Third-year ISAE-SUPAERO engineering students
Research Area: Neuro & AI
December, 2020

Experimentation, measurements and brain-computer interface

Introduction to statistics for experimentation

Nicolas Drougard[1]

[1]ISAE-SUPAERO DCAS, Toulouse, FRANCE

nicolas.drougard@isae-supaero.fr

Outline

References

- [GS20] X. Gendre and F. Simatos, *Probabilités et statistique*, Tronc Commun Scientifique 1A, Formation Ingénieur ISAE-SUPAERO, 2020.
- [HMS20] G. Haine, D. Matignon, and M. Salaün, *Mathématiques déterministes*, Tronc Commun Scientifique 1A, Formation Ingénieur ISAE-SUPAERO, 2020.
- [Wik] Wikistat, *Statistique et machine learning de statisticien à data scientist*, web page **here**.

Outline

What are statistics?

Statistics or statistical data

Set of data observed on a specific phenomenon.

Statistical methods

A set of scientific methods for:

- **Planning** the collection of observations
  Which data? How much data? Questions?
- **Analyze** & **interpret** a large volume of data to extract relevant information.

Useful material [Wik]

Applications

- **economy:** economic forecasts, market studies, *etc.*
- **imaging:** shape detection, image classification, denoising, image reconstruction, *etc.*
- **agriculture:** crop yields, experiments with new strains, *etc.*
- **biology:** concentration of given molecules in the human body, evolution of species, *etc.*
- **engineering:** quality control, robotics, security, logistics, *etc.*
- **medicine, pharmacology**: epidemiology, experimentation of new treatments, medical imaging, *etc.*
- **psycho-physiological studies:** behavioural studies, neuroimaging cognitive, mental state detection, brain-computer interfaces *etc.*

Methodology

Population & Sample

- **Population:** set of all possible observations that can be made about a phenomenon.
- **Sample:** subset of the population.

Steps in a statistical analysis

1. Planning an experiment and collecting data
   - *representative sample, sample size,*
   - *relevance of the variables selected.*
2. descriptive statistics
   - *data formatting and description.*
3. Inferential statistics
   - *probable results on the phenomenes described by the data.*

Main stages

An essential first step in any statistical analysis:

Descriptive (or exploratory) statistics

- describes the observations made on a sample,
- methods for synthesizing data:
  - relevant numerical summaries
  - graphic representations.

Conclusions valid only for the sample: no generalization to the population!

Inferential statistics (or mathematical statistics)

- extrapolate the observed results to the general population
- inductive approach to generalize the results
- is based on the theory of probability.

Probabilistic modeling

Probability Theory

- provides a mathematical model of random phenomena
- allows you to quantify chance and predict the frequency of occurrence of events by calculating probabilities.

This is the **underlying model** of the realization of the experiment.

A deductive approach

- from model to experience,
- allows the properties of samples from a population to be studied using a probabilistic model.

Probabilistic modeling

Random Variables

- **quantitative** variable: real-valued variable
  ex: age, height, salary, heart rate
- **qualitative** (categorical) variable: indicates which category the individual belongs to
  ex: gender, opinion, task type

Variables

Quantitative variables

- **continuous** variable: any value that cannot be counted in a given interval
  ex: height, weight, income, duration of a flight, etc.
- **discrete** variable: discrete values that can be counted (often integers), *ex: number of failures per month of a machine, number of aircraft accidents during the year, etc.*

Variables

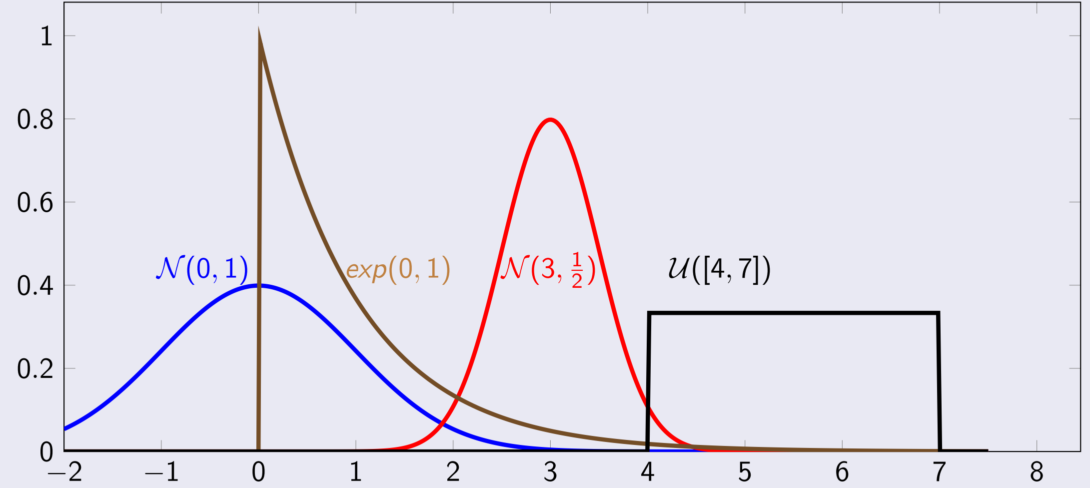
Qualitative variables

- **ordinal** variable: distinct categories in order, without being able to quantify the distance between them
  ex: physical condition (poor, average, good),
  task type (easy, normal, hard)
- **nominal** variable: separate non-ordered categories to which a name can be assigned
  ex: gender, color

Outline

Probability distribution

Density (continuous variable)



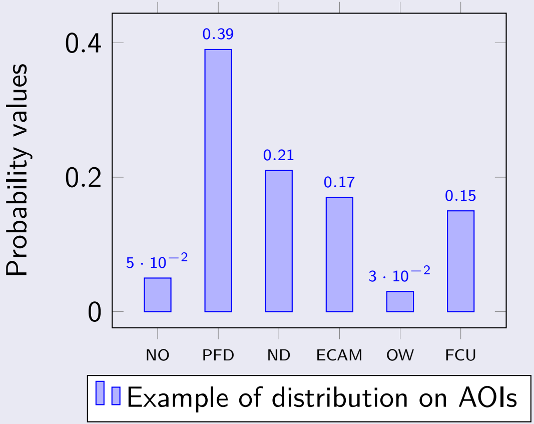
Density of famous distributions

- **Uniform** $\mathcal{U}(a, b)$ (with $a < b$): $\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$
  ex: the quantity of a liquid in a tank, current position of an elevator, etc.
- **Normal, gaussian** $\mathcal{N}(\mu, \sigma^2)$: $\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  ex: sum of many independent processes (such as measurement errors).
  Note that $\mathcal{N}(0, 1)$ is called the **standard normal distribution**.
- **Exponential** $\mathcal{E}(\lambda)$: $\lambda e^{-\lambda x}$
  ex: memoryless durations whose mean is λ.

Probability distribution

Other variables

For the other types of variables (discrete, ordinal, nominal) it is sufficient to define a probability for each event



AOI (Area Of Interest):
1 Primary Flight Display (PFD)
2 Navigation Display (ND)
3 Electronic Centralized Aircraft Monitoring (ECAM)
4 Out of Window (OW)
5 Flight Control Unit (FCU)
Example from *Lounis & all, 2018.*

- **Bernoulli** $Ber(p)$: $0 \leqslant p \leqslant 1$, $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$
  ex: flip a regular coin ($p = 0.5$).
- **Binomial** $Bin(n, p)$: $\forall k \in \{0, \ldots, n\}$, $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
  ex: sum of the results of n flips of coin.
- **Poisson** $\mathcal{P}(\lambda)$: $\forall k \in \{0, \ldots, n\}$ $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
  ex: number of events occurring in a fixed interval of time (durations between events have an exponential distribution).

Notations

- Random variable $X \sim P_X$ with values in $\mathcal{X}$.
- Random dataset $D_n = \{ X_i \}_{i=1}^n$.
- Independant $X_i \Rightarrow D_n \sim P_X^n$.
- Consider a real random variable $\mathcal{X} = \mathbb{R}$.
- Cumulative Distribution Function (CDF): $F_X(t) = \mathbb{P}(X \leqslant t)$.
- Empirical CDF: $F_{X,n}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{ X_i \leqslant t \}}$.
- $\forall t \in \mathbb{R}$, $F_{X,n}(t) \xrightarrow[n \to +\infty]{a.s.} F_X(t)$ (Strong Law of Large Numbers SLLN, see Theorem 3.1 [GS20]).
- Quantile function: $F_X^{-1}(s) = \inf \{ t \in \mathbb{R} \mid F(t) \geqslant s \}$.

A preferred distribution

A measurement is often based on a sum of i.i.d. samples. The Central Limit Theorem (*cf.* Theorem 3.14 [GS20]) provides us with a distribution that should be a good candidate: the normal distribution $\mathcal{N}\left(\mu, \sigma^2\right)$.

