BUSINESS STATISTICS & ANALYTICS (IDSS21030)

ESSEC Business School

Bachelor Period
2015-2016

Contents

The course

There are very few things which we know, which are not capable of being reduc'd to a Mathematical Reasoning : and when they cannot, it's a sign our knowledge of them is very small and confus'd; and where a mathematical reasoning can be had, it's as great a folly to make use of any other, as to grope for a thing in the dark, when you have a candle standing by you.

John Arbuthnot, 1692,

Preface, Of the Laws of chance

This course aims at introducing the main concepts, methods and strategies of analysis needed for a better understanding of the basic statistical procedures commonly used in business and economics. Statistics supports the management activity by means of two main approaches:

1. **Data Exploration:** when the purpose is to describe, by either visualizing or summarizing, a mass of data in order to better understand the phenomenon under study; such methods are frequently used either as tools for a first approach to the domain or as instruments for synthesising the data into information. In such a case, graphical representations and statistical indices are computed and interpreted for summarising the features of the data and extracting the information relevant to the decision making process.
2. **Data Modelling:** when the purpose is to build up a mathematical model for an observed reality. In such a case, the parameters of the mathematical model are estimated according to an optimizing statistical criterion. Then, the estimated parameters help in understanding the strength and kind of the relationships between observed phenomena so as to explain the past, understand the present and eventually forecast the future. A simple model, called *the linear regression model* is introduced.

The presentation of methods belonging to both approaches is accompanied by the discussion of real examples so as to make the understanding of concepts and statistical methods easier and directly connected to business and related fields where they provide support to decision making processes.

Communication taking place mostly in English nowadays, all students must master statistical tools in this language. This is why these notes are in English.

Remark 1 (How to read and study?) *We want to stress that while reading these notes, the important aspect to keep in mind is that you must* ***understand*** *the methodology and philosophy of statistics rather than simply remember the formulas. For instance, try to understand why the Central Limit Theorem is the key result that allows to do any statistics and how it is used to center and scale random variables to yield an asymptotic standard normal. Once you have understood this, you have mastered the main tool of the course!*
The handout may seem long, but once you have understood the concepts, the theoretical content is actually limited. Many elements are just repetitions and applications of the same methodology in different contexts. This is why it is important to understand *rather than simply* memorize *(in particular since you are allowed one sheet of handwritten notes and formulas in the exam).*

Introduction

You have to know how to deal with numbers,
otherwise you should not do business.
And numbers naturally come with uncertainty.
That is why Statistics is so important to your business.
(G. Tiao, Int. Conf. on Business & Indust. Stat. , Aug.07)

What we mean by statistics and analytics

Definition 1 *Statistics : science of collecting, organizing, presenting, analyzing and interpreting data.*

Used as a decision tool in many disciplines such as financial analysis, econometrics, auditing, production and operations

including services improvement, and marketing research.

Examples of applications:

- Accounting: auditors use statistical sampling procedures when conducting audits for their clients.
- Economics: economists use statistical information in making forecasts about the future of the economy or some aspect of it.
- Finance: financial advisors use a variety of statistical information, including price-earnings ratios and dividend yields, to guide their investment recommendations.
- Law: nowadays lawyers need to understand their probability of winning a case. Comparisons with previous similar cases, where the lawyers need controlling for the differences between all cases, are becoming systematic.
- Marketing: electronic point-of-sale scanners at retail checkout counters are being used to collect data for a variety of marketing research applications.
- Production: a variety of statistical quality control charts are used to monitor the output of a production process and the conformance to standards

Which tools to help to take good decisions?

- Exploratory Analysis or Descriptive Statistics: one or multidimensional exploratory approaches directly on the data , presented under summary calculations, graphs, charts and tables; study in the multivariate case of the linear relations between variables that can also allow to select the most

pertinent variables to explain another variable.

- Inferential Statistics: mathematical methods to generalize results obtained on a sample to results on the population; consists of setting a probabilistic model (with unknown parameters) when considering the data as realizations of random variables and estimating the parameters from the data (with related tools as confidence intervals and tests).

Some vocabulary

Definition 2 *A **population** is the collection of all items under consideration, people or objects, or of interest for the purpose of the analysis.*
*A **sample** is a portion of the population selected for analysis; it needs to be representative to be able to deduce results from the sample on the whole population.*
*A **parameter** is a summary measure that describes a characteristic of the population.*

Inference:

- act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former
- act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty

A statistic is a summary measure computed from a sample. It is also used as a tool to draw inference.

Overview

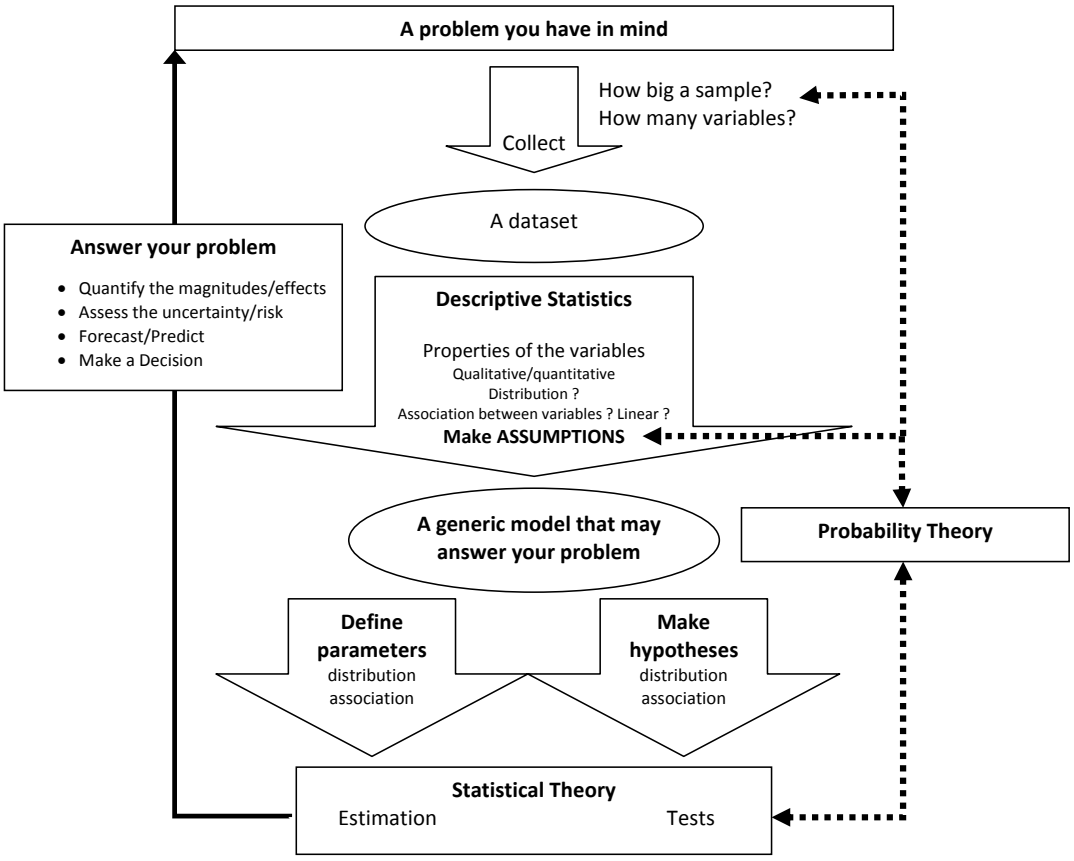
1. In a deterministic setting, working directly on the data,
   (a) we propose a descriptive analysis in the univariate case (chapter 1) as well as in the bivariate case (chapter 2);
   (b) we then extend to the beginning of *unsupervized* classification, i.e. techniques that help discover similarities in the dataset. We study in particular the idea of Clustering.
   (c) In this multidimensional case, the selection of the variables used to explain linearly another one can be done with the help of another exploratory approach: the Principal Component Analysis (PCA or Analyse en Composantes Principales, ACP, in French).
2. In a probabilistic setting, the data are seen as realizations of random variables;

(a) this allows to build a general probabilistic model in view of further understanding a phenomenon and forecasting.
(b) To fit the model (with unknown parameters) on the data, we will need to estimate the parameters thanks to mathematical methods of the inferential statistic, (chapter 4);
6.
(c) Closely related to the estimation problem is the notion of statistical testing, studied in chapter 5.

3. Finally, we study a very useful tool, the linear regression model, which allows to analyze how one or several variables influence another. It also helps making predictions.

The following graph presents an overview of the statistical methodology that

we will analyze in this book.



References

There are many books on *Business Statistics* in the library that those who have no prior knowledge of Statistics. American textbooks are usually less technical than their French counterparts.

A good reference used in French Engineering schools is: Saporta, G (2011). *Probabilités, Analyse des Données et Statistique*, 3e ed. Technip.

For those who have never followed a course in statistics or probabilites, a good start is to refer to any of the “Business Statistics” books in the library.

We recommend that you use books as they are a good source of examples and exercises that you should attempt for practice (although we provide some at the end of the handout and in the chapters).

Here are a few.

1. Anderson, D.R., D.J. Sweeney and T.A.Williams (2006), *Modern Business Statistics*, South-Western / Thomson Learning.
2. Casella, G. and Berger, R.L. (2002) *Statistical Inference.* 2nd ed. Duxbury. (**)
3. Cottrell M., Genon-Catalot V. et Duhamel C., (2005), *Exercices de Probabilités, Troisième éditon*, Paris : Cassini.
4. Curwin J. and R. Slater (2004), *Quantitative Methods for Business Decisions*, Intl Thomson Business Press
5. Droesbeke, J.-J. (1997), *Eléments de statistique*, Ed. de l'Université de Bruxelles et Ed. Ellipses.
6. Foata, D., J. Franchi and A. Fuchs, (2012), *Calcul des probabilités,* 3e ed. Dunod.
7. Fraser, C. (2009), *Business Statistics for Competitive Advantage with Excel 2007. Basics, Model Building and Cases*, Springer.
8. Grimmett, G. and D.R. Stirzaker (2001), *Probability and random processes*, Oxford University Press. (**)
9. Goldfarb, B. and C. Pardoux, (2007), Introduction à la méthode statistique.
10. Lindsey, J.K. (2004), *Introduction to applied statistics: a modelling approach*, Oxford University Press.
11. McClave, J.T., P.G. Benson and T. Sincich (2005), *Statistics for Business and Economics*, Upper Saddle River, NJ: Prentice Hall.
12. Mendenhall, W., Reinmuth, J. E. and R. J. Beaver (1993), *Statistics for Management and Economics*.
13. Moore, D. S. and George .P. McCabe (2006), *Introduction to the Practice of Statistics*, Fifth Edition, New York : W.H. Freeman.
14. Newbold, P., William L. Carlson, Betty M. Thorne (2003), *Statistics for Business and Economics.*
15. Py, Bernard (2010), *La statistique sans formule mathématique*, Pearson Education.
16. Ross, S. (2005), *Introduction to Orobability and Statistics for Engineers and Scientists*, 2nd ed. Elsevier. (**)
17. Ross, S. (2007), *Introduction to Probability Models,*8th Edition, Elsevier, available as an e-book. (*)
18. Tenenhaus, M. (2007), *Statistique: méthodes pour Décrire, Expliquer et*

Prévoir, Editions Dunod.
19. George R. Terrell (2000),*Mathematical Statistics: a unified introduction*, Springer Verlag.
20. Vidal, A. (2010), *Statistique descriptive et inférentielle avec Excel – Approche par l'exemple*, 2ème édition., Editions PU Rennes, Collection : Didact Statistiques.
21. Wonnacott, T. and R. Wonnacott (1991), *Statistique en economie, gestion, sciences et medecine*, Economica.

Softwares

Students are required to use Excel and the XLStat software packages to carry out calculations and perform empirical work. Excel is a common spreadsheet. Spreadsheets are the starting point for most data analysis. While the limitations of Excel become obvious quickly, knowledge of Excel is useful at all levels of statistics and econometrics. You can either load the built-in Excel Data Analysis macro or use the **XLStat** add-in. **In this course we will use XLStat which you can download at ESSEC (unfortunately, there is a problem in the current version with MacOS)**

For those who have a more advanced background or an interest in computing, you may also use some of the following:

- Matlab is a programming language with built in support for matrices and sophisticated graphics and is commonly used in the academic community and outside. SciLab is the freeware version.
- Eviews is a menu driven software environment designed for econometrics with an emphasis on financial econometrics.
- R. Is a statistical programming language with a graphical output device. It is freely available and highly influential in statistical science. There is a lot of free software in its library.
- Stata. This is popular with quite a large number of microeconometricians (in accounting and control, microeconomics, management, operations management, corporate finance...)
- SAS is used a lot in the industry.

PART I

Descriptive Analysis & Analytics

Chapter 1
Data Visualization & Univariate Exploration

This chapter mostly recalls what you already know. In terms of Data Visualization (or *Data Viz*), we provide below the strict minimum. In addition to the capacities of Excel and XLStat, we recommend the software Tableau, for which there are numerous online videos that help you learn it and practice it by yourself with your own datasets.

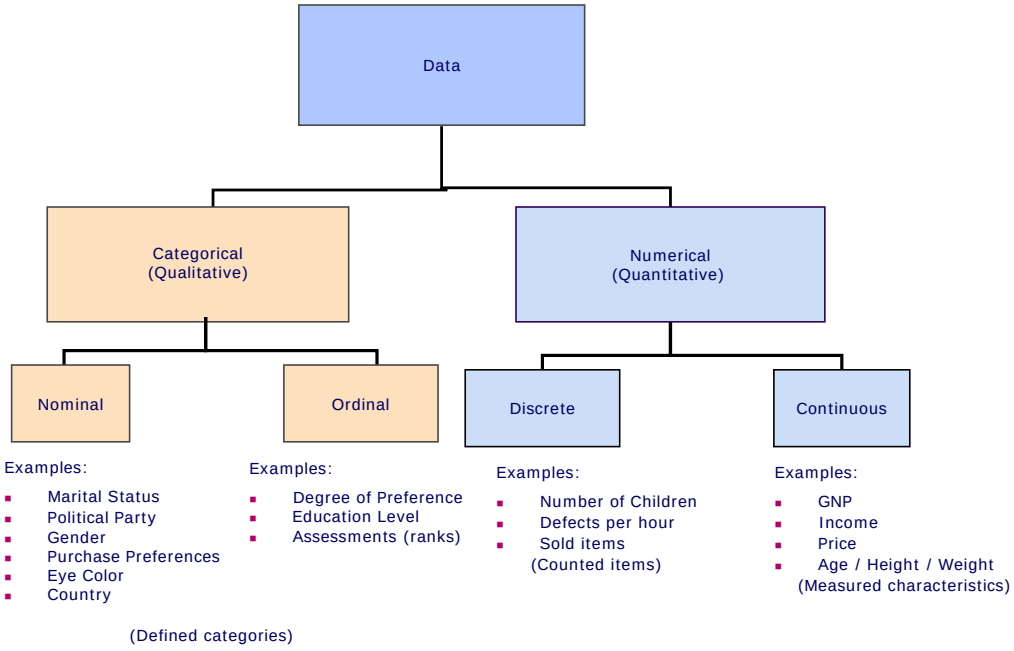
1.1 Tabular and Graphical Procedures

- The techniques of Univariate Exploratory Analysis consist of simple arithmetic (numerical methods) and easy-to-draw pictures (graphical procedures) that can be used to describe and summarize data quickly and consistently, by referring to one variable at a time.
- The choice of statistical methods depends on the type of data you have to study.

Exploratory Analysis

- Summarizing Qualitative (Categorical) Data:
  - ► Frequency Distribution
  - ► Relative or Percent Frequency Distribution
  - ► Bar Graphs and Pie Charts
- Summarizing Quantitative (Numerical) Data:
  - ► Frequency Distribution
  - ► Relative of Percent Frequency Distributions

- Histogram
- Cumulative Frequency Distributions



1.1.1 Qualitative Data

1.1.1.1 Frequency Distribution

Definition 3 *A **frequency distribution** is a tabular summary of data showing the frequency (also called "count"), that is the number of items in each of several nonoverlapping classes, or the number of statistical units showing a specific category of a categorical variable.*

The objective is to **provide insights** about the data that cannot be quickly obtained by looking only at the original data.

Example 1 (The "Marada Inn") *Example on qualitative data:*
Guests staying at Marada Inn were asked to rate the quality of their accommodations as being excellent, above average, average, below average, *or* poor. *The ratings provided by a sample of 20 quests are shown below as a series of*

categories:

Below Average	Average	Above Average
Above Average	Above Average	Above Average
Above Average	Below Average	Below Average
Average	Poor	Poor
Above Average	Excellent	Above Average
Average	Above Average	Average
Above Average	Average	

Frequency Distribution for the "Marada Inn":

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

1.1.1.2 Relative Frequency Distribution

Definition 4 *The **relative frequency** of a class is the fraction or proportion (multiplied by 100 in case you want a percent) of the total number of data items belonging to the class.*

A **relative (or percent) frequency distribution** is a tabular summary of a set

of data showing the relative (or percent) frequency for each class.

Rating	Relative Frequency	Percent Frequency
Poor	.10	10
Below Average	.15	15
Average	.25	25
Above Average	.45	45
Excellent	.05	5
Total	1.00	100

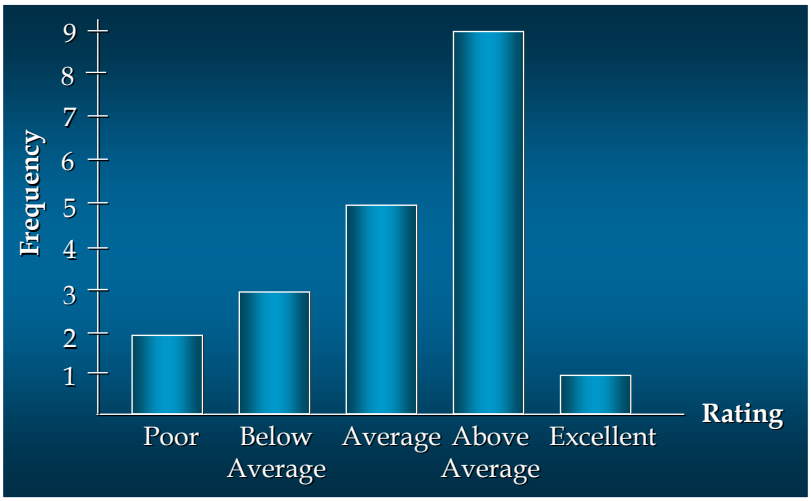
1.1.1.3 Bar Graph and Pie Charts for Qualitative Data

Bar Graph

Definition 5 *A **bar graph** is a graphical device for depicting qualitative data that have been summarized in a frequency, relative frequency, or percent frequency distribution.*

- *On the horizontal axis we specify the labels that are used for each of the classes.*
- *A **frequency, relative frequency**, or **percent frequency** scale can be used for the vertical axis.*
- *Using a bar of fixed width drawn above each class label, we extend the height appropriately.*
- *The **bars are separated** to emphasize the fact that each class is a separate category.*

Example on qualitative data: The "Marada Inn"

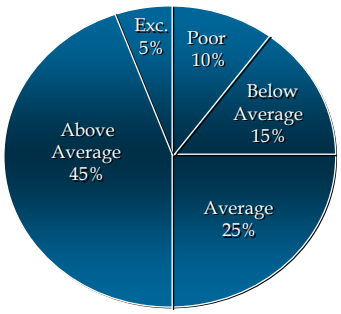


The pie chart

The pie chart is a commonly used graphical device for presenting relative frequency distributions for qualitative data.

