to a probability distribution F_θ, where θ is/are (a) parameter(s) which characterize(s) the distribution ((μ,σ^2) for a normal distribution, p for a Bernoulli,...).You want to estimate, thanks to your sample, the values of (some of) this/these parameter(s).

In practice: random sample

To answer your question, a sample $\{x_1,\ldots,x_n\}$ of n observations is available to you.

Each of this observation comes from your population which is distributed according to a probability distribution F_θ, where θ is/are (a) parameter(s) which characterize(s) the distribution ((μ,σ^2) for a normal distribution, p for a Bernoulli,...).You want to estimate, thanks to your sample, the values of (some of) this/these parameter(s).

If your sample has been **collected properly**, you can see it as a realisation of random vector $(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are independent random variables such that, for all $1 \leq j \leq n$, $X_j \sim F_\theta$.

Estimation and estimator

From your sample $\{x_1,\ldots,x_n\}$, you want to deduce an estimation of the parameter(s) θ: a real number $\widehat{\theta}$ for which you hope that $|\theta - \widehat{\theta}|$ is small.

Estimation and estimator

From your sample $\{x_1,\dots,x_n\}$, you want to deduce an estimation of the parameter(s) θ: a real number $\widehat{\theta}$ for which you hope that $|\theta-\widehat{\theta}|$ is small.

In order to insure good statistical properties for this estimation, we define an **estimator** as a function

$$G:\mathbb{R}^n \to \Theta : (X_1,\dots,X_n)\mapsto G(X_1,\dots,X_n),$$

where Θ is the set of all possible values for θ.

Estimation and estimator

From your sample $\{x_1,\ldots,x_n\}$, you want to deduce an estimation of the parameter(s) θ: a real number $\widehat{\theta}$ for which you hope that $|\theta-\widehat{\theta}|$ is small.

In order to insure good statistical properties for this estimation, we define an **estimator** as a function

$$G:\mathbb{R}^n\to\Theta:(X_1,\ldots,X_n)\mapsto G(X_1,\ldots,X_n),$$

where Θ is the set of all possible values for θ.

An estimator is a random variable. The realisation of this random variable on your sample is your estimation:

$$\widehat{\theta}=G(x_1,\ldots,x_n).$$

Good properties

- bias :

$$b_G(\theta) := \mathbb{E}_\theta[G(X_1,\ldots,X_n)] - \theta$$

Good properties

- bias :

$$b_G(\theta) := \mathbb{E}_\theta[G(X_1, \ldots, X_n)] - \theta$$

Goal: $b_G(\theta) = 0$, for all $\theta \in \Theta$

Good properties

- bias :

$$b_G(\theta) := \mathbb{E}_\theta[G(X_1, \ldots, X_n)] - \theta$$

Goal: $b_G(\theta) = 0$, for all $\theta \in \Theta$ (or at least $\lim_{n\to+\infty} b_G(\theta) = 0$)

Good properties

- bias :

$$b_G(\theta) := \mathbb{E}_\theta[G(X_1,\ldots,X_n)] - \theta$$

Goal: $b_G(\theta) = 0$, for all $\theta \in \Theta$ (or at least $\lim_{n\to+\infty} b_G(\theta) = 0$)

- Mean square error :

$$\mathrm{MSE}_G(\theta) = \mathbb{E}_\theta[(G(X_1,\ldots,X_n) - \theta)^2].$$

Good properties

- bias :

$$b_G(\theta) := \mathbb{E}_\theta[G(X_1,\ldots,X_n)] - \theta$$

Goal: $b_G(\theta) = 0$, for all $\theta \in \Theta$ (or at least $\lim_{n\to+\infty} b_G(\theta) = 0$)

- Mean square error :

$$\mathsf{MSE}_G(\theta) = \mathbb{E}_\theta[(G(X_1,\ldots,X_n) - \theta)^2].$$

Goal: $\mathsf{MSE}_G(\theta)$ to be as small as possible.

- Sufficiency :

The distribution of $(X_1,\ldots,X_n)$ knowing that $G(X_1,\ldots,X_n) = g$ does not depend on θ.

Good properties

- bias :

$$b_G(\theta) := \mathbb{E}_\theta[G(X_1,\dots,X_n)] - \theta$$

Goal: $b_G(\theta) = 0$, for all $\theta \in \Theta$ (or at least $\lim_{n\to+\infty} b_G(\theta) = 0$)

- Mean square error :

$$\mathrm{MSE}_G(\theta) = \mathbb{E}_\theta[(G(X_1,\dots,X_n) - \theta)^2].$$

Goal: $\mathrm{MSE}_G(\theta)$ to be as small as possible.