Assume that

- *the data we manipulate are real numbers* ($X \in \mathbb{R}$)
- *they come from a normal distribution*
- *whose variance is known to be* σ^2

the random dataset: independent and identically distributed (i.i.d) random variables

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

Goal: we would like to have more information on the population mean.

Is it greater/lower/equal to some $\mu_0 \in \mathbb{R}$*?*

Distribution of the empirical average

The distribution of the empirical mean, $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is also a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

- $\mathbb{E}\left[\overline{X_n}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$ [linear expectation]
- $Var\left[\overline{X_n}\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$ [variance property for i.i.d. variables]

Transformation to a standard normal distribution

- If $\mathbb{E}\left[X\right] = \mu$, then $\mathbb{E}\left[X - \mu\right] = 0$.
- If $Var\left[X\right] = \sigma^2$, then $Var\left[\frac{X}{\sigma}\right] = \frac{1}{\sigma^2} Var\left[X\right] = 1$.

Using these results, we know that

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

We want to be able to say that
"the mean (of the population of interest) μ is **lower** than μ_0"
only when the risk of being wrong is low.

A good idea is to look if $\overline{X_n}$ is low enough to limit the risk of being wrong:

decision

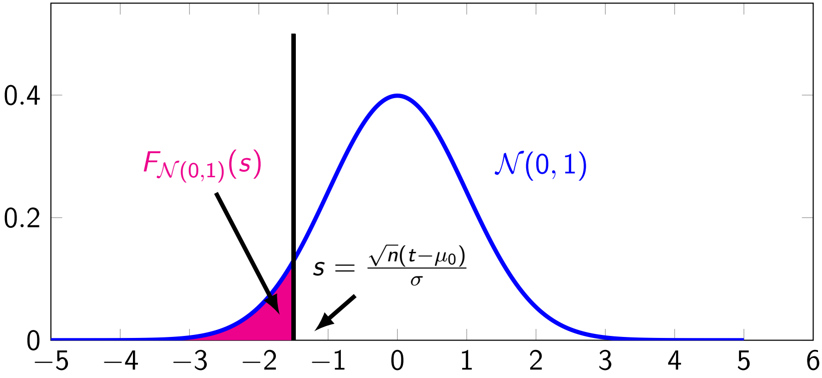
- " Not sure if μ is **lower** than μ_0 or not" (H_0 accepted) when $\overline{X_n} > t$,
- " Almost certain that μ is **lower** than μ_0" (H_0 rejected) when $\overline{X_n} \leqslant t$.

But... what is the value of $t \in \mathbb{R}$?

Suppose that $\mu > \mu_0$: the risk of being wrong is the probability that $\overline{X_n} \leqslant t$, denoted by α (and usually equal to 0.1, 0.05 or 0.01).

Risk, or probability to be wrong

$\alpha(\mu) = \mathbb{P}\left(\overline{X_n} \leqslant t\right) = \mathbb{P}\left(\frac{\sqrt{n}(\overline{X_n}-\mu)}{\sigma} \leqslant \frac{\sqrt{n}(t-\mu)}{\sigma}\right) = F_{\mathcal{N}(0,1)}(\frac{\sqrt{n}(t-\mu)}{\sigma})$ where $F_{\mathcal{N}(0,1)}$ is the cumulative density function of $\mathcal{N}(0,1)$: $F_{\mathcal{N}(0,1)}(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{x^2}{2}}dx$.

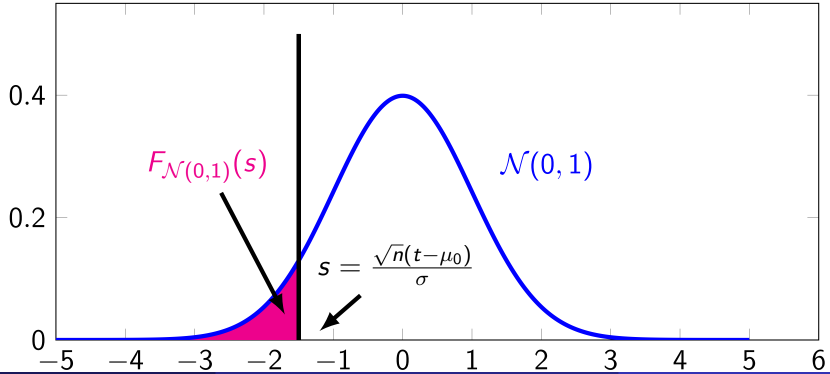


Risk, or probability to be wrong

$\alpha = sup_{\mu > \mu_0}\alpha(\mu) = sup_{\mu > \mu_0} F_{\mathcal{N}(0,1)}(\frac{\sqrt{n}(t-\mu)}{\sigma}) = F_{\mathcal{N}(0,1)}(\frac{\sqrt{n}(t-\mu_0)}{\sigma})$ since $F_{\mathcal{N}(0,1)}$ is an increasing function.

$\rightarrow \frac{\sqrt{n}(t-\mu_0)}{\sigma} = F^{-1}_{\mathcal{N}(0,1)}(\alpha) = s$

$\rightarrow t = \frac{\sigma}{\sqrt{n}}s + \mu_0$: using t to decide, if $\mu > \mu_0$ (H_0), the probability to be wrong is α.



Some vocabulary

Definitions

- **Statistical hypothesis:** a statement about the parameters describing a population (not a sample).
- **Null hypothesis** H_0: a hypothesis associated with a contradiction to a theory one would like to prove.
- **Alternative hypothesis** H_1: a hypothesis associated with a theory one would like to prove.

In the previous example, "$H_0 : \mu > \mu_0$" is the null hypothesis and "$H_1 : \mu \leqslant \mu_0$" is the alternative hypothesis.

Definitions

- **Statistical test:** a procedure whose inputs are samples and whose result is a hypothesis.
- **Test statistic:** a function of the variables in the dataset, denoted by S.
- **Region of acceptance:** the set of values of the test statistic for which we fail to reject the null hypothesis.
- **Region of rejection / Critical region:** the set of values of the test statistic for which the null hypothesis is rejected. This region, depending on the threshold t, is denoted by $\mathcal{R}(t)$.
- **Critical value:** the threshold value delimiting the regions of acceptance and rejection for the test statistic.

In the previous example, the test was to calculate the statistic $S = \overline{X_n}$ and see if it was lower or greater than t. The region of acceptance was $]t, +\infty[$, the region of rejection $]-\infty, t]$, and the critical value was t.

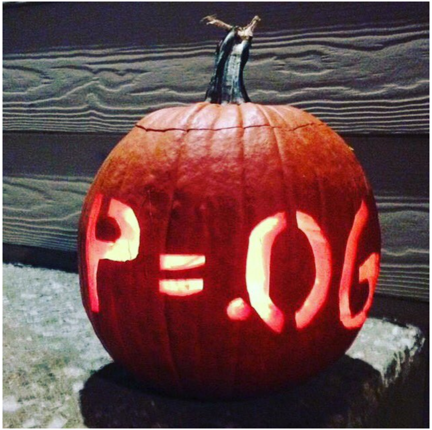
Definitions

- **p-value:** the probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.

Recall that $\alpha = \mathbb{P}_{H_0}(S \in \mathcal{R}(t))$. Given the realized/observed statistics s, it is the smallest risk for which the null hypothesis is rejected:

$$p = \inf_{\substack{t \in \mathbb{R} \\ \text{such that} \\ s \in \mathcal{R}(t)}} \mathbb{P}_{H_0}(S \in \mathcal{R}(t)).$$

In the previous example, $p = \mathbb{P}_{\mu > \mu_0}\left(\overline{X_n} < \overline{x_n}\right)$. If $p > \alpha$, the test does not reject H_0. Usually $\alpha = 0.05$.



Exercice

create the stat. test with

- "$H_0 : \mu < \mu_0$" "$H_1 : \mu \geqslant \mu_0$"?
- "$H_0 : \mu = \mu_0$" "$H_1 : \mu \neq \mu_0$"?
- unknown variance?
    - scipy.stats: `ttest_1samp`

Two sets $(X_1, \ldots, X_{n_1})$ & $(Y_1, \ldots, Y_{n_2})$. We don't know the variance σ^2.
Is mean difference statistically significant ?

Suppose that means of the populations are equal (H_0).

Then, $S = \dfrac{(\overline{X_n} - \overline{Y_n})}{\widehat{\sigma_M}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim Student(n_1 + n_2 - 2) = t(n_1 + n_2 - 2)$.

That is, with

- $\overline{X_n} = \frac{1}{n_1} \sum_{i=1}^{n} X_i$, $\overline{Y_n} = \frac{1}{n_2} \sum_{i=1}^{n} Y_i$,
- $\widehat{\sigma_M} = \sqrt{\frac{(n_1-1)v_1+(n_2-1)v_2}{n_1+n_2-2}}$, with $v_1 = \frac{1}{n-1} \sum_{i=1}^{n}(X_i - \overline{X_n})^2$ and $v_2 = \frac{1}{n-1} \sum_{i=1}^{n}(Y_i - \overline{Y_n})^2$.