- First draw a circle; then use the relative frequencies subdivide the circle into sectors that correspond to the relative frequency for each class.
- Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume .25(360) = 90 degrees of the circle.

Example 2 (the "Marada Inn")



- *One-half of the surveyed customers gave Marada a quality rating of "above average" or "excellent" (looking at the left side of the pie). This might please*

the manager.

- *For each customer who gave an "excellent" rating, there were two customers who gave a "poor" rating (looking at the top of the pie). This should displease the manager.*

1.1.2 Quantitative data

1.1.2.1 Frequency Distribution and Relative Frequency Distribution

- Need to select **Number of Classes** (k):
  - ► There exist several empirical rules
  - ► The minimum number of classes depends on the number of observed units but also on the data distribution
  - ► Sturges' rule: $k = 1 + 3.3 \log_{10}(N)$ where N is the number of observed units (approximate k to the next integer)
- **Class Width** (use classes of equal width):

$$\frac{Largest\ Data\ Value - Smallest\ Data\ Value}{Number\ of\ Classes}$$

Example 3 (The "Hudson Auto Repair") ***Example on quantitative data:*** *The manager of "Hudson Auto" would like to get a better picture of the distribution of costs for engine tune-up parts. A sample of 50 customer invoices has been taken and the costs of parts, rounded to the nearest dollar, are listed below as a series of individual values:*

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

If we choose six classes, the Approximate Class Width :

$$(109 - 52)/6 = 9.5 \approx 10$$

Cost ($)	Frequency	Relative Frequency	Percent Frequency
50-59	2	.04	4
60-69	13	.26	26
70-79	16	.32	32
80-89	7	.14	14
90-99	7	.14	14
100-109	5	.10	10
Total	50	1.00	100

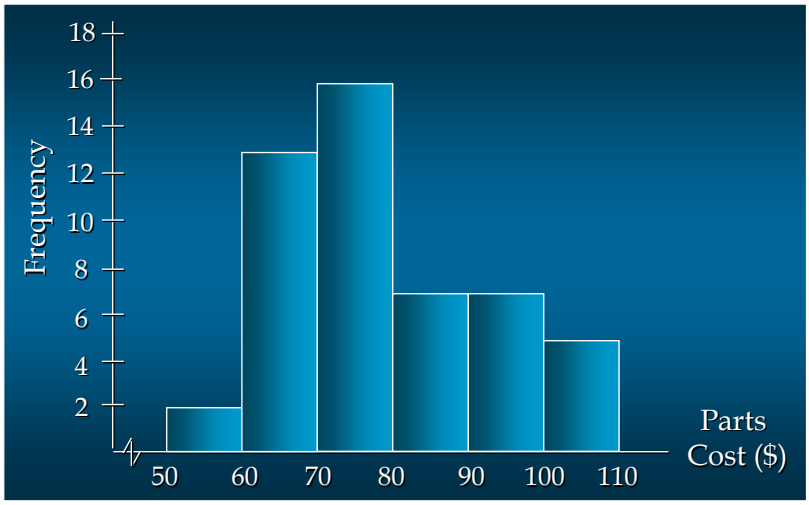
- ***Insights Gained from the Percent Frequency Distribution:***

  - ► *Only 4% of the parts costs are in the $50-59 class.*
  - ► *30% of the parts costs are under $70.*
  - ► *The greatest percentage (32% or almost one-third) of the parts costs are in the $70-79 class.*
  - ► *10% of the parts costs are $100 or more.*

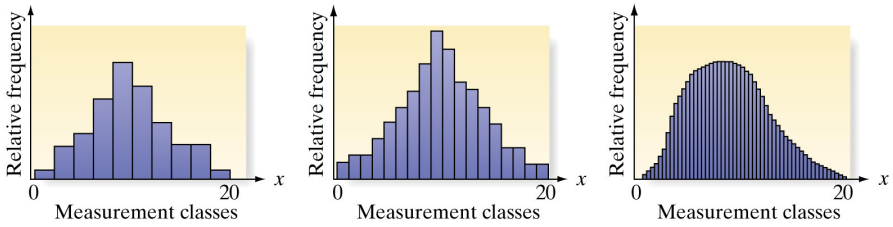
1.1.2.2 Histogram for Quantitative Data

- A common graphical presentation of quantitative data is a **histogram.**
- The classes of the variable of interest are placed on the horizontal axis and the frequency, relative frequency, or percent frequency is placed on the vertical axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has **no natural separation between rectangles of adjacent classes.**

Application to the "Hudson Auto Repair" example



- Choose the number of classes by stretching/shrinking the class width:



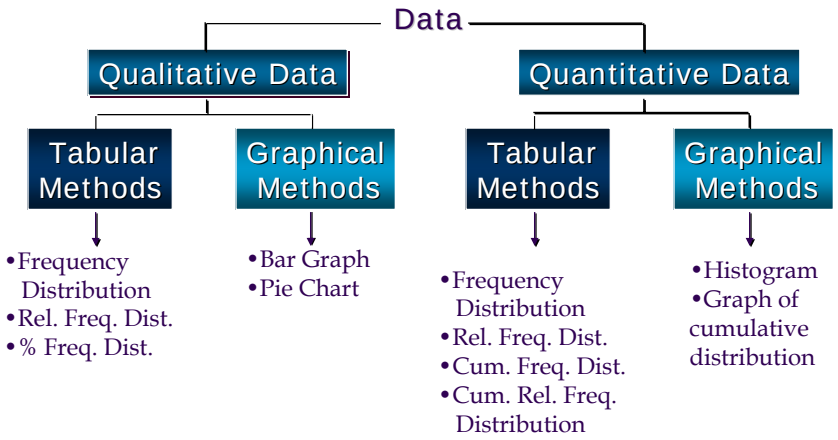
  - ► The proportion of the total area under the histogram that falls in an interval of the horizontal axis is equal to the relative frequency of measurements falling in the same interval.
  - ► As the number of measurements in a data set is increased, you can obtain a better data description by decreasing the width of the class intervals. When the class intervals become small enough, a relative frequency histogram will (for all practical purposes) appear as a smooth curve.

1.1.2.3 Cumulative Distribution

Definition 6 *The **cumulative (relative, percent) frequency distribution** shows*

the number (proportion, percentage) of items with values less than or equal to the upper limit of each class.

1.1.3 Conclusion: Synopsis of the Tabular and Graphical Procedures for Univariate Exploratory Analysis



In practice, numerous softwares propose methods of visualization for data. In the course, you have

1.2 Descriptive Parameters

1.2.1 Measures of location

1.2.1.1 Mean

- The mean of a data set is the arithmetic average of all the data values.
- If the data constitute a sample with size n, the mean is a “statistic” and is denoted as:

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

if the data constitutes the whole population of size N, the mean is a “parameter” and is denoted as:

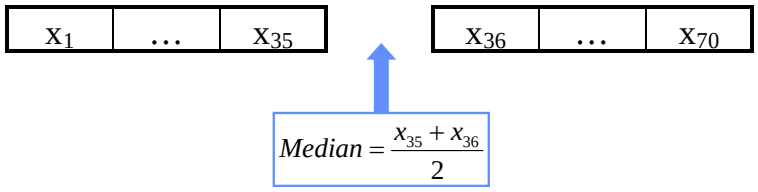
$$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$$

1.2.1.2 Median

- A few extremely large (small) data values can inflate (deflate) the mean as it is sensitive to the presence extreme values. In such a case, the median is more "robust".
- The **median** of a data set is the middle number when **the data items are arranged in ascending order**: $x_1 \leq x_2 \leq .... \leq x_n$, the median splits your ordered data set into two equal parts (two halves)
  - ► For an *odd* number of observations, the median is the middle value.
  - ► For an *even* number of observations, the median is the average of the two middle values.

Application to the "Apartment Rents":

- In the "Apartment Rents" example there is an even number of items (70).
- Median = 475.



1.2.1.3 Mode

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two (bimodal data distribution) or more (multimodal data distribution) different values.
  - ► The mode is the only measure of location for a nominal categorical variable.
  - ► For an ordinal categorical variable, the median may be computed by referring to the frequency distribution.

Application to the "Apartment Rents": 450 is the Mode as it occurs most frequently (7 times).

1.2.1.4 Percentiles and Quartiles

- A percentile (or, more generally, a quantile) provides information about how the data are spread over the interval from the smallest value to the largest

value.

- The p**th percentile** of a data set is a value such that at least p percent of the items take on this value or less and at least $(100 - p)$ percent of the items take on this value or more.
  - ► Arrange the data in ascending order.
  - ► Compute index i, the position of the pth percentile:

$$i = (p/100)n$$

  - ► If i is not an integer, round up. The p-th percentile is the value in the ith position.
  - ► If i is an integer, the p-th percentile is the average of the values in positions i and $i + 1$.

Application to the "Apartment Rents":

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

1. What is the 90th Percentile?

$$i = (p/100)n = (90/100)70 = 63$$

Averaging the 63rd and 64th data values:

$$90th\ Percentile = (580 + 590)/2 = 585$$

2. What is the median? (Note that : 50th percentile = Median)

$$Position\ of\ the\ Median = (percentile/100)n = (50/100)70 = 35$$

Averaging the 35th and 36th data values:

$$Median = (475 + 475)/2 = 475$$

- Quartiles
  - ► Quartiles are specific percentiles
  - ► First Quartile = 25th Percentile
  - ► Second Quartile = 50th Percentile = Median
  - ► Third Quartile = 75th Percentile

Application to the "Apartment Rents":

1. What is the 3rd quartile? Third quartile = 75th percentile

$$i = (p/100)n = (75/100)70 = 52.5 = 53$$

$$Third\ quartile = 525$$

Five-Number Summary for the Apartment Rents example:

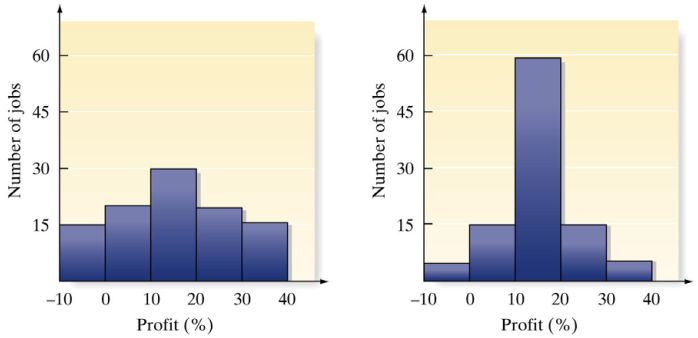
$$\begin{aligned} Lowest\ Value &= 425 \quad First\ Quartile = 445 \\ Median &= 475 \\ Third\ Quartile &= 525 \quad Largest\ Value = 615 \end{aligned}$$

1.2.1.5 Considerations on Measures of Location

- The choice of the most suitable measure of location depends on the type and features of the distribution;
- The "best" measure of location does not exist. It is extremely important to evaluate the differences between measures as they provide a further insight on the distribution (e.g. the shape);
- Some useful guidelines in real practice:
  - ► The mode is useful when you need to "minimize the displeased ones", i.e. in those situations where the agreement/consensus and the number of single units mean a lot for the decisions. Briefly speaking, the mode is a tool of governance;
  - ► The median minimizes the overall costs and is robust to extreme values. Therefore, the median is a useful tool for those decisions implying high costs in extreme situations;
  - ► The mean is the data barycenter and proposes a value that equally distributes the total amount among the statistical units. As a consequence, it provides decisions where, given in the same number, the extremes have a higher weight than the central values. Therefore, the mean is a measure of general equilibrium.

1.2.2 Measures of Variability

- Most often, measures of location are not enough to summarize a distribution thus providing only a partial description of a quantitative data set.
- For example, suppose we are comparing the profit margin per construction job (as a percentage of the total bid price) for 100 constructions jobs for each of two cost estimators working for a large construction company. The histograms for the two sets are shown below:



a. Cost estimator A b. Cost estimator B

- Both data sets are symmetric with equal means, medians and modes.
- The frequency distributions are still different: cost estimator A has profit margins spread with almost equal relative frequency over the measurement classes (higher incidence of high profit margins with higher risk), while cost estimator B has profit margins clustered about the center of the distribution (more stable).

1.2.2.1 Range and Interquartile Range

- The *range* of a data set is the **difference between the largest and smallest data values** (measurements).
  - ► It is the *simplest measure* of variability.
  - ► It loses *sensitivity* when data sets are large.
  - ► It *ignores how the values distribute* between the two extreme values, i.e. it does not tell all about variability.

For the "Apartment Rents" example:

Range = largest value - smallest value

Range = 615 - 425 = 190

- Interquartile Range (IQR): the *interquartile range* of a data set is the difference between the third quartile and the first quartile.
  - ► It is the range for the middle 50% of the data.
  - ► It overcomes the sensitivity to extreme data values.

For the "Apartment Rents" example:

3rd Quartile (Q3)= 525 1st Quartile (Q1) = 445

IQR = Q3 – Q1 = 525 – 445 = 80

1.2.2.2 Variance and standard deviation

Variance

- The variance is a measure of variability that utilizes all the data.
- It measures the spread of data around the mean as it is based on the difference between the value of each observation and the mean.
  - ► The variance is the **average of the squared differences** between each data value and the mean.
  - ► If the data set is a sample, the variance is denoted by s^2 (or s_x^2 to denote the variable it refers to)

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

Remark: If the data set consists of the whole population, the variance is denoted by σ^2:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}$$

- Why squared differences?
  - ► The sum of the deviations from the mean is always 0.
  - ► Data values further away from the mean contribute more to variability.
  - ► Interpretation problem: its measurement units are the square of the observed units.

For the "Apartment Rents" example:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = 2,996.16$$

Standard deviation

- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily comparable, than the variance, to the mean.
- If the data set is a sample, the standard deviation is denoted s

$$s = \sqrt{s^2}$$

- If the data set is a population, the standard deviation is denoted σ (sigma).

$$\sigma = \sqrt{\sigma^2}$$

For the "Apartment Rents" example:

$$s = \sqrt{s^2} = \sqrt{2996.16} = 54.74$$

1.2.2.3 Coefficient of Variation (CV)

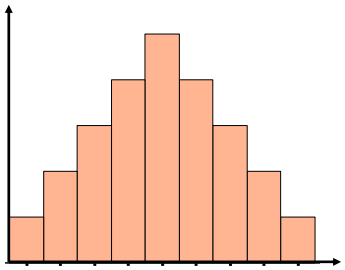
- The coefficient of variation indicates how large the standard deviation is in relation to the mean (in absolute value)
  ► it is a **risk coefficient**.
- It allows to compare the variability between two distributions even when they have different means.
- If the data set is a sample, the coefficient of variation is computed as follows:

$$\mathsf{CV} = \frac{s}{|\overline{x}|} \times 100$$

For the "Apartment Rents" example:

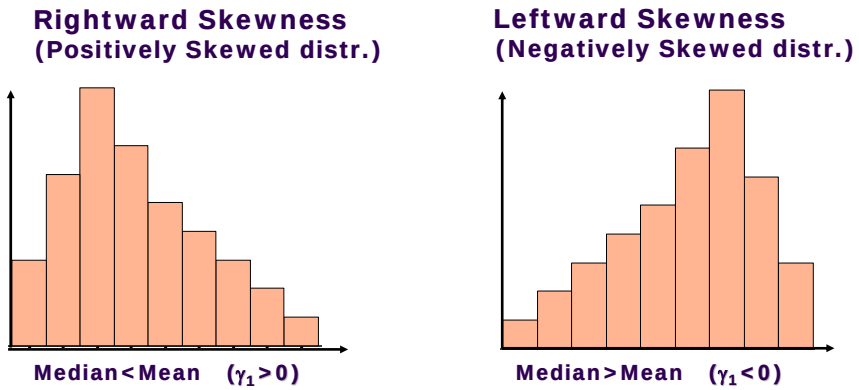
$$\frac{s}{|\overline{x}|} \times 100 = \frac{54.74}{490.80} \times 100 = 11.15$$

1.2.3 Shape of a Distribution



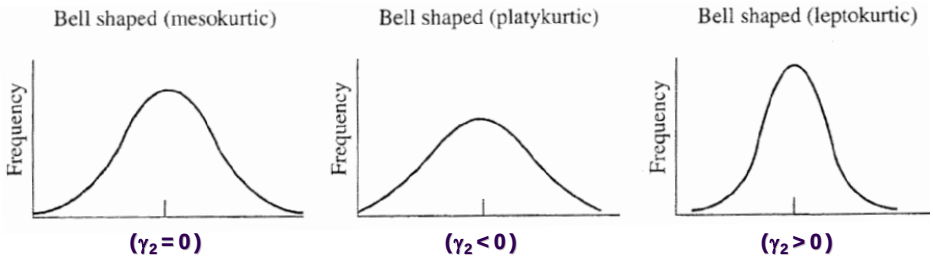
- In case of a unimodal distribution, a comparison between the mean and the median provides information on the shape of the distribution.
- (Mean = Median) $\rightarrow$ symmetric distribution ($\gamma_1 = 0$), where the **Fisher Asymmetry index** is defined by

$$\gamma_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^3}{\sigma^3}$$



- **Kurtosis:** The kurtosis is a measure of the "peakedness" of the distribution having the Normal distribution as a reference. Higher kurtosis means less frequent extreme deviations from the mean (centre of the distribution) as

opposed to more frequent modestly-sized deviations.



The Fisher Kurtosis index is defined as

$$\gamma_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^4}{\sigma^4} - 3$$

Kurtosis is the property of being peaked (γ_2>0, leptokurtic), flat (γ_2<0, platykurtic), or somewhere in between (e.g. γ_2=0, mesokurtic, in case of the Normal "bell-shaped" distribution).