- Sufficiency :

The distribution of $(X_1,\dots,X_n)$ knowing that $G(X_1,\dots,X_n) = g$ does not depend on θ.

Interpretation: the estimator uses all the information available in the sample.

Good properties

- Robustness.

Good properties

- Robustness.
- Maximum likelihood.

Good properties

- Robustness.
- Maximum likelihood.
- ...

Good properties

- Robustness.
- Maximum likelihood.
- ...

Example

When we want to estimate the expectation, the sample mean

$$\overline{X}_n := \frac{1}{n} \sum_{j=1}^{n} X_j$$

is unbiased, its mean-square error goes to 0 with n, is not robust. It is sufficient and the maximum likelihood estimator for the normal distribution.

Back to CLT

In many practical situations, the law of X is unknown. So, how do we choose a good estimator?

Back to CLT

In many practical situations, the law of X is unknown. So, how do we choose a good estimator?

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x\right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

Back to CLT

In many practical situations, the law of X is unknown. So, how do we choose a good estimator?

Central Limit Theorem

Let $(X_j)_{j\in\mathbb{N}^*}$ be a family of independent and identically distributed random variables of expectation μ and variance σ^2. For all $x \in \mathbb{R}$, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq x\right) = \Phi(x),$$

where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

Many statistical methods focusing on the expectation used $\overline{X}_n$ as estimator. These methods generally consider $\overline{X}_n$ as normally distributed but required that n is big enough.

How good is this approximation ?

How good is this approximation ?

It is still a very active field of research in probability!

How good is this approximation ?

It is still a very active field of research in probability!
It involves:

- various notion of “distance” between random variable;

How good is this approximation ?

It is still a very active field of research in probability!
It involves:

- various notion of “distance” between random variable;
- generalizations of CLT;

How good is this approximation ?

It is still a very active field of research in probability!
It involves:

- various notion of "distance" between random variable;
- generalizations of CLT;
- Stein method;

How good is this approximation ?

It is still a very active field of research in probability!
It involves:

- various notion of "distance" between random variable;
- generalizations of CLT;
- Stein method;
- Malliavin calculus;

How good is this approximation ?

It is still a very active field of research in probability!
It involves:

- various notion of “distance” between random variable;
- generalizations of CLT;
- Stein method;
- Malliavin calculus;
- ...

How good is this approximation ?

It is still a very active field of research in probability!
It involves:

- various notion of "distance" between random variable;
- generalizations of CLT;
- Stein method;
- Malliavin calculus;
- ...

In practice, it is generally admitted that $n > 30$ is good !

Asymptotic or exact method ?

Various statistical methods based on $\overline{X}_n$ and the normal distribution are a bit different whether you can actually assume that X is normally distributed or not.

Asymptotic or exact method ?

Various statistical methods based on $\overline{X}_n$ and the normal distribution are a bit different whether you can actually assume that X is normally distributed or not.

If X is normally distributed, the method is exact.

Asymptotic or exact method ?

Various statistical methods based on $\overline{X}_n$ and the normal distribution are a bit different whether you can actually assume that X is normally distributed or not.

If X is normally distributed, the method is exact.

Otherwise, the method is asymptotic.

Asymptotic or exact method ?

Various statistical methods based on $\overline{X}_n$ and the normal distribution are a bit different whether you can actually assume that X is normally distributed or not.

If X is normally distributed, the method is exact.

Otherwise, the method is asymptotic.

It is important to be able to quickly checked if normality is a good assumption or not.

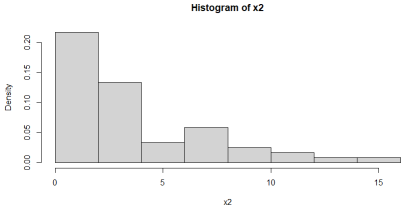
Some tricks to test normality

Some tricks to test normality

- Shape of the histogram

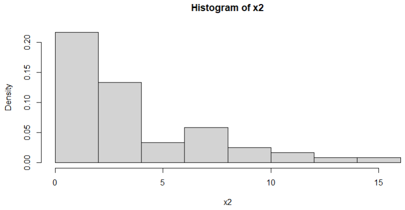
Some tricks to test normality

- Shape of the histogram



Some tricks to test normality

- Shape of the histogram



⚠The normal distribution is NOT the only one to present a bell curve ⚠

Some tricks to test normality

- qQ-plot

Quantiles

Let $\alpha \in]0,1[$, we say that

- q_α is a α-quantile of the random variable X if

$$\mathbb{P}(X < q_\alpha) \leq \alpha \text{ and } \mathbb{P}(X \leq q_\alpha) \geq \alpha;$$