S is distributed according to a Student law with parameter (degree of freedom) $n_1 + n_2 - 2$.

If the realisation of S, denoted by s, is too far from 0, we can conclude that theoretical means (means of the population) are not equals. ▶ scipy.stats: `ttest_ind`

Null hypothesis is H_0 : "means of both datasets are equal". If the realisation of S is too far from 0, we can conclude that theoretical means (means of the population) are not equals.
Rejection region: $\mathcal{R}(t) = \{ s \in \mathbb{R} \mid \; |s| > t \}$.

p value

$$
\begin{aligned}
p \quad &\overset{def}{=} \inf_{\substack{t\in\mathbb{R} \\ \text{such that} \\ s\in\mathcal{R}(t)}} \mathbb{P}_{H_0}\Big(S \in \mathcal{R}(t)\Big) \overset{\substack{null \\ hypoth.}}{=} \inf_{\substack{t\in\mathbb{R} \\ \text{such that} \\ s\in\mathcal{R}(t)}} \mathbb{P}_{H_0}\Big(|S| > t\Big) \\
&\overset{\substack{t\mapsto\mathbb{P}_{H_0}(\,|S|>t\,) \\ decreasing \\ function}}{=} \lim_{t\to|s|} \mathbb{P}_{H_0}\Big(|S| > t\Big) \overset{\substack{continuous \\ cum.dens.func.}}{=} \mathbb{P}_{H_0}\Big(|S| > |s|\Big) \\
&= \mathbb{P}_{H_0}\Big(S > |s|\Big) + \mathbb{P}_{H_0}\Big(S < -|s|\Big) = 2\mathbb{P}_{H_0}\Big(S > |s|\Big) \\
&= 2\Big(1 - F_{t(n_1+n_2-2)}(|s|)\Big).
\end{aligned}
$$

Outline

Let $d_n = \{y_1, \dots, y_n\}$ be a dataset, realization/observation of the random vector Y:

$$\begin{array}{ccccccccc} \boldsymbol{Y} & = & \boldsymbol{\mu} & + & \boldsymbol{\varepsilon} & = & \mathbb{X}\,\boldsymbol{\theta} & + & \boldsymbol{\varepsilon} \in \mathbb{R}^n \\ \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} & = & \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} & + & \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} & = & \begin{pmatrix} x_{1,1} & \dots & x_{1,d} \\ x_{2,1} & \dots & x_{2,d} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,d} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} & + & \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \end{array}$$

Definition of the terms of this model

- the random vector $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, **gaussian random noise** with variance $\sigma^2 \in \mathbb{R}_+^*$, and **independent and identically distributed** (i.i.d) components $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$,
- the constant **parameter vector** $\theta \in \mathbb{R}^d$, with $d < n$,
- the constant **injective linear function** $\mathbb{X} \in \mathbb{R}^{n \times d}$,
- the constant **mean vector** $\mu = \mathbb{X}\theta \in V$,
- the d-**dimensional vector subspace** $V = \mathrm{Im}(\mathbb{X}) \subset \mathbb{R}^n$ (possible mean vectors).

General Linear Model

Some explanations:

- $\mathbb{X} \in \mathbb{R}^{n \times d}$ defines $V = \mathrm{Im}(\mathbb{X})$,
  i.e. expert **hypotheses on the possible mean vectors** μ.
- $\theta \in \mathbb{R}^{d}$ **parameterizes** the possible mean vectors μ,
  these parameters are identifiable since $\mathbb{X}$ is injective.
- As we'll see, many classical statistical studies are based on this model.

Maximum Likelihood Estimator of (μ, σ^2)

The Maximum Likelihood Estimator (MLE) of $(\mu, \sigma^2) \in \mathbb{R}^n \times \mathbb{R}_+^*$ is $\left(\widehat{\mu_n} = \Pi_V Y, \widehat{\sigma_n^2}\right)$ where

- Π_V is the **orthogonal projection matrix** on the d-dimensional vector subspace $V \subset \mathbb{R}^n$ containing the possible mean vectors,
- the **variance estimator** $\widehat{\sigma_n^2} = \frac{1}{n} \| Y - \Pi_V Y \|^2 = \frac{1}{n} \sum_{i=1}^n [Y_i - (\Pi_V Y)_i]^2$, is the squared euclidean distance between Y and the projection $\Pi_V Y$ divided by n.

Proof – 1

Since $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ $\forall i \in \{1, \ldots, n\}$ and $\{Y_i\}_{i=1}^n$ i.i.d, the density of $Y_i \in \mathbb{R}$ is $f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right)$ and the likelihood $\mathcal{L}$, that depends on the parameters $(\mu, \sigma^2) \in \mathbb{R}^n \times \mathbb{R}_+^*$ and the random vector Y, is then $\mathcal{L}(\mu, \sigma^2, Y)$

$$= \prod_{i=1}^n f_{Y_i}(Y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \mu_i)^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu_i)^2\right).$$

Proof – 2

The likelihood is $\mathcal{L}(\mu, \sigma^2, Y) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\|Y-\mu\|^2\right)$.

- Let's fix $\sigma^2 \in \mathbb{R}_+^*$: maximum reached for $\mu \in V$ that minimizes $\|Y-\mu\|$, *i.e.* $\widehat{\mu_n} = \Pi_V Y$ (*cf.* minimal distance to a vector subspace, Corollaries 7.6. & 7.7. in [HMS20]).
- Now, the maximum is reached for σ^2 that maximizes $\log\left(\mathcal{L}\left(\Pi_V Y, \sigma^2, Y\right)\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|Y - \Pi_V Y\|^2$.

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}\log\left(\mathcal{L}\left(\Pi_V Y, \sigma^2, Y\right)\right) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y-\Pi_V Y\|^2 \\
\Rightarrow \quad \frac{\partial}{\partial\sigma^2}\log\left(\mathcal{L}\left(\Pi_V Y, \sigma^2, Y\right)\right) \geqslant 0 \quad \Leftrightarrow & \quad \sigma^2 \leqslant \frac{1}{n}\|Y-\Pi_V Y\|^2 \\
\Rightarrow \quad \frac{\partial}{\partial\sigma^2}\log\left(\mathcal{L}\left(\Pi_V Y, \sigma^2, Y\right)\right) = 0 \quad \Leftrightarrow & \quad \sigma^2 = \frac{1}{n}\|Y-\Pi_V Y\|^2.
\end{aligned}$$

Finally, $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2 = \frac{1}{n}\|Y - \widehat{\mu_n}\|^2$. $\square$

Remark: $\mathbb{E}_{\mu,\sigma^2}\left[\Pi_V Y\right] = \Pi_V \mathbb{E}_{\mu,\sigma^2}\left[Y\right] = \Pi_V \mu = \mu$ since $\mu \in V$: $\Pi_V Y$ is unbiased.

General Linear Model

Result: orthogonal projection matrix Π_V expressed with $\mathbb{X}$

If $\mathbb{X} \in \mathbb{R}^{n\times d}$ is injective, then

- $\mathbb{X}^T\mathbb{X}$ is invertible,
- $\Pi_V = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$.

Proof

• $\mathbb{X}^T\mathbb{X}\theta = 0 \implies \theta^T\mathbb{X}^T\mathbb{X}\theta = 0 \implies (\mathbb{X}\theta)^T\mathbb{X}\theta = 0 \implies \|\mathbb{X}\theta\|^2 = 0 \implies \mathbb{X}\theta = 0$
$\xRightarrow{\mathbb{X} \text{ injective}} \theta = 0 \implies \ker(\mathbb{X}^T\mathbb{X}) = \{0\} \implies \mathbb{X}^T\mathbb{X} \in \mathbb{R}^{d\times d}$ is injective $\implies \mathbb{X}^T\mathbb{X}$ is invertible.

• Now, $\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right)^T = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \implies \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ is symetric.
$\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right)\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right) = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \implies \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ is idempotent.
Symetric + Idempotent $\Rightarrow \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ orthogonal projection matrix.

• $\mathbb{X}$ on the left $\Rightarrow \mathrm{Im}\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right) \subseteq V \overset{def}{=} \mathrm{Im}(\mathbb{X})$. • $\forall y \in V$, $\exists! \theta \in \mathbb{R}^d$ such that
$\mathbb{X}\theta = y$: $\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\theta = \mathbb{X}\theta = y \Rightarrow V \subseteq \mathrm{Im}\left(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right)$. □

Maximum Likelihood Estimator of (θ, σ^2)

If $\mathbb{X} \in \mathbb{R}^{n \times d}$ is injective, the Maximum Likelihood Estimator (MLE) of $(\theta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+^*$ is $\left(\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1} X^T Y, \widehat{\sigma_n^2}\right)$ where the variance estimator is still

$$\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2 = \frac{1}{n}\left\|Y - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y\right\|^2.$$

Proof

The likelihood is $\mathcal{L}(\theta, \sigma^2, Y) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\|Y - \mathbb{X}\theta\|^2\right)$.