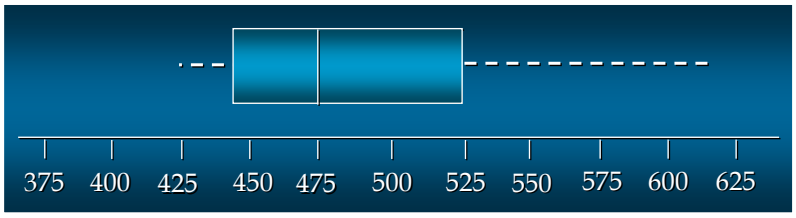
1.2.4 Box Plot

- A box is drawn with its ends located at the first and third quartiles.
- A vertical line is drawn in the box at the location of the median (another vertical line might be drawn at the location of the mean).
- Limits are located (not drawn) using the interquartile range (IQR):
  - ► The lower limit is located at 1.5(IQR) below Q1.
  - ► The upper limit is located at 1.5(IQR) above Q3.
  - ► Data outside these limits are considered as anomalous (unexpected) observations: **outliers**.
- Whiskers (dashed lines) are drawn from the ends of the box to the smallest and largest data values inside the limits.
- The locations of each outlier is shown with the symbol $*$

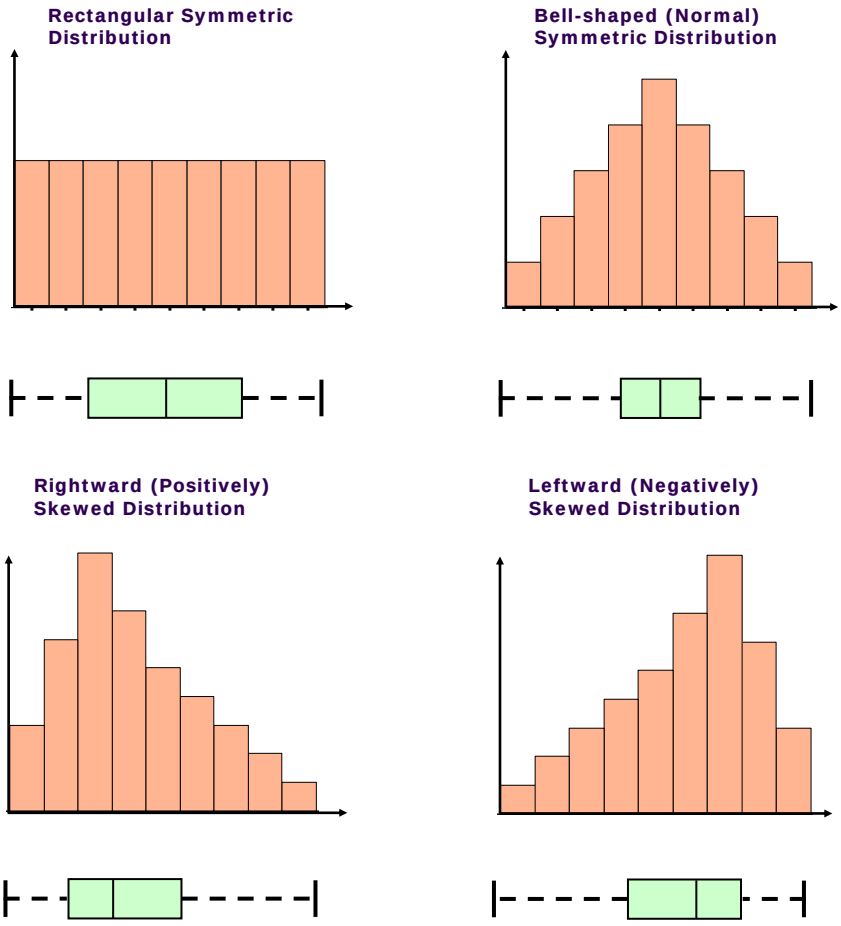
Application to the "Apartment Rents" example:

Simple Box Plot
Lower Limit: Q1 – 1.5(IQR) = 445 – 1.5(80) = 325
Upper Limit: Q3 + 1.5(IQR) = 525 + 1.5(80) = 645

There are no outliers.



1.2.4.1 Box Plot & Histogram



Chapter 2
Bivariate Exploratory Analysis

2.1 Descriptive bivariate statistics

- So far we have focused on methods that are used to summarize the data for *one variable at a time*.
- Often, in real practice, you are interested in tabular and graphical methods that will help understand *the relationship between two variables*.
- **Crosstabulation** and a **scatter diagram** are two methods (tabular and graphical, respectively) for summarizing the data for two (or more) variables simultaneously.
- **Association indices** are numerical methods for measuring the strength of relationship taking into account the type of data and the nature of relationships.

2.2 Crosstabulation

Crosstabulation is a tabular method for summarizing the data for two variables simultaneously. It can be used when:

- One variable is qualitative and the other is quantitative (mixed table)
- Both variables are qualitative (contingency table)
- Both variables are quantitative (correlation table)

The left and top margin labels define the classes for the quantitative variables or the categories of categorical variables.

2.2.1 An example of correlation table (labels in 1000€):

The table contains information about

- A variable X, by row, with k categories (classes);
- A variable Y, by column, with h categories (classes);

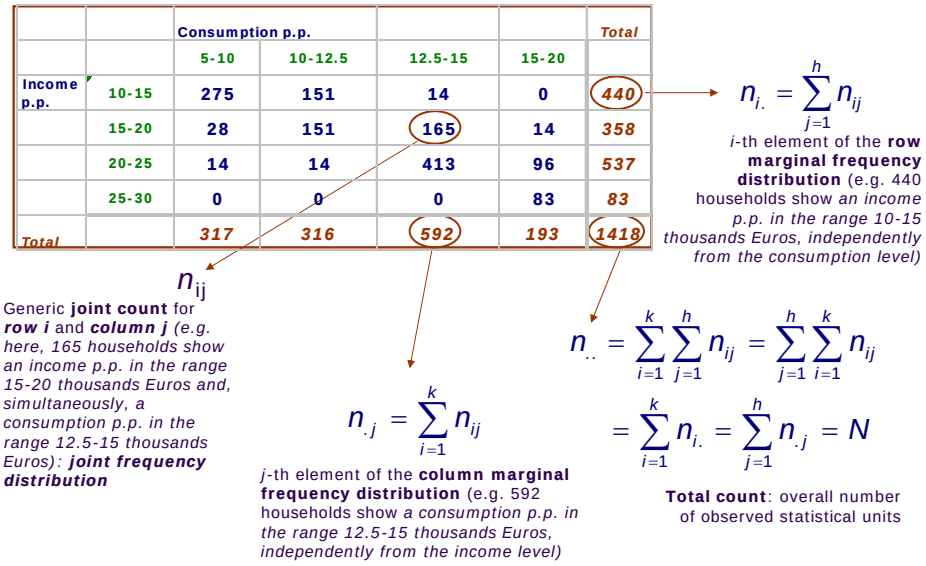
where two types of distributions are represented:

- **Marginals**
  - ► A marginal distribution for X;
  - ► A marginal distribution for Y;

- **Conditionals**
  - ► k distributions of Y conditioned by the categories of X;
  - ► h distributions of X conditioned by the categories of Y;
- This together give us the **joint distribution** of X and Y.

		Consumption p.p.				Total
		5-10	10-12.5	12.5-15	15-20	
Income p.p.	10-15	275	151	14	0	440
	15-20	28	151	165	14	358
	20-25	14	14	413	96	537
	25-30	0	0	0	83	83
Total		317	316	592	193	1418

Notation and Interpretation



2.2.2 Computing Means in Crosstabs

2.2.2.1 Mean of the row variable

		Y				
		40-50	50-65	65-75	75-95	Tot
X	150-160	57	52	4	0	113
	160-170	53	147	24	1	225
	170-175	5	138	61	6	210
	175-180	0	46	116	23	185
	180-200	0	0	15	52	67
	Tot	115	383	220	82	800

$$M(X) = \frac{1}{N}\sum_{i=1}^{k} x_i \cdot n_{i.} = \frac{155.0 \times 113 + 165.0 \times 225 + 172.5 \times 210 + 177.5 \times 185 + 190.0 \times 67}{800}$$

$$= \frac{136433}{800} = 170.54$$

This is a **weighted mean** of the X variable where the central value (x_i) of each class is weighted by the corresponding **row marginal frequency** ($n_{i.}$).

2.2.2.2 Mean of the column variable

Example: Consumption *(Y)* and Income *(X)* of 800 people in hundreds €

		Y				
		40-50	50-65	65-75	75-95	Tot
X	150-160	57	52	4	0	113
	160-170	53	147	24	1	225
	170-175	5	138	61	6	210
	175-180	0	46	116	23	185
	180-200	0	0	15	52	67
	Tot	115	383	220	82	800

$$M(Y) = \frac{1}{N}\sum_{j=1}^{h} y_j \cdot n_{.j} = \frac{45.0 \times 115 + 57.5 \times 383 + 70.0 \times 220 + 85.0 \times 82}{800}$$

$$= \frac{49568}{800} = 61.96$$

This is a **weighted mean** of the Y variable where the central value (y_j) of each class is weighted by the corresponding **column marginal frequency** ($n_{.j}$)

2.2.2.3 Conditional Means $Y|x_i$

Example: Consumption *(Y)* and Income *(X)* of 800 people in hundreds €

		Y				
		40-50	50-65	65-75	75-95	Tot
X	150-160	57	52	4	0	113
	160-170	53	147	24	1	225
	170-175	5	138	61	6	210
	175-180	0	46	116	23	185
	180-200	0	0	15	52	67
	Tot	115	383	220	82	800

$$M(Y|X=x_1) = \frac{1}{n_{1.}} \sum_{j=1}^{h} y_j \cdot n_{1j} = \frac{45.0 \times 57 + 57.5 \times 52 + 70.0 \times 4 + 85.0 \times 0}{113}$$

$$= \frac{5835}{113} = 51.64$$

2.2.2.4 Conditional Means $X|y_j$

Example: Consumption *(Y)* and Income *(X)* of 800 people in hundreds €

		Y				
		40-50	50-65	65-75	75-95	Tot
X	150-160	57	52	4	0	113
	160-170	53	147	24	1	225
	170-175	5	138	61	6	210
	175-180	0	46	116	23	185
	180-200	0	0	15	52	67
	Tot	115	383	220	82	800

$$M(X|Y=y_2) = \frac{1}{n_{.2}} \sum_{i=1}^{k} x_i \cdot n_{i2} = \frac{155.0 \times 52 + 165.0 \times 147 + 172.5 \times 138 + 177.5 \times 46 + 190.0 \times 0}{383}$$

$$= \frac{64285}{383} = 167.85$$

2.3 Measuring the Association in a Contingency Table

2.3.1 Distributional Independence

Observed Values and row% (conditional distributions)

			Performance			Total
			Low	Medium	High	
Fund	A	Freq.	13	33	38	84
		%	15.5%	39.3%	45.2%	100.0%
	B	Freq.	38	102	40	180
		%	21.1%	56.7%	22.2%	100.0%
	C	Freq.	90	45	20	155
		%	58.1%	29.0%	12.9%	100.0%
Total		Freq.	141	180	98	419
		%	33.7%	43.0%	23.4%	100.0%

- **Distributional (Absolute) Independence condition:**

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{i'j}}{n_{i'.}} = \frac{n_{.j}}{n_{..}}$$

  ► **Example:** the row percentages in the first cell of each row shall all be equal to 33.7% while they are 15.5%, 21.1% and 58.1%, respectively.

- **Theoretical frequencies can be computed under the assumption of independence** (in case of independence)

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n_{..}} \Rightarrow \widehat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

  where $n = n_{..}$

  ► **Example:** the count in the first cell of the first row is 13 while it shall be $(84 \times 141)/419 = 28.3$

- Remark: We could work just as well on the columns as we have done on the rows.

2.3.2 Computation of the χ^2 index

Observed Values, row % and theoretical frequencies

			Performance			Total
			Low	Medium	High	
Fund	A	Obs. Freq.	13	33	38	84
		Theoret. Freq.	28.3	36.1	19.6	84.0
		%	15.5%	39.3%	45.2%	100.0%
	B	Obs. Freq.	38	102	40	180
		Theoret. Freq.	60.6	77.3	42.1	180.0
		%	21.1%	56.7%	22.2%	100.0%
	C	Obs. Freq.	90	45	20	155
		Theoret. Freq.	52.2	66.6	36.3	155.0
		%	58.1%	29.0%	12.9%	100.0%
Total		Obs. Freq.	141	180	98	419
		Theoret. Freq.	141.0	180.0	98.0	419.0
		%	33.7%	43.0%	23.4%	100.0%

The Chi-square index χ^2 (khi-deux in French)

$$\begin{aligned}\chi^2 &= \sum_i \sum_j \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} \\ &= \frac{(13 - 28.3)^2}{28.3} + \frac{(33 - 36.1)^2}{36.1} + ... + \frac{(20 - 36.3)^2}{36.1} \\ &= 83.780\end{aligned}$$

- ► $\chi^2 = 0$ in case of in case of absolute independence
- ► $\chi^2 > 0$ in case of association (the stronger the association, the higher the index)

Remark 2 *A more precise interpretation will be given in the chapter on Hypothesis Testing*

2.4 Measuring the Linear Dependence in a Correlation Table

2.4.1 Two quantitative variables

The distribution of Income per person and Consumption per person in the 103 Italian districts is presented below.

			Consumption p.p. (in 1000€)				Total
			5-10	10-12.5	12.5-15	15-20	
Icnome p.p. (in 1000€)	9-15	freq.	20	11	1		32
		%	62.5%	34.4%	3.1%		100.0%
	15-20	freq.	2	11	12	1	26
		%	7.7%	42.3%	46.2%	3.8%	100.0%
	20-25	freq.	1	1	30	7	39
		%	2.6%	2.6%	76.9%	17.9%	100.0%
	25-30	freq.				6	6
		%				100.0%	100.0%
Total		freq.	23	23	43	14	103
		%	22.3%	22.3%	41.7%	13.6%	100.0%

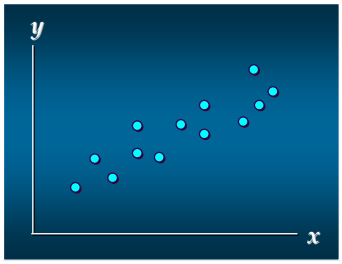
If the χ^2 index is computed to study the association between Income and Consumption, only the frequency distributions are taken into account. Therefore, the information provided by the numerical nature of the variables is lost. In case of numerical variables (correlation table) a graphical representation and a measure of the linear association between the two variables may be considered.

2.4.2 Scatter Diagram

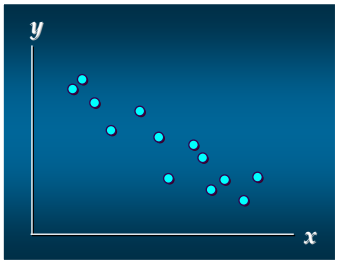
- A scatter diagram (or scatterplot) is a graphical presentation of the relationship between two quantitative variables.
- One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.
- Each plotted point relates to an observed unit.
- The general pattern of the plotted points suggests the overall relationship between the variables.

Types of relationships:

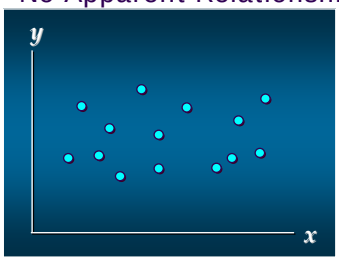
■ A Positive Relationship



■ A Negative Relationship



■ No Apparent Relationship



2.4.3 Covariance

- The covariance is a measure of the linear association between two variables.
  - ► Positive values indicate a positive (concordance) relationship.
  - ► Negative values indicate a negative (discordance) relationship.

- Computation
  - ► If the data set is a sample, the covariance between X and Y is denoted by:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

  - ► If the data set is the whole population (of N individuals), the covariance

is denoted by:

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \mu_x) \left(y_i - \mu_y\right)}{N}$$

where μ denotes the expectation (the mean over the whole population)

Remark: We work in general with a subsample of the whole population so we use s_{xy} which is an estimate of the population parameter σ_{xy}. The reason we divide by $n - 1$ and not by n will be clear when we study the properties of estimators. It is due to the fact that when we do not observe the whole population, we need to compute first the means which are themselves estimates of the expectations. It is said that we thus lose "one degree of freedom".

- The interpretation of covariance is difficult.
  - ► Its measurement unit is the product of the measurement units of the two variables.
  - ► Its absolute value can not be interpreted because its maximum is not known.
- This problem can be solved by normalizing the index so that its possible values are comprised within a known interval.