- Q_α is an empirical α-quantile of the data set $\{x_1, \ldots, x_n\}$ if

$$\frac{\#\{1 \leq j \leq n : x_j < Q_\alpha\}}{n} \leq \alpha \text{ and } \frac{\#\{1 \leq j \leq n : x_j \leq Q_\alpha\}}{n} \geq \alpha$$

Some tricks to test normality

- qQ-plot

Quantiles

Let $\alpha \in]0,1[$, we say that

- q_α is a α-quantile of the random variable X if

$$\mathbb{P}(X < q_\alpha) \leq \alpha \text{ and } \mathbb{P}(X \leq q_\alpha) \geq \alpha;$$

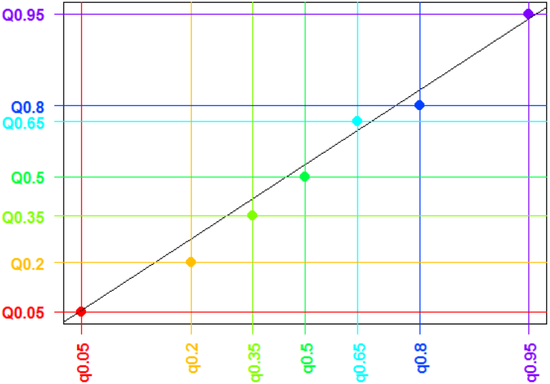
- Q_α is an empirical α-quantile of the data set $\{x_1, \ldots, x_n\}$ if

$$\frac{\#\{1 \leq j \leq n : x_j < Q_\alpha\}}{n} \leq \alpha \text{ and } \frac{\#\{1 \leq j \leq n : x_j \leq Q_\alpha\}}{n} \geq \alpha$$

If $\{x_1, \ldots, x_n\}$ is a realisation of $X_1, \ldots, X_n$, i.i.d. with the law of X, for all α, q_α and Q_α should be close (if n is large enough).

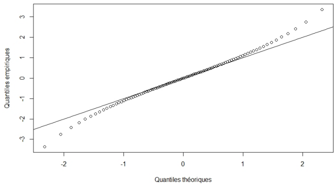
Some tricks to test normality

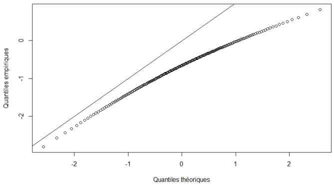
- qQ-plot

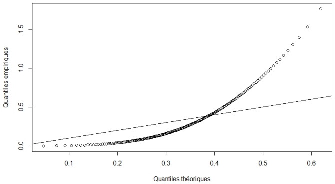


Some tricks to test normality

- qQ-plot







Choose the good method

- when available, an exact method is always better than an asymptotic method;

Choose the good method

- when available, an exact method is always better than an asymptotic method;
- the histogram comparison and the qQ-plot can be used with any laws of probability;

Choose the good method

- when available, an exact method is always better than an asymptotic method;
- the histogram comparison and the qQ-plot can be used with any laws of probability;
- other tests for distributions are defined in the literature.

$\Rightarrow$ Try to identify the good law of probability to use as a model.

Choose the good method

- when available, an exact method is always better than an asymptotic method;
- the histogram comparison and the qQ-plot can be used with any laws of probability;
- other tests for distributions are defined in the literature.

$\Rightarrow$ Try to identify the good law of probability to use as a model.
$\Rightarrow$ Deduce the good method to use with this law.

Context

You have a hypothesis H_0 for your population and you want to test whether this hypothesis seems reasonable or not.

Context

You have a hypothesis H_0 for your population and you want to test whether this hypothesis seems reasonable or not.

As already said, this hypothesis can often be stated as a affirmation concerning the value of the parameter characterising the distribution of your variable (expectation, variance,...).

Context

You have a hypothesis H_0 for your population and you want to test whether this hypothesis seems reasonable or not.

As already said, this hypothesis can often be stated as a affirmation concerning the value of the parameter characterising the distribution of your variable (expectation, variance,...).