- Let's fix $\sigma^2 \in \mathbb{R}_+^*$: maximum reached for $\theta \in \mathbb{R}^d$ that minimizes $\|Y - \mathbb{X}\theta\|$, *i.e.* when $\mathbb{X}\theta = \Pi_V Y$ *i.e.* when $\mathbb{X}\theta = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y \xRightarrow{\mathbb{X} injective} \widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$.
- Now, the maximum is reached for σ^2 that maximizes $\log\left(\mathcal{L}\left(\widehat{\theta_n}, \sigma^2, Y\right)\right)$ $= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left\|Y - \mathbb{X}\widehat{\theta_n}\right\|^2 = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|Y - \Pi_V Y\|^2$.

  So $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2 = \frac{1}{n}\left\|Y - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y\right\|^2$ (*cf.* proof slide 32 & Π_V expr.).

  □

General Linear Model (GLM)

Random variables $Y_i \in \mathbb{R}$ with **means** μ_i and **noise vector** $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.

Injective matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ with **image** $V \stackrel{def}{=} \operatorname{Im}(\mathbb{X})$, and **parameters** $\theta \in \mathbb{R}^d$.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \dots & x_{1,d} \\ x_{2,1} & \dots & x_{2,d} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,d} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbb{X}\,\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Maximum Likelihood Estimators (MLEs)

- MLE of the **mean** $\widehat{\mu_n} = \Pi_V Y = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y = \mathbb{X}\widehat{\theta_n}$.
- MLE of the **parameters** $\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$.
- MLE of the **variance** $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - (\Pi_V Y)_i \right)^2$.

Under the GLM hypotheses, we know from now the expressions of

- the MLE of the mean vector $\widehat{\mu_n} \in \mathbb{R}^n$,
- the MLE of the parameter vector $\widehat{\theta_n} \in \mathbb{R}^d$,
- and the MLE of the variance $\widehat{\sigma_n^2} \in \mathbb{R}_+^*$,

as functions of $\mathbb{X}$ & Y.

Next question

What are the distributions of these random quantities ?

Reminders

Some famous probability distributions

- A **chi-squared** distribution with k degrees of freedom, denoted by $\mathcal{X}^2(k)$, is the distribution of $\|X\|^2$ with $X \sim \mathcal{N}(0, I_k)$, *i.e.* $\sum_{i=1}^{k} X_i^2 \sim \mathcal{X}^2(k)$, if $X_i \in \mathcal{N}(0,1)$ i.i.d. (*cf.* Def. 4.3 of [GS20]).
- A **Student** distribution with k degrees of freedom, denoted by $t(k)$, is the distribution of $\frac{X}{\sqrt{Y/k}}$ with $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{X}^2(k)$ and $X \perp\!\!\!\perp Y$.
- A **Fisher** distribution with k_1 and k_2 degrees of freedom, denoted by $\mathcal{F}(k_1, k_2)$, is the distribution of $\frac{X/k_1}{Y/k_2}$ with $X \sim \mathcal{X}^2(k_1)$, $Y \sim \mathcal{X}^2(k_2)$ and $X \perp\!\!\!\perp Y$.

These distributions have known continuous densities, hence their cumulative density functions and their quantile functions can be computed.

Cochran Theorem

Let us consider

- a centered random vector $X \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$, with variance $\sigma^2 > 0$,
- $\{E_i\}_{i=1}^{p}$ orthogonal vector subspaces such that $\bigoplus_{i=1}^{p} E_i = \mathbb{R}^n$ and $E_i \perp E_j \; \forall i \neq j$,
- $\Pi_{E_i} \in \mathbb{R}^{n \times n}$ the orthogonal projection on E_i.

Then,

1. **distribution of projections**: $\Pi_{E_i} X \sim \mathcal{N}\left(0, \sigma^2 \Pi_{E_i}\right) \; \forall i \in \{1, \ldots, p\}$.
   independance: $\forall 1 \leqslant i, j \leqslant p$, $\Pi_{E_i} X \perp\!\!\!\perp \Pi_{E_j} X$, *i.e.* $\forall 1 \leqslant k, l \leqslant n$, $(\Pi_{E_i} X)_k \perp\!\!\!\perp (\Pi_{E_j} X)_l$.
2. **distribution of sum of squares**: $\frac{1}{\sigma^2} \|\Pi_{E_i} X\|^2 \sim \mathcal{X}^2\Big(dim(E_i)\Big)$, $\forall i \in \{1, \ldots, p\}$.
   independance: $\forall 1 \leqslant i, j \leqslant p$, $\left\|\Pi_{E_i} X\right\|^2 \perp\!\!\!\perp \left\|\Pi_{E_j} X\right\|^2$.

Cochran Theorem

Random vector $X \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$, $\{E_i\}_{i=1}^p$ orthogonal vector subspaces with $\bigoplus_{i=1}^p E_i = \mathbb{R}^n$, $\Pi_{E_i} \in \mathbb{R}^{n\times n}$ orthogonal projection on E_i.

1. $\Pi_{E_i} X \sim \mathcal{N}\left(0, \sigma^2 \Pi_{E_i}\right)$ and $\Pi_{E_i} X \perp\!\!\!\perp \Pi_{E_j} X \quad \forall (i,j) \in \{1, \dots, p\}^2$.
2. $\frac{1}{\sigma^2}\left\|\Pi_{E_i} X\right\|^2 \sim \mathcal{X}^2\left(dim(E_i)\right)$ and $\left\|\Pi_{E_i} X\right\|^2 \perp\!\!\!\perp \left\|\Pi_{E_j} X\right\|^2, \quad \forall (i,j) \in \{1, \dots, p\}^2$.

Proof –1

Orthonormal basis of $\mathbb{R}^n$: $\{\overbrace{e_1, e_2, \dots, e_{n_1}}^{\text{basis of } E_1}, \overbrace{e_{n_1+1}, \dots, e_{n_1+n_2}}^{\text{basis of } E_2}, \dots, \overbrace{\dots, e_{n-1}, e_n}^{\text{basis of } E_p}\}$.

$\Rightarrow \langle e_i, e_j \rangle = \mathbb{1}_{\{i=j\}}\ \forall 1 \leqslant i,j \leqslant n$. Let's define $M = \begin{pmatrix} e_1^T \\ \vdots \\ e_n^T \end{pmatrix} = \begin{pmatrix} (e_1)_1 & \dots & (e_1)_n \\ \vdots & & \vdots \\ (e_n)_1 & \dots & (e_n)_n \end{pmatrix}$.

• MX is a gaussian vector as a linear transformation of the gaussian vector X (*cf.* Def. 4.2 of [GS20]). • $\mathbb{E}[MX] = M\mathbb{E}[X] = 0$. • $Var(MX) = MVar(X)M^T = \sigma^2 MM^T = \sigma^2 I_n$.

Cochran Theorem

Random vector $X \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$, $\{E_i\}_{i=1}^p$ orthogonal vector subspaces with $\bigoplus_{i=1}^p E_i = \mathbb{R}^n$, $\Pi_{E_i} \in \mathbb{R}^{n \times n}$ orthogonal projection on E_i.

1. $\Pi_{E_i} X \sim \mathcal{N}\left(0, \sigma^2 \Pi_{E_i}\right)$ and $\Pi_{E_i} X \perp\!\!\!\perp \Pi_{E_j} X \quad \forall (i,j) \in \{1, \ldots, p\}^2$.
2. $\frac{1}{\sigma^2} \|\Pi_{E_i} X\|^2 \sim \mathcal{X}^2\left(dim(E_i)\right)$ and $\left\|\Pi_{E_i} X\right\|^2 \perp\!\!\!\perp \left\|\Pi_{E_j} X\right\|^2, \quad \forall (i,j) \in \{1, \ldots, p\}^2$.

Proof –2

- MX gaussian, • $\mathbb{E}[MX] = 0$, • $Var(MX) = \sigma^2 I_n \overset{Def.}{\Longrightarrow} MX = \begin{pmatrix} \langle e_1, X \rangle \\ \vdots \\ \langle e_n, X \rangle \end{pmatrix} \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.
- MX gaussian, • $Var(MX)$ diagonal $\xrightarrow{Prop.4.2 \text{ [GS20]}}$ $(MX)_i = \langle e_i, X \rangle \perp\!\!\!\perp (MX)_j$, $\forall i \neq j$.
- $\Pi_{E_i} X = \sum_{k | e_k \in E_i} \langle e_k, X \rangle e_k$, • $\{k \mid e_k \in E_i\} \cap \{k \mid e_k \in E_j\} = \emptyset \Longrightarrow$

$\Big\{ \quad \Pi_{E_i} X \perp\!\!\!\perp \Pi_{E_j} X$

Cochran Theorem

Random vector $X \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$, $\{E_i\}_{i=1}^p$ orthogonal vector subspaces with $\bigoplus_{i=1}^p E_i = \mathbb{R}^n$, $\Pi_{E_i} \in \mathbb{R}^{n\times n}$ orthogonal projection on E_i.

