LEPL1109 Statistics & Data science: part I.

Lecturer: Donatien Hainaut

UCL, Institute of Statistics, Biostatistics and Actuarial Sciences

Academic year 2024-2025

Roadmap

Lecture 1	Lecture 1.1 Back to probabilities Lecture 1.2 Independence & linear dependence Lecture 1.3 Normal random variable and central limit theorem	⇨	Self-learning 1: descriptive statistics and first data visualization tools
Lecture 2	Lecture 2.1 Estimation Lecture 2.2 Methods of moments Lecture 2.3 Likelihood maximization	⇨	Self-learning 2: Other fundamental random variables Self-learning 3: Simulations & Bootstrapping
Lecture 3	Lecture 3.1 Empirical mean and standard deviations: properties Lecture 3.2 Hypothesis testing, 1 population	⇨	Self-Learning 4: Properties of $\bar{X}$ and S^2, 2 populations Self learning 5: Hypothesis testing, 2 populations
Lecture 4	Lecture 4.1 Linear regression Lecture 4.2 Properties of regression Coefficients Lecture 4.3 ANOVA		

Roadmap

Semaine	Dates			Cours Magistral 2h/cours Lundi 2pm-4pm	Hackathons	TP
1	16/09/2024	au	20/09/2024	Stat 1		X
2	23/09/2024	au	27/09/2024	Stat 2		X
3	30/09/2024	au	04/10/2024	Stat 3		X
4	07/10/2024	au	11/10/2024	Stat 4	X (H1)	
5	14/10/2024	au	18/10/2024	Stat 5		X
6	21/10/2024	au	25/10/2024	Q&A H1		X
7	28/10/2024	au	01/11/2024	suspension des cours		
8	04/11/2024	au	08/11/2024	Data 1, & Data 2 vdd 8/11, à 16h15		X (TP1)
9	11/11/2024	au	15/11/2024	férié le 11/11		X (TP2)
10	18/11/2024	au	22/11/2024		X (H2)	
11	25/11/2024	au	29/11/2024	Data 3		X (TP3)
12	02/12/2024	au	06/12/2024	Data 4		X (drill questions exam)
13	09/12/2024	au	13/12/2024		X (H3)	
14	16/12/2024	au	20/12/2024	Data 5		

Roadmap

	Monday	Tuesday	Wednesday	Thursday	Friday
8:30 - 10:30		Serie 1	Serie 3	Serie 6	Serie 9
10:45 - 12:45			Serie 4	Serie 7	
14:00 - 16:00	Lecture		Serie 5	Serie 8	Serie 10
16:15 - 18:15		Serie 2	Serie 11		

TP's and Hackathons: REGISTER to a group on Moodle!

Hackathons

- Hackathons are small pedagogical group works in Python (Jupyter notebook to download).
- 3 activities of this type are planned during the semester : 1 in statistic and 2 in data.
- The sessions of exercises during these weeks are Q&A about hackathons
- Hackathons are done by groups of 5 to 6 students

Lecture 1.1 Back to probabilities

Random variable

This chapter reviews some concepts of probability, used later in statistics.

- If we consider an experiment, the set of possible results ω of this experiment is noted Ω
- We associate to each event A of the sample space Ω a measure of probability, $P(A)$, such that $P(A) \in [0, 1]$ and $P(\Omega) = 1$.

A **Random variable** is a measurable function from the sample space Ω to the set of real numbers $\mathbb{R}$, $X : \Omega \rightarrow \mathbb{R}$. We denote it by a capital letter and its realization by a lowercase letter $(X(\omega) = x)$.

Random variable

- The set of all possible values of a r.v. X is the state space (or the **range** of X) of the r.v. : $\{x \in \mathbb{R} \mid x = X(\omega)\,,\ \omega \in \Omega\}$.
- A r.v. is called **discrete** if its state space has a finite or countable number of elements.
- A r.v. is called **continuous** if it takes arbitrary real values between a minimum and a maximum. The state space is either an interval of $\mathbb{R}$, or $\mathbb{R}$.
- **Example**: In the context of a chemical experiment, let consider the r.v. X = time of reaction in ms (one thousandth of a second) after application of a stimulant. In the normal conditions, this time varies between 263 ms and 613 ms. So $P(263 \leq X \leq 613) = 1$ and $\mathrm{Range}(X) = [263;\ 613]$.

Probability distribution

The probability mass function (pmf) $p(x)$ of a discrete random variable X is a function that associates to all values x of X a probability $P(X = x)$:

$$p(x) \quad = \quad P(X = x)\,.$$

If the range of X is $R = \{x_1, x_2, ..\}$ then $p(x) > 0$ if $x \in R$ and $p(x) = 0$ if $x \notin R$. Furthermore $\sum_i p(x_i) = 1$.

Once the pmf is known, we calculate the probability that $P(X \in I)$ where I is a discrete subset of $\mathbb{R}$ as

$$P(X \in I) \quad = \quad \sum_{i\,:\,x_i \in I} p(x_i)$$

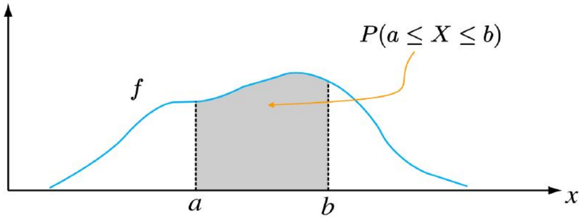
Probability density function (pdf)

If the random variable is **continuous**, we cannot enumerate mass probabilities for all possible values of x.

Definition:

Let X be a continuous random variable. The **probability density function (pdf)** of X is the function $f(x) \geq 0$ such that for any $I \subset \mathbb{R}$, we have that

$$P(X \in I) \quad = \quad \int_I f(x)dx$$



Cumulative distribution function (cdf)

Definition:

The cumulative distribution function (cdf), $F(x)$, of a random variable (discrete or continuous) X indicates for each possible value of x the probability that X takes the value equal or less than x.

$$F(x) = P(X \leq x)$$

In the discrete case, if the state space (range) of $X = \{x_1, x_2, ...\}$ then

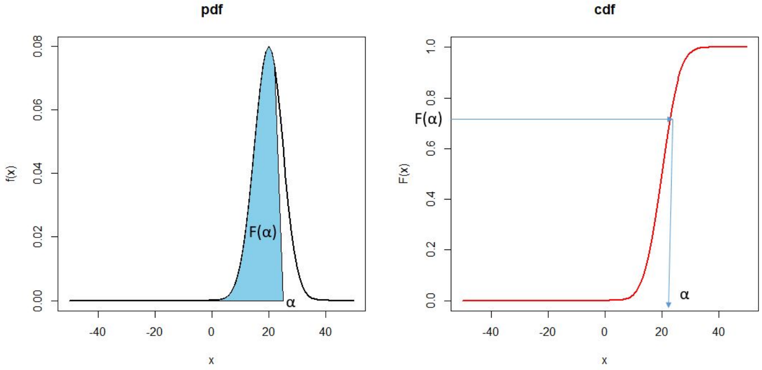
$$F(x) = \sum_{i : x_i \leq x} p(x_i)$$

In the continuous case,

$$F(x) = \int_{-\infty}^{x} f(u) du$$

Link between the pdf & cdf

The pdf is the derivative of the cdf: $f(x) = \frac{d}{dx}F(x)$.



Furthermore, given that the integral of the pdf $f(x)$ is a probability, we must have $\int_{-\infty}^{\infty} f(x)dx = P(X \in range(X)) = 1$.

Mathematical expectation

The expectation of a r.v. X is noted $\mathbb{E}(X) = \mu_X$. If X is a discrete r.v.

$$\mu_X = \sum_{x \in range(X)} x\, p(x)\,.$$

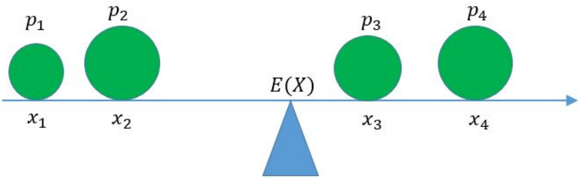
If X is a continuous r.v. with a density $f(x)$, the expectation of X is equal to

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

We will see later that if we observe n outcomes of a random variable x_i, the empirical mean $\bar{x} = \frac{\sum_{i=1:n} x_i}{n}$ converges to the expected value μ_X(see "Law of Large Numbers") Numpy: `np.mean(X)`

Mathematical expectation

- If X is a discrete r.v., its expectation μ_X is the center of gravity of points



- **Property:** Let X and Y be two random variables then

$$
\begin{aligned}
\mathbb{E}(X+Y) &= \mathbb{E}(X)+\mathbb{E}(Y) \\
\mathbb{E}(aX+b) &= a\mathbb{E}(X)+b
\end{aligned}
$$

Mathematical expectation

The expected value of a real-valued function $h(.)$ of a discrete r.v. X is equal to

$$\mathbb{E}\left(h(X)\right) = \sum_{x \in range(X)} h(x)\, p(x)\,.$$

If X is a continuous r.v. with a density $f(x)$, the expectation of $h(X)$ is equal to

$$\mathbb{E}\left(h(X)\right) = \int_{-\infty}^{+\infty} h(x)\, f(x) dx$$

Important remark: in general we do not have (check it on a example)

$$\mathbb{E}\left(h(X)\right) \neq h\left(\mathbb{E}(x)\right)$$
$$\mathbb{E}\left(XY\right) \neq \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$$

Notice that the k^{th} moment of X is defined by $\mathbb{E}\left(X^k\right)$.

Variance

The variance of a random variable X is denoted by

$$\mathbb{V}ar(X) = \mathbb{V}(X) = \sigma_X^2$$

(discrete or continuous) and is defined as

$$\sigma_X^2 \quad = \quad \mathbb{E}\left((X - \mu_X)^2\right)$$

The **standard deviation** is the square root of the variance $\sigma_X = \sqrt{\mathbb{V}(X)}$.

The standard deviation is a measure of dispersion of the probability function around its balancing point (the mean). When the variance is close to zero, outcomes of X are concentrated around the mean. On the contrary, if the variance is big, realizations of X may be very far from each other.

Variance

Properties of the variance. If X and Y are two r.v. (discrete or continuous) and $a, b \in \mathbb{R}$, then

- $\mathbb{V}ar(X) = \mathbb{E}\left(X^2\right) - \left(\mathbb{E}\left(X\right)\right)^2$
- $\mathbb{V}ar(a) = 0$
- $\mathbb{V}ar(a + bX) = b^2\, \mathbb{V}\left(X\right)$
- In general, $\mathbb{V}ar(X + Y) \neq \mathbb{V}\left(X\right) + \mathbb{V}(Y)$ except if X and Y are independent

- The proofs are let to the reader and do not present any difficulty.

- If we observe n outcomes of a random variable x_i, the empirical variance $s^2 = \frac{\sum_{i=1:n}(x_i - \bar{x})^2}{n-1}$ is used as estimator of σ_X^2 (details provided later) Numpy: np.std(X)

Law of Large Numbers (LLN)

Law of Large Numbers: Let $X_{i=1,\ldots,n}$ be a sequence of uncorrelated r.v.'s with the same expectation $\mu_X = \mathbb{E}(X_i)$ and variance σ_X^2. When $n \to \infty$, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to μ_X:

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} P(|\bar{X}_n - \mu_X| \geq \epsilon) \quad = \quad 0\,.$$

We denote this by $\bar{X}_n \xrightarrow{p} \mu_X$.

Formal proof: Expectations/Variances are additive, and $\mathbb{V}(aY) = a^2\mathbb{V}(Y)$. Then $\mathbb{E}\left(\bar{X}_n\right) = \mu_X$ and

$$\mathbb{V}\left(\bar{X}_n\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^n X_i\right) = \frac{\sigma_X^2}{n}$$

The variance tends to zero when $n \to \infty$.

Quantiles of a distribution

Quantile of a distribution Let p be a probability between 0 and 1 and X be a r.v;. The number q_p satisfying the relation

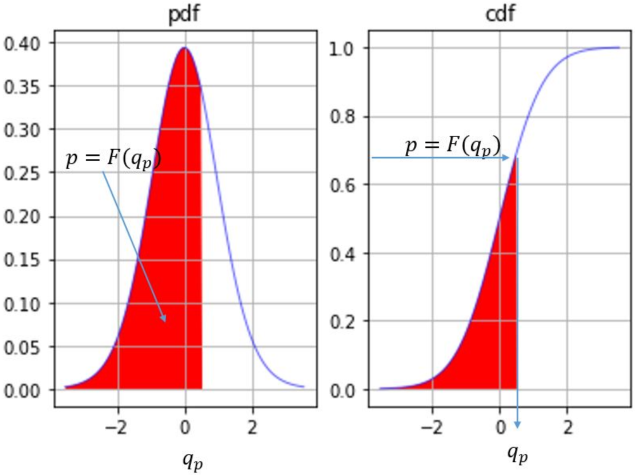
$$P(X \leq q_p) = p$$

is the quantile of order p for X. If X is continuous and if its cdf $F(x)$ is invertible then $q_p = F^{-1}(p)$.

If $p = 5\%$ then X is below $q_{5\%}$ with a probability of 5% and X is above $q_{5\%}$ with a probability of 95%. The quantile is used in the banking/insurance industry as measure of risk (Value At Risk).

Quantiles of a distribution

Python scipy.stats: sc.norm.ppf(0.01,loc=μ,scale=σ)



Quantiles of a distribution

Example

- Let the r.v. X be the operating life (in days) until failure of a certain device.
- Assume that the density of X is

$$f(x) = \frac{1}{50} e^{-\frac{x}{50}} I_{[0,\infty)}(x)$$

- To calculate the quantile of order p of X, we must solve the following equation for q

$$p = \frac{1}{50} \int_0^q e^{-\frac{x}{50}} dx$$

- The solution is given by

$$q_p = -50 \ln(1-p)$$

- Thus, in 5% of case the device operating time will be smaller than 2.56 days!

Function of random variables

Linear transformation. Let X be a continuous r.v. with density $f_X(x)$. The density (pdf) of $Y = a + bX$ is given by

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right)$$

Function of random variables

Proof. Assume that $b < 0$ By definition of Y,

$$\begin{aligned} P(Y \leq y) &= P(a + bX \leq y) \\ &= P(X \geq \frac{y-a}{b}) \quad \text{because b<0} \\ &= \int_{\frac{y-a}{b}}^{+\infty} f_X(x)\, dx \end{aligned}$$

Let $z = a + bx$ then $x = \frac{z-a}{b}$ and $dx = \frac{1}{b} dz$. The upper bound becomes $z_{up} = a + b(\infty) = -\infty$ and $z_{low} = y$

$$\begin{aligned} P(Y \leq y) &= \int_{y}^{-\infty} \frac{1}{b} f_X(\frac{z-a}{b})\, dz \\ &= -\int_{-\infty}^{y} \frac{1}{b} f_X(\frac{z-a}{b})\, dz \end{aligned}$$

and we conclude as $b < 0$, **end.** **In python, scipy.stats: loc = a and scale = b**

Self-Learning 1: Descriptive statistics and first data visualization tools

Introduction

Some vocabulary:

- **Population**: collection of all possible outcomes, responses, measurements, or counts that are of interest.
- **Sample**: a subset of population.
- **Descriptive statistics**: branch of statistics that involves the organization, summarization and display of data.

Download files “DescriptiveStatistics.py” and “DurationData.csv”. Run the Python script and make sure that the path is well the one of the directory containing the csv file.

Introduction

Example: supply chain

- We consider an automotive manufactoring chain with a succession of workstations:

Some parts of motors are bolted together by a team of workers. We measure the time spent by team of 3 and 5 workers on 1000 motors (sample size $n = 1000$).

Introduction

- We have the following table of durations in minutes:

	Duration	
Motor #	3 Workers, X	5 Workers, Y
1	24.23	15.23
$\vdots$	$\vdots$	$\vdots$
1000	18.5	8.56

- The time spent by teams of 3 and 5 workers are random quantities respectively denoted by X and Y.