You collect data and according to these data you want to decide whether you should

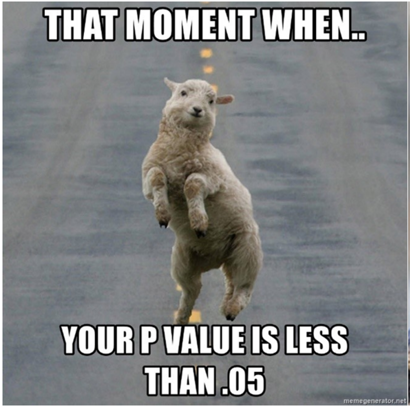
- reject your hypothesis H_0;
- not reject your hypothesis H_0.

⚠ you never accept H_0 ⚠

Naive use of p-value

Naive use of p-value

Naive use of p-value



Naive use of p-value

Hypothesis testing: review of the process

Hypothesis testing: review of the process

1. Determine Θ

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

		Reality	
		H_0 is true	H_1 is true
Decision taken	Do not reject H_0	Good!	Wrong!
	Reject H_0	Wrong!	Good!

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

		Reality	
		H_0 is true	H_1 is true
Decision taken	Do not reject H_0	Good!	Type II error
	Reject H_0	Type I error	Good!

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

		Reality	
		H_0 is true	H_1 is true
Decision taken	Do not reject H_0	Good!	Type II error
	Reject H_0	Type I error	Good!

Type I error and type II error have antagonistic behaviours, one can not guarantee to limit both types together below a certain level.

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

Hypothesis testing methods are built by first choosing a significance level α and imposing

$$\mathbb{P}(\text{Reject } H_0 | H_0 \text{ is true}) \leq \alpha.$$

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

Hypothesis testing methods are built by first choosing a significance level α and imposing

$$\mathbb{P}(\text{Reject } H_0 | H_0 \text{ is true}) \leq \alpha.$$

Among all possible tests satisfying this property, we work, if possible, with one for which

$$\mathbb{P}(\text{Don't Reject } H_0 | H_1 \text{ is true})$$

is minimal

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.
2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

How to choose H_0 and H_1 ?

It means that you ONLY perfectly master the type I error ! It must guide your choice of H_0 and H_1.

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.

It means that you ONLY perfectly master the type I error ! It must guide your choice of H_0 and H_1.

Remark: the condition $p \leq 0,05$ means that we impose

$$\mathbb{P}(\text{Reject } H_0 | H_0 \text{ is true}) \leq 0,05.$$

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test
   - A test statistic $T(X_1, \ldots, X_n)$;

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test
   - A test statistic $T(X_1, \ldots, X_n)$;
   - The significance level α;

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test
   - A test statistic $T(X_1, \ldots, X_n)$;
   - The significance level α;
   - A critical region $\mathscr{R}$.

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test
   - A test statistic $T(X_1, \ldots, X_n)$;
   - The significance level α;
   - A critical region $\mathscr{R}$.

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test
   - A test statistic $T(X_1, \ldots, X_n)$;
   - The significance level α;
   - A critical region $\mathscr{R}$.

Reject $H_0 \Leftrightarrow T \in \mathscr{R}$

Hypothesis testing: review of the process

1. Determine Θ, the set of all possible values for your parameter(s) θ.

2. Formulate your hypotheses: write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and determine
   - the null hypothesis: $H_0 : \theta \in \Theta_0$;
   - the alternative hypothesis: $H_1 : \theta \in \Theta_1$.
3. Determine the ingredients of your test
   - A test statistic $T(X_1, \ldots, X_n)$;
   - The significance level α;
   - A critical region $\mathscr{R}$.

Reject $H_0 \Leftrightarrow T \in \mathscr{R}$:

$$\mathbb{P}(T \in \mathscr{R} | H_0 \text{ is true}) \leq \alpha$$

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu,\sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow$ $H_1 : \mu \neq \mu_0$ (two-sided test).

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow$ $H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0"

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)
    - "the mean can only be lower than or equal to μ_0" $\Rightarrow H_1 : \mu < \mu_0$ (left one-sided test)

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)
    - "the mean can only be lower than or equal to μ_0" $\Rightarrow H_1 : \mu < \mu_0$ (left one-sided test)
- Do you know σ^2?