1. $\Pi_{E_i} X \sim \mathcal{N}\left(0, \sigma^2 \Pi_{E_i}\right)$ and $\Pi_{E_i} X \perp\!\!\!\perp \Pi_{E_j} X \quad \forall (i,j) \in \{1, \ldots, p\}^2$.
2. $\frac{1}{\sigma^2} \|\Pi_{E_i} X\|^2 \sim \mathcal{X}^2\Big(dim(E_i)\Big)$ and $\Big\|\Pi_{E_i} X\Big\|^2 \perp\!\!\!\perp \Big\|\Pi_{E_j} X\Big\|^2, \quad \forall (i,j) \in \{1, \ldots, p\}^2$.

Proof –3

Independence results are shown. Now,

- $\Pi_{E_i} X$ is also a **gaussian vector**, with **mean** $\mathbb{E}\left[\Pi_{E_i} X\right] = \Pi_{E_i} \mathbb{E}\left[X\right] = 0$, and **variance** $Var(\Pi_{E_i} X) = \Pi_{E_i} Var(X) \Pi_{E_i}^T = \sigma^2 I_n \Pi_{E_i} \Pi_{E_i}^T = \sigma^2 \Pi_{E_i}$ (indeed Π_{E_i} is symetric & idempotent) $\implies \Pi_{E_i} X \sim \mathcal{N}\left(0, \sigma^2 \Pi_{E_i}\right)$.
- $\frac{1}{\sigma^2} \|\Pi_{E_i} X\|^2 = \frac{1}{\sigma^2} \Big\langle \sum_{k|e_k \in E_i} \langle e_k, X\rangle e_k, \sum_{k|e_k \in E_i} \langle e_k, X\rangle e_k \Big\rangle = \frac{1}{\sigma^2} \sum_{k|e_k \in E_i} \left(\langle e_k, X\rangle\right)^2$ $= \sum_{k|e_k \in E_i} \left(\frac{\langle e_k, X\rangle}{\sigma}\right)^2$. Since $\frac{\langle e_k, X\rangle}{\sigma} \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$, $\frac{1}{\sigma^2} \|\Pi_{E_i} X\|^2 \sim \mathcal{X}^2\left(\#\{k | e_k \in E_i\}\right)$. $\square$

General Linear Model

Back to the estimation problem: $Y = \mu + \varepsilon$,
with mean $\mu \in V$, vector subspace $V \subset \mathbb{R}^n$, $dim(V) = d$, noise $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ and $\sigma^2 > 0$.

Distribution of the MLE of (μ, σ^2)

Let $(\widehat{\mu_n}, \widehat{\sigma_n^2})$ be the MLE of (μ, σ^2), *i.e.* $\widehat{\mu_n} = \Pi_V Y$ and $\sigma_n^2 = \frac{1}{n} \|Y - \Pi_V Y\|^2$.

- $\widehat{\mu_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$,
- $\widehat{\mu_n} \sim \mathcal{N}\left(\mu, \sigma^2 \Pi_V\right)$,
- $\frac{n\widehat{\sigma_n^2}}{\sigma^2} \sim \mathcal{X}^2(n-d)$.

Proof – 1

Consider $V^\perp$ the orthogonal complement of V: $V \perp V^\perp$ and $V \bigoplus V^\perp = \mathbb{R}^n$.

- $\widehat{\mu_n} = \Pi_V Y = \Pi_V(Y-\mu) + \Pi_V \mu = \Pi_V(Y-\mu) + \mu$ (because $\mu \in V$),
- $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2 = \frac{1}{n}\|\Pi_{V^\perp} Y\|^2 = \frac{1}{n}\|\Pi_{V^\perp}(Y-\mu) + \Pi_{V^\perp}\mu\|^2 = \frac{1}{n}\|\Pi_{V^\perp}(Y-\mu)\|^2$.

We now apply the Cochran Theorem on $Y - \mu = \varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ with $V = E_1$ & $V^\perp = E_2$.

Back to the estimation problem: $Y = \mu + \varepsilon$,
with mean $\mu \in V$, vector subspace $V \subset \mathbb{R}^n$, $dim(V) = d$, noise $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ and $\sigma^2 > 0$.

Distribution of the MLE of (μ, σ^2)

Let $(\widehat{\mu_n}, \widehat{\sigma_n^2})$ be the MLE of (μ, σ^2), *i.e.* $\widehat{\mu_n} = \Pi_V Y$ and $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2$.

- $\widehat{\mu_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$,
- $\widehat{\mu_n} \sim \mathcal{N}\left(\mu, \sigma^2 \Pi_V\right)$,
- $\frac{n\widehat{\sigma_n^2}}{\sigma^2} \sim \mathcal{X}^2(n-d)$.

Proof – 2

Results of the Cochran theorem:

- $\Pi_V(Y-\mu) \perp\!\!\!\perp \Pi_{V^\perp}(Y-\mu)$. Since $\begin{cases} \widehat{\mu_n} = \Pi_V(Y-\mu) + \mu \\ \widehat{\sigma_n^2} = \frac{1}{n}\|\Pi_{V^\perp}(Y-\mu)\|^2 \end{cases}$, we get $\widehat{\mu_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$.
- $\Pi_V(Y-\mu) \sim \mathcal{N}\left(0, \sigma^2 \Pi_V\right) \implies \widehat{\mu_n} = \Pi_V(Y-\mu) + \mu \sim \mathcal{N}\left(\mu, \sigma^2 \Pi_V\right)$.
- $\frac{1}{\sigma^2}\|\Pi_{V^\perp} Y\|^2 \sim \mathcal{X}^2\left(dim(V^\perp)\right) \implies \frac{n\widehat{\sigma_n^2}}{\sigma^2} \sim \mathcal{X}^2(n-d)$.

$\square$

General Linear Model

Now, consider the estimation problem: $Y = \mathbb{X}\theta + \varepsilon$,
with injective matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$, $\theta \in \mathbb{R}^d$, $V = \mathrm{Im}(\mathbb{X})$, noise $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ and $\sigma^2 > 0$.

Distribution of the MLE of (θ, σ^2)

Let $(\widehat{\theta_n}, \widehat{\sigma_n^2})$ be the MLE of (θ, σ^2), *i.e.* $\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$ and $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2$.

- $\widehat{\theta_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$,
- $\widehat{\theta_n} \sim \mathcal{N}\left(\theta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\right)$, and still
- $\frac{n\widehat{\sigma_n^2}}{\sigma^2} \sim \mathcal{X}^2(n-d)$.

Proof

- $\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbb{X})\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\widehat{\mu_n}$. As a measurable function of $\widehat{\mu_n}$, $\widehat{\theta_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$.
- $\widehat{\theta_n} \overset{def}{=} (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$, gaussian vector, as a linear transformation of the gauss. vector Y.

$$\mathbb{E}\left[\widehat{\theta_n}\right] = \mathbb{E}\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y\right] = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{E}[Y] = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\theta = \theta.$$

$$Var(\widehat{\theta_n}) = Var\Big((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y\Big) = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Var\Big(Y\Big)\left[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\right]^T$$
$$= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\sigma^2 I_n\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}. \qquad \square$$

Summary: MLE distributions

General Linear Model (GLM)

Random variables $Y_i \in \mathbb{R}$ with **means** μ_i and **noise vector** $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.

Injective matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ with **image** $V \stackrel{def}{=} \mathrm{Im}(\mathbb{X})$, and **parameters** $\theta \in \mathbb{R}^d$.

$$\boldsymbol{Y} \quad = \quad \boldsymbol{\mu} \quad + \quad \boldsymbol{\varepsilon} \quad = \quad \mathbb{X}\,\boldsymbol{\theta} \quad + \quad \boldsymbol{\varepsilon} \quad \in \mathbb{R}^n$$

MLEs and distributions

- MLE of the **mean** $\widehat{\mu_n} = \Pi_V Y = \mathbb{X}\widehat{\theta_n}$, with distribution $\widehat{\mu_n} \sim \mathcal{N}\left(\mu, \sigma^2 \Pi_V\right)$.
- MLE of the **parameters** $\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$, with distr. $\widehat{\theta_n} \sim \mathcal{N}\left(\theta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\right)$.
- MLE of the **variance** $\widehat{\sigma_n^2} = \frac{1}{n}\|Y - \Pi_V Y\|^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - (\Pi_V Y)_i\right)^2$,

  with distribution $\frac{n\widehat{\sigma_n^2}}{\sigma^2} \sim \mathcal{X}^2(n-d)$.
- **independences** $\widehat{\mu_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$ and $\widehat{\theta_n} \perp\!\!\!\perp \widehat{\sigma_n^2}$.