Remark: The conventional notation is to denote σ_{xy} (or s_{xy}) the covariance but σ_x^2 (s_x^2) the variance, this is because

$$\sigma_{xx} = \sigma_x^2$$

2.4.4 Linear Correlation Coefficient

- The simple linear correlation coefficient can take on values between -1 and +1.
  - ► Values near -1 indicate a strong negative (descending) linear relationship.
  - ► Values near +1 indicate a strong positive (ascending) linear relationship.
- If the data set is a sample, the linear correlation coefficient between X and Y is the covariance scaled by the product of respected standard deviations of X and Y :

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

- If the available data set is the whole population, the linear correlation

coefficient is:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Remark: The Cauchy-Schwarz inequality implies that the correlation is less than unity in absolute value:

$$\textit{Cauchy-Schwarz:} \quad |\sigma_{xy}| \leq \sigma_x \sigma_y$$

hence

$$|\rho_{xy}| = \left| \frac{\sigma_{xy}}{\sigma_x \sigma_y} \right| = \frac{|\sigma_{xy}|}{\sigma_x \sigma_y} \leq 1$$

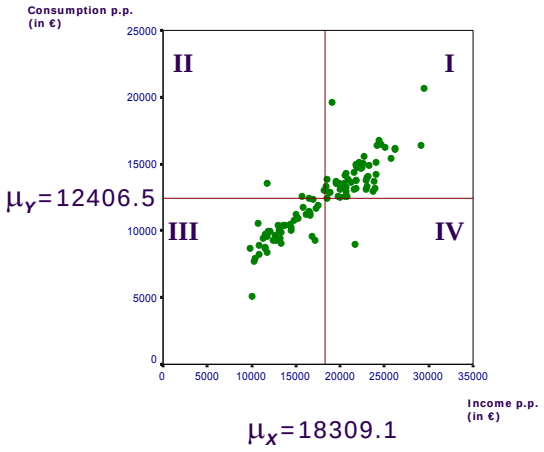
i.e.

$$-1 \leq \rho_{xy} \leq 1$$

2.4.4.1 Application to the "Income $\times$ Consumption" example

- Data Description: The distribution of Income per person and Consumption per person in the 103 Italian districts (i.e. the whole population is available: $N = 103$)

	Income p.p.	Consump. p.p
AG	9864.3	8676.5
AL	20555.0	14150.9
AN	22310.9	14667.4
AO	24583.4	16475.0
AP	18179.3	13014.7
AQ	16165.1	11258.8
AR	19780.3	12601.6
AT	18540.8	12395.0
AV	13169.7	9451.2
BA	14770.7	10794.0
BG	20606.6	12963.1
BI	23963.6	13169.7
BL	20606.6	14305.9
BN	12653.2	9657.7
BO	29489.7	20658.3
BR	11775.2	13531.2
BS	21639.5	8986.4
BZ	24170.2	16371.7
:	:	:



- Computation of the covariance and the correlation

	Income p.p.	Consumption p.p.	X-μ_X	Y-μ_Y	(X-μ_X)(Y-μ_Y)
AG	9864.3	8676.5	-8444.8	-3730.0	31499375.2
AL	20555.0	14150.9	2245.8	1744.4	3917682.2
AN	22310.9	14667.4	4001.8	2260.9	9047555.7
AO	24583.4	16475.0	6274.20	4068.48	25526434.96
AP	18179.3	13014.7	-129.9	608.2	-78986.8
AQ	16165.1	11258.8	-2144.0	-1147.7	2460805.0
AR	19780.3	12601.6	1471.2	195.1	286948.1
AT	18540.8	12395.0	231.7	-11.5	-2671.6
AV	13169.7	9451.2	-5139.5	-2955.3	15188948.7
BA	14770.7	10794.0	-3538.5	-1612.5	5705975.1
BG	20606.6	12963.1	2297.5	556.6	1278709.2
BI	23963.6	13169.7	5654.5	763.2	4315210.1
BL	20606.6	14305.9	2297.5	1899.4	4363739.1
BN	12653.2	9657.7	-5656.0	-2748.8	15546829.9
BO	29489.7	20658.3	11180.5	8251.8	92259330.2
BR	11775.2	13531.2	-6533.9	1124.7	-7348534.5
BS	21639.5	8986.4	3330.4	-3420.1	-11390443.6
BZ	24170.2	16371.7	5861.0	3965.2	23240086.3
:	:	:			

The means and standard deviations are

$$\mu_x = 18309.1; \quad \mu_y = 12406.5$$
$$\sigma_x = 4805.30; \quad \sigma_y = 2618.95$$

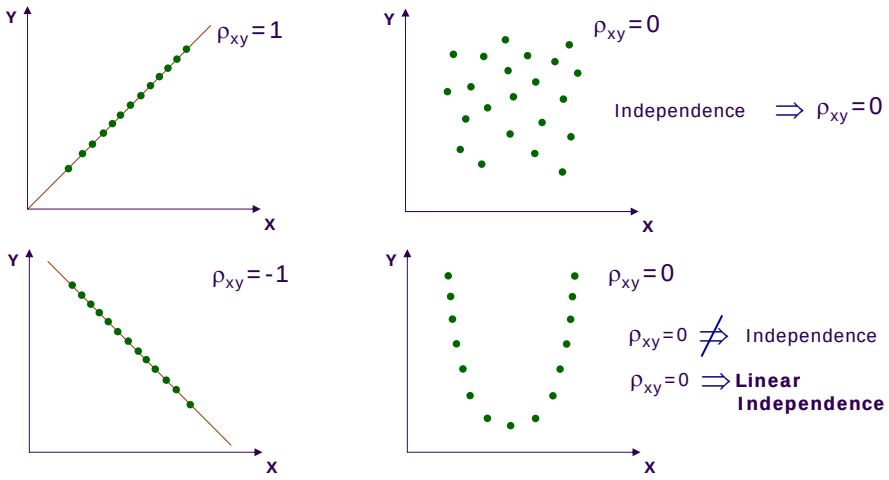
The covariance is

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \mu_x)\left(y_i - \mu_y\right)}{N} = 10957749.08$$

and the linear correlation coefficient is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{10957749.08}{4805.30 \times 2618.95}$$
$$= 0.87$$

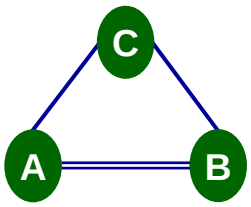
2.4.5 Interpreting Correlation



2.4.6 Spurious Correlation

Spurious correlation may exist between variables that are not logically related to each other.

Example.



- The price of a vegetable sold at the market in London and the water height of the Thames river.
- The number of French tourists landing in Rome and the number of ice-creams sold in the Italian capital.

Partial correlation coefficient:

$$r_{AB.C} = \frac{r_{AB} - r_{AC} \cdot r_{BC}}{\sqrt{\left(1 - r_{AC}^2\right) \cdot \left(1 - r_{BC}^2\right)}}$$

Chapter 3
Elements of Business Analytics

This chapter presents an introduction to some techniques used in the context of Business Analytics.

3.1 Cluster Analysis

Cluster Analysis is a standard method in Business Analytics to understand the client base and perform a market segmentation. Clustering refers to the practice of gathering a set of objects and separating them into groups of similar objects. By exploring these different groups, determining how similar and how different they are, one can get a lot of information about a set of data.

Clustering is used in general when there is so much data that it is not possible for an individual to group the individuals by simply looking at the data, so there is a need for a systematic methodology. Revealing relationships in large datasets is useful across industries, e.g. recommending films based on the habits of clients (think of how Netflix gained so much market share in the U.S.), identifying crime hotspots in urban area (as in NYC in the 1990s) or grouping return-related financial investments to ensure diversification in a portfolio.

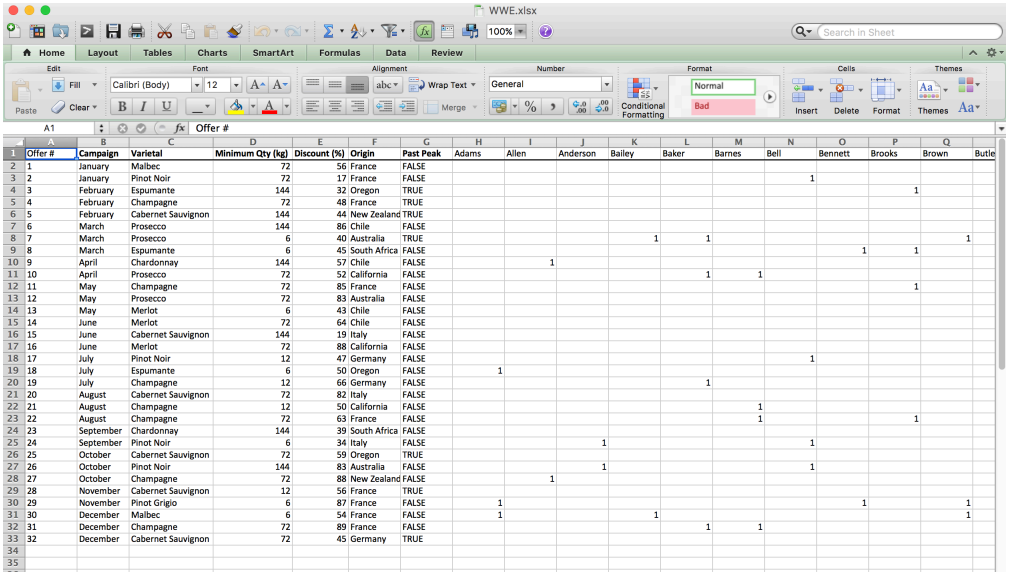
We now explore the most common method of clustering called k*-means clustering*, this is not the most fancy method but it is easy to understand and to implement. In the following, we will introduce the methodology via an example.

Example 4 *In the running example of this chapter, we are going to consider the company Wholesale Wine Emporium (WWE) who imports wine bought wholesale to the U.S. and then sells it to individual retailers. Retailers are told of each offer via email and they decide about ordering it. The business works well, but WWE thinks that by understand the tastes and needs of its client base, it can serve it better. In particular, if one could segment the clients according to their interests, one could customize the newsletters better and improve the sales.*
The dataset WWE is available on the course website and contains to spreadsheets:

- *Spreadsheet #1* $\mathrm{Offer\ information}$ *lists all the offers made over the last year to the clients. It details the date (month), the wine variety, the minim quantity sold, at which discount, the origin of the wine and finally an information about whether the wine has reached its quality peak.*

- *Spreadsheet #2* Transactions *lists all the offers taken by customers.*

The first task is to consolidate *the tables into a unique matrix. For this we want to add to each offer in Spreasheet #1 the names of the customers who took up this offer. We use the function* Pivot *in Excel, with offers as rows and customers as columns, based on the* Transactions *sheet. Then we copy/paste* Pivot into Transactions *and create a new sheet called* Matrix.



Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown	Butle
1	January	Malbec	72	56	France	FALSE											
2	January	Pinot Noir	72	17	France	FALSE							1				
3	February	Espumante	144	32	Oregon	TRUE									1		
4	February	Champagne	72	48	France	TRUE											
5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE											
6	March	Prosecco	144	86	Chile	FALSE											
7	March	Prosecco	6	40	Australia	TRUE				1	1						1
8	March	Espumante	6	45	South Africa	FALSE								1	1		
9	April	Chardonnay	144	57	Chile	FALSE		1									
10	April	Prosecco	72	52	California	FALSE					1	1					
11	May	Champagne	72	85	France	FALSE									1		
12	May	Prosecco	72	83	Australia	FALSE											
13	May	Merlot	6	43	Chile	FALSE											
14	June	Merlot	72	64	Chile	FALSE											
15	June	Cabernet Sauvignon	144	19	Italy	FALSE											
16	June	Merlot	72	88	California	FALSE											
17	July	Pinot Noir	12	47	Germany	FALSE							1				
18	July	Espumante	6	50	Oregon	FALSE	1										
19	July	Champagne	12	66	Germany	FALSE					1						
20	August	Cabernet Sauvignon	72	82	Italy	FALSE											
21	August	Champagne	12	50	California	FALSE						1					
22	August	Champagne	72	63	France	FALSE						1			1		
23	September	Chardonnay	144	39	South Africa	FALSE											
24	September	Pinot Noir	6	34	Italy	FALSE			1				1				
25	October	Cabernet Sauvignon	72	59	Oregon	TRUE											
26	October	Pinot Noir	144	83	Australia	FALSE			1				1				
27	October	Champagne	72	88	New Zealand	FALSE		1									
28	November	Cabernet Sauvignon	12	56	France	TRUE											
29	November	Pinot Grigio	6	87	France	FALSE	1							1		1	
30	December	Malbec	6	54	France	FALSE	1			1						1	
31	December	Champagne	72	89	France	FALSE					1	1					
32	December	Cabernet Sauvignon	72	45	Germany	TRUE											

Standardizing the data: In the running example of this chapter, the data we work with are binary (either the customer buys or she doesn't) but in general this is not the case. Hence for comparability, the data are standardized, i.e. the variables are demeaned and scaled by the standard deviation as in the following transform:

$$x_i \longrightarrow \frac{x_i - \overline{x}}{s_x}$$

The next task is to choose the number of clusters k. The principle is that the number should be small, so that we really group the customers into a few identifiable types. We will revisit this issue later, but for now, we start with $k = 4$ clusters.

3.1.1 A metric for the distance

The main input into the algoritm for k-mean clustering is the choice of distance. The idea of k-mean clustering is that we are going to choose some cluster centers as centers of gravity of a small group of individuals. The center of gravity is

defined such that it is *central*, i.e. to minimize the total distance between the center of gravity and the individuals belonging to the cluster.

Hence, we want to find the coordinates (the position) of each center of gravity (each *mean* among the k means). The position is defined in terms of the rows in the matrix (or the columns). But are we more interested in rows or in columns?

Example 5 *In the example above, the rows are the offers and the columns are the characteristics of these offers (who bought them, in particular).*

- *If we consider the position of the clusters in terms of offers/transactions (*rows*), it means we try to group the customers in terms of their interests.*
- *If we consider the position of the clusters in terms of customers (*$\mathrm{columns}$*), it means we try to group the offers by similarity in terms of the customers who bought them.*

In the following, we are going to group the customers in terms of their interests, so next time we have an offer, we can target customers who are more likely to be interested. Clustering by columns will not serve the same purpose here (although related).

How to measure the distance between a customer and a cluster? We have to remember that in the dataset, the only information we have about customers is the list of offers they accepted, hence if there are P customers and N offers, each customer p can be represented by a vector

$$x_p = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} \leftarrow \text{accepted offer 1} \\ \leftarrow \text{refused offer 2} \\ \leftarrow \text{refused offer 3} \\ \vdots \\ \leftarrow \text{accepted offer } N \end{matrix} = \begin{bmatrix} x_{1,p} \\ x_{2,p} \\ x_{3,p} \\ \vdots \\ x_{N,p} \end{bmatrix}$$

Now assume that the center of gravity of the cluster c, for $c = 1, ..., k$ is

$$\mu_c = \begin{bmatrix} \mu_{1,c} \\ \mu_{2,c} \\ \mu_{3,c} \\ \vdots \\ \mu_{N,c} \end{bmatrix}$$

Then the distance between customer p and cluster c is the distance between x_p and μ_c in terms of Euclidian distance, i.e.

$$d\left(x_p, \mu_c\right) = \sqrt{\sum_{i=1}^{N} \left(x_{i,p} - \mu_{i,c}\right)^2} = \|x_p - \mu_c\|$$

and hence customer p is assigned to cluster c if the distance between x_p and μ_c is smaller than the distance from x_p to all other μ_{c^*}, for $c^* = 1, ..., k$; i.e.

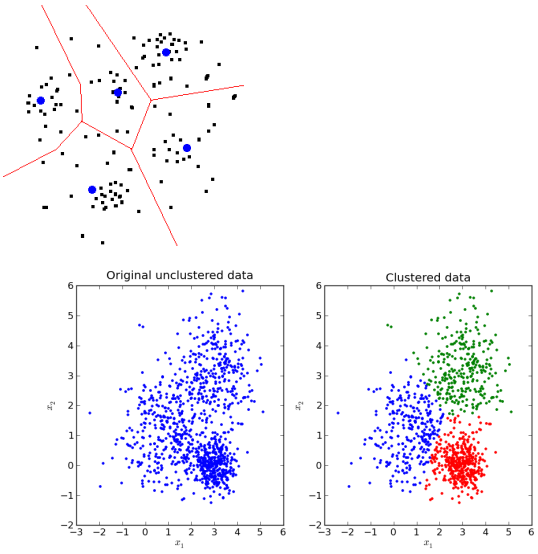
$$d\left(x_p, c\right) = \min_{c^*=1,...,k} d\left(x_p, \mu_{c^*}\right)$$

The solution to this optimization problem is highly nonlinear and difficult. Fortunately, it is available in XLStat which minimizes the total (squared) distance to clusters:

$$\left(\mu_1, ..., \mu_k\right) = \underset{\left(\mu_1^*,...,\mu_k^*\right)}{argmin} \sum_{p=1}^{P} d^2\left(x_p, c\right) = \underset{\left(\mu_1^*,...,\mu_k^*\right)}{argmin} \sum_{p=1}^{P} \min_{c^*=1,...,k} d^2\left(x_p, \mu_{c^*}^*\right)$$

Here the coordinates of the cluster centers must all be between 0 and 1 since the all the variables are binary and either zero and unity.

The graphs below present examples of clustering



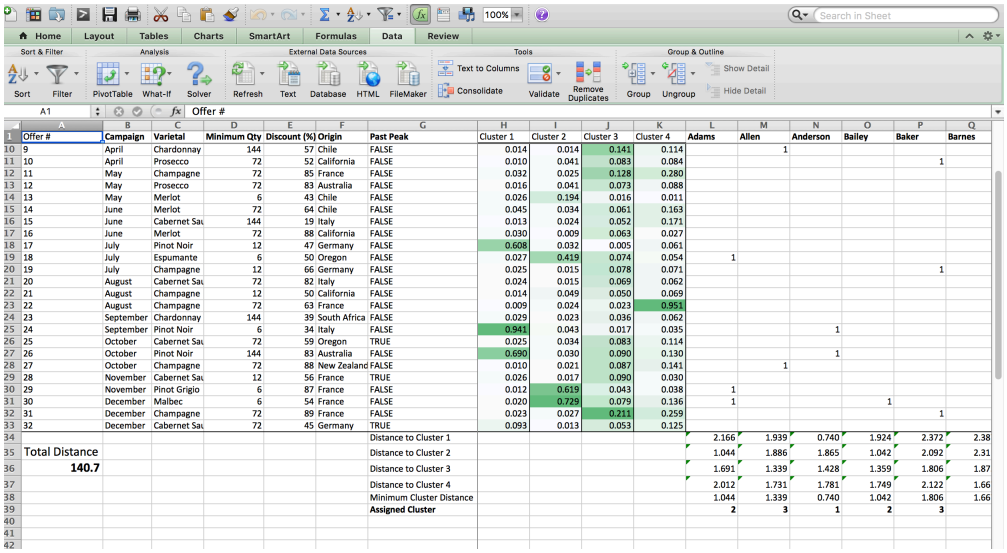
3.1.2 Making sense of the results

Once the algorithm provides the optimal cluster centers, the modeler still has to

make sense of them. This is the interpretation stage, which is the interesting part. Each cluster center is given *coordinates* in the space of the row variable (in the running example: the various offers). Notice that since algorithms are used when solving the optimization problem, it may happen that one does not always get the exact same results.

- The first way to analyze the output is to look, for each cluster, what are the highest coordinates and see to what row (*offer*) characteristics they correspond. Often, it is easy to find a pattern.