- Each observation in the sample is a realization of this random variable. The i^{th} realizations are denoted by x_i and y_i for $i = 1, ..., 1000$. These records are stored in the file "DurationData.csv".

Mean, Median, Quantiles

Which descriptive statistics could we use to analyze this dataset? We naturally think to the mean and the median in order to give a central value for the data set.

- The **SAMPLE MEAN** is the average value for a set of n observations. This is the center of gravity of data.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The **SAMPLE MEDIAN** is the value $q_{0.5}$ such that 50% of realizations are below $q_{0.5}$ and 50% are above $q_{0.5}$. The sample median is also called a quantile of order 0.5.

Mean, Median, Quantiles

The concept of median can be extended by the notion of sample quantile. For a given $p \in (0, 1)$, the p^{th} sample quantile, noted q_p, is a value such that a proportion p of the sample is smaller that q_p and a proportion $1 - p$ of observations is larger than q_p.

- **SAMPLE QUANTILE** : Given a set of realizations $x_1, \ldots, x_n$ we define quantiles as follows:
  - We sort the value by ascending order:

    $$x_{(1)} \leq \ldots \leq x_{(n)},$$

    these values are called the order statistics of the original sample.
  - q_p is the order statistics

    $$q_p = x_{(1+(n-1)p)}$$

Mean, Median, Quantiles

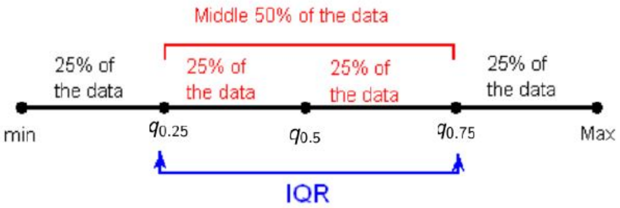
Example: supply chain ,

```
In [22]: data = pd.read_csv("durationData.csv",sep=';' )
    ...: stat = data.describe()
    ...: print(stat)
    ...:
          workers3     workers5
count  1000.000000  1000.000000
mean     19.897557     9.893154
std       5.080540     6.336737
min       2.371155     0.384120
25%      16.512532     5.388633
50%      19.799504     8.560589
75%      23.121497    13.201465
max      37.203430    43.190108
```

We import the data stored in a "csv" file in a dataframe (library "pandas", read_csv()) and we use the command describe(). Other command, numpy : np.mean(), np.std(), np.corrcoef(), np.quantile(x,q)

RANGE, IQR, OUTLIERS

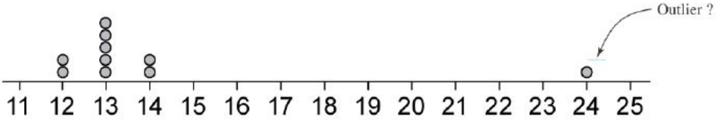
- The **RANGE** is the difference between the maximum and the minimum: $x_{max} - x_{min}$
- The **INTER-QUARTILE RANGE** (IQR) is the difference between symmetric quartiles, e.g. $IQR = q_{0.75} - q_{0.25}$



Python command: library "scipy.stat": iqr()

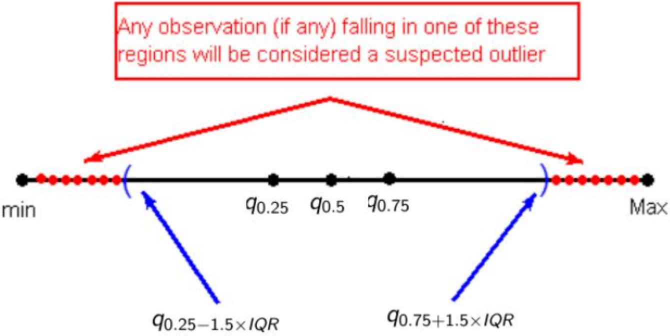
RANGE, IQR, OUTLIERS

- **OUTLIERS :** observations that appear to be far away from the rest of the data set. OUTLIERS may be due to errors of measurement or of encoding.
- In this case, it should be corrected or removed from the sample.



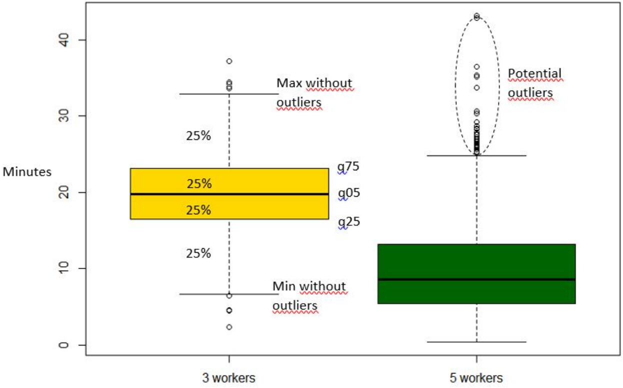
RANGE, IQR, OUTLIERS

An observation is considered as a suspected outlier if the value is below $q_{0.25} - 1.5 \times IQR$ or above $q_{0.75} + 1.5 \times IQR$



Graphical analysis

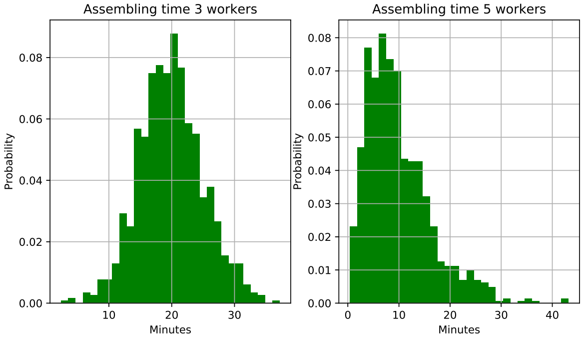
- The **BOXPLOT** graphically represents the distribution of observations by reporting the following statistics: min, $q_{0.25}$, $q_{0.50}$, $q_{0.75}$ and max. It also reports suspected outliers (IQR criterion).



Python command, library "matplotlib.pyplot": boxplot(.) ; title(.)

Graphical analysis

- The **HISTOGRAM** : This tool aims to vizualize the distribution of numerical observations. An histogram breaks the range of values into intervals and counts the proportion of observations in each bin. This is the empirical pdf or pmf.



Python command, library "matplotlib.pyplot": hist() , subplot()

Variance and standard deviation

- The mean provides a central value for the data set. How do we measure the dispersion/spread around this average?
- Solution, the **sample VARIANCE**. We sum up the quadratic spread between observations and the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- the **sample STANDARD DEVIATION** is the square root of the variance: $s = \sqrt{s^2}$ (Numpy command: std(,ddof=1) , mean(), median())

	Duration	
	3 Workers, X	5 Workers, Y
mean	19.89	9.89
median	19.79	8.56
Std. Dev.	5.08	6.34

Lecture 1.2 Independence & linear dependence

Independent random variables

Two r.v. X and Y are independent ($X \perp\!\!\!\perp Y$) if for every A and $B \in \mathbb{R}$, we have

$$P(X \in A\,,\ Y \in B) = P(X \in A)P(Y \in B)$$

Consequence of independence:

- $p(x, y) = p_X(x)p_Y(y)$ when X and Y are discrete
- $f(x, y) = f_X(x)f_Y(y)$ when X and Y are continuous

Example of independent r.v.:

- X: variation of the BEL 20 stocks index,
- Y: temperatures in Louvain-La-Neuve

Covariance

When two random variables are not independent, how can we measure the degree of dependence between them? The linear dependence is measured by the covariance.

Let X and Y be random variables. The **covariance** between X and Y is

$$\begin{aligned}\sigma_{XY} &= \mathbb{C}ov(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) \\ &= \mu_{XY} - \mu_X\,\mu_Y\end{aligned}$$

The covariance can be thought of as the mean of matches and mismatches among the pair (X, Y). The covariance is positive when the matches outweigh the mismatches and is negative when the mismatches outweigh matches.

Covariance

- If we observe n outcomes $(x_k)_{k=1,\ldots,n}$ and $(y_k)_{k=1,\ldots,n}$ of r.v.'s X and Y. The empirical covariance is estimated as follows

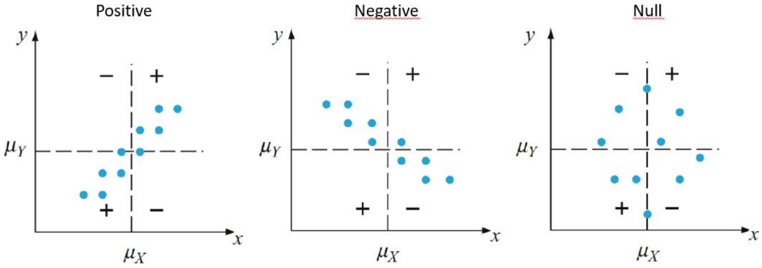
$$\sigma_{XY} \approx s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Covariance command in Python : numpy, np.cov(m , rowvar=True)

Covariance

Three types of dependence:

- $\mathbb{C}(X, Y) > 0$: X and Y move in the same direction
- $\mathbb{C}(X, Y) < 0$: X and Y move in opposite directions



Covariance

From the definition of covariance, we can deduce the following properties

- $\mathbb{C}(X, Y) = \mathbb{C}(Y, X)$
- $\mathbb{C}(X, X) = \mathbb{V}(X)$ and $\mathbb{C}(a, X) = 0$ for all $a \in \mathbb{R}$
- if $X \perp\!\!\!\perp Y$ then $\mathbb{C}(X, Y) = 0$ (X and Y are uncorrelated)
- $\mathbb{C}(aX + bY, Z) = a\mathbb{C}(X, Z) + b\mathbb{C}(Y, Z)$
- $\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathbb{C}(X, Y)$

Attention if $\mathbb{C}(X, Y) = 0 \nRightarrow X \perp\!\!\!\perp Y$!!! Example, $Y = X^2$ with $\mathbb{C}(X, Y) = 0$:

		X		
		-1	0	1
Y	0	0	1/3	0
	1	1/3	0	1/3

Correlation

What type of dependence does the correlation measure? The covariance is a measure of **LINEAR** dependence between two r.v.

To explain this, we define a scaled measure of dependence. Reason? The size of the covariance depends upon the scale of variables $\mathbb{C}(aX, Y) = a\mathbb{C}(X, Y)$.

Let X and Y be random variables. The **correlation** between X and Y is defined as

$$\rho_{XY} = \frac{\mathbb{C}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Correlation

According the Cauchy Schwarz inequality, we have that $|\sigma_{XY}| \leq \sigma_X \sigma_Y$ with equality if and only if $Y = aX + b$. Then the correlation has the following properties:

- ρ_{XY} is scale invariant
- $-1 \leq \rho_{XY} \leq 1$
- If $\rho_{XY} = 1$ then $Y = a + bX$ with $b \in \mathbb{R}^+$ (Y proportional to X)
- If $\rho_{XY} = -1$ then $Y = a + bX$ with $b \in \mathbb{R}^-$(Y inversely proportional to X)

The correlation is a measure of the linear dependence between two r.v. easier to interpret than σ_{XY} because it is "unit-free" and in $[-1, 1]$.

Correlation

- If we observe n outcomes $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ of r.v. X and Y. The correlation is estimated by the empirical correlation:

$$\rho_{XY} \approx r_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Python command: numpy, np.corrcoef(x,y).

Lecture 1.3 Normal random variable and Central Limit Theorem

Normal distribution

A r.v. X follows a **Normal distribution** if its density is given by the following function

$$f(x) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ are parameters. We note $X \sim \mathcal{N}(\mu, \sigma^2)$.

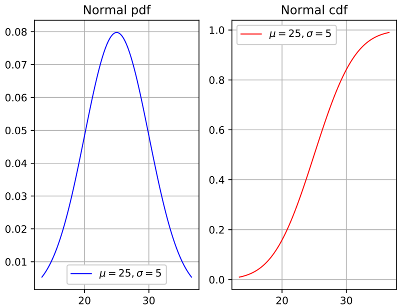
This is used for modeling e.g. the noise in signal processing. The parameters are easy to interpret:

$$\mathbb{E}(X) = \mu \quad , \quad \mathbb{V}(X) = \sigma^2$$

$$\underbrace{m_X(t)}_{(*)} := \mathbb{E}\left(e^{tX}\right) \quad = \quad \exp\left(\mu t + \frac{1}{2}t^2\sigma^2\right)$$

(*) moment generating function $\left.\frac{\partial^n m_X(t)}{\partial t^n}\right|_{t=0} = \mathbb{E}(X^n)$

Normal distribution



See file **statisticalDistributionPlot.py**.
Python: scipy.stats: command norm.pdf(x,loc=μ,scale=σ), norm.cdf(x,loc=μ,scale=σ) , and quantiles with norm.ppf(x,loc=μ,scale=σ)

Standardization

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ then for any a and $b \neq 0$ then $a + bX$ is a normal r.v.

$$\mathcal{N}\left(a + b\mu\,,\; b^2\sigma^2\right)$$

If $X \sim \mathcal{N}\left(\mu_x, \sigma_x^2\right)$ and $Y \sim \mathcal{N}\left(\mu_y, \sigma_y^2\right)$ and covariance is $\sigma_{x,y} = \mathbb{C}\left(X, Y\right)$

$$aX + bY \sim \mathcal{N}\left(a\mu_x + b\mu_y\,,\; a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\sigma_{x,y}\right)$$

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ then $\frac{X-\mu}{\sigma} \sim Z = \mathcal{N}\left(0, 1\right)$ (**standardization**) and

$$P\left(X \leq x\right) \quad = \quad P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

Central Limit Theorem (CLT)

Central Limit Theorem. Let $X_1, ..., X_n$ be a sequence of iid (*) r.v.'s with $\mathbb{E}(X_i) = \mu$ and $\mathbb{V}(X_i) = \sigma^2$. As $n \to \infty$

$$Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \quad \overset{d}{\to} \quad Z \sim \mathcal{N}(0, 1)$$

i.e.

$$P(Z_n \leq z) \overset{n\to\infty}{\longrightarrow} P(Z \leq z) \, \forall z$$

Interpretation: for large n, whatever the distribution of X_i, the distribution of the sample mean $\bar{X}_n$ and of the sum $S_n = \sum_{i=1}^n X_i$ may be approached by a Normal distribution

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$S_n \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$$

(*) iid: independent identically distributed

Central Limit Theorem (CLT)

Example 2: The CLT may be used to approach the binomial distribution $Bi(n, p)$. Let $Y_n \sim Bi(n, p)$. Y_n is a sum of n independent Bernoulli r.v. X_i ($X_i = 1$ with probability p and zero otherwise):

$$Y_n \quad = \quad X_1 + \ldots + X_n \, .$$

We have that $\mathbb{E}\,(X_i) = p$ and $\mathbb{V}(X_i) = p\,(1-p)$. Let $\hat{p}_n = \frac{Y_n}{n}$. By the CLT we have that

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \quad = \quad \frac{Y_n - np}{\sqrt{np\,(1-p)}} \xrightarrow{d} \mathcal{N}\,(0,1)$$

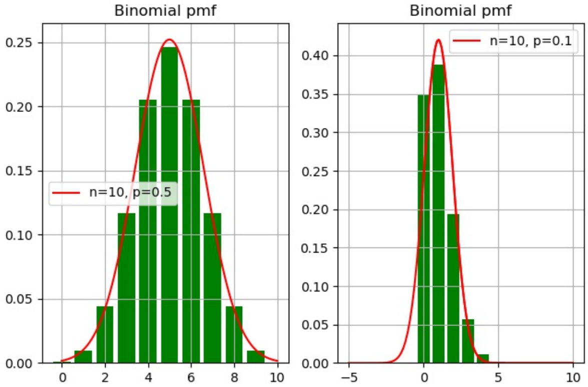
Equivalent formulation:

$$Bi(n,p) \quad \approx \quad \mathcal{N}\,(np\,,\ np(1-p))$$

This approximation will be rather good if p (and $1-p$) is not too small and n sufficiently large : $n > 9\frac{max(p,1-p)}{min(p,1-p)}$.