Under the GLM hypotheses, we know from now the distributions of the maximum likelihood estimators $\widehat{\mu_n} \in \mathbb{R}^n$, $\widehat{\theta_n} \in \mathbb{R}^d$, and $\widehat{\sigma_n^2} \in \mathbb{R}_+^*$.

Next question

Could these distributions help us to test hypotheses on the model's parameters/means ?

Linear Hypothesis Tests: F-test

Likelihood Ratio Test

Matrix $\mathbb{X}^{n\times d}$ is injective, so $V = \mathrm{Im}(\mathbb{X}) \overset{def}{=} \mathbb{X}(\mathbb{R}^d)$ is d-dimensional.
Let $\Theta_0 \subset \mathbb{R}^d$ be a p-dimensional vector subpace of the parameter space $\mathbb{R}^d$. Let us define the p-dimensional vector subspace $W = \mathbb{X}(\Theta_0) \subset V$, image of Θ using the linear transf. $\mathbb{X}$.
The likelihood-ratio test of

$$H_0 : \mu \in W \quad \text{against} \quad H_1 : \mu \notin W$$
$$\Updownarrow$$
$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \notin \Theta_0$$

with significance level (or risk) $\alpha \in]0,1[$, rejects H_0 when

$$\frac{(n-d)\,\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\,\|Y - \Pi_V Y\|^2} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha),$$

where $F^{-1}_{\mathcal{F}(k_1,k_2)}(1-\alpha)$ is the quantile of order $1-\alpha$ of the Fisher distribution $\mathcal{F}(k_1,k_2)$.

Likelihood Ratio Test: F-test

Vector subpaces ■ $V = \mathbb{X}(\mathbb{R}^d)$, $dim(V) = d$, ■ $\Theta_0 \subset \mathbb{R}^d$, $dim(\Theta_0) = p$,
and ■ $W = \mathbb{X}(\Theta_0) \subset V$, $dim(W) = p$.

The likelihood-ratio test of $H_0 : \mu \in W$ $(\theta \in \Theta_0)$ against $H_1 : \mu \notin W$ $(\theta \notin \Theta_0)$
with significance level $\alpha > 0$, rejects H_0 when $\frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha)$.

Proof – 1

The Likelihood-ratio test rejects H_0 when the likelihood ratio $\rho = \frac{\sup_{H_1} \mathcal{L}(\mu,\sigma^2,Y)}{\sup_{H_0} \mathcal{L}(\mu,\sigma^2,Y)}$ is too high.

$\sup_{H_0} \mathcal{L}(\mu,\sigma^2,Y) = \sup_{(\mu,\sigma^2)\in W\times\mathbb{R}^*_+} \mathcal{L}(\mu,\sigma^2,Y) = \mathcal{L}\left(\Pi_W Y, \frac{1}{n}\|Y - \Pi_W Y\|^2, Y\right)$ using MLEs in slide 31 with W. So, since the likelihood is $\mathcal{L}(\mu,\sigma^2,Y) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\|Y-\mu\|^2\right)$,

$$\sup_{H_0} \mathcal{L}(\mu,\sigma^2,Y) = \left(\frac{1}{\frac{2\pi}{n}\|Y - \Pi_W Y\|^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\|Y - \Pi_W Y\|^2}{\frac{2}{n}\|Y - \Pi_W Y\|^2}\right) = \frac{\left(\frac{n}{2\pi}\right)^{\frac{n}{2}}}{\|Y - \Pi_W Y\|^n} e^{-\frac{n}{2}}.$$

Likelihood Ratio Test: F-test

Vector subpaces ▪ $V = \mathbb{X}(\mathbb{R}^d)$, $dim(V) = d$, ▪ $\Theta_0 \subset \mathbb{R}^d$, $dim(\Theta_0) = p$,
and ▪ $W = \mathbb{X}(\Theta_0) \subset V$, $dim(W) = p$.

The likelihood-ratio test of $H_0 : \mu \in W$ $(\theta \in \Theta_0)$ against $H_1 : \mu \notin W$ $(\theta \notin \Theta_0)$ with significance level $\alpha > 0$, rejects H_0 when $\frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha)$.

Proof – 2

- Likelihood ratio $\rho = \frac{\sup_{H_1} \mathcal{L}(\mu,\sigma^2,Y)}{\sup_{H_0} \mathcal{L}(\mu,\sigma^2,Y)}$ with $\sup_{H_0} \mathcal{L}(\mu,\sigma^2,Y) = \frac{\left(\frac{n}{2\pi}\right)^{\frac{n}{2}}}{\|Y-\Pi_W Y\|^n} e^{-\frac{n}{2}}$.
- Now consider the numerator $\sup_{H_1} \mathcal{L}(\mu,\sigma^2,Y) = \sup_{(\mu,\sigma^2)\in(V\setminus W)\times\mathbb{R}_+^*} \mathcal{L}(\mu,\sigma^2,Y)$.

  Let $\{\overbrace{e_1, e_2, \ldots, e_p}^{\text{basis of } W}, e_{p+1}, \ldots, e_d\}$ be an orthonormal basis of V. As Π_V is an orthogonal projection and as $\langle Y, e_d\rangle = \langle Y-\mu, e_d\rangle + \langle \mu, e_d\rangle \sim \mathcal{N}\left(\langle \mu, e_d\rangle, \sigma^2\right)$ (see slide 40),
  $\mathbb{P}(\Pi_V Y \in W) = \mathbb{P}(\forall i > p, \langle \Pi_V Y, e_i\rangle = 0) = \mathbb{P}(\forall i > p, \langle Y, \Pi_V e_i\rangle = 0)$
  $= \mathbb{P}(\forall i > p, \langle Y, e_i\rangle = 0) \leqslant \mathbb{P}(\langle Y, e_d\rangle = 0) = 0 \implies \Pi_V Y \notin W$ almost surely (a.s.).

Likelihood Ratio Test: F-test

Vector subpaces ■ $V = \mathbb{X}(\mathbb{R}^d)$, $dim(V) = d$, ■ $\Theta_0 \subset \mathbb{R}^d$, $dim(\Theta_0) = p$,
and ■ $W = \mathbb{X}(\Theta_0) \subset V$, $dim(W) = p$.

The likelihood-ratio test of $H_0 : \mu \in W$ $(\theta \in \Theta_0)$ against $H_1 : \mu \notin W$ $(\theta \notin \Theta_0)$ with significance level $\alpha > 0$, rejects H_0 when $\frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha)$.

Proof – 3

- Likelihood ratio $\rho = \frac{\sup_{H_1} \mathcal{L}(\mu,\sigma^2,Y)}{\sup_{H_0} \mathcal{L}(\mu,\sigma^2,Y)}$ with $\sup\limits_{H_0} \mathcal{L}(\mu,\sigma^2,Y) = \frac{\left(\frac{n}{2\pi}\right)^{\frac{n}{2}}}{\|Y - \Pi_W Y\|^n} e^{-\frac{n}{2}}$.
- Since $\Pi_V Y \notin W$ a.s., and since $\Pi_V Y$ maximizes $\mathcal{L}(\mu,\sigma^2,Y)$ for $\mu \in V$ (see slide 32), we have $\sup\limits_{H_1} \mathcal{L}(\mu,\sigma^2,Y) = \sup\limits_{(\mu,\sigma^2)\in(V\setminus W)\times\mathbb{R}^*_+} \mathcal{L}(\mu,\sigma^2,Y) = \mathcal{L}\left(\Pi_V Y, \frac{1}{n}\|Y - \Pi_V Y\|^2, Y\right)$

$$= \frac{\left(\frac{n}{2\pi}\right)^{\frac{n}{2}}}{\|Y - \Pi_V Y\|^n} e^{-\frac{n}{2}} \implies \rho = \frac{\|Y - \Pi_W Y\|^n}{\|Y - \Pi_V Y\|^n} = \left(\frac{\|Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2}\right)^{\frac{n}{2}} = \left(\frac{\|Y - \Pi_V Y + \Pi_V Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2}\right)^{\frac{n}{2}}.$$

Likelihood Ratio Test: F-test

Vector subpaces ■ $V = \mathbb{X}(\mathbb{R}^d)$, $dim(V) = d$, ■ $\Theta_0 \subset \mathbb{R}^d$, $dim(\Theta_0) = p$,
and ■ $W = \mathbb{X}(\Theta_0) \subset V$, $dim(W) = p$.