Example 6 *The Excel output provides examples of the coordinates of the cluster centers and the assignment of the customers to their closest cluster.*

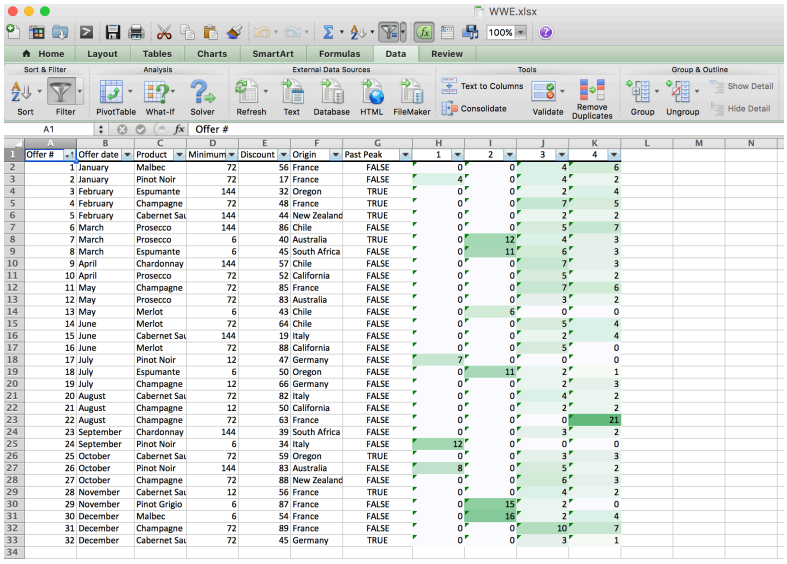


Offer #	Campaign	Varietal	Minimum Qty	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen	Anderson	Bailey	Baker	Barnes
9	April	Chardonnay	144	57	Chile	FALSE	0.014	0.014	0.141	0.114		1				
10	April	Prosecco	72	52	California	FALSE	0.010	0.041	0.083	0.084					1	
11	May	Champagne	72	85	France	FALSE	0.032	0.025	0.128	0.280						
12	May	Prosecco	72	83	Australia	FALSE	0.016	0.041	0.073	0.088						
13	May	Merlot	6	43	Chile	FALSE	0.026	0.194	0.016	0.011						
14	June	Merlot	72	64	Chile	FALSE	0.045	0.034	0.061	0.163						
15	June	Cabernet Sau	144	19	Italy	FALSE	0.013	0.024	0.052	0.171						
16	June	Merlot	72	88	California	FALSE	0.030	0.009	0.063	0.027						
17	July	Pinot Noir	12	47	Germany	FALSE	0.608	0.032	0.005	0.061						
18	July	Espumante	6	50	Oregon	FALSE	0.027	0.419	0.074	0.054	1					
19	July	Champagne	12	66	Germany	FALSE	0.025	0.015	0.078	0.071					1	
20	August	Cabernet Sau	72	82	Italy	FALSE	0.024	0.015	0.069	0.062						
21	August	Champagne	12	50	California	FALSE	0.014	0.049	0.050	0.069						
22	August	Champagne	72	63	France	FALSE	0.009	0.024	0.023	0.951						
23	September	Chardonnay	144	39	South Africa	FALSE	0.029	0.023	0.036	0.062						
24	September	Pinot Noir	6	34	Italy	FALSE	0.941	0.043	0.017	0.035			1			
25	October	Cabernet Sau	72	59	Oregon	TRUE	0.025	0.034	0.083	0.114						
26	October	Pinot Noir	144	83	Australia	FALSE	0.690	0.030	0.090	0.130			1			
27	October	Champagne	72	88	New Zealand	FALSE	0.010	0.021	0.087	0.141		1				
28	November	Cabernet Sau	12	56	France	TRUE	0.026	0.017	0.090	0.030						
29	November	Pinot Grigio	6	87	France	FALSE	0.012	0.619	0.043	0.038	1					
30	December	Malbec	6	54	France	FALSE	0.020	0.729	0.079	0.136	1			1		
31	December	Champagne	72	89	France	FALSE	0.023	0.027	0.211	0.259					1	
32	December	Cabernet Sau	72	45	Germany	TRUE	0.093	0.013	0.053	0.125						
						Distance to Cluster 1					2.166	1.939	0.740	1.924	2.372	2.38
Total Distance						Distance to Cluster 2					1.044	1.886	1.865	1.042	2.092	2.31
140.7						Distance to Cluster 3					1.691	1.339	1.428	1.359	1.806	1.87
						Distance to Cluster 4					2.012	1.731	1.781	1.749	2.122	1.66
						Minimum Cluster Distance					1.044	1.339	0.740	1.042	1.806	1.66
						Assigned Cluster					**2**	**3**	**1**	**2**	**3**	

The idea then is to look at the coordinates of the cluster center and see if the highest coordinates correspond to some interesting characteristics of the corresponding offer.
In this example, we notice that for Cluster 1, the highest coordinates all correspond to Pinot Noir*, so this seems to be a defining characteristic. For Cluster 2, we notice that all the mean has large coordinates for offers that include small quantities, so in this cluster, customers clearly do not need to buy in bulk to accept an offer.*
Yet, overall, it is not so easy to determine the patterns for all clusters.

- When the coordinates do not suffice to interpret the cluster, one can also look at which individuals are assigned to each cluster. When we do not possess anyadditional information about these individuals, we need to resort to the characteristics of the rows. For instance, we can count how many individuals from each cluster are concerned with each row.

Example 7 *In the example above, we can compute the top deals for each cluster. This means looking for each offer how many members of each cluster accepted it. Then we can look at the offers that were accepted by most members in each cluster. The following graph shows the results.*



Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4
1	January	Malbec	72	56	France	FALSE	0	0	4	6
2	January	Pinot Noir	72	17	France	FALSE	4	0	4	2
3	February	Espumante	144	32	Oregon	TRUE	0	0	2	4
4	February	Champagne	72	48	France	TRUE	0	0	7	5
5	February	Cabernet Sau	144	44	New Zealand	TRUE	0	0	2	2
6	March	Prosecco	144	86	Chile	FALSE	0	0	5	7
7	March	Prosecco	6	40	Australia	TRUE	0	12	4	3
8	March	Espumante	6	45	South Africa	FALSE	0	11	6	3
9	April	Chardonnay	144	57	Chile	FALSE	0	0	7	3
10	April	Prosecco	72	52	California	FALSE	0	0	5	2
11	May	Champagne	72	85	France	FALSE	0	0	7	6
12	May	Prosecco	72	83	Australia	FALSE	0	0	3	2
13	May	Merlot	6	43	Chile	FALSE	0	6	0	0
14	June	Merlot	72	64	Chile	FALSE	0	0	5	4
15	June	Cabernet Sau	144	19	Italy	FALSE	0	0	2	4
16	June	Merlot	72	88	California	FALSE	0	0	5	0
17	July	Pinot Noir	12	47	Germany	FALSE	7	0	0	0
18	July	Espumante	6	50	Oregon	FALSE	0	11	2	1
19	July	Champagne	12	66	Germany	FALSE	0	0	2	3
20	August	Cabernet Sau	72	82	Italy	FALSE	0	0	4	2
21	August	Champagne	12	50	California	FALSE	0	0	2	2
22	August	Champagne	72	63	France	FALSE	0	0	0	21
23	September	Chardonnay	144	39	South Africa	FALSE	0	0	3	2
24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0	0
25	October	Cabernet Sau	72	59	Oregon	TRUE	0	0	3	3
26	October	Pinot Noir	144	83	Australia	FALSE	8	0	5	2
27	October	Champagne	72	88	New Zealand	FALSE	0	0	6	3
28	November	Cabernet Sau	12	56	France	TRUE	0	0	4	2
29	November	Pinot Grigio	6	87	France	FALSE	0	15	2	0
30	December	Malbec	6	54	France	FALSE	0	16	2	4
31	December	Champagne	72	89	France	FALSE	0	0	10	7
32	December	Cabernet Sau	72	45	Germany	TRUE	0	0	3	1

We see that indeed the top deals for Cluster 1 correspond to Pinot noir. For Cluster 3, the evidence is mixed, but there seems to be a pattern towards bubbly wines (espumante or champagne), preferably French and with a good bargain. As for Cluster 4, it seems to correspond to one specific deal, August Champagne. There is quite an overlap between Clusters 3 and 4.

3.1.3 Silhouette

The questions that steps in mind from the previous example is that maybe $k = 4$ is not a good choice. Fortunately, there exist techniques for assessing whether a particular value of k is a good idea. This is called the *silhouette*. The silhouette is a score that is assigned to the clusters. The questions is the following:

Are individuals within a cluster much closer to each other than they are to individuals in other clusters?

The silhouette of an individual is formally computed as follows:

$$\frac{\text{average distance to individuals in the closest cluster} \; - \; \text{average distance within the cluster}}{\text{maximum of the two averages above}}$$

The denominator above ensures that

$$\text{silhouette} \in (-1, 1)$$

If the silhouette is close to unity, then the matching seems good since the distance with individuals within the cluster is much smaller than the distance to those in the next cluster. When the silhouette is close to zero, then the two clusters appear equally good to the individuals within them. When the silhouette is negative, then clearly some individuals would be better placed in another cluster.

Example 8 *In the example above, the silhouette ranges from -0.22 to 0.60, with an average of 0.15, which is rather close to zero.*

Things become a little more interpretable when looking at $k = 5$. But still, it appears that only two clusters are really identified. In reality the problem with this dataset is that it is binary (either individual took the offer, or did not) and it is much more instructive to know that an individual was interested in a deal than to know that they were not. Indeed, an individual may want to buy a lot of champagne but for some reason they had a lot in stock, so they didn't take a deal. Hence, we are not sure why they did not buy (i.e. we don't know whether they didn't buy because they weren't interested or whether it is due to another reason), whereas we know when they buy that they are interested.
There exist better metrics for the distance that the Euclidian distance when we deal with binary datasets, such as spherical k-means, but this is beyond this course.

3.2 Principal Component Analysis

Beyond k-means clustering, Principal Component Analysis (PCA) is a key feature of product positioning and data summarizing. It is widely used in Marketing, consulting, strategy...

Basically, the idea is the following:

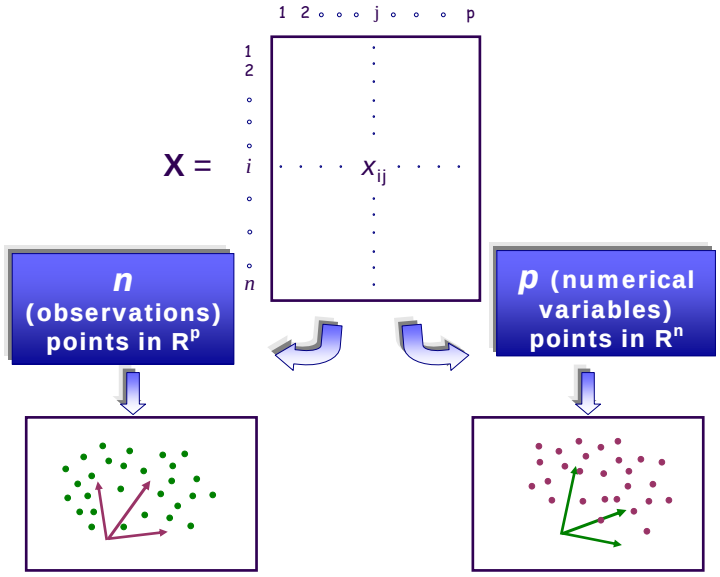
- When dealing with a sample of n observations but with a large number p of variables that are too numerous we construct new variables (the principal components or factors) as linear combination of the original variables.
- We choose the linear combinations in the same way as if we were simply rotating the axes in a graph.
- The new variables are constructed so that they are as close as possible to the observations and can explain as much as possible of their variability (called *inertia*).

The following notes are here to provide an introduction.

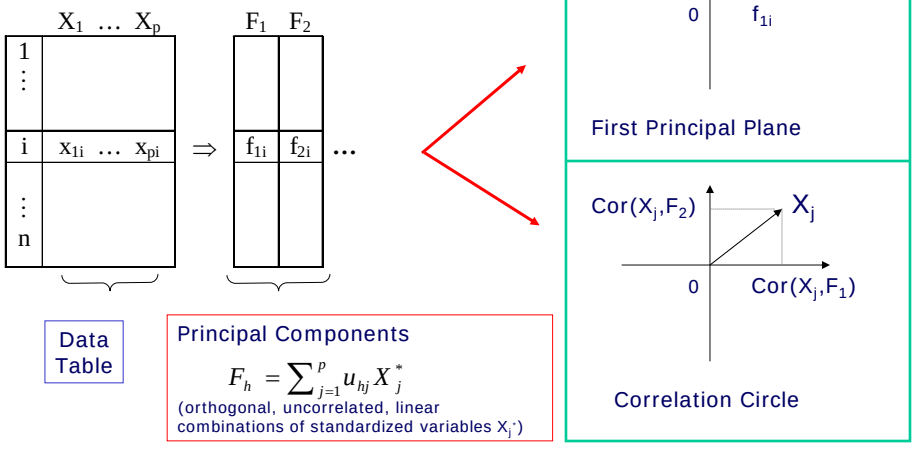
3.2.1 Multidimensional Data Analysis

- It is a set of techniques and methods with the aim of visualizing and interpreting the structure of large datasets,
- has the objective of : showing the latent structure underlying the analyzed system.
- This is done by means of a dimensionality reduction of the variable or the observation representation space (onto a best approximation lower dimensional space) so as the extracted structural information may be considered as being optimal with respect to a fixed statistical criterion that aims at minimizing the information loss.

3.2.1.1 Data Matrix, Projections and Visualization Rules



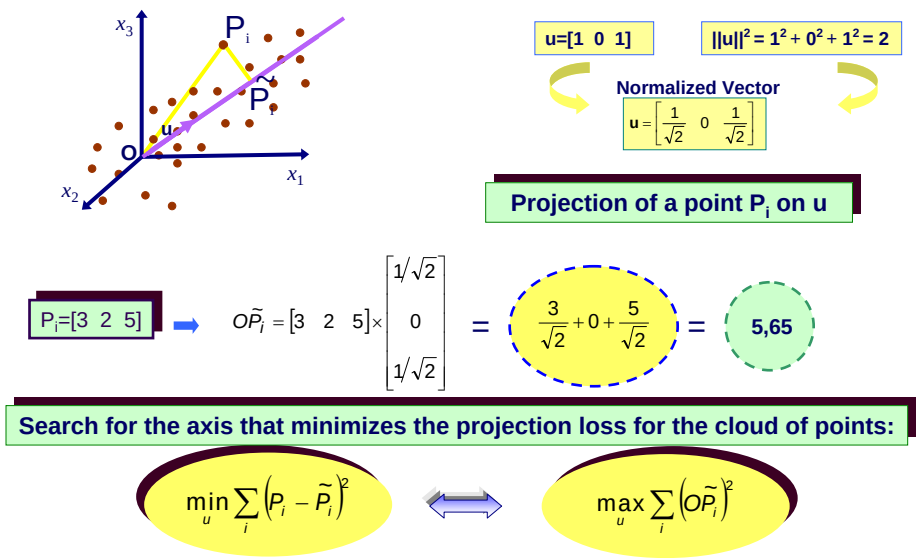
a) Data visualization onto a lower dimensional space



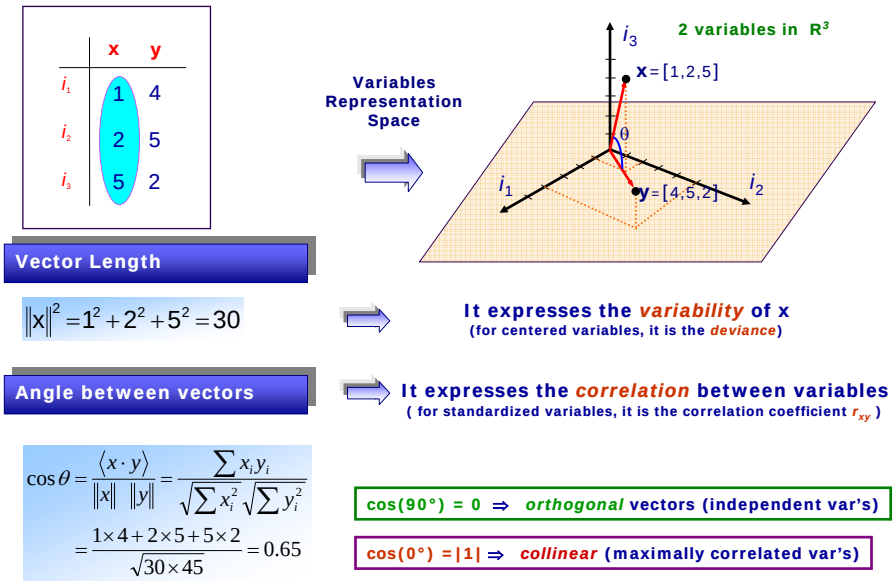
The weights u_{hj} are unknown and depend from the optimized statistical criterion.

The weights u_{hj} are unknown and depend from the optimized statistical criterion.