Central Limit Theorem (CLT)

Approximation of a Binomial law $Bi(n, p)$ by a $\mathcal{N}(np\,,\ n(1-p))$ (see Python file TCLillustration.py) :



Self learning 2 Other fundamental random variables

(see also Appendix 1)

Chi-square distribution

- The **chi-square** (χ^2) distribution with n degrees of freedom is the distribution of a sum of the squares of n independent standard normal $\mathcal{N}(0,1)$ r.v.

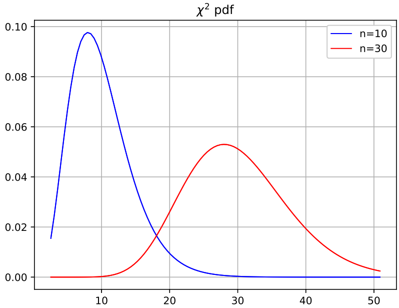
A r.v. X defined on $\mathbb{R}^+$ follows a χ^2-**distribution** of parameters n, when its density (pdf) is given by

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

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ (Gamma function).

- Then $\mathbb{E}(X) = n$ and $\mathbb{V}(X) = 2n$.
- The relation between the χ^2 and normal r.v.'s will play an important role in the second part of the course (statistical inference: hypothesis testing).

Chi-square distribution



See file **statisticalDistributionPlot.py**.
Python: scipy.stats: command chi2.pdf(x,df=n), chi2.cdf(x,df=n) , and quantiles with chi2.ppf(x,df=n)

Student's T

We will conclude this section by introducing two distributions playing an important role in statistical inference.

T distribution. Let Z and Y be two independent r.v. $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_n$ then

$$T_n = \frac{Z}{\sqrt{Y/n}}$$

is the **Student'T r.v.** with n degrees of freedom. Its pdf is

$$f_{T_n}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad t \in \mathbb{R}$$

Fisher-Snedecor

F distribution. Let X and Y be two independent χ^2 r.v. with n_1 and n_2 degrees of freedom. Then

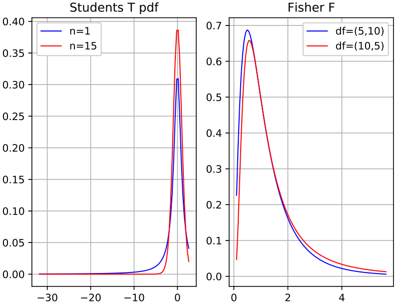
$$F_{n_1,n_2} = \frac{X/n_1}{Y/n_2}$$

is distributed as **a Fisher-Snedecor distribution r.v.** $\mathcal{F}(n_1, n_2)$ with n_1 and n_2 degrees of freedom. Its pdf is

$$f_{F_{n_1,n_2}}(z) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} z^{\left(\frac{n_1}{2}-1\right)}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}} \; t \in \mathbb{R}$$

Properties: if $F_{n_1,n_2} \sim \mathcal{F}(n_1, n_2)$ then $1/F_{n_1,n_2} \sim \mathcal{F}(n_2, n_1)$.

Student's T & Fischer S.



See file **statisticalDistributionPlot.py**.
Python: scipy.stats: command
t.pdf(x,df=n),
t.cdf(x,df=n) ,
f.pdf(x,dfn=n_1, dfd=n_2) ,
f.cdf(x,dfn=n_1, dfd=n_2)

Lecture 2.1 Estimation

Introduction

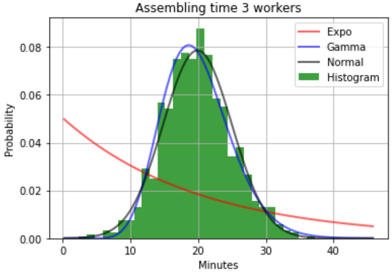
Suppose that a random variable X of interest to an experimenter has a probability distribution with density $f(x \mid \theta)$ where $\theta \in \Theta$ is a vector of parameters.

To learn about X, the experimenter proceeds by obtaining repeated data (sample values) of the random variable X, written $x_1, x_2, \ldots, x_n$. These are the observed values (realizations) of a set of n random variables X_1, X_2,..., X_n.

The collection of random variables X_1, X_2,..., X_n is called a random sample (of size n) if **(i)** they have the same probability distribution $f(x \mid \theta)$ and **(ii)** are mutually independent

Introduction

Example 1: 1000 assembling times of a mechanical devices with 3 workers. Each duration is a realization of a random variable. Each measure x_i is the result of a trial X_i, independent from others.



Which candidate distribution for modeling a duration? Compare histogram (emp. pdf) with pdf with the same range: here continuous r.v. : Exponential or Gamma random variables (defined on $\mathbb{R}^+$) or eventually a normal (approximation since the range is $\mathbb{R}$).

Estimator

If we assume that $X_i \sim N(\mu, \sigma)$ then $\theta = (\mu, \sigma) \in \Theta = \mathbb{R}^2_+$, **how do we estimate θ?**

An estimator of θ, generically denoted by $\widehat{\theta}$ is any function $h(.)$ of the random sample:

$$\widehat{\theta} = h(X_1, ..., X_n) \in \Theta$$

used to estimate θ. As θ is unknown, $\widehat{\theta}$ gives an approximation.

An estimate of θ, is an observed value of this estimator calculated from the observed sample, $x_1, ..., x_n$:

$$\widehat{\theta}_{obs} = h(x_1, ..., x_n) \in \Theta$$

Estimator

The estimator is a function of n random variables and **therefore is also a random variable**!

$\widehat{\theta}$ is and unbiased estimator of θ if

$$\mathbb{E}\left(\widehat{\theta}\right) = \mathbb{E}\left(h(X_1, ..., X_n)\right) = \theta.$$

The bias is the difference between the expectation and the real unknown value:

$$B(\widehat{\theta}) = \mathbb{E}\left(\widehat{\theta}\right) - \theta\,.$$

The Mean square error (MSE) measures the average error:

$$MSE\left(\widehat{\theta}\right) = \mathbb{E}\left(\left(\widehat{\theta} - \theta\right)^2\right)$$

Estimator

BIAS-VARIANCE decomposition of the MSE:

$$MSE\left(\widehat{\theta}\right) = B(\widehat{\theta})^2 + \mathbb{V}\left(\widehat{\theta}\right) \ .$$

Proof:

$$\begin{aligned} MSE\left(\widehat{\theta}\right) &= \mathbb{E}\left(\left(\widehat{\theta}-\theta\right)^2\right) = \mathbb{E}\left(\widehat{\theta^2}\right) - 2\theta\mathbb{E}\left(\widehat{\theta}\right) + \theta^2 \\ &= \underbrace{\mathbb{E}\left(\widehat{\theta^2}\right) - \mathbb{E}\left(\widehat{\theta}\right)^2}_{\mathbb{V}(\widehat{\theta})} + \underbrace{\mathbb{E}\left(\widehat{\theta}\right)^2 - 2\theta\mathbb{E}\left(\widehat{\theta}\right) + \theta^2}_{B(\widehat{\theta})^2} \end{aligned}$$

Conclusion: the best estimator should have the lowest bias and variance!

Lecture 2.2 Method of Moments

Moments matching

We observe $x_1, ..., x_n$, realizations of $X_{i=1:n}$. We think that $X_{i=1:n}$ have the same pdf $f(x|\theta)$ as X. In order to estimate $\theta \in \mathbb{R}^d$, we match the d moments

$$\mu_k(\theta) = \mathbb{E}\left((X)^k\right) \quad k = 1, ..., d$$

with the d empirical moments (r.v. noted M_k, realizations: m_k)

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad k = 1, ..., d$$

The estimator of moments $\widehat{\theta}$ is solution of a system with d equations:

$$\mu_k(\widehat{\theta}) = M_k \quad k = 1, ..., d.$$

$$\mu_k(\widehat{\theta}_{obs}) = m_k \quad k = 1, ..., d.$$

Moments matching

Example 1.1: we observe $x_1, \dots, x_n$ realizations of $X \sim expo(\beta)$, i.e. pdf $f(x \mid \beta) = \frac{1}{\beta} e^{-\frac{1}{\beta}x}$. According to Appendix 1, $\mathbb{E}(X) = \beta$ and $M_1 = \bar{X} = \frac{1}{n} \sum_{i=1:n} X_i$ then

$$\widehat{\beta} = \bar{X} \quad , \quad \widehat{\beta}_{obs} = \bar{x} \, .$$

Example 1.2: we observe $x_1, \dots, x_n$ realizations of $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \sigma^2$. If there are only 2 parameters, we match the empirical mean and variance with theoretical mean and variance.

$$\widehat{\mu} = \bar{X} \quad , \quad \widehat{\mu}_{obs} = \bar{x} \, .$$

$$\widehat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1} \quad , \quad \widehat{\sigma}^2_{obs} = s^2 \, .$$

Moments matching

Example 1.3: we observe $x_1, ..., x_n$ realizations of $X \sim Gamma\,(\alpha, \beta)$, $\mathbb{E}\,(X) = \alpha\beta$ and $\mathbb{V}\,(X) = \alpha\beta^2$.

$$\left\{ \begin{array}{c} \widehat{\alpha}\widehat{\beta} = \bar{X} \\ \widehat{\alpha}\widehat{\beta}^2 = S^2 \end{array} \right. \quad \Longleftrightarrow \quad \left\{ \begin{array}{c} \widehat{\alpha} = \frac{\bar{X}^2}{S^2} \\ \widehat{\beta} = \frac{S^2}{\bar{X}} \end{array} \right.$$

For a Python illustration, see the file "MomentMatching.py"

Example 2: we perform n experiments, success $X_i = 1$ with proba p, otherwise $X_i = 0$. Realizations of $X \sim Be\,(p)$, $\mathbb{E}\,(X) = p$ then:

$$\widehat{p} = \bar{X} \quad , \quad \widehat{p}_{obs} = \bar{x}\,.$$

Moments matching

Question: are these estimates reliable? We see that they are based on $\bar{X}$ and S^2. First answer:

$$\boxed{\bar{X} \text{ is an unbiased estimator of } \mathbb{E}(X).}$$

Proof: evident since $\mathbb{E}\left(\bar{X}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X) = \mathbb{E}(X)$. Therefore in examples 1.1 and 1.2, $\widehat{\beta}$ and $\widehat{\mu}$ are unbiaised estimators.

Second answer: based on the CLT stating that whatever the distribution of $X_{i=1:n} \sim X$,

$$\bar{X} \sim \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

then the higher is n, the lower is the variance of $\bar{X}$ around $\mathbb{E}(X)$ (but $\mathbb{E}(X)$ and $\mathbb{V}(X)$ are unknown in practice)

Moments matching

We cannot say a lot about the properties of the moment estimators. In general they are "consistent", i.e. the value of the estimator $\widehat{\theta} = h(X_1, ..., X_n)$ converges to the true value when $n \to \infty$.

We do not know much about their variance, except that their variance is larger than or equal to the variance of the maximum likelihood estimators (introduced later).

To summarize, the moment estimators are easy to construct but do not always possess the best statistical properties.

A much better approach than moment matching to find an estimator is the likelihood maximization.

Lecture 2.3 Likelihood Maximization

Likelihood Maximization

Let us consider a random sample $X_{1:n} \sim X$. We think that X has a pdf $f(x|\theta)$ where θ is the vector of unknown parameters. Let us recall that

$$P\left(x \leq X \leq x + dx\right) \quad \approx \quad f(x|\theta)\, dx\,.$$

Since the $X_{i:n}$ are independent, the probability to observe realizations $x_{1:n}$ of $X_{1:n}$ is:

$$\begin{aligned}
&P\left(x_1 \leq X_1 \leq x_1 + dx, \ldots, x_n \leq X_n \leq x_n + dx\right) \\
&\quad = P\left(x_1 \leq X_1 \leq x_1 + dx\right) \ldots P\left(x_n \leq X_n \leq x_n + dx\right) \\
&\quad = \prod_{k=1}^{n} f(x_k|\theta)\, dx\,.
\end{aligned}$$

Likelihood Maximization

The probability that the observed sample has been generated by the model is then proportional ($\propto$) to the likelihood function:

$$L(x_1, \ldots, x_n|\theta) := \prod_{k=1}^{n} f(x_k|\theta)$$

The maximum likelihood estimator (**MLE**) of θ is the value which maximises the likelihood of the observed sample

$$\widehat{\theta} = \arg\max_{\theta} L(x_1, \ldots, x_n|\theta)$$

In practice $\widehat{\theta}$ is found by deriving w.r.t. θ the log-likelihood function $l(.) = \ln L\,(.)$ i.e.

$$l(x_1, \ldots, x_n|\theta) = \sum_{k=1}^{n} \ln\left(f(x_k|\theta)\right) \; .$$

Likelihood Maximization

Example 1: If $X_{i:n}$ are *expo*(β) then $f(x|\beta) = \frac{1}{\beta} e^{-\frac{1}{\beta}x}$ and

$$
\begin{aligned}
L(\beta) &= \beta^{-n} e^{-\frac{1}{\beta}(x_1+\ldots+x_n)} \\
l(\beta) &= -n \ln(\beta) - \frac{1}{\beta}(x_1 + \ldots + x_n)
\end{aligned}
$$

Then

$$
\frac{\partial l}{\partial \beta} = -n\frac{1}{\beta} + \frac{1}{\beta^2}(x_1 + \ldots + x_n) = 0
$$

And we retrieve here the estimator by moment matching (not always the case!):

$$
\widehat{\beta}_{obs} = \bar{x} \quad , \quad \widehat{\beta} = \bar{X} \, .
$$

When we test several models, we select the one with the highest log-likelihood.

Likelihood Maximization

Statistical properties of the MLE :

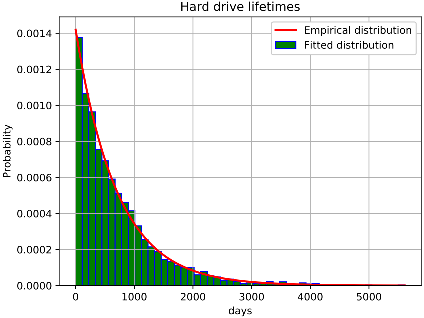
- A MLE is asymptotically without bias, asymptotically normal.
- A MLE is of minimum variance :

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \to \mathcal{N}(0, v(\theta))$$

with asymptotic variance $v(\theta) = \lim_{n\to\infty} \mathbb{V}\left(\sqrt{n}\,\widehat{\theta}\right)$ to be shown smallest possible of all (asymptotically) unbiased estimators of θ.

Likelihood Maximization

Example : see file bootstrapping.py. We have a dataset about 2987 lifetimes of hard-drives (in days). We want to model the HDD lifetime by an exponential distribution, *expo* (β).



We use the scipy command expon.fit(data=..., scale=...). Python finds $\widehat{\beta}$ by ML and a localization parameter (loc) such that $X - loc \sim Expo(\beta)$.
We find $\beta = 704$, loc=0 and the log-likelihood is -22573.

Likelihood Maximization, discrete r.v.

The maximum likelihood estimator (**MLE**) for a discrete random variable

$$\widehat{\theta} = \arg\max_{\theta} l(x_1, \dots, x_n|\theta)$$

where $l(.)$ is the sum of log of **pmf** :

$$l(x_1, \dots, x_n|\theta) = \sum_{i=1}^{n} \ln\left(p(x_k|\theta)\right) .$$

Example 2: If $X_{i:n} \in \{0, 1\}$ are $Ber(p)$, $P(X = x) = p^x(1-p)^{1-x}$ and

$$l(p) = \sum_{i=1}^{n} \ln\left(p^{x_i}(1-p)^{1-x_i}\right)$$

Since $\frac{\partial l}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} x_i - (n - \sum_{i=1}^{n} x_i)\frac{1}{1-p} = 0$ then $\widehat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Likelihood Maximization

For discrete r.v. we replace the pdf by the pmf.