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)
    - "the mean can only be lower than or equal to μ_0" $\Rightarrow H_1 : \mu < \mu_0$ (left one-sided test)
- Do you know σ^2?
  - YES

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)
    - "the mean can only be lower than or equal to μ_0" $\Rightarrow H_1 : \mu < \mu_0$ (left one-sided test)
- Do you know σ^2?
  - YES use the statistic $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
  - NO

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)
    - "the mean can only be lower than or equal to μ_0" $\Rightarrow H_1 : \mu < \mu_0$ (left one-sided test)
- Do you know σ^2?
  - YES use the statistic $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
  - NO estimate σ^2 using $\widetilde{S_n^2} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2$

Example: mean of normally distributed population

You study a population X for which you can assume $X \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test if $\mu = \mu_0$ is a reasonable hypothesis for your population: $H_0 : \mu = \mu_0$

- Do you privilege a side for your alternative hypothesis?
  - NO $\Rightarrow H_1 : \mu \neq \mu_0$ (two-sided test).
  - YES
    - "the mean can only be greater than or equal to μ_0" $\Rightarrow H_1 : \mu > \mu_0$ (right one-sided test)
    - "the mean can only be lower than or equal to μ_0" $\Rightarrow H_1 : \mu < \mu_0$ (left one-sided test)
- Do you know σ^2?
  - YES use the statistic $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
  - NO estimate σ^2 using $\widetilde{S_n^2} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2$and use the statistic $\frac{\overline{X}_n - \mu}{\frac{\widetilde{S_n}}{\sqrt{n}}} \sim t_{n-1}$ (Student law with $n-1$ degrees of freedom).

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α.

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α. We first consider the right one-sided test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0.$$

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α. We first consider the right one-sided test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0.$$

Intuition: an empirical mean on the sample $\{x_1, \ldots, x_n\}$ significantly greater than μ_0 must lead to reject H_0.

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α. We first consider the right one-sided test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0.$$

Intuition: an empirical mean on the sample $\{x_1, \ldots, x_n\}$ significantly greater than μ_0 must lead to reject H_0.$\Rightarrow$ The critical area should be of the form $[\mu_0 + C, +\infty[$.

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α. We first consider the right one-sided test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0.$$

Intuition: an empirical mean on the sample $\{x_1, \ldots, x_n\}$ significantly greater than μ_0 must lead to reject H_0.$\Rightarrow$ The critical area should be of the form $[\mu_0 + C, +\infty[$.

How to choose the value of C?

- if σ^2 is known and if H_0 is true, $\overline{X}_n \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α. We first consider the right one-sided test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0.$$

Intuition: an empirical mean on the sample $\{x_1, \ldots, x_n\}$ significantly greater than μ_0 must lead to reject H_0.$\Rightarrow$ The critical area should be of the form $[\mu_0 + C, +\infty[$.

How to choose the value of C?

- if σ^2 is known and if H_0 is true, $\overline{X}_n \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n}) \Leftrightarrow \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. Therefore, if $z_{1-\alpha}$ is the quantile of order $1-\alpha$ of the $\mathcal{N}(0,1)$ law,

$$\mathbb{P}\left(\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha} \middle| H_0 \text{ is true} \right) = \alpha$$

Example: mean of normally distributed population

Let us determine the critical region $\mathscr{R}$ with a significance level α. We first consider the right one-sided test

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0.$$

Intuition: an empirical mean on the sample $\{x_1, \ldots, x_n\}$ significantly greater than μ_0 must lead to reject H_0. $\Rightarrow$ The critical area should be of the form $[\mu_0 + C, +\infty[$.

How to choose the value of C?

- if σ^2 is known and if H_0 is true, $\overline{X}_n \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n}) \Leftrightarrow \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. Therefore, if $z_{1-\alpha}$ is the quantile of order $1-\alpha$ of the $\mathcal{N}(0,1)$ law,

$$\mathbb{P}\left(\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha} \middle| H_0 \text{ is true}\right) = \alpha \Leftrightarrow \mathbb{P}\left(\overline{X}_n \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \middle| H_0 \text{ is true}\right) = \alpha$$

Example: mean of normally distributed population

- if σ^2 is unknown, we use the quantile $t_{n-1,1-\alpha}$ of the Student law with $n-1$ degrees of freedom

$$\mathbb{P}\left(\overline{X}_n \geq \mu_0 + t_{n-1,1-\alpha}\frac{\widetilde{S_n}}{\sqrt{n}} \middle| H_0 \text{ is true}\right) = \alpha.$$

Example: mean of normally distributed population

- if σ^2 is unknown, we use the quantile $t_{n-1,1-\alpha}$ of the Student law with $n-1$ degrees of freedom

$$\mathbb{P}\left(\overline{X}_n \geq \mu_0 + t_{n-1,1-\alpha}\frac{\widetilde{S_n}}{\sqrt{n}}\middle| H_0 \text{ is true}\right) = \alpha.$$

For the left one-sided test, we "reverse the inequalities":

- if σ^2 is known

$$\mathscr{R} = (-\infty, \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}];$$