3.2.1.2 Projection of a Point on an Axis



3.2.1.3 Geometrical Representation of Statistical Variables



3.2.2 Principal Component Analysis (PCA)

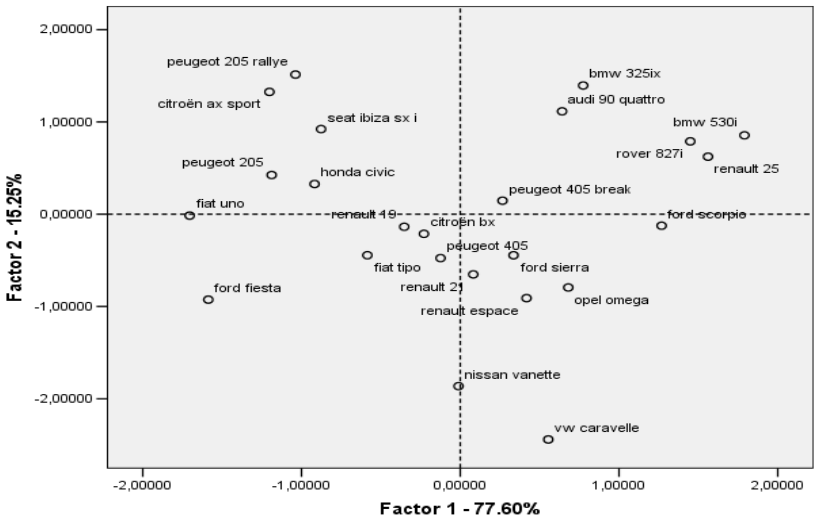
- PCA is a multivariate descriptive or exploratory method
- PCA has the objective of:
  - reducing the original variables into a lower number (which?) of orthogonal (non correlated) synthesized variables (factors)
  - visualizing correlations among the original variables and between these variables and the factors
  - visualizing proximities among statistical units
- Examples of PCA uses:
  - construction of scores (that summarize behaviors of n statistical units on a set of p variables: customers' profiles in a bank, financial performance of enterprises, purchase behaviors, life quality in a country, customer satisfaction for a brand, intelligence quotient, etc.)
  - before a multiple regression, to solve the multicollinearity problem (it replaces variables with orthogonal factors)
  - tandem analysis: in a data matrix, it reduces the number of "columns" (variables) before clustering the "rows" (statistical units)

3.2.2.1 Example on “Product Positioning”

- The dataset

Model	Displacem.	Power	Speed	Weight	Length	Width
Honda Civic	1396	90	174	850	369	166
Renault 19	1721	92	180	965	415	169
Fiat Tipo	1580	83	170	970	395	170
Peugeot 405	1769	90	180	1080	440	169
Renault 21	2068	88	180	1135	446	170
Citroën BX	1769	90	182	1060	424	168
BMW 530i	2986	188	226	1510	472	175
Rover 827i	2675	177	222	1365	469	175
Renault 25	2548	182	226	1350	471	180
Opel Omega	1998	122	190	1255	473	177
Peugeot 405 Break	1905	125	194	1120	439	171
Ford Sierra	1993	115	185	1190	451	172
BMW 325ix	2494	171	208	1300	432	164
Audi 90 Quattro	1994	160	214	1220	439	169
Ford Scorpio	2933	150	200	1345	466	176
Renault Espace	1995	120	177	1265	436	177
Nissan Vanette	1952	87	144	1430	436	169
VW Caravelle	2109	112	149	1320	457	184
Ford Fiesta	1117	50	135	810	371	162
Fiat Uno	1116	58	145	780	364	155
Peugeot 205	1580	80	159	880	370	156
Peugeot 205 Rallye	1294	103	189	805	370	157
Seat Ibiza SXI	1461	100	181	925	363	161
Citroën AX Sport	1294	95	184	730	350	160

- Descriptive Statistics & Simple Correlation Coefficients



Descriptive Statistics

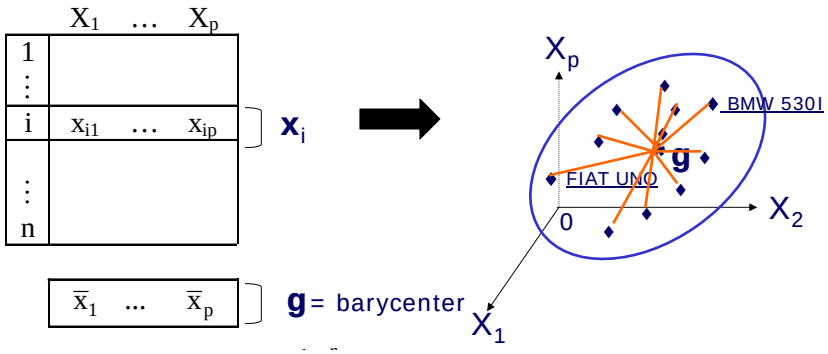
	N	Mean	Std. Deviation	Variance
Cylindrée	24	1906.12	527.91	278687.6
Puissance	24	113.67	38.78	1504.232
Vitesse	24	183.08	25.22	635.819
Poids	24	1110.83	230.29	53034.058
Longueur	24	421.58	41.34	1709.036
Largeur	24	168.83	7.65	58.580

Remark:
In this output (from the SPSS software), variances and standard deviations are computed as unbiased modified estimators, i.e. « dividing by n-1 ». However, in PCA, most often data are considered to be a population to explore rather than a sample. In XLSTAT users are allowed to choose whether to divide by « n » or « n-1 ». In this chapter, we show the SPSS approach where « n-1 » is chosen.

	Correlation					
	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Cylindrée	1.000	.861	.693	.905	.864	.709
Puissance	.861	1.000	.894	.746	.689	.552
Vitesse	.693	.894	1.000	.491	.532	.363
Poids	.905	.746	.491	1.000	.917	.791
Longueur	.864	.689	.532	.917	1.000	.864
Largeur	.709	.552	.363	.791	.864	1.000

3.2.2.2 Total Inertia

- Cloud N, of n points in $\mathbb{R}^p$



Total inertia of the cloud is computed as the sum of square Euclidian distances,

$d(x_i, g)$, between each observation and the barycenter (center of gravity)

$$\begin{aligned} I(N, \mathbf{g}) &= \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g) \\ &= \sum_{j=1}^{p} \sigma_j^2 \end{aligned}$$

It is the sum of variances.

Application to the *Product Positioning* example:

$$\sum_{j=1}^{p} \sigma_j^2 = 278,687.6 + ... + 58.58 = 335,629.325$$

3.2.2.3 Data Standardization

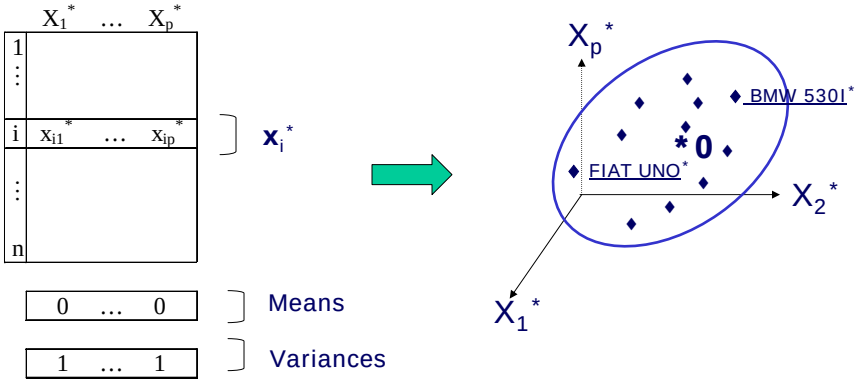
In order to make the visualization easier as well as to avoid the *problem of different measurement units*, raw data are replaced by standardized (centered and reduced) data (here, the unbiased estimator for the standard deviation is used):

$$\begin{aligned} x_1^* &= \frac{x_1 - \overline{x}_1}{s_1} \\ &\vdots \\ x_p^* &= \frac{x_p - \overline{x}_p}{s_p} \end{aligned}$$

so the $x_1^*, ..., x_p^*$ have zero mean and unit standard deviation.

Standardized Data: new cloud denoted N^*

Total Inertia of standardized data (X)*



$N^* = \{x_1^*, \ldots, x_i^*, \ldots, x_n^*\}$

Barycentre: $\mathbf{g}^* = \mathbf{0}$; Total Inertia: $I(N^*, \mathbf{0}) = p$

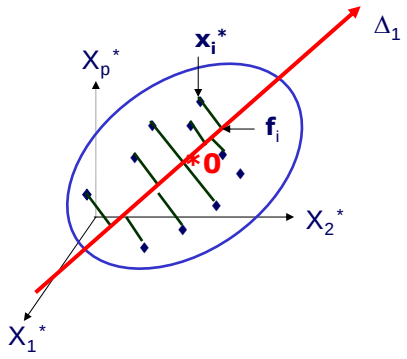
Application to «Product Positioning»:

Total Inertia = Number of variables = $p = 6$

3.2.2.4 Objectives of PCA

- **Objective 1:** We search for the axis Δ_1 fitting as well as possible (in terms of projections, not residuals as in regression) the points in the cloud of standardized variables N^* (**Best fit criterion**).
- **Objective 2:** We search for the axis Δ_1 in the direction of maximum dispersion of cloud N^* (**Maximum variability criterion**).

3.2.2.5 Methods for PCA



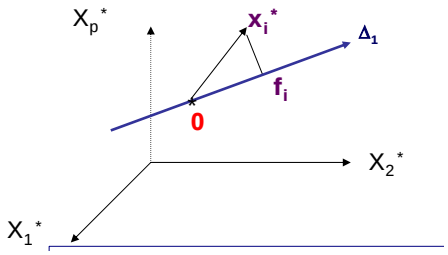
- Method 1 for the best fit criterion: We aim at minimizing the inertia of cloud N^* with respect to the axis Δ_1 (or First principal component, F_1)

$$I\left(N^*, \Delta_1\right) = \frac{1}{n}\sum_{i=1}^{n} d^2\left(x_i^*, f_i\right)$$

- Method 2 for the maximum variability criterion: We aim at maximizing the inertia of cloud N^* projected on the axis Δ_1 (i.e. the $f_1, ..., f_n$)

$$I\left(\{f_1, ..., f_n\}, 0\right) = \frac{1}{n}\sum_{i=1}^{n} d^2\left(f_i, 0\right)$$

- Objectives 1 & 2 are both attained simultaneously:



From

$$d^2\left(x_i^*, 0\right) = d^2\left(x_i^*, f_i\right) + d^2\left(f_i, 0\right)$$

we have

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} d^2\left(x_i^*, 0\right)}_{\text{Total inertia: } p} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} d^2\left(f_i, 0\right)}_{\substack{\text{Inertia explained by } \Delta_1 \\ \text{(maximized retained information)}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} d^2\left(x_i^*, f_i\right)}_{\substack{\text{Residual inertia} \\ \text{(minimized information loss)}}}$$

3.2.2.6 Features of the First Principal Axis

- First-order solution
  - Axis Δ_1 encounters the barycenter 0 of the cloud of points N^*.
  - Axis Δ_1 is spanned by the normalized vector $\mathbf{u}_1$, eigenvector of the correlation matrix $R = {}^tX^*X^*$ associated to its biggest eigenvalue λ_1:

$$ {}^tX^*X^*\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

    where ${}^tX^*$ is the transposed of X^*, the matrix of standardized variables scaled by $(n-1)^{-1/2}$.
  - The inertia explained by Δ_1 is equal to λ_1.
  - The portion of inertia explained by the first principal axis Δ_1 is equal to λ_1/p because:

$$\sum_{j=1}^{p} \lambda_j = p = \text{Total inertia}$$

- Eigenvalues λ_h and Eigenvectors $\mathbf{u}_h$

NUMBER	EIGENVALUES	PERCENT.	CUMULAT. PERCENT.
1	4.6560	77.60	77.60
2	0.9152	15.25	92.85
3	0.2404	4.01	96.86
4	0.1027	1.71	98.57
5	0.0647	1.08	99.65
6	0.0210	0.35	100.00

Component Score Coefficient Matrix

	Component					
	1	2	3	4	5	6
Cylindrée	.206	.036	-.819	-.156	-3.141	-.075
Puissance	.192	.440	-.081	-1.528	1.206	-3.879
Vitesse	.159	.693	.754	.998	.028	3.109
Poids	.199	-.267	-.988	-.384	1.859	3.632
Longueur	.199	-.309	.090	2.221	.653	-3.021
Largeur	.175	-.500	1.389	-1.140	-.517	.821

- Application to «Product Positioning»:
  - Total Inertia = 6
  - Inertia explained by the first principal axis = λ_1 = 4.656
  - The first axis is given by

$$\begin{aligned}\mathbf{F}_1 &= u_{11}\mathbf{X}_1^* + ... + u_{16}\mathbf{X}_6^* \\ &= .206 \times Cylindrée^* + .192 \times Puissance^* + .159 \times Vitesse^* + \ldots + .175 \times Largeur^*\end{aligned}$$

  - Portion of inertia explained by the first principal axis:

$$\frac{\lambda_1}{p} = \frac{4.656}{6} = 0.776$$

- The first principal component explains 77.6% of the total variability.

3.2.2.7 Features of the Second Principal Axis

- Second-order solution
  - We search for a second principal axis Δ_2 being orthogonal to Δ_1 and fitting as good as possible the cloud of points.
  - This new axis encounters the barycenter 0 and is spanned by the normalized vector $\mathbf{u}_2$, eigenvector of the correlation matrix R associated to its second largest eigenvalue λ_2.

- ► The second principal component $\mathbf{F}_2$ is defined by projecting the points on this second principal axis.
- ► The second principal component $\mathbf{F}_2$ is centred, with variance λ_2, and is non correlated to the first principal component $\mathbf{F}_1$

3.2.2.8 Optimal Number of Axes

There are several rules for choosing how many factors we should retain: we denote the number of factors we choose as h

- Percentage of Explained Variance: choose the minimum h that satisfies

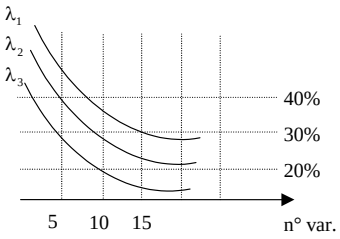
$$\tau_h = \frac{\lambda_1 + ... + \lambda_h}{\lambda_1 + ... + \lambda_p} \geq 0.75$$

- Kaiser's rule (eigenvalue one): choose the minimum h that satisfies

$$\lambda_1 + ... + \lambda_h > 1$$

- Scree test: choose h which sees the biggest drop in eigenvalues between λ_h and λ_{h+1}

- Use an Abacus, in general the percentage of explained variance behaves as follows:



so we can set the minimum τ_h according to the number of variables we observe.

3.2.2.9 Global Quality of the Analysis

- Total Inertia = Total Variance = p
- Portion of explained variance by the first principal component = λ_1/p

- Portion of explained variance by the second principal component = λ_2/p
- Portion of explained variance by the first two principal components = $(\lambda_1 + \lambda_2)/p$
- And so on for the following dimensions...

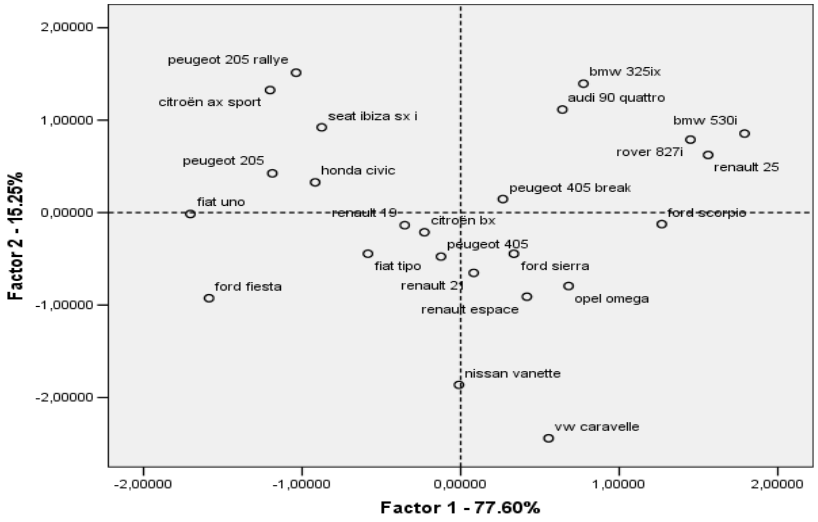
3.2.2.10 Standardized Principal Components: scores of cars

Facteurs centrés-réduits

	Modèle	Facteur 1	Facteur 2	Facteur 3	Facteur 4	Facteur 5	Facteur 6
1	Honda Civic	-.918	.327	1.059	-1.244	-.455	.745
2	Renault 19	-.353	-.136	.882	.651	-.868	.012
3	Fiat Tipo	-.585	-.444	.936	-.580	-.663	1.347
4	Peugeot 405	-.124	-.476	.372	1.867	.107	.192
5	Renault 21	.082	-.652	-.129	1.939	-1.263	.885
6	Citroën BX	-.229	-.212	.302	1.269	-.237	1.185
7	BMW 530i	1.790	.855	-1.029	-.425	-.464	.976
8	Rover 827i	1.449	.789	-.028	.023	-.178	-.441
9	Renault 25	1.560	.623	1.254	-.590	.311	-.276
10	Opel Omega	.680	-.794	1.022	1.224	1.144	-.602
11	Peugeot 405 Break	.266	.146	.696	.584	.574	-.683
12	Ford Sierra	.335	-.445	.219	.974	.416	-.471
13	BMW 325ix	.774	1.394	-1.952	-.482	.331	-1.041
14	Audi 90 quattro	.641	1.115	.292	.101	2.097	-.368
15	Ford Scorpio	1.268	-.125	-.770	-.137	-2.853	-.479
16	Renault Espace	.417	-.909	.519	-1.215	.582	.857
17	Nissan Vanette	-.011	-1.863	-2.493	-.293	1.646	1.840
18	VW Caravelle	.554	-2.440	.601	-2.048	-.073	-1.728
19	Ford Fiesta	-1.587	-.925	-.142	-.360	-.104	-1.229
20	Fiat Uno	-1.705	-.015	-1.014	.437	.282	-1.508
21	Peugeot 205	-1.188	.425	-1.595	-.006	-.945	-.803
22	Peugeot 205 rallye	-1.038	1.513	.201	.337	.833	-.439
23	Seat Ibiza sxi	-.877	.922	-.095	-1.083	.325	1.684
24	Citroën AX sport	-1.202	1.325	.891	-.942	-.545	.345
Mean		.000	.000	.000	.000	.000	.000
Std. Dev.		1.000	1.000	1.000	1.000	1.000	1.000

3.2.3 Representation on the Factorial Plan

3.2.3.1 Observations



Interpretation Rules:

- **Proximities between observations** on the factorial plan are interpreted as similar behaviors on the observed variables:
  - ► e.g. “Peugeot 205 rallye” and “Citroën ax sport” (upper left corner of the representation) are very close to each other and show very similar profiles on the six variables while they are very far from “Nissan vanette” and “VW caravelle” (at the bottom of the representation) thus showing different profiles from these two car models. The latter cars are, instead, very similar to each other.
- The observations may be compared if the **quality of representation** on the factorial plan is good enough.
- If the factorial plan is “too crowded”, then **project the barycenters of groups** of observations defined by grouping according with categories of a nominal variable.