Example 3: If $X_{i:n}$ are $Po(\lambda)$ then $f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$ and

$$L(\lambda) = e^{-n\lambda}\frac{\lambda^{(x_1+\ldots+x_n)}}{x_1!\ldots x_n!}$$
$$l(\lambda) = -n\lambda - (x_1 + \ldots + x_n)\ln\lambda - \ln(x_1!\ldots x_n!)$$

Then

$$\frac{\partial l}{\partial \lambda} = -n + \frac{1}{\lambda}(x_1 + \ldots + x_n) = 0$$

and

$$\widehat{\lambda}_{obs} = \bar{x} \quad , \quad \widehat{\lambda} = \bar{X}\,.$$

Self-Learning 3 Simulations & Bootstrapping

Introduction

- The variance of an estimator measures its reliability: the higher is the variance, the lower we should be confident in using this estimate.
- The variance of an estimator may in some rare (simple) cases be determined analytically.
- There exists a powerful numerical alternative, called "bootstrapping" and based on simulations.

Bootstrapping

1) Simulate M new likely data sets from the available one.
2) Estimate $\widehat{\theta}_k$ for $k = 1, ..., M$ on each of these data sets.
3) Compute the mean and variance of the series of $\left(\widehat{\theta}_k\right)_{k=1:M}$.

Simulations of random numbers

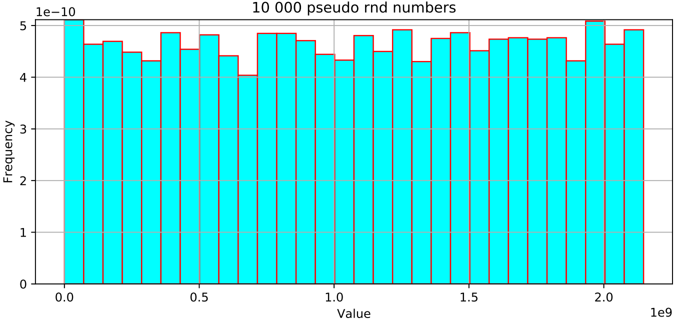
- To perform boostrapping, we need to generate random numbers.
- A pseudo-random number generator (PRNG) is an algorithm for generating a sequence of numbers whose properties approximate the properties of sequences of uniform random numbers. The Linear Congruential Generator (LCG):

$$X_{n+1} = (aX_n + c) \mod m$$

- $X = \{X_0, ...X_m\}$ is a sequence of pseudo random numbers $\in [0, m]$
  - m is the modulus (Ansi C/C++ 2^{31})
  - a is the multiplier (Ansi C/C++ 1103515245)
  - c is the increment (Ansi C/C++ 12345)
  - X_0 is the **seed** value

Simulation of random numbers?

- Histogram of 10 000 pseudo-random numbers generated by the LCG algorithm. See file "randomNumbers.py" (e.g. uniform.rvs(loc=0,scale=1,size=10000))



Utility? Simulations of manufacturing chain, financial derivatives pricing, risk management simulations and so on...

Simulation of continuous r.v.

Consider a continuous cdf $F(x)$ which is invertible. Let X be a r.v. defined

$$X = F^{-1}(U)$$

where U is a continuous uniform r.v. on $[0,1]$. Then X has $F(x)$ for cdf.

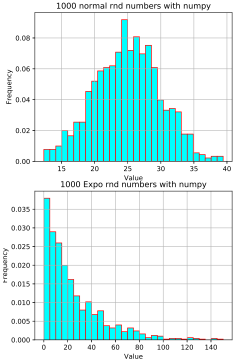
Proof:

$$\begin{aligned} P(X \leq x) &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) \\ &= F(x) \end{aligned}$$

Then to simulate X, we simulate a sample $\{u_1, ..., u_n\}$ from $U \sim uni([0,1])$ with e.g. the Linear Congruential Generator. A sample of X is then given by computed $\{F^{-1}(u_1), ..., F^{-1}(u_n)\}$.

Simulation of continuous r.v.

- In Python, we have the following random numbers generators:



Continuous

- uniform.rvs(loc=0,scale=1,size=10000)
- norm.rvs(loc=μ,scale=σ,size=10000),
- t.rvs(df=n,size=10000)
- chi2.rvs(df=n,size=10000),
- gamma.rvs(a=α,scale=β,size=10000)
- expon.rvs(scale=β, size=1000)

Discrete

- binom.rvs(n=10, p=0.2, size=10000)
- geom.rvs(p=0.2, size=10000)
- poisson.rvs(mu=λ, size=10000)

Bootstrapping

- Bootstrapping methods are methods based on resampling to substitute complex calculations by Monte Carlo simulations.
- Let X be a random variable whose the cumulative distribution is noted $F(.)$, which belongs to a parametric family of distribution $\{F_\theta \; : \; \theta \in \Theta\}$.
- We observe a sample of n observations of X, $\mathrm{X} = (X_1, \dots X_n)$. An occurence of this vector is noted $(x_1, \dots, x_n)$. The empirical cumulative distribution is noted

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{]-\infty, x]}(X_i)$$

- We denote by $\widehat{\theta}(\mathrm{X})$ an estimator of θ,calculated with X.

Bootstrapping

- A bootstrapped sample is obtained by resampling. Let $\mathrm{X}^* = (X_1^* \dots X_n^*)$ such that $P(X_i^* = X_j) = \frac{1}{n}$ for $1 \leq i, j \leq n$. One occurence $x^* = (x_1^*, \dots x_n^*)$ is a sequence of n draws from the initial sample with replacement.

- $\widehat{\theta}(\mathrm{X}^*)$ is called bootstrap replication of $\widehat{\theta}(\mathrm{X})$. The number of possible bootstrap sample is equal to $C_{2n-1}^n = \frac{(2n-1)!}{n!(n-1)!}$.

- Variance is obtained by drawing M bootstrapped sample: $\boldsymbol{X}^{*m} = (X_1^{*m}, \dots X_n^{*m})$ $m = 1 \dots M$.

Bootstrapping

θ is estimated by

$$\bar{\theta}^* = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}(\mathrm{X}^{*m})$$

where $\widehat{\theta}(\mathrm{X}^{*m})$ is the bootstrap replication of $\widehat{\theta}(\mathrm{X})$.

The empirical variance of the estimator is ;

$$S^2\left(\widehat{\theta}\right) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\theta}(\mathrm{X}^{*m}) - \bar{\theta}^*\right)^2$$

Bootstrapping

The sampling distribution of the estimator, $H(x) = P\left(\widehat{\theta}(\mathrm{X}) \leq x\right)$, is the empirical cdf of $\left(\widehat{\theta}(\mathrm{X}^{*m})\right)_{m=1,\ldots,M}$

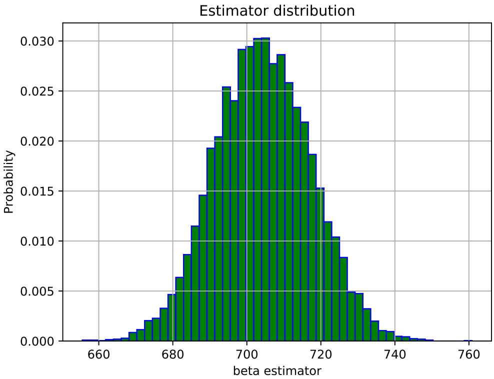
$$H^{boot}(x) = \frac{1}{M} \sum_{m=1}^{M} 1_{\{\widehat{\theta}(\mathrm{X}^{*m}) < x\}}$$

A confidence interval at the level α, for θ is obtained from the sampling distribution of the estimator:
$\left[\underbrace{H_{boot}^{-1}(\alpha/2)}_{\theta_{low}}, \underbrace{H_{boot}^{-1}(1-\alpha/2)}_{\theta_{up}}\right]$. This is the interval such that:

$$P\left(\theta_{low} \leq \theta \leq \theta_{up}\right) = \alpha$$

Bootstrapping

Example : see file bootstrapping.py. We have a dataset about 2987 lifetimes of hard-drives (in days). We fit an *expo*(β).



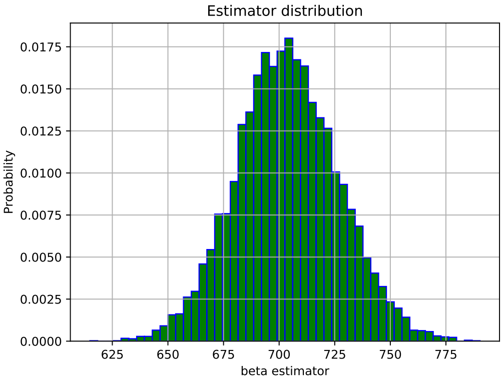

```
Beta estimate :704.38
Beta standard deviation :12.92
Beta 2,5%  quantile :679.0
Beta 97,5% quantile :730.0
```

The 95% confidence interval is such that :

$$P(679 \leq \beta \leq 730) = 95\%$$

Bootstrapping

Example (con't): see file bootstrapping.py. We use only 1000 records instead of 2987.




```
Beta estimate :703.77
Beta standard deviation :23.05
Beta 2,5%  quantile :659.0
Beta 97,5% quantile :750.0
```

Less data... then more uncertainty about the real value of β. The 95% confidence interval is bigger than with the full dataset:

$$P(659 \leq \beta \leq 750) = 95\%$$

Lecture 3.1 Empirical mean and standard deviations: properties

Properties of $\bar{X}$ and S^2

Are estimators reliable ? We have seen that most of them use the empirical mean $\bar{X}$ and variance S^2. What are the properties of these statistics?

Let us consider n random variables $X_{i=1:n} \sim X$ and denote $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(V)$. The CLT states that whatever the distribution of $X_{i=1:n}$, the empirical mean tends to a normal

$$\bar{X} \sim N\left(\mu\, , \frac{\sigma^2}{n}\right) \quad \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Properties of $\bar{X}$ and S^2

A Confidence interval for μ at level $1-\alpha$ (e.g. α=5%) is an interval $[\mu_L, \mu_U]$ such that μ is in this interval with a probability $1-\alpha$.

In this case, $\bar{X}$ is an estimator of μ and σ^2 is known. Since $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$ we infer that

$$P\left(z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

But $-z_{1-\alpha/2} = z_{\alpha/2}$. The $1-\alpha$ confidence interval for μ is then

$$\left[\bar{X} - z_{1-\alpha/2}\sqrt{\sigma^2/n}\,;\, \bar{X} + z_{1-\alpha/2}\sqrt{\sigma^2/n}\right]$$

Properties of $\bar{X}$ and S^2

Example 1: a) A sample of $n = 40$ Gamma r.v. X with a mean $\mu = 100$ and variance $\sigma^2 = 25$. What is the probability that the sample mean $\bar{X}$ is between 98.4505 and 101.5495?

$$\begin{aligned}
&P\left(98.4505 \leq \bar{X} \leq 101.5495\right) \\
&\approx P\left(\frac{98.4505 - 100}{\sqrt{25/40}} \leq Z \leq \frac{101.5495 - 100}{\sqrt{25/40}}\right) \\
&= P(-1.96 \leq Z \leq 1.96) \\
&= P(Z \leq 1.96) - P(Z \leq -1.96) = 0.95
\end{aligned}$$

b) If we do not know μ, can we find a 95% confidence interval ($\alpha = 5\%$) for μ? If $\bar{X} = 99.89$ and $Z_{97.5\%} = 1.96$ then

$$\mu \in \left[99.89 - 1.96\sqrt{\frac{25}{40}}\,;\, 99.89 + 1.96\sqrt{\frac{25}{40}}\right] = [98.34\,;\, 101.44]$$

Numerical solution: see Python file XbarAndS2examples.py

Properties of $\bar{X}$ and S^2

1. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ is an unbiased estimator of $\mathbb{V}(X)$.
2. If $X_i \sim \mathcal{N}(\mu, \sigma^2)$, S^2 and $\bar{X}$ are independent and then

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1} \qquad (1)$$

Proof. (for information). We denote by $SST = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$, the sum of squared total deviations and $\boldsymbol{X} = (X_1, ...X_n)^\top$. Let $\boldsymbol{I}_n$ be $n \times n$ identity matrix and $\boldsymbol{J}_n$ is the $n \times n$ matrix of ones. The SST is rewritten under matrix form:

$$SST = \boldsymbol{X}^\top\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n\right)\boldsymbol{X}$$

The matrix $\boldsymbol{M} = \left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n\right)$ is idempotent, $M^2 = M$, and symmetric, $M = M^\top$.

Properties of $\bar{X}$ and S^2

Proof (cont'd). The trace of $\boldsymbol{M}$ is equal to $tr(\boldsymbol{M}) = n - 1$. We next use the properties that eigenvalues of idempotent matrix must be equal to zero or one. To prove this, let ϕ the normalized eigenvector of $\boldsymbol{M}$ with eigenvalue λ: $\boldsymbol{M}\phi = \lambda\phi$ then

$$\underbrace{\boldsymbol{M}\boldsymbol{M}}_{=M}\phi = \lambda \underbrace{\boldsymbol{M}\phi}_{\lambda\phi} \quad \Rightarrow \quad \lambda\phi = \lambda^2\phi$$

and $\lambda = 0$ or 1. The trace of $\boldsymbol{M}$ is also the sum of eigenvalues. Then $\boldsymbol{M}$ has $n - 1$ eigenvalues equal to 1 and one equal to zero. The spectral decomposition of $\boldsymbol{M}$ is $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top$ where $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}_n$ because $\boldsymbol{M}$ is symmetric and

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I}_{n-1} & 0_{(n-1)\times 1} \\ 0_{1\times(n-1)} & 0 \end{pmatrix}$$

We infer that $\left(D\boldsymbol{A}^\top\boldsymbol{X}\right)_n = 0$. Since $\boldsymbol{D} = \boldsymbol{D}^\top$ and $\boldsymbol{D} = \boldsymbol{D}\boldsymbol{D}$:

$$SST = \boldsymbol{X}^\top\boldsymbol{M}\boldsymbol{X} = \left(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{D}\right)\left(\boldsymbol{D}^\top\boldsymbol{A}^\top\boldsymbol{X}\right)$$

Properties of $\bar{X}$ and S^2

Proof (cont'd). We note $\boldsymbol{e} = (1, .., 1)^\top$. Since $\boldsymbol{X} \sim \mathcal{N}(\mu \boldsymbol{e}, \sigma^2 \boldsymbol{I}_n)$, using standard normal theory:

$$\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{X} \sim \mathcal{N}\left(\mu \boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{e}, \sigma^2 \boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{D}\right) \sim \mathcal{N}\left(\mu \boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{e}, \sigma^2 \boldsymbol{D}\right)$$

showing that components of $\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{X}$ with $\left(\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{X}\right)_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots n-1$ and $\left(\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{X}\right)_n = 0$. Then $\left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top \boldsymbol{X}\right)/\sigma^2$ is χ^2_{n-1} with expectation equal to $n-1$:

$$\mathbb{E}\left(\left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top \boldsymbol{X}\right)/\sigma^2\right) \quad = \quad n-1\,.$$

Therefore the estimator is unbiased:

$$\mathbb{E}\left(S^2\right) = \mathbb{E}\left(\frac{\left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top \boldsymbol{X}\right)}{n-1}\right) \quad = \quad \sigma^2\,.$$

end

Properties of $\bar{X}$ and S^2

A Confidence interval for σ^2 at level $1-\alpha$ (e.g. α=5%) is an interval $[\sigma_L^2, \sigma_U^2]$ such that σ^2 is in this interval with a probability $1-\alpha$.