- if σ^2 is unknown

$$\mathscr{R} = (-\infty, \mu_0 - t_{n-1,1-\alpha}\frac{\widetilde{S_n}}{\sqrt{n}}].$$

Example: mean of normally distributed population

For a two-sided test, we symmetrize the reasoning around μ_0:

Example: mean of normally distributed population

For a two-sided test, we symmetrize the reasoning around μ_0:

- if σ^2 is known

$$\mathscr{R} = (-\infty, \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}] \cup [\mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, +\infty);$$

- if σ^2 is unknown

$$\mathscr{R} = (-\infty, \mu_0 - t_{n-1,1-\frac{\alpha}{2}} \frac{\widetilde{S_n}}{\sqrt{n}}] \cup [\mu_0 + t_{n-1,1-\frac{\alpha}{2}} \frac{\widetilde{S_n}}{\sqrt{n}}, +\infty).$$

Mean of normally distributed population: summary

- Right one-sided

σ^2	Decision area
known	μ_0 $\quad$ $\mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
unknown	μ_0 $\quad$ $\mu_0 + \frac{\tilde{s}_n}{\sqrt{n}} t_{n-1,1-\alpha}$

Mean of normally distributed population: summary

- Left one-sided

σ^2	Decision area
known	$\mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$ $\quad$ μ_0
unknown	$\mu_0 - \frac{\widetilde{s}_n}{\sqrt{n}} t_{n-1,1-\alpha}$ $\quad$ μ_0

Mean of normally distributed population: summary

- Two-sided

σ^2	Decision area		
known	$\mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$	μ_0	$\mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$
unknown	$\mu_0 - \frac{\widetilde{s}_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}$	μ_0	$\mu_0 + \frac{\widetilde{s}_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}$

Here comes the p-value

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

Here comes the p-value

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

If σ^2 is known

Test	Direction of the alternative hypothesis	p-value
Right one-sided	To the right of μ_0	$\mathbb{P}(\overline{X}_n \geq \overline{x}_n \| H_0$ is true)
Lest one-sided	To the left of μ_0	$\mathbb{P}(\overline{X}_n \leq \overline{x}_n \| H_0$ is true)
Two-sided	Both side of μ_0, symmetrically	$\mathbb{P}(\|\overline{X}_n - \mu_0\| \geq \|\overline{x}_n - \mu_0\| \| H_0$ is true)

Here comes the p-value

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

If σ^2 is unknown

Test	Direction of the alternative hypothesis	p-value
Right one-sided	To the right of μ_0	$\mathbb{P}(\frac{\overline{X}_n-\mu_0}{\frac{\widetilde{S}_n}{\sqrt{n}}} \geq \frac{\overline{x}_n-\mu_0}{\frac{\widetilde{s}_n}{\sqrt{n}}} \mid H_0 \text{ is true})$
Lest one-sided	To the left of μ_0	$\mathbb{P}(\frac{\overline{X}_n-\mu_0}{\frac{\widetilde{S}_n}{\sqrt{n}}} \leq \frac{\overline{x}_n-\mu_0}{\frac{\widetilde{s}_n}{\sqrt{n}}} \mid H_0 \text{ is true})$
Two-sided	Both side of μ_0, symmetrically	$\mathbb{P}(\left\lvert\frac{\overline{X}_n-\mu_0}{\frac{\widetilde{S}_n}{\sqrt{n}}}\right\rvert \geq \left\lvert\frac{\overline{x}_n-\mu_0}{\frac{\widetilde{s}_n}{\sqrt{n}}}\right\rvert \mid H_0 \text{ is true})$

Here comes the p-value

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

In any case, we found that

$$p-\text{value}=\alpha \Leftrightarrow \mathbb{P}(\text{Reject } H_0|H_0 \text{ is true})=\alpha.$$

Here comes the p-value

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

In any case, we found that

$$p-\text{value} = \alpha \Leftrightarrow \mathbb{P}(\text{Reject } H_0|H_0 \text{ is true}) = \alpha.$$

Major advantage: you don't need to fix a level of significance before the experiment. The p-value gives you the level on which you reject or not the null hypothesis.

Data snooping

The p-value is computed differently according to the test you are considering.

Data snooping

The p-value is computed differently according to the test you are considering.
In particular, a one-sided test reject H_0"earlier" in its direction than the two-sided test.

Data snooping

The p-value is computed differently according to the test you are considering.
In particular, a one-sided test reject H_0"earlier" in its direction than the two-sided test.