3.2.3.2 Coordinates of variables on $\mathbf{F}_h$

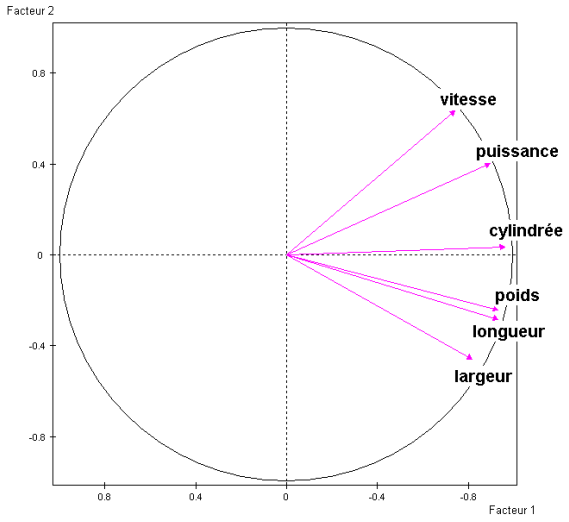
Correlations with the first 5 (out of 6) principal components

VARIABLES	CORRELATIONS				
	1	2	3	4	5
Cylindrée	0.96	0.03	-0.20	-0.02	0.20
Puissance	0.89	0.40	-0.02	-0.16	-0.08
Vitesse	0.74	0.63	0.18	0.10	0.00
Poids	0.93	-0.24	-0.24	-0.04	-0.12
Longueur	0.93	-0.28	0.02	0.23	-0.04
Largeur	0.81	-0.46	0.33	-0.12	0.03

Example

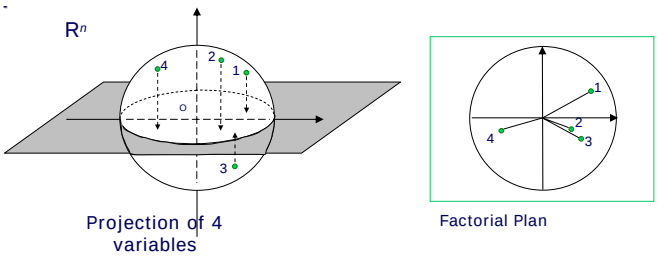
$$\begin{aligned} \text{corr}\,(Cylindrée, \mathbf{F}_1) &= 0.96 \\ &= \text{coordinate of the variable } \textit{cylindrée} \text{ on the first factor } \mathbf{F}_1 \end{aligned}$$

Variables (Loading Plot)



3.2.3.3 Representation of the Variables

- Interpretation of the Correlation Circle



  - ► An axis is an “artificial variable” as it is a linear combination of p observed variables
  - ► The size of variable-axis correlations allows a better understanding of the axis meaning (provided by more strongly correlated variables) and helps the interpretation
  - ► “SIZE” Effect: if most variables are positively correlated among them (as in our application), this means they are simultaneously strong or weak -> the 1st axis will oppose “big” values to “small” values thus providing a quantitative discrimination along the horizontal axis (in our application, we expect to have “big” cars on the right as they somehow show high values simultaneously on all 6 variables. Conversely, “small” cars will be on the left. Intermediate situations will occur in between.)
  - ► “SHAPE” Effect: the successive axes (starting from the 2nd axis) usually discriminate in qualitative terms and oppose groups of variables with different meaning. In our application, we expect “sport” cars at the top (positive side of axis 2) as they are mainly characterized by power and speed while “family” cars will be at the bottom (negative side of axis 2) as they are mainly characterized by width and length.
- Supplementary Points: Observations and Variables
  - ► If an observation (or a set) is far from the cloud of points on the factorial plan, then its position is considered to be atypical. After provisionally removing this observation (or the set) from the data, run PCA again and then project the removed observation(s) a posteriori on the obtained axes.
  - ► If some of the observed variables are not directly linked with the themes studied in the data, they can be projected afterwards without directly participating to the search for the axes:
    - ∗ quantitative variables: compute variable-factors correlations
    - ∗ categorical variables: compute the average observation for each category and project the “average” observations

- Choose the active and supplementary observations/variables at the very beginning of the analysis before running PCA.

PART II

Statistical Inference

Chapter 4
Point Estimation

4.1 Statistical models

4.1.1 How to handle a problem?

- We have n observations (data) $x_1, \cdots, x_n$ of an experience/phenomenon (in economics or marketing or ...).
  - ► We decide to consider this experience as random and we propose a probabilistic modelling of such a phenomenon; it means in practice to introduce a variable of interest X in the population. X is called the parent r.v. and has a unknown distribution, called the parent distribution.
  - ► We decide to consider the data as realizations of **i.i.d. (independently and identically distributed) random variables.** $X_1, \cdots, X_n$ with the same distribution as the parent one: $x_i = X_i(\omega)$, $1 \leq i \leq n$.
- We build a statistical model, i.e. we choose a family of distributions $\mathcal{P} = (\mathsf{P}_\theta, \theta \in \Theta \subset \mathbb{R}^k)$ depending on an (unknown) parameter θ (which can be multidimensional) such that the true parent distribution belongs to the family $\mathcal{P}$.
  The choice of $\mathcal{P}$ is made **on the basis of past experiences or theory, or the researcher's own assumptions**.
- Then, from the data collected on the sample, we have to estimate the true value of the unknown parameter θ or to test a hypothesis on it.

- In this chapter we will focus on the notion of estimators:
  - ► What is an estimator ?
  - ► How to build an estimator? with which method ?

- ► How to evaluate the quality of an estimator ?
- In subsequent chapters, we will see methods to check the validity of our model

4.2 Estimators

4.2.1 Definitions

Let $(X_1, \cdots, X_n)$ be a n-sample with parent law $\mathcal{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$.

Definition 7 *An **estimator** of θ is a **random variable.** $T_n = h_n(X_1, \cdots, X_n)$, with $h_n : \mathbb{R}^n \to \Theta \subset \mathbb{R}^k$ ($k \geq 1$) some function.*

Definition 8 *Definition. A realization $h_n(X_1(\omega), \cdots, X_n(\omega))$ of the (r.v.) estimator $T_n = h_n(X_1, \cdots, X_n)$ is called an **estimate** of θ.*

For instance, $h_n(x_1, ..., x_n)$ is an estimate of θ calculated from the data $(x_1, \cdots, x_n)$.

Example 9 *Coin tossing: we want to estimate the proportion of "head" obtained when playing 25 times (see gauchers.xls).*
So we consider $n = 25$ independent trials $X_1, \cdots, X_n$ with Bernoulli distribution $\mathcal{B}(p) = \mathcal{B}(1, p)$, where only 2 results are possible, denoted 1 (for "head") and 0 (for "tail"), with probability p and $1 - p$ respectively.
Our goal: to estimate p (proportion of "head").

- *The family of Bernoulli distributions is defined with the real parameter $p \in]0, 1[$ of the parent distribution $\mathcal{B}(p)$.*
- *Since $\mathsf{E}(X) = p$ when $X \sim \mathcal{B}(p)$, a natural estimator of p would be the empirical mean denoted $\overline{X}_n$ and defined by*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

  where the X_i's have the same cdf as the parent r.v. X.
  Note that $\overline{X}_n$ is well defined as $h_n(X_1, ..., X_n)$.
- *An estimate of p built from the observations would then be:*

$$\overline{x}_n = h_n(x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad x_i \in \{0, 1\}.$$

- *Numerical application: we build various samples of size 25. We compute the empirical mean associated with each sample and obtain different estimates of* p.

	Pile ou Face Réalisations de plusieurs échantillons de taille 25									
N	cas 1	cas 2	cas 3	cas 4	cas 5	cas 6	cas 7	cas 8	cas 9	cas 10
1	0	1	0	0	0	1	0	0	1	0
2	1	1	0	0	0	1	0	1	1	1
3	1	0	0	1	0	1	0	1	0	0
4	0	0	0	0	0	0	0	1	0	0
5	0	1	1	0	0	1	1	0	1	1
6	1	1	1	1	1	1	0	1	0	0
7	1	1	1	1	1	1	0	0	0	1
8	0	1	1	0	1	0	1	0	0	1
9	0	1	0	1	1	1	1	1	0	0
10	1	1	0	0	1	1	1	0	0	1
11	1	0	1	1	1	1	1	1	1	1
12	1	1	1	0	1	1	1	0	0	0
13	1	1	1	0	1	0	1	1	1	1
14	0	1	0	1	0	1	1	1	0	1
15	0	1	0	1	0	1	1	0	1	0
16	0	1	0	0	1	0	1	0	1	0
17	0	1	1	1	1	0	0	0	0	0
18	1	0	0	1	0	1	0	0	0	0
19	1	1	1	1	1	1	0	1	0	1
20	0	0	0	1	1	0	0	0	0	1
21	1	1	0	0	0	1	0	1	0	0
22	1	0	0	1	1	0	1	0	0	1
23	0	1	0	1	0	0	1	0	1	0
24	1	0	1	1	0	1	0	0	0	1
25	0	0	0	1	1	0	0	1	1	0
Somme	13	17	10	15	14	16	12	11	9	12
Moyenne	0,52	0,68	0,40	0,60	0,56	0,64	0,48	0,44	0,36	0,48

Note that the estimates are close to the theoretical value $p = 0.5$ *of our model.*

- *Remark: Here we limit our study to the case of the "point" estimate of a parameter* θ*. A more general problem would be the estimate of* $g(\theta)$*, where* $g : \Theta \longrightarrow \Theta' \subset \mathbb{R}^{k'}$*,* $k' \leqslant k$*. Note that the point estimate corresponds to the case where* $g = Id$*.*

4.2.2 Bias of an estimator

Estimates of a parameter θ provide only approximated values of the true (unknown) value, say θ_0, of θ.

Therefore we ask at least for the estimates to be distributed around θ_0. To do so, we would like to impose that the estimator $T_n = h_n(X_1, \cdots, X_n)$ of θ satisfies the condition

$$\mathsf{E}_{\theta_0}[T_n] = \theta_0,$$

where E_{θ_0} is used to denote that the expectation is computed under the assumption that the true parameter is indeed θ_0. The difference

$$b_n(T_n, \theta_0) \ := \ \mathsf{E}_{\theta_0}[T_n] - \theta_0$$

is called the bias of the estimator T_n.

Definition 9 *An estimator $T_n = h_n(X_1, \cdots, X_n)$ of θ is called unbiased if $\forall n,\ b_n(T_n, \theta_0) = 0$.*
T_n is an asymptotically unbiased estimator of θ if

$$\forall \theta_0 \in \Theta, \quad \lim_{n\to\infty} b_n(T_n, \theta_0) = 0 \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathsf{E}_{\theta_0}[T_n] = \theta_0.$$

In applications, if it is not possible/easy to built a unbiased estimator of some parameter, we will at least look for an asymptotically unbiased one; in this case, it will be necessary to work on large samples.

Let us consider our previous example of the coin tossing. *Question: Is the estimator $\bar{X}_n$ of p unbiased for the true value p_0 ?*

We have

$$E_{p_0}\left[\overline{X}_n\right] = E_{p_0}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E_{p_0}[X_i] = \frac{1}{n}(n \times p_0) = p_0,$$

which implies that $\overline{X}_n$ is an unbiased estimator of p_0.

Remark 3 *In what follows, the true value θ_0 will be written as θ since often no confusion in the meaning is possible.*

4.2.3 Convergence of an estimator

Another request we can make on the r.v. estimator of a parameter is that this r.v. tends to the true value of the parameter at least in probability.

Definition 10 *An estimator T_n of θ is said **convergent in probability** to the (true) value of the parameter θ or **consistent** for θ if the sequence of r.v. $(T_n)_{n\in\mathbb{N}^*}$ converges in probability to θ, i.e. for all $\theta \in \Theta$, for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathsf{P}_\theta[|T_n - \theta| \leq \varepsilon] = 1 \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathsf{P}_\theta[|T_n - \theta| > \varepsilon] = 0$$

where P_θ denotes the distribution associated with the parameter θ.

We will then work with large samples to get a better information on the parameter itself.

One easy way to check about the convergence of an estimator to the value of the parameter it estimates, is to use the Bienaymé-Chebyshev inequality.

Let T_n be an unbiased estimator of θ, such that $\lim_{n\to\infty} V(T_n) = 0$, then the sequence $(T_n)_{n\in N^*}$ converges in probability to θ.

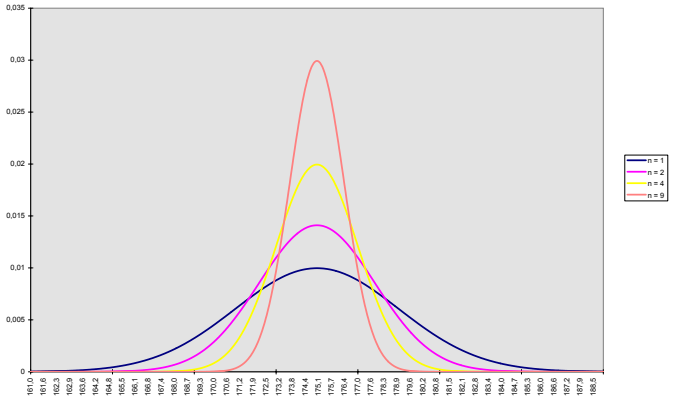


Figure 4.1Density of the estimator T_n of male adult height as a function of the sample size n.

Example 10 *Consider a population of male adults. Suppose we estimated the parameter giving the mean height of this population. Let us represent the density function of the r.v. estimator T_n for different sample sizes n as in figure 4.1.*
We notice that the density function of T_n appears more concentrated around the true (unknown) value of the parameter when n is increasing.

4.2.4 Quadratic Risk

To compare estimators of a given parameter θ, there exist various criteria that a statistician can use. In this course, we will only present one of these criteria, namely the **Quadratic Risk.**

Definition 11 *Let T_n be an estimator of θ. The Quadratic risk of T_n w.r.t. θ, denoted $R(T_n, \theta)$, is given by*

$$\begin{aligned} R(T_n,\theta) &= \mathsf{E}_\theta[(T_n-\theta)^2] = \mathsf{Var}_\theta(T_n) + (\theta - \mathsf{E}_\theta(T_n))^2 \\ &= \mathsf{Var}_\theta(T_n) + b_n^2(T_n,\theta). \end{aligned}$$

Consequence: to compare two unbiased estimators, it will be enough to compare their variances, the better one will be the one with the smaller variance (less spread around its mean).

4.2.5 Asymptotic behavior

To understand the behavior of estimators in large samples (hence asymptotically),

we need three theorems.

First, the Law of Large Numbers tells us that the average tends to the expectation. This is useful when we want to estimate a parameter.

Theorem 1 *(*Weak*) **Law of large numbers (LLN)***

*Let $(X_i)_{i\geq 1}$ be a sequence of real **i.i.d.** random variables with finite mean: $|\mathsf{E}[X_1]| < \infty$. Then we have*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathsf{E}[X_1] \quad as\ n \to \infty.$$

Notice that we can apply the previous LLN to the square for instance: if $\mathsf{E}[X_1^2]$ exists then, as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathsf{E}[X_1^2]$$

This is a very powerful result since it shows that even if the variance does not exist, the sample mean converges to the expectation.

Now the Central Limit theorem is essential as it tells us that whatever the distribution of (X_i) , the Gaussian distribution arises naturally when we focus on the sample average. So asymptotically (i.e. in large samples), we do not need to know the distribution of X_i to know how the mean behaves!

Theorem 2 (Lindberg-Lévy Central Limit Theorem or C.L.T.) *Let $(X_i)_{i\geq 1}$ be a sequence of $\mathbb{R}$-valued random variables, with i.i.d. components and finite first two moments, $\mathsf{E}[X_1] = \mu_X \in \mathbb{R}$ and variance σ_X^2. Then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_X\right) \xrightarrow{d} \mathcal{N}(0, \sigma_X^2) \quad as \quad n \to \infty.$$

The CLT states that approximately for n large $\sqrt{n}\left(\overline{X}_n - \mu_X\right) \sim \mathcal{N}(0, \sigma_X^2)$ so

$$\overline{X}_n - \mu_X \overset{approx}{\sim} \mathcal{N}(0, \frac{\sigma_X^2}{n})$$
$$\overline{X}_n \overset{approx}{\sim} \mathcal{N}(\mu_X, \frac{\sigma_X^2}{n})$$

The following result, known as Slutsky's formula is useful because their remains an issue in the previous CLT: to know the asymptotic distribution of $\sqrt{n}\left(\overline{X}_n - \mu_X\right)$, we need to know σ_X^2... but in fact if we have an estimator of σ_X^2, say $\widehat{\sigma}_X^2$, then we can use it. (We will see what we properly mean by *estimator* in next chapter).

Theorem 3 (Slutsky's formula) *Let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be sequences of real random variables.*
Suppose that, as $n \to \infty$, $X_n \xrightarrow{d} X$ with X real r.v. (r.r.v.) and $Y_n \xrightarrow{\mathsf{P}} c$ with c some constant. Then, for any real-valued continuous function h,

$$h(X_n, Y_n) \xrightarrow{d} h(X, c) \quad as\ n \to \infty.$$

To see how Slutsky's formula can be used consider the case where X_i is scalar with variance σ_X^2. Then the CLT says that

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma_X^2)$$

so in fact

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

Assume that you use the usual estimator of the variance $S_X^2 \xrightarrow{p} \sigma_X^2$ and let the function $h(a,b) = a/\sqrt{b}$ then since

$$\begin{cases} \sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} X \sim \mathcal{N}(0, \sigma_X^2) \\ S_X^2 \xrightarrow{p} \sigma_X^2 \end{cases}$$

it follows that

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_X} = h\left(\sqrt{n}\left(\overline{X}_n - \mu\right), S_X^2\right) \xrightarrow{d} h\left(X, \sigma_X^2\right) \sim \mathcal{N}(0,1)$$

i.e. we know that asymptotically

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

this will be most useful throughout the course!

4.3 Estimation methods

4.3.1 Empirical estimators

Let consider a n-sample $(X_1, \cdots, X_n)$ defined on a probability space $(\Omega, \mathcal{A}, \mathsf{P})$, with an unknown parent cdf F defined on $\mathbb{R}$,

$$F(t) \stackrel{def}{=} \mathsf{P}[X \leq t], \quad \forall t \in \mathbb{R}.$$

Goal: to estimate the cdf F, then the moments of X (when they exist), and also the median or more generally the quantiles.

How? by an empirical method, i.e. using the observations directly, without any hypothesis;

$\nearrow$ advantage of such method: easy and very intuitive; gives already an idea of what to expect;

$\searrow$ disadvantage: provides estimators with often poor qualities, which are often used as preliminary estimators. Estimators of quantities (as e.g. moments)which may not even exist.

4.3.1.1 Empirical distribution function

Definition 12 *Let $(X_1, \cdots, X_n)$ be a n-sample defined on a probability space $(\Omega, \mathcal{A}, \mathsf{P})$. We define the empirical distribution function F_n associated with $(X_1, \cdots, X_n)$ as the random (i.e. defined on Ω) function*

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathsf{1}_{]-\infty;x]}(X_i) = \frac{\#\{X_i \leq x\}}{n}$$

where $\mathsf{1}_{Set}(x)$ is the indicator variable that takes the value 1 if $x \in Set$ and zero otherwise;
$\#\{X_i \leq x\}$ refers to the number of X_i (for $1 \leq i \leq n$) that are smaller than x.
(Examples of the empirical distribution function are on the course website.)

In practice, we like to express the empirical cdf with the order statistics $(X_{1:n}, \cdots, X_{n:n})$ associated with the n-sample, where $X_{1:n} \leq X_{2:n} \cdots \leq X_{n:n}$ (with possible multiplicities in case of identical values). Then we obtain

$$F_n(x) = \frac{i}{n} \text{ if } x \in [X_{i:n}, X_{i+1:n}[, \quad \text{with } X_{0:n} := -\infty \text{ and } X_{n+1:n} := +\infty,$$

or

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < X_{1:n} \\ i/n & \text{if} \quad x \in [X_{i:n}, X_{i+1:n}[\quad (1 \leq i \leq n-1) \\ 1 & \text{if} \quad x \geq X_{n:n} \end{cases}$$

- Convergence:
  Notice that the definition of F_n implies that, for fixed t, the r.v. $nF_n(x) = \sum_{i=1}^{n} 1_{]-\infty;x]}(X_i)$ has a binomial distribution $\mathcal{B}(n, F(x))$, where F denotes the unknown parent cdf.
- Consequence: since we have seen how to estimate the parameter p of a Binomial distribution $\mathcal{B}(n, p)$, we can propose an estimate of $F(x)$ and say that

$$F_n(t) \text{ is an unbiased estimator of } F(x), \text{ for each fixed } x.$$

Moreover, it can be proved that

$$\sqrt{n}\left(F_n(x) - F(x)\right) \xrightarrow{d} \mathcal{N}\left(0, F(x)[1 - F(x)]\right), \quad as \quad n \to \infty.$$

4.3.1.2 Empirical quantiles

The quantiles associated with the empirical cdf, defined by

$$x_n(p) := \inf\{x | F_n(x) \geq p\}, \quad with \quad 0 < p < 1.$$

are called empirical quantiles.
Since the empirical cdf is constant between 2 observations (by definition), then the empirical quantiles are observations.