If $X \sim N(\mu, \sigma^2)$, S^2 is an estimator of σ. Since $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ we infer that

$$P\left(\chi^2_{n-1\,,\,\alpha/2} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{n-1\,,\,1-\alpha/2}\right) \;=\; 1-\alpha$$

The $1-\alpha$ confidence interval is then

$$\left[\frac{(n-1)}{\chi^2_{n-1\,,\,1-\alpha/2}}S^2\,;\;\frac{(n-1)}{\chi^2_{n-1\,,\,\alpha/2}}S^2\right]$$

Properties of $\bar{X}$ and S^2

Example 2 a) An i.i.d. sample of size $n = 100$ is drawn from a normal population with $\sigma^2 = 25$. If S^2 is used to estimate σ^2, what is the probability that the absolute error is greater than 5?

$$P\left(|S^2-\sigma^2|>5\right) = P\left((S^2-\sigma^2)>5\right) + P\left((S^2-\sigma^2)<-5\right)$$
$$= P\left(99\frac{S^2}{\sigma^2} > 99\left(\frac{5}{\sigma^2}+1\right)\right) + P\left(99\frac{S^2}{\sigma^2} < 99\left(1-\frac{5}{\sigma^2}\right)\right)$$
$$= P\left(\chi^2_{99} > 118.8\right) + P\left(\chi^2_{99} < 79.2\right) = 0.1568$$

b) We do not know σ. What is the 95% conf. interval ($\alpha = 5\%$) for σ? If $S^2 = 5.22^2$, $\chi^2_{99\ \alpha/2} = 73.36$, $\chi^2_{99\ 1-\alpha/2} = 128.42$:

$$\sigma \in \left[\sqrt{\frac{99}{\chi^2_{99\,,\,1-\alpha/2}}S^2}\,;\ \sqrt{\frac{99}{\chi^2_{99\,,\,\alpha/2}}S^2}\right] = [4.58\,;\,6.07]$$

Numerical solution: see Python file XbarAndS2examples.py

Properties of $\bar{X}$ and S^2

A student's t r.v. is $= T_n = \frac{Y}{\sqrt{Z/n}}$ where $Z \sim \chi^2_n$ and $Y \sim N(0,1)$. Therefore:

If $X_i \sim \mathcal{N}(\mu, \sigma^2)$ then the following ratio is a Student's T r.v.

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \tag{2}$$

Proof: from previous results, $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Furthermore, $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$. We conclude that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Big/ \left(\sqrt{\frac{(n-1)}{(n-1)}\frac{S^2}{\sigma^2}}\right) \sim t_{n-1}.$$

Properties of $\bar{X}$ and S^2

A Confidence interval for μ at level $1-\alpha$ (e.g. α=5%) is an interval $[\mu_L, \mu_U]$ such that μ is in this interval with a probability $1-\alpha$.

In this case, $\bar{X}$ is an estimator of μ and σ is unknown:

$$P\left(t_{n-1\,,\,\alpha/2} \leq \frac{\bar{X}-\mu}{\sqrt{S^2/n}} \leq t_{n-1\,,\,1-\alpha/2}\right) \quad = \quad 1-\alpha$$

the Student's T is symmetric: $-t_{n-1\ \alpha/2} = t_{n-1\ 1-\alpha/2}$. The $1-\alpha$ confidence interval is then

$$\mu \in \left[\bar{X} - t_{n-1\,,\,1-\alpha/2}\sqrt{S^2/n}\,;\ \bar{X} + t_{n-1\,,\,1-\alpha/2}\sqrt{S^2/n}\right]$$

Properties of $\bar{X}$ and S^2

Example 3 An i.i.d. sample of size $n = 100$ is drawn from a normal population. We do not know $\sigma^2(= 10^2)$ and $\mu(=100)$.

b) What is the 95% confidence interval ($\alpha = 5\%$) for μ? If $\bar{X} = 100.91$, $S^2 = (10.97)^2$, $t_{99\,,1-\alpha/2} = 1.98$:

$$\mu \in \left[100.91 - 1.98\,\frac{10.97}{\sqrt{100}}\,;\,100.91 + 1.98\,\frac{10.97}{\sqrt{100}}\right] = [98.78\,;\,102.15]$$

Numerical solution: see Python file XbarAndS2examples.py

Properties of $\bar{X}$ and S^2

	Distribution of X	μ	σ^2	Statistics
$\bar{X}$	unknown and large n	ok	ok	$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$
S^2	$\mathcal{N}\left(\mu,\sigma^2\right)$	-	ok	$(n-1)\,\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$
$\bar{X}$	$\mathcal{N}\left(\mu,\sigma^2\right)$	ok	-	$\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$

This table is the key to understand Hypothesis testing.

Self-Learning 4: Properties of $\bar{X}$ and S^2, 2 populations

Properties of $\bar{X}$ and S^2, 2 populations

We consider two i.i.d. normal samples:

$$\boldsymbol{X}_1 = \{X_{1,1}, ..., X_{1,n_1}\} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right) \;\Rightarrow \bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$
$$\boldsymbol{X}_2 = \{X_{2,1}, ..., X_{2,n_2}\} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right) \;\Rightarrow \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

We note $\bar{X}_1 = \sum_{i=1:n} X_{1,i}/n_1$ and $\bar{X}_2 = \sum_{i=1:n} X_{2,i}/n_2$. If we remember the properties of normal r.v.,

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

and the normalized difference is

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1) \tag{3}$$

Properties of $\bar{X}$ and S^2, 2 populations

Unbiased estimator of σ_1^2 and σ_2^2 are

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n} \left(X_{1,i} - \bar{X}_1\right)^2 \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n} \left(X_{2,i} - \bar{X}_2\right)^2$$

The following ratio is a Fisher-Snedecor random variable:

$$\frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2} \quad \sim \quad F_{n_1 - 1\,,\, n_2 - 1} \tag{4}$$

Proof: $(n_1 - 1)\,\frac{S_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$ and $(n_2 - 1)\,\frac{S_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$ then

$$\frac{\frac{(n_1-1)}{n_1-1}\,\frac{S_1^2}{\sigma_1^2}}{\frac{(n_2-1)}{n_2-1}\,\frac{S_2^2}{\sigma_2^2}} \sim F_{n_1 - 1\,,\, n_2 - 1}$$

Properties of $\bar{X}$ and S^2, 2 populations

If the two populations have the same variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (but $\mu_1 \neq \mu_2$), an unbiased "pooled" estimator of this variance is

$$S_{pool}^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2},$$

i.e. $\mathbb{E}\left(S_{pool}^2\right) = \sigma^2$ and

$$(n_1 + n_2 - 2)\, \frac{S_{pool}^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}\,. \tag{5}$$

Proof: exactly the same as the proof for 1 population.

Properties of $\bar{X}$ and S^2, 2 populations

A student's t r.v. is $= T_n = \frac{Z}{\sqrt{Y/n}}$ where $Z \sim \chi^2_n$ and $Y \sim N(0,1)$. Therefore:

If the two populations have the same variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (but $\mu_1 \neq \mu_2$)

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \, . \qquad (6)$$

Proof: Direct consequence from previous results, $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$ and $(n_1 + n_2 - 2)\frac{S^2_{pool}}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$.

Properties of $\bar{X}$ and S^2, 2 normal populations

		μ_1, μ_2	σ_1^2, σ_2^2	Statistics
$\bar{X}_1 - \bar{X}_2$	$\sigma_1^2 \neq \sigma_2^2$	ok	ok	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$
S_{pool}^2	$\sigma_1^2 = \sigma_2^2$	-	ok	$(n_1 + n_2 - 2)\frac{S_{pool}^2}{\sigma^2} \sim \chi_{n_1+n_2-2}^2$.
S_1^2, S_2^2	$\sigma_1^2 \neq \sigma_2^2$	-	ok	$\frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2} \sim F_{n_1-1, n_2-1}$
$\bar{X}_1 - \bar{X}_2$	$\sigma_1^2 = \sigma_2^2$	ok	-	$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

This table is the key to understand Hypothesis testing for 2 populations

Lecture 3.2 Hypothesis testing, 1 population

Introduction

Definition: a hypothesis is a claim (assertion) about a parameter θ of a random sample.

The **null hypothesis** is denoted by $H_0 = \theta \in \Theta_0$ and states the assertion to be tested. The alternative hypothesis is $H_1 = \theta \in \Theta_1$.

Example: 25-30 years users of Facebook spend per day, on average μ_0 hours on the website.

We note by μ the average time per day spent on Facebook for these users, H_0 is

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

But μ is not observable... we must estimate it .

Introduction

Based on some observations which are sampled from the probability distribution, we will make our decision of accepting or rejecting H_0. This procedure is called a **statistical test**.

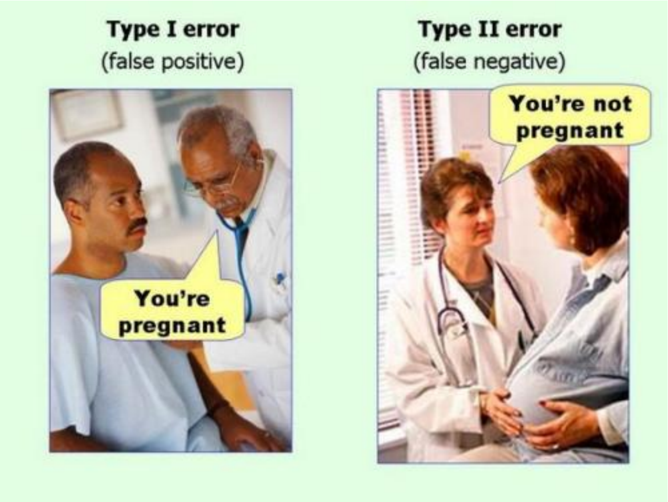
The decision whether to reject the null H_0 is based on the sample. There are 4 possible cases summarized in the following table :

	Reality	
Decision	H_0 is true	H_1 is true
Accept H_0	Correct	Type 2 error
Reject H_0	Type 1 error	Correct

The probability of making a type I error is called the **significance level** of the test and is usually denoted as α: $P(\text{ Type 1 error}) = \alpha$

Introduction

$$H_0: \text{ you are not pregnant}$$



Introduction

The decision to reject the null hypothesis is based on a test statistic, noted $T(.)$. Given a sample $\boldsymbol{X} = \{X_1, ..., X_n\}$ of observations, $T(\boldsymbol{X}) : \mathbb{R}^n \to \mathbb{R}$. $T(\boldsymbol{X})$ is such that its distribution is known.

1) For a choosen α, we determine a rejection region $R_\alpha \subset \mathbb{R}$, such that $P($ Type 1 error$) = \alpha$.

2) We calculate $t = T(\boldsymbol{x})$ the observed value of $T(\boldsymbol{X})$.

3) Decision:

- If $t \in R_\alpha$ then reject H_0.
- if $t \notin R_\alpha$ then do not reject H_0.

Single mean test, σ^2 unknown

We consider a i.i.d. sample $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown variance. We test

$$H_0 \; : \; \mu = \mu_0$$

against 3 alternatives:

$$\begin{aligned} a) \quad H_1 &: \quad \mu > \mu_0 \\ b) \quad H_1 &: \quad \mu < \mu_0 \\ c) \quad H_1 &: \quad \mu \neq \mu_0 \end{aligned}$$

From equation (2), a good choice for the test statistic is the Student's t:

$$T(\boldsymbol{X}) = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

Single mean test, σ^2 unknown

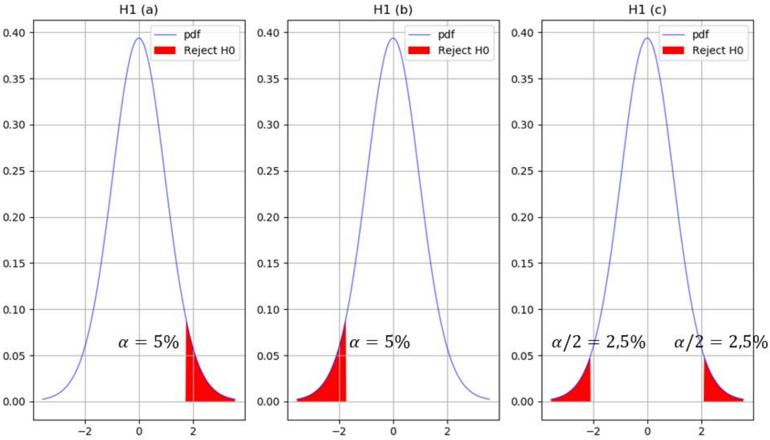
If $t = T(\boldsymbol{x})$ is the observed value of the Student's statistic is "too far" from the mean of the student's distribution, it is likely that H_0 is false.

We first find a critical value c under the assumption that H_0 is true:

$$
\begin{aligned}
&a) \qquad P(T(\boldsymbol{X}) > c \,|\, H_0 \textit{ is true}) = \alpha \\
&b) \qquad P(T(\boldsymbol{X}) < c \,|\, H_0 \textit{ is true}) = \alpha \\
&c) \qquad P(T(\boldsymbol{X}) < -c \text{ or } T(\boldsymbol{X}) > c \,|\, H_0 \textit{ is true}) = \alpha
\end{aligned}
$$

The critical values are the α, $1-\alpha$ or $(\alpha/2\,,\,1-\alpha/2)$ quantiles of the Student's t (c and $-c$ because t is symmetric)

Single mean test, σ^2 unknown



E.g. : $H_1\,(a)$, $\mu > \mu_0$: if $T(\boldsymbol{x})$ is in the red area, a smaller and then more likely statistics maybe obtained with a higher μ_0 => we reject H_0.

Single mean test, σ^2 unknown

We reject H_0 at the level α if $T(\boldsymbol{x}) = \sqrt{n}\frac{\bar{x}-\mu_0}{s}$

$$
\begin{array}{llcll}
a) & H_1 & : & \mu > \mu_0 & T(\boldsymbol{x}) > t_{n-1\ 1-\alpha} \\
b) & H_1 & : & \mu < \mu_0 & T(\boldsymbol{x}) < t_{n-1\ \alpha} \\
c) & H_1 & : & \mu \neq \mu_0 & T(\boldsymbol{x}) < t_{n-1\ \alpha/2} \text{ or } T(\boldsymbol{x}) > t_{n-1\ 1-\alpha/2}
\end{array}
$$

Single mean test, σ^2 unknown

Example 1. SampleMeanTest.py A bottle filling machine is set to fill bottles with soft drink to a volume of 500 ml. The actual volume is known to follow a normal distribution. The manufacturer believes the machine does not work correctly. A sample of 20 bottles is taken and the volume of liquid inside is measured.

test	Volume
1	484.11
	...
	502.85
	...
19	449.08
20	489.27

$H_0 : \mu = 500$ v.s. $H_1 : \mu \neq 500$ Easy to program, package scipy.stats:

```
In [44]: Tx= (Stat.mean-500)/np.sqrt(Stat.variance/n)
    ...: # we compare it to percentiles of a t distribution
    ...: alpha = 0.05
    ...: t_l   = sc.t.ppf(q=alpha/2,df=n-1)
    ...: t_u   = sc.t.ppf(q=1-alpha/2,df=n-1)
    ...: #we see that Tx is in the 2.5% and 97.5% interval of the
Student's t
    ...: print([Tx,t_l,t_u])
[-1.5204626102079255, -2.0930240544082634, 2.093024054408263]
```

alternative: ttest_1samp(), We do not reject H_0 for $\alpha = 5\%$.