The likelihood-ratio test of $H_0 : \mu \in W$ $(\theta \in \Theta_0)$ against $H_1 : \mu \notin W$ $(\theta \notin \Theta_0)$ with significance level $\alpha > 0$, rejects H_0 when $\frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha)$.

Proof – 4

Likelihood ratio $\rho = \left(\frac{\|Y - \Pi_V Y + \Pi_V Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2}\right)^{\frac{n}{2}}$ with $Y - \Pi_V Y \in V^\perp$ & $\Pi_V Y - \Pi_W Y \in V$.

Using Pythagore theorem, $\rho = \left(\frac{\|Y - \Pi_V Y\|^2 + \|\Pi_V Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2}\right)^{\frac{n}{2}} = \left(1 + \frac{\|\Pi_V Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2}\right)^{\frac{n}{2}}$.

Now, under $H_0 : \mu \in W$, $\|\Pi_V Y - \Pi_W Y\|^2 = \|\Pi_V(Y-\mu) - \Pi_W(Y-\mu) + \Pi_V\mu - \Pi_W\mu\|^2$
$= \|\Pi_V(Y-\mu) - \Pi_W(Y-\mu) + \mu - \mu\|^2 = \|\Pi_{V\cap W^\perp}(Y-\mu)\|^2$.

Note that $(V \cap W^\perp) \perp V^\perp$. Cochran theorem $\Rightarrow$ • $\frac{\left\|\Pi_{V\cap W^\perp}(Y-\mu)\right\|^2}{\sigma^2} \sim \mathcal{X}^2(d-p)$.

• $\frac{\|Y - \Pi_V Y\|^2}{\sigma^2} = \frac{\left\|\Pi_{V^\perp} Y\right\|^2}{\sigma^2} = \frac{\left\|\Pi_{V^\perp}(Y-\mu)\right\|^2}{\sigma^2} \perp\!\!\!\perp \frac{\left\|\Pi_{V\cap W^\perp}(Y-\mu)\right\|^2}{\sigma^2}$ • $\frac{\|Y - \Pi_V Y\|^2}{\sigma^2} \sim \mathcal{X}^2(n-d)$.

General Linear Model

Likelihood Ratio Test: F-test

Vector subpaces ■ $V = \mathbb{X}(\mathbb{R}^d)$, $dim(V) = d$, ■ $\Theta_0 \subset \mathbb{R}^d$, $dim(\Theta_0) = p$,
and ■ $W = \mathbb{X}(\Theta_0) \subset V$, $dim(W) = p$.

The likelihood-ratio test of $H_0 : \mu \in W$ $(\theta \in \Theta_0)$ against $H_1 : \mu \notin W$ $(\theta \notin \Theta_0)$
with significance level $\alpha > 0$, rejects H_0 when $\frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha)$.

Proof – 5

LR $\rho = \left(1 + \frac{\|\Pi_V Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2}\right)^{\frac{n}{2}} = g\left(\frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2}\right)$ with g increasing fct. Under H_0,
$\frac{\|\Pi_V Y - \Pi_W Y\|^2}{\sigma^2} \sim \mathcal{X}^2(d-p)$, $\frac{\|Y - \Pi_V Y\|^2}{\sigma^2} \sim \mathcal{X}^2(n-d)$ and $\|Y - \Pi_V Y\|^2 \perp\!\!\!\perp \|\Pi_V Y - \Pi_W Y\|^2$.
Definition of Fisher distribution (slide 37) $\Longrightarrow S \overset{def}{=} \frac{(n-d)\|\Pi_V Y - \Pi_W Y\|^2}{(d-p)\|Y - \Pi_V Y\|^2} \sim \mathcal{F}(d-p, n-d)$.
For a test with significance α, we want t (or rather $g^{-1}(t)$) such that $\alpha = \mathbb{P}_{H_0}(\text{reject } H_0)$
$= \mathbb{P}_{H_0}(\rho > t) = \mathbb{P}_{H_0}(S > g^{-1}(t)) = 1 - \mathbb{P}_{H_0}(S < g^{-1}(t)) = 1 - F_{\mathcal{F}(d-p,n-d)}(g^{-1}(t))$.
We can conlude $g^{-1}(t) = F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha)$. $\square$

Summary: Linear Hypotheses Testing

General Linear Model (GLM)

Random variables $Y_i \in \mathbb{R}$ with **means** μ_i and **noise vector** $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.
Injective matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ with **image** $V \stackrel{def}{=} \mathrm{Im}(\mathbb{X})$, and **parameters** $\theta \in \mathbb{R}^d$.

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbb{X}\,\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Likelihood-ratio test or F-test

$\Theta_0 \subset \mathbb{R}^d$ p-dimensional vector subpace of $\mathbb{R}^d$, and $W = \mathbb{X}(\Theta_0) \subset V$.
The Likelihood-ratio test of $H_0 : \mu \in W$ against $H_1 : \mu \notin W$, $\Leftrightarrow$ $H_0 : \theta \in \Theta_0$ *vs* $H_1 : \theta \notin \Theta_0$, with significance level $\alpha \in]0,1[$, rejects H_0 when

$$S = \frac{(n-d)\left\|\Pi_V Y - \Pi_W Y\right\|^2}{(d-p)\left\|Y - \Pi_V Y\right\|^2} = \frac{(n-d)\left\|\Pi_V Y - \Pi_W Y\right\|^2}{(d-p)n\widehat{\sigma_n^2}} > F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha),$$

where $F^{-1}_{\mathcal{F}(k_1,k_2)}(1-\alpha)$ is the quantile of order $1-\alpha$ of the Fisher distribution $\mathcal{F}(k_1,k_2)$.

Particular cases

- Y function of quantitative variables: Linear regression
- Y function of categorical variables: ANalysis Of VAriance (ANOVA)

Suppose that, $\forall i \in \{1, \ldots, n\}$, the i^{th} row of $\mathbb{X}$ contains d values, $\left\{ x_i^{(1)}, \ldots, x_i^{(d)} \right\}$, that could explain linearly Y_i:

$$Y = \mathbb{X}\theta = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & \ldots & x_1^{(d)} \\ x_2^{(1)} & \ldots & x_2^{(d)} \\ \vdots & & \vdots \\ x_n^{(1)} & \ldots & x_n^{(d)} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- the MLE is the linear least square parameters $\widehat{\theta_n} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$.
- confidence interval: $\left\{ \theta_0 \in \mathbb{R}^d \,\middle|\, H_0 : \theta = \theta_0 \text{ accepted } \right\}$
$$= \left\{ \theta_0 \in \mathbb{R}^d \,\middle|\, \frac{(n-d)\|\Pi_V Y - \mathbb{X}\theta_0\|^2}{(d-p)\|Y - \Pi_V Y\|^2} \leqslant F^{-1}_{\mathcal{F}(d-p,n-d)}(1-\alpha) \right\}$$
- Test if values whose indices are in I are necessary for the regression: $H_0 : \theta \in \Theta_0$ with $\Theta_0 = \{\theta \mid \theta_i = 0 \; \forall i \in I\}$, i.e. the linear span of vectors $(\underset{(1)}{0}, \ldots, \underset{(i)}{1}, \ldots, \underset{(d)}{0})^T \; \forall i \notin I$.

Outline

ANalysis Of VAriance (ANOVA)

Model

$Y_{ij} = m_i + \varepsilon_{ij}, \qquad \forall i \leqslant d,\ \forall j \leqslant n_i$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d, $m_i \in \mathbb{R}$ and $\sum_{i=1}^{d} n_i = n$.

e.g. d conditions/treatments, for each condition i, n_i living beings under condition i (only!), Y_{ij} = measurement on the j^{th} living beings of the i^{th} condition group.

In the framework of GLM

$$Y = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ Y_{2,1} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{d,n_d} \end{pmatrix} = \begin{pmatrix} m_1 \\ \vdots \\ m_1 \\ m_2 \\ \vdots \\ m_2 \\ \vdots \\ m_d \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n_1} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,n_2} \\ \vdots \\ \varepsilon_{d,n_d} \end{pmatrix} = m_1 \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix} + m_2 \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + m_d \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n_1} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,n_2} \\ \vdots \\ \varepsilon_{d,n_d} \end{pmatrix}$$

In the framework of GLM

$$Y = \mu + \varepsilon = \mathbb{X}\theta = m_1 e_1 + m_2 e_2 + \ldots + m_d e_d + \varepsilon$$

with $\forall i \in \{1, \ldots, d\}$, $e_i \in \mathbb{R}^n$ such that $\forall j \in \{1, \ldots, n_i\}$, $(e_i)_j = \mathbb{1}_{\{\mu_j = m_i\}}$.