σ^2	Key values
known	$\mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$, $\mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$, μ_0, $\mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$, $\mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$
unknown	$\mu_0 - \frac{\tilde{s}_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}$, $\mu_0 - \frac{\tilde{s}_n}{\sqrt{n}} t_{n-1,1-\alpha}$, μ_0, $\mu_0 + \frac{\tilde{s}_n}{\sqrt{n}} t_{n-1,1-\alpha}$, $\mu_0 + \frac{\tilde{s}_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}$

Data snooping

The p-value is computed differently according to the test you are considering.
In particular, a one-sided test reject H_0"earlier" in its direction than the two-sided test.
It is absolutely FORBIDDEN to start a test as two-sided and then, according to the data collected, to go on as a one-sided test! This illegal practice is named "data snooping".

Use the correct definition !

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

Use the correct definition !

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

The p-value gives an indication on how well the data are coherent with the null hypothesis.

Use the correct definition !

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

The p-value gives an indication on how well the data are coherent with the null hypothesis.

- the p-value is not the probability that the null hypothesis is true;

Use the correct definition !

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

The p-value gives an indication on how well the data are coherent with the null hypothesis.

- the p-value is not the probability that the null hypothesis is true;
- the p-value is not the probability that the null hypothesis is true in view of the value observed for the test-statistic

Use the correct definition !

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

The p-value gives an indication on how well the data are coherent with the null hypothesis.

- the p-value is not the probability that the null hypothesis is true;
- the p-value is not the probability that the null hypothesis is true in view of the value observed for the test-statistic:

$$\mathbb{P}(x|H_0) \neq \mathbb{P}(H_0|x)$$

Use the correct definition !

Definition

The p-value is the probability , under the assumption that the null hypothesis is correct, of obtaining test results at least as extreme, in the direction of the alternative hypothesis, as the result actually observed.

The p-value gives an indication on how well the data are coherent with the null hypothesis.

- the p-value is not the probability that the null hypothesis is true;
- the p-value is not the probability that the null hypothesis is true in view of the value observed for the test-statistic:

$$\mathbb{P}(x|H_0) \neq \mathbb{P}(H_0|x)$$

by Bayes formula:

$$\mathbb{P}(x|H_0) = \frac{\mathbb{P}(H_0|x)\,\mathbb{P}(x)}{\mathbb{P}(H_0)}$$

Investigating $\mathbb{P}(H_0|x)$

Chebyshev's inequality

If X is a random variable with $\mathbb{E}[x] = \mu$ and $\mathrm{Var}[X] = \sigma^2$, for all $r > 0$,

$$\mathbb{P}(|X - \mu| \geq r\sigma) \leq \frac{1}{r^2}$$

Investigating $\mathbb{P}(H_0|x)$

Chebyshev's inequality

If X is a random variable with $\mathbb{E}[x] = \mu$ and $\text{Var}[X] = \sigma^2$, for all $r > 0$,

$$\mathbb{P}(|X - \mu| \geq r\sigma) \leq \frac{1}{r^2}$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$ we even have

- $\mathbb{P}(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0,6827$;
- $\mathbb{P}(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0,9545$;
- $\mathbb{P}(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0,9973$.

Investigating $\mathbb{P}(H_0|x)$

Chebyshev's inequality

If X is a random variable with $\mathbb{E}[x] = \mu$ and $\text{Var}[X] = \sigma^2$, for all $r > 0$,

$$\mathbb{P}(|X - \mu| \geq r\sigma) \leq \frac{1}{r^2}$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$ we even have

- $\mathbb{P}(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0,6827$;
- $\mathbb{P}(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0,9545$;
- $\mathbb{P}(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0,9973$.

David Colquhoun: “if you want to maintain your ratio of false discovery below 5%, you should use the $68 - 95 - 99$ rule or the threshold $0,001$”.

Statement from the American Statistical Society

Statement from the American Statistical Society

- p-values can indicate how incompatible the data are with a specified statistical model.