Single variance test

We consider a i.i.d. sample $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown variance. We test

$$H_0 \;:\; \sigma^2 = \sigma_0^2$$

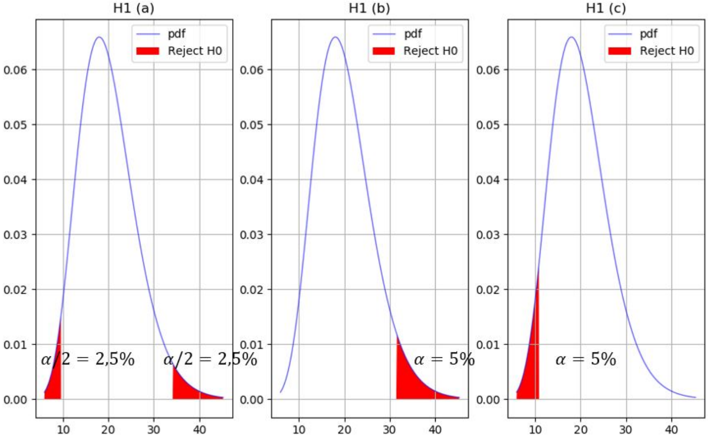
against 3 alternatives:

$$\begin{array}{llll} a) & H_1 & : & \sigma^2 \neq \sigma_0^2 \\ b) & H_1 & : & \sigma^2 > \sigma_0^2 \\ c) & H_1 & : & \sigma^2 < \sigma_0^2 \end{array}$$

From Equation (1), a good choice for the test statistic is the χ^2 test:

$$T(\boldsymbol{X}) = (n-1)\,\frac{S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Single variance test



E.g. : $H_1\,(c)$, $\sigma^2 < \sigma_0^2$: if $T(\boldsymbol{x})$ is in the red area, a higher and then more likely statistics maybe obtained with a smaller σ_0 => we reject H_0.

Single variance test

We reject H_0 at the level α if $T(\boldsymbol{x}) = (n-1)\,\frac{s^2}{\sigma_0^2}$

a)	$H_1 : \sigma^2 \neq \sigma_0^2$	$T(\boldsymbol{x}) < \chi^2_{n-1\ \alpha/2}$ or $T(\boldsymbol{x}) > \chi^2_{n-1\ 1-\alpha/2}$
b)	$H_1 : \sigma^2 > \sigma_0^2$	$T(\boldsymbol{x}) > \chi^2_{n-1\ 1-\alpha}$
c)	$H_1 : \sigma^2 < \sigma_0^2$	$T(\boldsymbol{x}) < \chi^2_{n-1\ \alpha}$

Single variance test

Example 1 cont'd. SampleVarianceTest.py The quality control dept. requires that the st. dev. of soda volumes is not higher than 20ml to avoid complaints from customers.

test	Volume
1	484.11
	...
	502.85
	...
19	449.08
20	489.27

One sided sample variance test: $H_0 : \sigma = 20$, $H_1 : \sigma > 20$. No function available but easy to program:

```
In [2]:
   ...:
   ...:
   ...: sg0=20
   ...: Tx= (n-1)*Stat.variance/sg0**2
   ...: # we compare it to percentiles of a chi2 distribution
   ...: alpha = 0.05
   ...: t_u   = sc.chi2.ppf(q=1-alpha,df=n-1)
   ...: #we see that Tx is in the  95% interval
   ...: #of a chi-square
   ...: print([Tx,t_u])
[29.199522237499995, 30.14352720564616]
```

We do not reject $H_0 : \sigma^2 = 20^2$ for $\alpha = 5\%$

P-value

> **P-value** : It is the smallest level of significance for which the data indicate rejection of the null hypothesis.

Let $T(\boldsymbol{X})$ be a test statistic such that small values of T give evidence that H_0 is wrong. For a given sample $\boldsymbol{x}$, the p-value is:

$$p(\boldsymbol{x}) = P\left(T(\boldsymbol{X}) < T(\boldsymbol{x}) \,|\, H_0 \text{ is true}\right)$$

Let $T(\boldsymbol{X})$ be a test statistic such that high values of T give evidence that H_0 is wrong. The p-value is:

$$p(\boldsymbol{x}) = P\left(T(\boldsymbol{X}) > T(\boldsymbol{x}) \,|\, H_0 \text{ is true}\right)$$

Let $T(\boldsymbol{X})$ be a test statistic **symmetric around zero** such that high and low values of T give evidence that H_0 is wrong. The p-value is:

$$p(\boldsymbol{x}) = 2P\left(T(\boldsymbol{X}) > |T(\boldsymbol{x})| \,|\, H_0 \text{ is true}\right)$$

P-value

A small p-value indicates that H_0 is very unlikely. A high p-value informs us that H_0 is likely.

Example: we consider a i.i.d. sample $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown variance. We test

$$H_0 \ : \ \mu = \mu_0$$

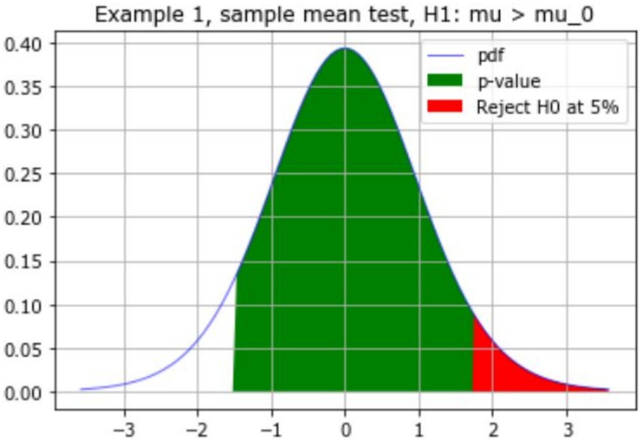
against $H_1 \ : \mu > \mu_0$. The p-value is such that $p(\boldsymbol{x}) = P(t_{n-1} > T(\boldsymbol{x}))$ where $T(\boldsymbol{x}) = \sqrt{n}\frac{\bar{x}-\mu_0}{s}$. For a confidence level, α, we reject H_0 if

$$p < \alpha \ \Leftrightarrow \ T(\boldsymbol{x}) > t_{n-1,\alpha} \, .$$

Rejecting a null hypothesis at level α using the critical region method is equivalent to rejecting H_0 when $p(x) < \alpha$.

P-value

Example 1. $H_0 : \mu = 500$ v.s. $H_1 : \mu > 500$. $T(\boldsymbol{x}) = -1.5204$ and $p = 0.9275$.



P-value

Example 1 cont'd. SampleMeanTest.py and SampleVarianceTest.py.

p-value for the 2 sided T test ($H_1 : \mu \neq 500$) on the average volume $p = 2\,P(t_{n-1} > |\,T(\boldsymbol{x})|)$

```
In [264]: pval = 2 * (1-sc.t.cdf(np.abs(Tx),df=n-1))
     ...: print(pval)
0.14486225283259024
```

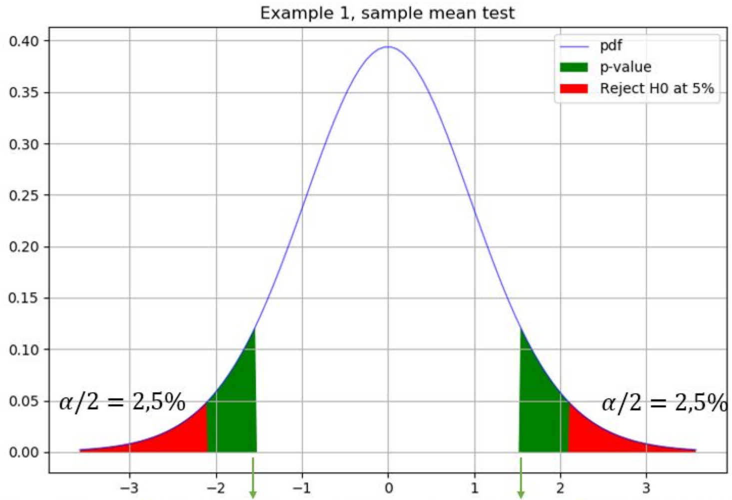
p-value>5%. We do not reject $H_0 : \ Vol. = 500 \ (\alpha = 5\%)$

p-value for the one sided sample variance test. p-values: $p = P(\chi^2_{n-1} > T(\boldsymbol{x}))$

```
In [275]: pval = 1-sc.chi2.cdf(Tx,df=n-1)
     ...: print(pval)
     ...:
0.06291083863033275
```

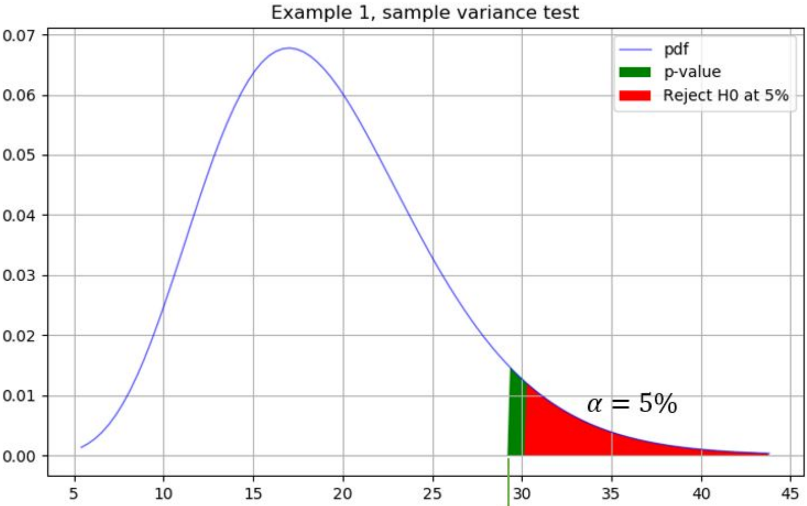
p-value>5%. We do not reject $H_0 : \sigma^2 = 20^2$ for $\alpha = 5\%$

P-value



$$p - value = P(t_{19} < T(x) = -1{,}52) \quad + \quad P(t_{19} > |T(x)| = 1{,}52) = 14{,}48\%$$

P-value



$$p - value = P(\chi^2_{19} > T(x) = 29{,}190) = 6{,}26\%$$

Self learning 5: Hypothesis testing, 2 populations

Comparison of 2 means

We consider two i.i.d 2 populations with the **same variance**:

$$\boldsymbol{X}_1 = \{X_{1,1}, ..., X_{1,n}\} \sim \mathcal{N}\left(\mu_1, \sigma^2\right) \Rightarrow \bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$$
$$\boldsymbol{X}_2 = \{X_{2,1}, ..., X_{2,n}\} \sim \mathcal{N}\left(\mu_2, \sigma^2\right) \Rightarrow \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$$

We test if the 2 samples have the same means, or more generally :

$$H_0 \ : \ \mu_1 - \mu_2 = \delta$$

Against

$$a) \ H_1 \ : \ \mu_1 - \mu_2 > \delta$$
$$b) \ H_1 \ : \ \mu_1 - \mu_2 < \delta$$
$$c) \ H_1 \ : \ \mu_1 - \mu_2 \neq \delta$$

2 cases: σ known and unknown.

Comparison of 2 means

Case 1 σ known: If we remember the properties of normal r.v.

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

From equation (3), we use as statistics of test:

$$T(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1) \tag{7}$$

We reject $H_0 : \mu_1 - \mu_2 = \delta$ at the level α if

a)	H_1 :	$\mu_1 - \mu_2 > \delta$	$T(\boldsymbol{x}_1, \boldsymbol{x}_2) > z_{1-\alpha}$
b)	H_1 :	$\mu_1 - \mu_2 < \delta$	$T(\boldsymbol{x}_1, \boldsymbol{x}_2) < z_{\alpha}$
c)	H_1 :	$\mu_1 - \mu_2 \neq \delta$	$T(\boldsymbol{x}_1, \boldsymbol{x}_2) < z_{\alpha/2}$ or $T(\boldsymbol{x}_1, \boldsymbol{x}_2) > z_{1-\alpha/2}$

where e.g. z_α is the percentile of a $Z \sim \mathcal{N}(0,1)$ $(P(Z \leq z_\alpha) = \alpha)$

Comparison of 2 means

Case 2 σ unknown: From equation (6), we use as statistics of test:

$$T(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{S_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \tag{8}$$

We reject $H_0 : \mu_1 - \mu_2 = \delta$ at the level α if

a) $H_1 : \mu_1 - \mu_2 > \delta$ $\quad T(\boldsymbol{x}_1, \boldsymbol{x}_2) > t_{n_1+n_2-2\,1-\alpha}$

b) $H_1 : \mu_1 - \mu_2 < \delta$ $\quad T(\boldsymbol{x}_1, \boldsymbol{x}_2) < t_{n_1+n_2-2\,\alpha}$

c) $H_1 : \mu_1 - \mu_2 \neq \delta$ $\quad \begin{cases} T(\boldsymbol{x}_1, \boldsymbol{x}_2) < t_{n_1+n_2-2\,\alpha/2} & \text{or} \\ T(\boldsymbol{x}_1, \boldsymbol{x}_2) > t_{n_1+n_2-2\,1-\alpha/2} \end{cases}$

where e.g. $t_{n_1+n_2-2\,\alpha}$ is the α-percentile of a Student's T.

Comparison of 2 means

Example 1. ComparisonMeanTest.py Two machines fill bottles to a volume of 500 ml (normal distribution). The manufacturer believes that machines 1 and 2 fills different volumes. A sample of 2x20 bottles is taken and the volume inside is measured.

Volume Mach. 1	Volume Mach. 2
484.11	525.19
...	...
502.85	516.3
...	...
449.08	517.7
489.27	512.63

scipy.stats. $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$

```
In [314]: Xb1   = np.mean(X1)
     ...: Xb2   = np.mean(X2)
     ...: S1    = np.std(X1,ddof=1)
     ...: S2    = np.std(X2,ddof=1)
     ...: Spool = np.sqrt(((n-1)*S1**2+(n-1)*S2**2)/(n+n-2))
     ...: Tx    = (Xb1-Xb2)/(Spool*np.sqrt(1/n+1/n))
     ...: # we compare it to percentiles of a t distribution
     ...: alpha = 0.05
     ...: t_l   = sc.t.ppf(q=alpha/2,df=n+n-2)
     ...: t_u   = sc.t.ppf(q=1-alpha/2,df=n+n-2)
     ...: #we see that Tx is in the 2.5% and 97.5% interval of the
Student's t
     ...: print([Tx,t_l,t_u])
[-4.013925858585345, -2.0243941645751367, 2.024394164575136]

In [315]: pval = 2*sc.t.cdf(-np.abs(Tx),df=n+n-2)
     ...: print(pval)
0.0002709539978837068
```

p-value=0.02%<5%, We **reject** H_0 for $\alpha = 5\%$.
Other command: ttest_ind(X1,X2)

Comparison of 2 variances

We consider two i.i.d 2 populations with different variances:

$$\boldsymbol{X}_1 = \{X_{1,1}, ..., X_{1,n}\} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$$
$$\boldsymbol{X}_2 = \{X_{2,1}, ..., X_{2,n}\} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$$

We test if the 2 samples have the same variances i.e.:

$$H_0 \ : \ \sigma_1 = \sigma_2$$

Against

$$a)\ H_1 \ : \ \sigma_1 \neq \sigma_2$$
$$b)\ H_1 \ : \ \sigma_1 > \sigma_2$$
$$c)\ H_1 \ : \ \sigma_1 < \sigma_2$$

Comparison of 2 variances

From equation (4), we use as statistics of test (Fisher's test):

$$T(\mathbf{X}_1, \mathbf{X}_2) = \frac{S_1^2}{S_2^2} \quad \sim \quad F_{n_1-1\,,\,n_2-1} \tag{9}$$

We reject $H_0 : \sigma_1^2 = \sigma_2^2$ at the level α if

$$
\begin{array}{lllll}
a) & H_1 & : & \sigma_1 \neq \sigma_2 & \begin{cases} T(\mathbf{x}_1, \mathbf{x}_2) < F_{n_1-1, n_2-, 1\ \alpha/2} & \text{or} \\ T(\mathbf{x}_1, \mathbf{x}_2) > F_{n_1-1, n_2-, 1\ 1-\alpha/2} \end{cases} \\
b) & H_1 & : & \sigma_1 > \sigma_2 & T(\mathbf{x}_1, \mathbf{x}_2) > F_{n_1-1, n_2-, 1\ 1-\alpha} \\
c) & H_1 & : & \sigma_1 < \sigma_2 & T(\mathbf{x}_1, \mathbf{x}_2) < F_{n_1-1, n_2-, 1\ \alpha}
\end{array}
$$

where e.g. $F_{n_1-1, n_2-, 1\ \alpha}$ is the α-percentile of a Fisher.