- $\mathbb{X} = \begin{pmatrix} e_1 & e_2 & \ldots & e_d \end{pmatrix}$ and $\theta = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{pmatrix}$
- $\mu \in V = span(e_1, e_2, \ldots, e_d) \subset \mathbb{R}^n$
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

In the framework of GLM

$$Y = \mu + \varepsilon = \mathbb{X}\theta + \varepsilon = \begin{pmatrix} e_1 & e_2 & \dots & e_d \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{pmatrix} + \varepsilon$$

$$V = span\,(\, e_1, e_2, \dots, e_d\,) = span \left(\begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right)$$

Note that vectors $\{\, e_i \,\}_{i=1}^{d}$ are orthogonal.

$V = span\,(e_1, e_2, \dots, e_d)$. We compute the MLEs using that vectors $\{e_i\}_{i=1}^{d}$ are orthogonal:

- $\widehat{\mu_n} = \Pi_V Y = \sum_{i=1}^{d} \frac{\langle Y, e_i\rangle}{\|e_i\|^2} = \sum_{i=1}^{d} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \right) e_i = \sum_{i=1}^{d} Y_{i.} e_i$ with $Y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$
- $\widehat{\theta_n}$: we have $\mathbb{X}\widehat{\theta_n} = \widehat{\mu_n}$ *i.e.* $\sum_{i=1}^{d} \left(\widehat{\theta_n}\right)_i e_i = \sum_{i=1}^{d} Y_{i.} e_i \Rightarrow \widehat{\theta_n} = \begin{pmatrix} Y_{1.} \\ Y_{2.} \\ \vdots \\ Y_{d.} \end{pmatrix}$
- $\widehat{\sigma_n^2} = \frac{1}{n} \|Y - \Pi_V Y\|^2 = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} (Y_{ij} - (\Pi_V Y)_i)^2 = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i.})^2$

Model

$Y_{ij} = m_i + \varepsilon_{ij}, \qquad \forall i \leqslant d, \forall j \leqslant n_i$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d, $m_i \in \mathbb{R}$ and $\sum_{i=1}^{d} n_i = n$.

e.g. d conditions/treatments, for each condition i, n_i living beings under condition i (only!), Y_{ij} = measurement on the j^{th} living beings of the i^{th} condition group.

Could we test if conditions have effets?

ANOVA:

Linear Hypothesis test:

$$H_0 : m_1 = m_2 = \ldots = m_d \qquad \text{against} \qquad H_1 : \exists i \neq j \text{ s.t. } m_i \neq m_j$$

Likelihood-ratio test

We have simply $H_0 : \mu \in W = span\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right) = span(\mathbb{1}_n)$ where $\mathbb{1}_n = \sum_{i=1}^{d} e_i = \mathbb{1}_n$.

ANalysis Of VAriance (ANOVA)

Model

$Y_{ij} = m_i + \varepsilon_{ij}, \qquad \forall i \leqslant d, \forall j \leqslant n_i$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d, $m_i \in \mathbb{R}$ and $\sum_{i=1}^{d} n_i = n$.

ANOVA: test condition effect

Linear Hypothesis test: $H_0 : m_1 = m_2 = \ldots = m_d$ against $H_1 : \exists i \neq j$ s.t. $m_i \neq m_j$

Likelihood-ratio test

- $W = span(\mathbb{1}_n)$ where $\mathbb{1}_n = \sum_{i=1}^{d} e_i$, $dim(W) = 1$,
- $V = span(e_1, \ldots, e_d)$, $dim(V) = d$,
- $\Pi_V Y = \sum_{i=1}^{d} Y_{i.} e_i$,
- $\Pi_W Y = \frac{\langle \mathbb{1}_n, Y \rangle}{\|\mathbb{1}_n\|^2} \mathbb{1}_n = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} Y_{ij} \mathbb{1}_n = Y_{..} \mathbb{1}_n$.

$\Rightarrow$ The test with significance level α rejects H_0 when

$$S = \frac{(n-d)\,\|\Pi_W Y - \Pi_V Y\|^2}{(d-1)\,\|Y - \Pi_V Y\|^2} = \frac{(n-d)\sum_{i=1}^{d} n_i\,(Y_{i.} - Y_{..})^2}{(d-1)\sum_{i=1}^{d}\sum_{j=1}^{n_j}(Y_{ij} - Y_{i.})^2} > F^{-1}_{\mathcal{F}(d-1,n-d)}(1-\alpha).$$

ANalysis Of VAriance (ANOVA)

ANOVA with scipy:

- scipy.stats: `f_oneway`

Back to the Student test

Result: If $X \sim t(d)$ then $X^2 \sim \mathcal{F}(1, d)$.

$\Rightarrow$ When $d = 2$, an ANOVA is a Student t-test.

- scipy.stats: `ttest_ind`

Extensions for repeated measures

If variables are not independent, because of repeated measures, with different conditions, on the same living being,

- $d = 2$: paired t-test.
  - scipy.stats: `ttest_rel`
- $d > 2$: repeated-measure ANOVA.

Other extensions

- two-ways ANOVA, when two kind of conditions are considered.
- MANOVA when $Y_i \in \mathbb{R}^m$, $\forall i \in \{1, \ldots, n\}$.

Note on the issue of multiple comparison

- In the ANOVA framework, to discriminate which conditions is statistically different from the others, we would need multiple Student t-tests.
- However, when performing multiple tests on the same dataset, the chance of observing a rare event increases, just like erroneously rejecting.
- To compensate for that increase, Bonferroni method consist in dividing the risk (α) by the number of tests.
- Other methods exist: Dunnett, Scheffé, Tukey, ...

Bonferroni correction

Let us suppose we are doing m tests, with hypotheses $\{H_0^i\}_{i=1}^m$, statistics $\{S_i\}_{i=1}^m$, risks $\{\alpha_i\}_{i=1}^m$ and region of rejection $\{\mathcal{R}_i\}_{i=1}^m$.
Suppose also that $\cap_{i=1}^m H_0^i = H_0 \neq \emptyset$. Then,

$$\mathbb{P}_{H_0}\left(\cup_{i=1}^m \{S_i \in \mathcal{R}_i\}\right) \leqslant \sum_{i=1}^m \mathbb{P}_{H_0}(S_i \in \mathcal{R}_i) \leqslant \sum_{i=1}^m \alpha_i.$$

So if we want a global risk α and equal local risks α_i, we should use $\alpha_i = \frac{\alpha}{m}$.

Other useful tests

Non-parametric tests

A test is said non-parametric when no assumption is made on the distribution of the variables.

- Kruskal-Wallis test, non-parametric version of ANOVA
  - scipy.stats: `kruskal`
- Friedman test, non-parametric version of repeated-measure ANOVA
  - scipy.stats: `friedmanchisquare`
- Wilcoxon-Mann-Whitney, non-parametric version of t-test
  - scipy.stats: `mannwhitneyu`
- Wilcoxon test, non-parametric version of the paired t-test
  - scipy.stats: `wilcoxon`

Other useful tests

Permutation tests

H_0 : exchangeability of the conditions.

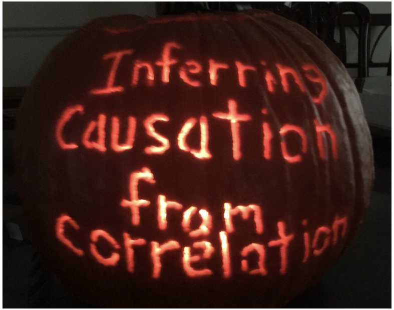
- compute the statistic s of the dataset from experimentation $y \in \mathbb{R}^n$.
- compute the smallest rejection region that rejects H_0 with statistic s, denoted by $\mathcal{R}_s$.
- generate $m \in \mathbb{N}$ datasets $\{ y_k^0 \}_{k=1}^m$ under H_0.
- compute the m statistics $\{ s_k^0 \}_{k=1}^m$.
- the empirical p-value is $\frac{\#\{ k \mid s_k^0 \in \mathcal{R}_s \}}{m}$.

- if $\mathcal{R}(t) = \{ s \in \mathbb{R} \mid s \leqslant t \}$, $\mathcal{R}_s = \mathcal{R}(s)$,
- if $\mathcal{R}(t) = \{ s \in \mathbb{R} \mid s \geqslant t \}$, $\mathcal{R}_s = \mathcal{R}(s)$,
- if $\mathcal{R}(t) = \{ s \in \mathbb{R} \mid |s| \geqslant t \}$, $\mathcal{R}_s = \mathcal{R}(|s|)$.

Other useful tests

Correlation tests

- Pearson test: the null hypothesis is a null correlation.
  - scipy.stats: `pearsonr`
- Spearman test: non-parametric version of the Pearson test.
  - scipy.stats: `spearmanr`



Institut Supérieur de l'Aéronautique et de l'Espace

10 avenue Édouard Belin – BP 54032

31055 Toulouse Cedex 4 – France

Phone: +33 5 61 33 80 80

www.isae-supaero.fr