Statement from the American Statistical Society

- p-values can indicate how incompatible the data are with a specified statistical model.
- p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

Statement from the American Statistical Society

- p-values can indicate how incompatible the data are with a specified statistical model.
- p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Statement from the American Statistical Society

- p-values can indicate how incompatible the data are with a specified statistical model.
- p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Statement from the American Statistical Society

- p-values can indicate how incompatible the data are with a specified statistical model.
- p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Fallacies

"Common Fallacies of Probability in Medical Context: A Simple Mathematical Exposition", Rufaidah Ali Rushdi and Muhammad Rushdi

Fallacies

"Common Fallacies of Probability in Medical Context: A Simple Mathematical Exposition", Rufaidah Ali Rushdi and Muhammad Rushdi

Fallacy	Fallacious Formula	Correct formula
Multiplication Fallacy	$P(A \cap B) = P(A)P(B)$, A and B are general	$P(A \cap B) = P(A\|B)P(B)$ $=P(B\|A)P(A)$ $P(A \cap B) = P(A)P(B)$, A and B are statistically independent
Addition Fallacy	$P(A \cup B) = P(A) + P(B)$, A and B are general	$P(A \cup B)$ $=P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B)$, A and B are mutually exclusive
Inverse Fallacy	$P(A\|B) = P(B\|A)$	$P(A\|B) = \frac{P(A)}{P(B)} P(B\|A)$
Conditional-Marginal Fallacy	$P(A\|B) = P(A)$ A and B are general	$P(A\|B) = P(A)$ A and B are statistically independent
Conjunction Fallacy	$P(A \cap B) > P(A)$ or $P(A \cap B) > P(B)$	$P(A \cap B) \leqslant min(P(A), P(B))$
The Appeal-to-Probability Fallacy	$\{P(A) > 0\} \Rightarrow$ $\{P(A) = 1\}$	$\{P(A) > 0\} \Rightarrow \{P(A) \in (0,1]\}$ trivially
The Base-Rate Neglect	$P(A) = P(\overline{A}) = 0.5$	P(A) is not necessarily equal to $P(\overline{A})$ Typically For medical applications when P(A) denotes disease prevalence $P(A) \ll 1$, $P(\overline{A}) \simeq 1$

Conditional probability

Conditional probability

$$\mathbb{P}(\text{Pope}|\text{Human}) = 1/7 \times 10^{-9}$$

Conditional probability

$$\mathbb{P}(\text{Pope}|\text{Human}) = 1/7 \times 10^{-9}$$

$$\mathbb{P}(\text{Human}|\text{Pope}) = 1$$

Conditional probability

$$\mathbb{P}(\text{Pope}|\text{Human}) = 1/7 \times 10^{-9}$$

$$\mathbb{P}(\text{Human}|\text{Pope}) = 1$$

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

Conditional probability

$$\mathbb{P}(\text{Pope}|\text{Human}) = 1/7 \times 10^{-9}$$

$$\mathbb{P}(\text{Human}|\text{Pope}) = 1$$

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

Less funny if it concerns the efficiency of a medical test, a treatment,...

Favourable event

We say that an event A is favourable to an other event B if

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

Favourable event

We say that an event A is favourable to an other event B if

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

British Home Office: "Among 1221 women murder victims between 1984 and 1988, 44% were killed by their husbands or lovers, 18% by other relatives, another 18% by friends or acquaintances. Only 14% were killed by strangers"

Favourable event

We say that an event A is favourable to an other event B if

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

British Home Office: "Among 1221 women murder victims between 1984 and 1988, 44% were killed by their husbands or lovers, 18% by other relatives, another 18% by friends or acquaintances. Only 14% were killed by strangers"
Do we conclude that marriage is favourable to murder?

Favourable event

We say that an event A is favourable to an other event B if

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

British Home Office: “Among 1221 women murder victims between 1984 and 1988, 44% were killed by their husbands or lovers, 18% by other relatives, another 18% by friends or acquaintances. Only 14% were killed by strangers”
Do we conclude that marriage is favourable to murder? NO.

Correlation

The correlation is used to measure if there exist a linear relationship between two variables.

Correlation

The correlation is used to measure if there exist a linear relationship between two variables.
With the data sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, we use the formula

$$r_{x,y} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}}.$$

Correlation

The correlation is used to measure if there exist a linear relationship between two variables.
With the data sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, we use the formula

$$r_{x,y} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}}.$$

which is the "observable" version

$$\frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

Correlation

The correlation is used to measure if there exist a linear relationship between two variables.
With the data sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, we use the formula

$$r_{x,y} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}}.$$

which is the "observable" version

$$\frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

It is then a *purely mathematical relationship*.

Correlation

The correlation is used to measure if there exist a linear relationship between two variables.
With the data sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$, we use the formula

$$r_{x,y} = \frac{\sum_{j=1}^{n} (x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n} (x_j - \overline{x})^2 \sum_{j=1}^{n} (y_j - \overline{y})^2}}.$$

which is the “observable” version

$$\frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

It is then a *purely mathematical relationship*. It does not mean that there is causal relation between your variable.

Be careful with Nicolas!