Comparison of 2 variances

Example 1. ComparisonVarianceTest.py Two machines fill bottles with soft drink to a volume of 500 ml. The actual volume is known to follow a normal distribution. Do filled volumes by machine 1 have a bigger variance than those of machine 2?

Volume Mach. 1	Volume Mach. 2
484.11	525.19
...	...
502.85	516.3
...	...
449.08	517.7
489.27	512.63

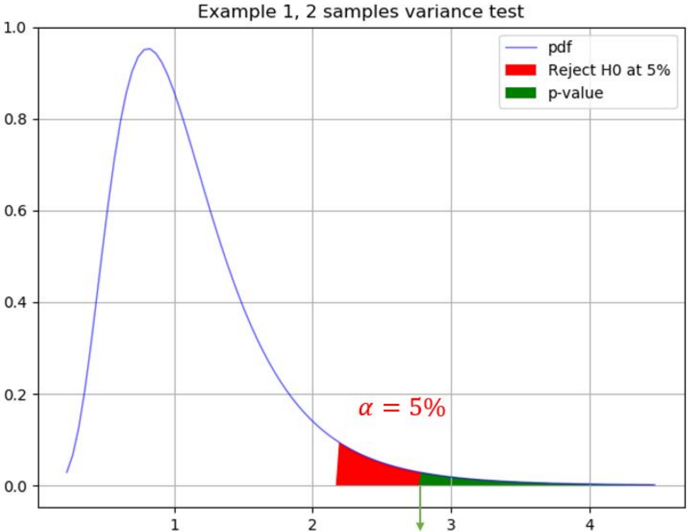
Package scipy.stats. $H_0 : \sigma_1 = \sigma_2$, $H_1 : \sigma_1 > \sigma_2$

```
In [56]: S1     = np.std(X1,ddof=1)
    ...: S2     = np.std(X2,ddof=1)
    ...: Tx      =S1**2/S2**2
    ...: # we compare it to percentiles of a t distribution
    ...: alpha = 0.05
    ...: f_u    = sc.f.ppf(q=1-alpha,dfn=n-1, dfd=n-1)
    ...: #we see that Tx is in the 2.5% and 97.5% interval of the
Student's t
    ...: print([Tx,f_u])
[2.767818462651005, 2.168251601406261]

In [57]: pval = 1-sc.f.cdf(Tx,dfn=n-1 , dfd=n-1)
    ...: print(pval)
    ...:
0.01594817166131013
```

p-value 1.59%<5%, We **reject** H_0 for $\alpha = 5\%$.
Other command: Bartlett(X1, X2) but it is a two-sided test, i.e. $H_1 : \sigma_1 \neq \sigma_2$!

Comparison of 2 variances



$$p - value = P\big(F_{19,19} > T(x) = 2{,}7681\big) = 1{,}59\%$$

Lecture 4.1 Linear regression

Linear regression

We observe n realizations Y_i that is related to k factors $(x_{i,1}, ..., x_{i,k})^\top$ for $i = 1, ..., n$. We postulate the following linear relation

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \epsilon_i \quad i = 1, ..., n \tag{10}$$

where $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. Matrix notations:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,k} \end{pmatrix}$$

$\boldsymbol{Y}$ is a n vector, $\boldsymbol{\beta}$ is a $k+1$ vector and X a $n \times (k+1)$ matrix.

Linear regression

The n vector of noises is noted $\boldsymbol{\epsilon}$. Notice that realizations of $\boldsymbol{Y}$ is a vector noted $\boldsymbol{y}$

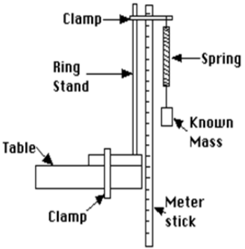
$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

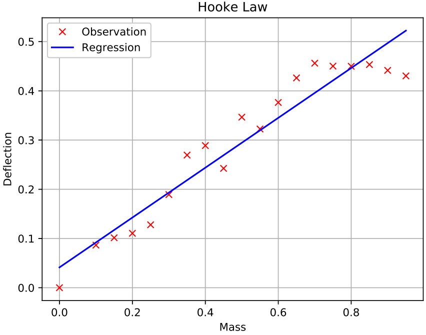
Eq. (10) is then reformulated as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

Linear regression

Example, Hooke's law experiments. A spring is loaded with a given weight M, and the resulting deflection D is measured. A linear relationship $D = \beta_1 M$ is expected. There are 19 records, mass in kilograms, and deflection in meters (see file springs.csv).





Linear regression

- The best $\widehat{Y}$ prediction of Y for a given value of $\boldsymbol{x} = (1, x_1, ..., x_k)^\top$ is:

$$\widehat{Y} = \mathbb{E}(Y \mid \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

- However $\boldsymbol{\beta}$ and σ^2 are unknown. We denote by $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ their estimates.
- How do we estimate the parameters $\boldsymbol{\beta}$ and σ^2 based on a data sample $(\boldsymbol{x}_i, y_i)$ for $i = 1, ..., n$?
- Methods of estimation: Likelihood maximization

Linear regression

We have $\theta = (\boldsymbol{\beta}, \sigma)$ and conditionally to $\boldsymbol{x}_i$, the response is normal:

$$Y_i | \boldsymbol{x}_i \sim \mathcal{N}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta} \,;\, \sigma^2\right) \qquad i = 1, \ldots, n$$

The estimate of θ maximizes the log-likelihood function:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\sigma^2}\right)\right) \qquad (11) \\ &= \sum_{i=1}^{n} \left(-\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\sigma^2}\right)\right) \\ &\propto -\underbrace{\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}_{(*)} \end{aligned}$$

Therefore, for a given $\sigma > 0$, the estimates of $\boldsymbol{\beta}$ minimizes (*).

Least Squares Minimization. The estimate $\widehat{\beta}$ of β minimizes the sum of squared errors (SSE):

$$\begin{aligned} \widehat{\beta} &= \arg_{\beta}\min\,(\boldsymbol{y}-\boldsymbol{X}\beta)^{\top}(\boldsymbol{y}-\boldsymbol{X}\beta) \qquad (12) \\ &= \arg_{\beta}\min \sum_{i=1}^{n}\left(y_i-\boldsymbol{x}_i^{\top}\beta\right)^2 \end{aligned}$$

which is:

$$\widehat{\beta} = \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}\,. \qquad (13)$$

Proof. Eq. (12) is a direct consequence of (*). If we derive it w.r.t. β and cancel the derivative:

$$2\boldsymbol{X}^{\top}(\boldsymbol{y}-\boldsymbol{X}\beta)=0 \qquad (14)$$

which admits Eq. (13) as solution.

Linear regression

The best prediction of Y_i for a vector of factor $\boldsymbol{x}_i$ is

$$\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\beta} \qquad i = 1, ..., n$$

or under matrix form, if $\widehat{\boldsymbol{y}}=(y_1, ..., y_n)^\top$, $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\beta}$. By construction

$$\widehat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top}\boldsymbol{y}$$
$$= \boldsymbol{H}\, y$$

where $\boldsymbol{H}$ is the hat matrix. This matrix is symmetric and idempotent, i.e. :

$$\boldsymbol{H} = \boldsymbol{H}^\top \qquad \boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}\,.$$

If the assumption of normality does not hold, the LSM estimators are not the ML estimators but we still can use the Least Squares Minimization.

Simple linear regression

Simple regression: Only one explanatory factor (simple regression, $k = 1$):

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Then

$$\left\{ \widehat{\beta}_0 \, , \, \widehat{\beta}_1 \right\} \quad = \quad \arg_{\beta_0,\beta_1} \min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \quad , \qquad (15)$$

Deriving Eq. (15) w.r.t. β_0 and β_1 and cancelling the differentials leads to

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Simple linear regression

Simple regression: The intercept $\widehat{\beta}_0$ and the slope, $\widehat{\beta}_1$, are:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Example, Hooke's law experiments. See regressionSpring.py. Either direct calculation or linregress(...). We find $\widehat{\beta}_1 = 0.506$ and $\widehat{\beta}_0 = 0.041$

```
In [75]: X = datn[:,0] ;   Y = datn[:,1]
    ...: Xb=np.mean(X) ;   Yb= np.mean(Y)
    ...: b1=np.sum((X-Xb)*(Y-Yb))/np.sum((X-Xb)**2)
    ...: b0=Yb-b1*Xb
    ...: # alternative scipy linregress(.)
    ...: slope, intercept, r_value, p_value, std_err =
sc.linregress(X,Y)
    ...: print(np.round([b0,intercept,b1,slope],3))
    ...:
[0.041 0.041 0.506 0.506]
```

Linear regression

- How do we estimate the quality of the model? We split the variance of observations into:
  - A variance explained by the model
  - A residual variance
- The higher is the variance explained by the model, the better is the goodness of fit.

Let us note $\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$ then

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SS\ Total} = \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{SS\ Error} + \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}_{SS\ Regression}$$

Proof. We have that $(y_i - \bar{y}) = (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y})$ and according to Eq. (14):

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} - \bar{y}\right)$$

$$= \left(\widehat{\beta}_0 - \bar{y}\right) \underbrace{\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)}_{\frac{1}{2}\frac{\partial\, SSE}{\partial \widehat{\beta}_0} = 0} + \sum_{j=1}^{k} \widehat{\beta}_j \underbrace{\sum_{i=1}^{n} \boldsymbol{x}_{i,j} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)}_{\frac{1}{2}\frac{\partial\, SSE}{\partial \widehat{\beta}_j} = 0} = 0$$

The $R^2 \in [0, 1]$ is the proportion of the variance explained by the model:

$$R^2 = \frac{SSR}{SST} \Leftrightarrow 1 - R^2 = \frac{SSE}{SST}.$$

The closer to unity, the better is the model.

Linear regression

Example, Hooke's law experiments. See regressionSpring.py. R^2 computed by linregress(...), $R^2 = 99.88\%$! Excellent fit.

Remark: Matrix notations of the SST, SSR, SSE. If $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{J}_n$ is the $n \times n$ matrix of ones:

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \boldsymbol{y}^\top \left(\boldsymbol{I}_n - \frac{1}{n} \boldsymbol{J}_n \right) \boldsymbol{y} \\
SSR &= \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 = \boldsymbol{y}^\top \left(\boldsymbol{H} - \frac{1}{n} \boldsymbol{J}_n \right) \boldsymbol{y} \\
SSE &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \boldsymbol{y}^\top \left(\boldsymbol{I}_n - \boldsymbol{H} \right) \boldsymbol{y}
\end{aligned}
$$

Lecture 4.2 Properties of regression coefficients

Linear regression

$\widehat{\boldsymbol{\beta}}$ is an unbiased Gaussian estimator of $\boldsymbol{\beta}$ (i.e. $\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \beta$):

$$\widehat{\boldsymbol{\beta}} \;\sim\; \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$$

Remark: here $\widehat{\boldsymbol{\beta}}$ is a multivariate normal, see appendix.

Proof 1) The estimator is a random variable $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ where $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Since a linear combination of Normal r.v.'s is Normal, $\widehat{\boldsymbol{\beta}}$ is normal. 2) Since $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\beta$, the estimator is unbiased:

$$\begin{aligned}
\mathbb{E}\left(\widehat{\beta}\right) &= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathbb{E}(\boldsymbol{Y}) \\
&= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{X}\beta = \beta .
\end{aligned}$$

Linear regression

2) Using standard linear algebra, we infer that:

$$\mathbb{V}\left(\widehat{\beta}\right) = \mathbb{V}\left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}\right)$$

$$= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathbb{V}(\boldsymbol{Y}) \boldsymbol{X} \left(\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)^\top = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

An unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-(k+1)} \underbrace{\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}\right)^\top \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}\right)}_{SSE}$$

and

$$(n-(k+1)) \frac{\widehat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-(k+1)}$$

is a chi-square random variable with $n-(k+1)$ degree of freedoms.

Linear regression (for information)

Proof: The trace of $\boldsymbol{H}$ is

$$\begin{aligned} tr\left(\boldsymbol{H}\right) &= tr\left(\boldsymbol{X}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\right) = tr\left(\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}\right) \\ &= tr\left(\boldsymbol{I}_{k+1}\right) = k+1\,. \end{aligned}$$

We next use the properties that eigenvalues of idempotent matrix ($\boldsymbol{H}^2 = \boldsymbol{H}$)must be equal to zero or one. To prove this, let ϕ the normalized eigenvector of $\boldsymbol{H}$ with eigenvalue λ: $\boldsymbol{H}\phi = \lambda\phi$ then

$$\underbrace{\boldsymbol{H}\boldsymbol{H}}_{=H}\phi = \lambda\underbrace{\boldsymbol{H}\phi}_{\lambda\phi} \quad\Rightarrow\quad \lambda\phi = \lambda^2\phi$$

and $\lambda = 0$ or 1. The matrix $\boldsymbol{I}_n - \boldsymbol{H}$ is also idempotent because:

$$\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\left(\boldsymbol{I}_n - \boldsymbol{H}\right) = \boldsymbol{I}_n - \boldsymbol{I}_n\boldsymbol{H} - \boldsymbol{H}\boldsymbol{I}_n + \boldsymbol{H}^2 = \boldsymbol{I}_n - \boldsymbol{H}$$

Linear regression (for information)

Proof (cont'd) Its trace is also the sum of eigenvalues:

$$tr\left(\boldsymbol{I}_n - \boldsymbol{H}\right) \quad = \quad tr\left(\boldsymbol{I}_n\right) - tr\left(\boldsymbol{H}\right) = n - (k+1)$$

then $\boldsymbol{I}_n - \boldsymbol{H}$ has $n-(k+1)$ eigenvalues equal to 1 and $(k+1)$ equal to zero. The spectral decomposition of $\boldsymbol{I}_n - \boldsymbol{H}$ is $\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top$ where $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}_n$ because $\boldsymbol{I}_n - \boldsymbol{H}$ is symmetric and

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I}_{n-k-1} & 0_{(n-k-1)\times(k+1)} \\ 0_{(k+1)\times(n-k-1)} & 0_{(k+1)\times(k+1)} \end{pmatrix}$$

Since $\boldsymbol{D} = \boldsymbol{D}^\top$ and $\boldsymbol{D} = \boldsymbol{D}\boldsymbol{D}$:

$$SSE = \boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{Y} = \left(\boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{D}\right)\left(\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{Y}\right)$$

Given that $\boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n\right)$, using standard normal theory:

$$\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{X}\beta, \sigma^2 \boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{D}\right) \sim \mathcal{N}\left(\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{X}\beta, \sigma^2 \boldsymbol{D}\right)$$

Linear regression (for information)

Proof (cont'd) showing that components of $\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{Y}$ are independent. Furthermore, $\left(\boldsymbol{D}^\top \boldsymbol{A}^\top \boldsymbol{Y}\right)_i \sim \mathcal{N}\left(0, \sigma^2\right)$ for $i = 1, \ldots, n-k-1$ and null otherwise. Then $\left(\boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top \boldsymbol{Y}\right)/\sigma^2$ is χ^2_{n-k-1} with expectation equal to $n-k-1$:

$$\mathbb{E}\left(\left(\boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top \boldsymbol{Y}\right)/\sigma^2\right) = n-k-1$$

Therefore the estimator is unbiased:

$$\mathbb{E}\left(\widehat{\sigma}^2\right) = \mathbb{E}\left(\frac{\left(\boldsymbol{Y}^\top \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top \boldsymbol{Y}\right)}{n-k-1}\right) = \sigma^2 .$$

end

Simple linear regression

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators i.e. $\mathbb{E}\left(\widehat{\beta}_0\right) = \beta_0$ and $\mathbb{E}\left(\widehat{\beta}_1\right) = \beta_1$. If $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$, their variances are

$$\mathbb{V}\left(\widehat{\beta}_0\right) = \frac{\sigma^2\,\overline{x^2}}{S_{xx}} \qquad \mathbb{V}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \quad \mathbb{C}\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \frac{-\sigma^2\bar{x}}{S_{xx}}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ (but σ^2 is unknown...)

Since $\epsilon \sim \mathcal{N}(0, \sigma^2)$ then

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2\,\overline{x^2}}{S_{xx}}\right) \quad \widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$(n-2)\,\frac{\widehat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$$

Simple Linear regression

In practice σ^2 is unknown and we replace it by $\widehat{\sigma}^2$: $\mathbb{V}\left(\widehat{\beta}_0\right) = \frac{\widehat{\sigma}^2\,\overline{x^2}}{S_{xx}}$ and $\mathbb{V}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}^2}{S_{xx}}$: consequence $\widehat{\beta}_0$ and $\widehat{\beta}_1$ become Student's T!

Example, Hooke's law experiments. See regressionSpring.py. Best package for linear regression: "statsmodels". Function OLS(Y,X).fit

	coef	std err
const	0.0413	0.021
x1	0.5064	0.036

- According to theory, β_0 should be zero... but we find $\widehat{\beta}_0 = 0.04$. However the standard deviation is high $\sqrt{\frac{\widehat{\sigma}^2\,\overline{x^2}}{S_{xx}}} = 0.02$ compared to $\widehat{\beta}_0$! $\widehat{\beta}_0$ is then very inaccurate.
- The standard deviation of $\widehat{\beta}_1$ is $\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}} = 0.02$, low compared to $\widehat{\beta}_1 = 0.5064$. $\widehat{\beta}_0$ is then reliable.

Test of the significance of linear regression

From properties of S^2, if $\beta_1 = ... = \beta_k = 0$,the normalized SST is a chi-square r.v. with $n-1$ d.f.

$$\frac{SST}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \sim \chi^2_{n-1}$$

From this section, the normalized SSE a chi-square r.v. with $n-(k+1)$ d.f.

$$\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 \sim \chi^2_{n-(k+1)}$$

Since $SST = SSE + SSR$ and as a χ^2_n r.v. is a sum of n r.v. :

$$\frac{SSR}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y}\right)^2 \sim \chi^2_k$$

Therefore, if $\beta_1 = ... = \beta_k = 0$, the next ratio is a Fisher r.v.:

$$F^* = \frac{SSR \,/\, k}{SSE \,/\, (n-k-1)} \sim F_{k\,,\,n-(k+1)}$$

Test of the significance of linear regression

If F^* (to interpret as explained variance on unexplained variance) is "too small" then the assumption of linearity between Y and $\boldsymbol{x}$ must be rejected.

Significance test of the regression:

$$H_0 \ : \ \beta_1 = \ldots = \beta_k = 0$$
$$H_1 \ : \beta_j \neq 0 \text{ for some } j \in \{1, .., k\}$$

with the statistics of test F^* :

$$F^* = \frac{MSR}{MSE} = \frac{SSR \ / \ k}{SSE \ / \ (n-k-1)} \quad \sim \quad F_{k\,,\,n-(k+1)}$$

Reject H_0 at a confidence level α (e.g. 5%)

- If $F^* > F_{k\,,\,n-(k+1)\,,\,1-\alpha}$
- if the p-value $p_{val} = P(F^* < F_{k\,,\,n-(k+1)})$ is lower than α

Significance of a simple linear regression

Example: Hooke's law experiments. In a simple framework, $Y_i = \beta_0 + \beta_1 \boldsymbol{x}_i + \epsilon_i$

	d.f.	Value	Mean of ...	F^*	p-value 5%
SSR	1 (k)	0.378	0.378	193.68	1.e-10
SSE	$17(n-2)$	0.033	0.0019		
SST	$18(n-1)$	0.411			

For $\alpha = 5\%$, $F_{1,17,1-\alpha} = 4.45$. Since $F^* > F_{1,17,0.95}$ and $p_{val} < 5\%$, we reject H_0.

See also regressionSpring.py. Package statsmodels,

```
F-statistic:                  193.7
Prob (F-statistic):        1.01e-10
```

Tests of regression coefficients

Let $c_{j,j}$ be the j^{th}diagonal element of $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$. Estimators $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, ..., \widehat{\beta}_k\right)^\top$ of linear regression coefficients are Student's T r.v.:

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{n-(k+1)}$$

Proof. $\widehat{\boldsymbol{\beta}}$ is a multivariate normal $\mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \ \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right)$ then

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \ c_{jj}}} \quad \sim \quad \mathcal{N}(0, 1)$$

The estimator $\widehat{\sigma}^2$ of variance σ^2 is $(n-(k+1))\,\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(k+1)}$. By definition of the Student's T, we conclude that

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{c_{jj}}} = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \ c_{jj}}} \Bigg/ \sqrt{\frac{n-(k+1)}{n-(k+1)}\frac{\widehat{\sigma}^2}{\sigma^2}} \sim t_{n-(k+1)}$$

Tests of regression coefficients

We use this result to test the significance of each β_j:

$$H_0 \ : \quad \beta_j = \beta_{j,0}$$

against 3 alternatives:

$$\begin{array}{lll} a) & H_1 \ : & \beta_j > \beta_{j,0} \\ b) & H_1 \ : & \beta_j < \beta_{j,0} \\ c) & H_1 \ : & \beta_j \neq \beta_{j,0} \end{array}$$

with the statistics of test $T_j^* = \frac{\widehat{\beta}_j - \beta_{j,0}}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{n-(k+1)}$. We reject H_0 at the level α if T_j^*

$$\begin{array}{lllll} a) & H_1 \ : & \beta_j > \beta_{j,0} & & T_j^* > t_{n-k-1\ 1-\alpha} \\ b) & H_1 \ : & \beta_j < \beta_{j,0} & & T_j^* < t_{n-k-1\ \alpha} \\ c) & H_1 \ : & \beta_j \neq \beta_{j,0} & & T_j^* < t_{n-k-1\ \alpha/2} \text{ or } T_j^* > t_{n-k-1\ 1-\alpha/2} \end{array}$$

Tests of regression coefficients

A Confidence interval for β_j at level $1-\alpha$ (e.g. α=5%) is an interval $[\beta_{j,L}, \beta_{j,U}]$ such that β_j is in this interval with a probability $1-\alpha$.

Since $\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{c_{jj}}} \sim t_{n-(k+1)}$ we infer that

$$P\left(t_{n-k-1\ \alpha/2} \leq \frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{c_{jj}}} \leq t_{n-k-1\ 1-\alpha/2}\right) \quad = \quad \alpha$$

The $1-\alpha$ confidence interval is then

$$\left[\widehat{\beta}_j + t_{n-k-1\ \alpha/2}\,\widehat{\sigma}\sqrt{c_{jj}}\,;\,\widehat{\beta}_j + t_{n-k-1\ 1-\alpha/2}\,\widehat{\sigma}\sqrt{c_{jj}}\right]$$

Or (the Student's T is symmetric: $-t_{n-1\ \alpha/2} = t_{n-1\ 1-\alpha/2}$)

$$\left[\widehat{\beta}_j - t_{n-k-1\ 1-\alpha/2}\,\widehat{\sigma}\sqrt{c_{jj}}\,;\,\widehat{\beta}_j + t_{n-k-1\ 1-\alpha/2}\,\widehat{\sigma}\sqrt{c_{jj}}\right]$$

Test of β: simple linear regression

When one explanatory factor (simple regression, $k = 1$), $Y_i = \beta_0 - \beta_1 x_i + \epsilon_i$, the coefficient $c_{0,0}$ and $c_{1,1}$ are easy to calculate:

$$c_{0,0} = \frac{\overline{x^2}}{S_{xx}} \qquad c_{1,1} = \frac{1}{S_{xx}}$$

where $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, and $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. We test

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_{1,0} \\ H_1 &: \quad \beta_1 \neq \beta_{1,0} \end{aligned}$$

with the statistics of test $T_1^* = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\widehat{\sigma}\sqrt{S_{xx}^{-1}}} \sim t_{n-2}$. The confidence interval is

$$\beta_1 \in \left[\widehat{\beta}_1 - t_{n-k-1\ 1-\alpha/2}\,\widehat{\sigma}\sqrt{S_{xx}^{-1}}\ ;\ \widehat{\beta}_1 + t_{n-k-1\ 1-\alpha/2}\,\widehat{\sigma}\sqrt{S_{xx}^{-1}}\right]$$

Test of β: simple linear regression

b) We test

$$\begin{aligned} H_0 &: \quad \beta_0 = \beta_{0,0} \\ H_1 &: \quad \beta_0 \neq \beta_{0,0} \end{aligned}$$

with the statistics of test $T_0^* = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\widehat{\sigma}\sqrt{\bar{x}^2 S_{xx}^{-1}}} \sim t_{n-2}$. The confidence interval is

$$\beta_0 \in \left[\widehat{\beta}_0 - t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{\bar{x}^2 S_{xx}^{-1}}\,;\,\widehat{\beta}_0 + t_{n-k-1\;1-\alpha/2}\,\widehat{\sigma}\sqrt{\bar{x}^2 S_{xx}^{-1}}\right]$$

Example: Hooke's law experiments. See file regressionSpring.py. The OLS(.) command from stats.model compute statistics, p-value and 95% confidence intervals.

Tests of regression coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0413	0.021	1.991	0.063	-0.002	0.085
x1	0.5064	0.036	13.917	0.000	0.430	0.583

For $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, the statistics is $T_1^* = 13.917$ and the p-value is $p = 2P(t_{n-2} > |T_1^*|) = 0$. Strong rejection of H_0. The 95% confidence interval is:

$$\beta_1 \in [0.430\,;\,0.583]$$

For $H_0 : \beta_0 = 0$, $H_1 : \beta_0 \neq 0$ the statistics is $T_0^* = 1.991$ and the p-value is $p = 2P(t_{n-2} > |T_0^*|) = 6.3\%$. Acceptation of H_0 for $\alpha = 5\%$. Notice that for $\alpha > 6.3\%$ we reject H_0. The 95% confidence interval is:

$$\beta_0 \in [-0.002\,;\,0.085]$$

Tests of regression coefficients

However, you can easily compute the statistics T_1^* and T_0^* p-values and confidence intervals directly with scipy and numpy:

```
#you can calculate the t stat yourself
Sxx   = sum((X-Xb)**2)                     #Sxx
X2b   = np.mean(X**2)                      #mean of X^2
Yhat  = results.predict(Xm)                #prediction
sghat = np.sqrt(sum((Y-Yhat)**2)/(n-2))    #estimate of sigma
# Test 1: beta1 =0 v.s. beta1<>0
T1    = b1/(sghat*np.sqrt(Sxx**-1))       #test statistics
pval1 = 2*(1-sc.t.cdf(abs(T1),df=n-2))    #p-value
#confidence interval for beta1
CI1   = b1+[-sghat*np.sqrt(Sxx**-1)*sc.t.ppf(q=0.975,df=n-2),  \
        +sghat*np.sqrt(Sxx**-1)*sc.t.ppf(q=0.975,df=n-2) ]
#very low pvalue , we reject H0 : b1=0 at 5%
# Test 2: beta0 =0 v.s. beta0<>0
T0    = b0/(sghat*np.sqrt(X2b*Sxx**-1))  #test statistics
pval0= 2*(1-sc.t.cdf(abs(T0),df=n-2))     #p-value
#confidence interval for beta0
CI0   = b0+[-sghat*np.sqrt(X2b*Sxx**-1)*sc.t.ppf(q=0.975,df=n-2),  \
            +sghat*np.sqrt(X2b*Sxx**-1)*sc.t.ppf(q=0.975,df=n-2) ]
```

Prediction interval, simple regression

For a model $Y = \beta_0 + \beta_1 X + \epsilon$, the prediction for an unobserved value $X = x_0$ is

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \,.$$

The **prediction interval** for Y_0 at level α is provided by

$$\left[\widehat{y}_0 - S_{pred}\, t_{n-2\,;\,1-\frac{\alpha}{2}} \;;\; \widehat{y}_0 + S_{pred}\, t_{n-2\,;\,1-\frac{\alpha}{2}}\right] \,,$$

where

$$S_{pred}^2 = \widehat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

This result is a direct consequence of previous propositions. A detailed proof is proposed in the last session of exercise.

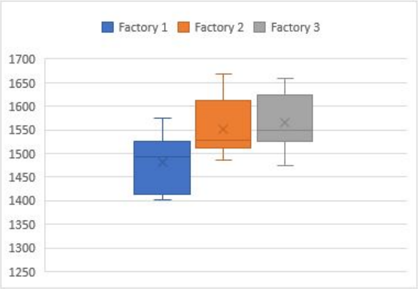
Lecture 4.3 Analysis of Variance (ANOVA)

1 factor ANOVA

Example : we want to test the resistance of a ceramic produced in three different factories: is it the same on average?

Data : 10 measurements per factory

Obs.	Factory 1	Factory 2	Factory 3
1	1495	1486	1572
2	1574	1610	1647
3	1401	1538	1615
4	1456	1515	1475
5	1418	1501	1521
6	1404	1522	1540
7	1517	1532	1544
8	1491	1617	1556
9	1501	1524	1659
10	1553	1669	1527
Mean	**1481**	**1551,4**	**1565,6**
S2	**3638,7**	**3542,3**	**3414,7**



1 factor ANOVA

If μ_1, μ_2 and μ_3 are respectively the mean resistance measured in each factory, we test (assumption normal dis. and same variance):

$$\begin{cases} H_0 : & \mu_1 = \mu_2 = \mu_3 \\ H_1 : & \exists\, i \neq j \,,\, \mu_i \neq \mu_j \end{cases} \tag{16}$$

We can think to do pairwise comparisons... but

- This is tedious when we have more than 3 sub-samples,
- The risk of type 1 error increases (rejecting H_0 that is true).

Alternative : test of significance of a categorical linear regression.

1 factor ANOVA

In many statistical applications, we study the influence of categorical factors on the behaviour of a variable of interest (the response) by an experiment.

Example of ceramic production : does the production site impact the resistance of the produced material?

The response, Y, is the resistance of ceramic and the factor is a categorical variable "factory", with 3 levels.

To reformulate this with a categorical regression we consider 2 binary (or "dummy") variables

$$X_1 = \begin{cases} 1 & \text{Fact. 1} \\ 0 & \text{else} \end{cases}, X_2 = \begin{cases} 1 & \text{Fact. 2} \\ 0 & \text{else} \end{cases}$$

pandas.get_dummies(data, drop_first=False, dtype=None)

1 factor ANOVA

"Dummified" dataset : $n = 30$ and $k = 2$,

i	y_i	$x_{i,1}$	$x_{i,2}$
1	1495	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	1553	1	0
11	1486	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
20	1669	0	1
21	1572	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
30	1527	0	0

Warning : do not introduce a third binary variables for the third factory! In this case, the matrix $\boldsymbol{X}$ is ill-posed !

1 factor ANOVA

Categorical linear regression : let $\epsilon \sim N(0, \sigma^2)$ then

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Testing eq. (16) is equivalent to test

$$\begin{cases} H_0 : & \beta_1 = \beta_2 = 0 \\ H_1 : & \exists\, i\,,\ \beta_i \neq 0 \end{cases} \tag{17}$$

We recognize the global F significance test (with $n = 30$, $k = 2$)!

Underlying assumptions : normality, same variance of sub-samples, independence

1 factor ANOVA

	d.f.	Value	Mean of ...	F^*	p-value 5%
SSR	2 (k)	41050	20525	5.81	0.00796
SSE	27($n-3$)	95361	3532		
SST	29($n-1$)	136411			

Conclusion: we reject the hypothesis of identical average hardness!

Remark: the assumption of equality of variances can be checked with the Bartlett's test (extension of the bivariate variance test).