Suppose that the observations are distinct, i.e. that the order statistics satisfy $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$, then we can write

$$x_n(p) \; = \; X_{i:n}, \quad if \quad \frac{i-1}{n} < p \leq \frac{i}{n}, \quad for\, 1 \leq i \leq n.$$

This definition remains true in the general case.

4.3.1.3 Empirical moments

- Definitions.

The empirical mean associated with the n-sample $(X_1, \cdots, X_n)$, usually denoted $\overline{X}_n$, is defined by

$$\overline{X}_n \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The empirical moment of order 2 associated with the n-sample $(X_1, \cdots, X_n)$ is defined by $\frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Note that this definition can be easily generalized to empirical moment of order k ($k \in \mathbb{N}^*$), as $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.

The empirical variance associated with the n-sample $(X_1, \cdots, X_n)$, which we will denote S_n^{*2} in this lecture notes (there exist any different notations such as $S_n'^2$) is defined by

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

We can easily check that

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

Remark 4 *Empirical moments can always be defined/computed, even if the parent distribution F doesn't have moments!*

- Empirical estimators of the moments.

We always consider the n-sample $(X_1, \cdots, X_n)$, with parent r.v. X and unknown cdf F.

By using the weak law of large numbers, we have: $if\ \mathsf{E}[X^k] < \infty$ $(k \in \mathbb{N}^*)$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{\mathsf{P}} \mathsf{E}[X^k].$$

- Empirical mean as an estimator of the mean of the parent r.v.

We can easily check that $\mathsf{E}(\overline{X}_n) = \mathsf{E}[X]$. It implies that the empirical mean $\overline{X}_n$ associated with the n-sample is an unbiased estimator of the (unknown) mean of the parent r.v. X.

Moreover, this estimator satisfies:

Theorem 4 *If* $\mathsf{E}[X^2]$ *exists, then*

$$\sqrt{n}\left(\overline{X}_n - \mathsf{E}[X]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad as \quad n \to \infty.$$

Since by assumption $\mathsf{E}[X^2] < +\infty$, then

$$\mathsf{Var}(\overline{X}_n) = \frac{1}{n}\mathsf{Var}(X_1) = \frac{1}{n}(\mathsf{E}[X^2] - (\mathsf{E}[X])^2) = \frac{\sigma^2}{n}.$$

Now applying the CLT provides the result.

In practice, to be able to use the previous result, we need to replace the unknown σ^2 by an estimator; so we consider the empirical variance S_n^{*2} which does converge in probability to the true (unknown) value σ^2. Then we obtain, by using properties on convergences (here the Slutsky theorem),

$$\sqrt{n}\frac{\overline{X}_n - \mathsf{E}[X]}{\sqrt{S_n^{*2}}} \xrightarrow{d} \mathcal{N}(0, 1), \quad as \quad n \to \infty.$$

- Empirical variance as an estimator of the variance of the parent r.v. X?

  Let us compute $\mathsf{E}(S_n^{*2})$. We have

$$\begin{aligned}
\mathsf{E}[S_n^{*2}] &= \frac{1}{n}\mathsf{E}\left(\sum_{i=1}^{n} X_i^2\right) - \mathsf{E}(\overline{X}_n^2) \\
&= \mathsf{E}[X^2] - \frac{1}{n^2}\left(n\mathsf{E}[X^2] + n(n-1)\mathsf{E}^2[X]\right) \\
&= \frac{n-1}{n}\mathsf{Var}(X).
\end{aligned}$$

Hence the empirical variance S_n^{*2} is a biased estimator of $\mathsf{Var}(X)$ (unless $X \overset{a.s.}{=}$ constant).

That is why we prefer to consider the modified unbiased estimator S_n^2 defined by

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

that we introduced in the descriptive statistics chapter.

The two estimators S_n^{*2} and S_n^2 are asymptotically equivalent:

$$\sqrt{n}(S_n^2 - S_n^{*2}) \xrightarrow{\mathsf{P}} 0 \quad as \quad n \to \infty.$$

Moreover, S_n^2 and S_n^{*2} have the same limit distribution.

Have a look at:
http://socr.stat.ucla.edu/htmls/SOCR_Distributions.html
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

4.3.2 (+) Exercise

A market survey involving young company managers is made to identify the type of computer matching their professional and personal needs. One question concerns the number of children they have, to check whether those with more children would most probably buy a personal computer. But, as shown in the following table, parents with many children were generally not ready to answer such a questionnaire because they were too busy.

	Total population		Sub-population of people ready to answer	
number x of children older than 5 years	frequency f	relative freq. f/N	frequency f	relative freq. f/N
0	20000	.40	6200	0.62
1	12000	.24	2100	0.21
2	10000	.20	1200	0.12
3	6000	.12	400	0.04
4	2000	.04	100	0.01
	N =50000	1	N =10000	1

Two types of samples are proposed:

1. one large sample involving 1000 managers interviewed without any follow up. Answering ratio was estimated to be 20% providing 200 answers as shown in the table.
2. a high quality sample involving 25 managers interviewed with expediting in order to obtain 25 answers (i.e. answering ratio 100%).

   ► Make reasonable assumptions about the population distribution and samples so that you can answer the following questions:

- Calculate the mean number μ of children in both populations.
- Which one of the two samples would provide an unbiased estimator of μ?
- Which one of the two samples would provide the lowest quadratic risk ?

Chapter 5
Hypothesis Testing

5.1 Definition

Example 11 *In a given store S, an average of 25 articles per week is sold. A recent survey, on a sample of 36 other stores belonging to the same chain, has given an average of 30 articles sold per week. We suppose the standard deviation to be known and equal to 8. The manager is thinking of hiring new sellers for store S. To make up her mind, she needs to find out whether 25 is statistically different from 30, i.e. whether store S is truly underperforming. Did it just happen by chance that the average was higher in that sample?*

A **statistical test** is a **decision process** allowing, on the basis of the observed sample (empirical evidence), a decision rule to decide whether an hypothesis concerning the population (called *null* hypothesis) is to be accepted or rejected in favour of an *alternative* hypothesis.

The tested hypotheses take the form:

- H_0 (**Null hypothesis):** that which is supposed to be true *a priori*;
- H_1 (**Alternative hypothesis**): that which is accepted if H_0 is rejected from the data.

In the example above: we introduce the parent r.v. X on $(\Omega, \mathcal{A}, \mathsf{P})$ which gives the weekly sales at a store and the sample of $n = 36$ r.v. $(X_i)_{i=1,\ldots,36}$. Let $\mu = \mathsf{E}[X]$, then

$$\mathsf{H}_0 : \mu = \mu_0 = 25$$

i.e. weekly sales at store S are representative of sales at all stores.

The alternative hypothesis (H_1) is that the true μ is higher than weekly sales at store S, i.e. that store truly is underperforming: this can be written as

$$\begin{aligned} \mathsf{H}_1 &: \mu = \mu_1 = 30 \text{ (Simple alternative hypothesis), or} \\ \mathsf{H}_1 &: \mu = \mu_1 > 25 \text{ (Unilateral, or one-tailed, test).} \end{aligned}$$

For more general questioning, we could also choose $\mathsf{H}_1 : \mu = \mu_1 \neq 25$ (Bilateral, or two-tailed, test) or $\mathsf{H}_1 : \mu = \mu_1 < 25$.

$\hookrightarrow$ The hypotheses that we test can be formulated as '*Is the sample compatible with a given model?*'

Among the two hypotheses, only one is true, but we will never know which one!

$\hookrightarrow$ We need to define a decision rule allowing to minimize the risk (probability) of making an error: which error?

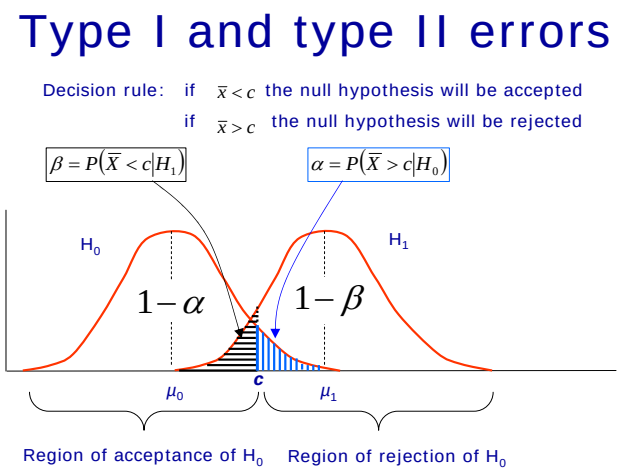
There are four outcomes from any decision, only two are right:

Manager's Decision (retained hypothesis)	Reality	
	H_0 is true	H_1 is true
Accept H_0	No error (prob.: $1-\alpha$)	Type II error (prob.: β)
Reject H_0, accept H_1	Type I error (prob.: α)	No error (prob.: $1-\beta$)

where α and β refer to probabilities of Type I and II errors respectively, i.e.

$$\alpha \overset{def}{=} \mathsf{P}\left(\text{reject } \mathsf{H}_0 \mid \mathsf{H}_0 \text{ is true}\right)$$
$$\beta \overset{def}{=} \mathsf{P}\left(\text{accept } \mathsf{H}_0 \mid \mathsf{H}_1 \text{ is true}\right)$$

α is called **size of the test**. $1-\beta$ is called **power of the test** as it is the probability to correctly reject H_0 when this hypothesis is wrong.



Densities for $\overline{X}_n$ statistic under the null and alternative hypotheses.

5.2 Methodology

Example 12 (continued) *In the previous example, assume that the manager fixes the hypotheses*

$$\begin{aligned} \mathsf{H}_0 &: \mu = \mu_0 \\ \mathsf{H}_1 &: \mu > \mu_0 \end{aligned}$$

and chooses a test size $\alpha = 5\%$. *Since* $\mu = \mathsf{E}[X]$, X *having an unknown cdf, the only way to let*

$$\mathsf{P}\left(\textit{erroneously reject } \mathsf{H}_0 \mid \mathsf{H}_0 \textit{ is true}\right) = \alpha$$

is to propose an estimator of μ, *e.g.* $\overline{X}_n$, *for which the asymptotic cdf is known (see chapter on point estimation).*
Using the CLT, as $n \to \infty$

$$\overline{X}_n \underset{\mathsf{H}_0}{\sim} \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

where the subscript H_0 *indicates that this is true* ***only if*** H_0 ***is true****. For* $n = 36 > 30$, *we can consider that* $\overline{X}_n$ *is approximately normally distributed. We proceed as for confidence intervals and define the r.v.*

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \underset{\mathsf{H}_0}{\sim} \mathcal{N}(0,1)$$

$\overline{X}_n$ *and* Z *are called* ***test statistics.*** *The natural corresponding rule will be to reject* H_0 *if* $\overline{X}_n >> \mu_0$, *which equivalently means* $Z >> 0$. *But,*

$$\mathsf{P}\left(Z > \widetilde{z}_{1-\alpha} \mid \mathsf{H}_0 \textit{ is true}\right) = \alpha$$

where the quantile $\widetilde{z}_{1-\alpha} = 1.65$ *for* $\alpha = 5\%$.
The ***decision rule*** *that the manager adopts is therefore:*

1. *Compute* $z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ *from the sample. Here it is* $z = \sqrt{36}(30-25)/8 = 3.75$
2. *Reject* H_0 *at the 5% size if* $z > 1.65$, *and accept it otherwise. Here* H_0 *is rejected.*

The manager concludes that the expectation of weekly sales is larger than 25 with probability 95%. She is confident that she needs to hire more sales assistants.

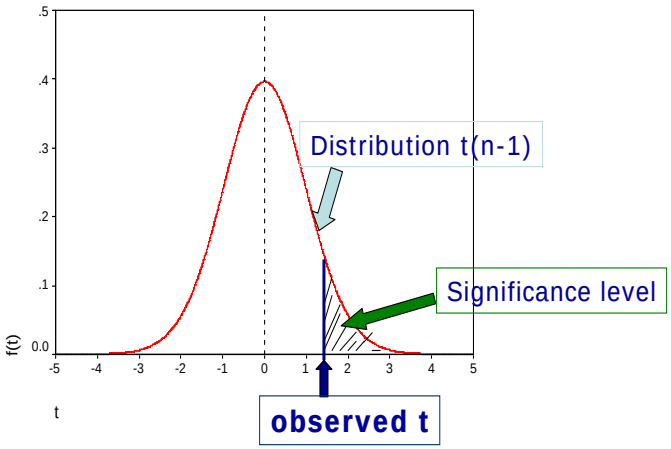
Remark: Notice that we can also express the decision rule in terms of $\overline{X}_n$ instead of Z since

$$\begin{aligned} Z > c &\Leftrightarrow \overline{X}_n > \mu_0 + c\sigma/\sqrt{n} \\ z > 1.65 &\Leftrightarrow \overline{x} > 27 \end{aligned}$$

- **More generally:** main testing steps
  (a) Define the null (H_0) and alternative (H_1) hypotheses
  (b) Choose the probability of Type I error (α) (depending on the study)
  (c) Build the test *Statistic* and derive its distribution (even if only asymptotically)
  (d) Define the *Decision Rule* and compute the threshold (say, c) associated to α
  (e) Select the sample
  (f) The test statistic is computed and its value in the sample compared to the fixed threshold (c)
  (g) The null hypothesis is accepted or rejected.

Definition 13 (p-value) *The significance level (SL), or p-value of an observed statistic z is the minimum (or infimum) value of α that leads to rejection of the null H_0. If z is a realization of Z, then,*

- *for a two-tailed test: p-value $= P(|Z| > |z|)$*
- *for a one-tailed (above) test: p-value $= P(Z > z)$. To compute the p-value, a table of quantiles or a statistical software is needed.*



Significance level, or *p-value*, for an observed statistic drawn from a T_{n-1} distribution, and corresponding

to a one-sided (above) test.

The decision rule for any test can be rewritten using the p-value as:

$$\text{Reject } \mathsf{H}_0 \text{ if the } p\text{-value} < \alpha$$

indeed, for a two-tailed test, for $\mathsf{P}(|Z| > c) = \alpha$

$$p\text{-value} < \alpha \quad \Leftrightarrow \quad \mathsf{P}(|Z| > |z|) < \mathsf{P}(|Z| > c) \quad \underset{\mathsf{H}_0}{\Leftrightarrow} \quad |z| > c$$

(it is clear graphically) and the latter is the standard decision rule. You can easily check that it is also valid for one-tailed tests.

5.3 Test on the mean when the distribution is unknown (large samples)

Let the n-sample $(X_i)_i$ with parent r.v. X on $(\Omega, \mathcal{A}, \mathsf{P})$ with $\mu = \mathsf{E}[X]$. We define the two-tailed statistical test

$$\mathsf{H}_0 : \mu = \mu_0 \text{ versus } \mathsf{H}_1 : \mu \neq \mu_0$$

We conduct the test with size α, and since the statistic

$$z = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{S} \underset{\mathsf{H}_0}{\longrightarrow} \mathcal{N}(0,1),$$

the decision rule is therefore:

Reject H_0 if $z < \widetilde{z}_{\alpha/2}$ or if $z > \widetilde{z}_{1-\alpha/2}$.

Since the latter is symmetric, the decision rule is:

$$\text{Reject } \mathsf{H}_0 \text{ if } |z| > \widetilde{z}_{1-\alpha/2}.$$

This method relies on the Central Limit Theorem, hence can only be applied when the sample size is large: $n > 30$.

5.4 Test on a proportion

Example 13 (Snack food survey) *A random sample of 50 consumers tried a new type of snack food. 29 consumers out of 50 did not like its taste.*

1. *Test the hypothesis* $\mathsf{H}_0 : \pi = \pi_0 = 0.5$ *against* $\mathsf{H}_1 : \pi > 0.5$, *where* π *is the proportion of customers who do not like the snack food in the population. You will use a test size of 5%*

2. *Report the observed significance level of your test (p-value).*

   Answer:

1. *introduce the r.v. X_i taking value 1 if individual i does not like the food, with probability π, and zero otherwise. Then $X_i \sim \mathcal{B}(\pi)$. It has been seen that $\widehat{\pi} = \overline{X}_n$ is an estimator of π. Since $n = 50 > 30$ and, under H_0, $n\pi_0(1-\pi_0) = 12.5 > 5$, we can approximate the distribution as:*

$$\sqrt{n}\frac{\widehat{\pi} - \pi_0}{\sqrt{\mathsf{Var}[X_i]}} \underset{\mathsf{H}_0}{\sim} \mathcal{N}(0,1).$$

   We need an estimator of $\mathsf{Var}[X_i]$ and could use the usual S_n^2.
   In fact, since $\mathsf{Var}[X_i] = \pi_0(1-\pi_0)$, then $\sqrt{n}\frac{\widehat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)}} \underset{\mathsf{H}_0}{\sim} \mathcal{N}(0,1)$.
   Hence, at size $\alpha = 0.05$, we reject H_0 if $\sqrt{n}\frac{\widehat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)}} > 1.65$.
   Here, $\sqrt{n}\frac{\overline{x}-\pi_0}{\sqrt{\pi_0(1-\pi_0)}} = 1.131 < 1.65 \Rightarrow$ we accept H_0.
2. *what is the significance level for $z = 1.131$? From a statistical table or software (e.g. Excel), we find the* p*-value of 13% from*

$$\text{p-}value = [1 - \mathsf{NORMSDIST}(1.131)] = 0.129$$

   as NORMSDIST *returns the cdf of the standard normal distribution and the test is one sided.*

Remark Test statistics are not unique. In the previous example we used $\sqrt{n}\frac{\widehat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)}}$, but notice that $\sqrt{n}\frac{\widehat{\pi}-\pi_0}{\sqrt{\widehat{\pi}(1-\widehat{\pi})}}$ and $\sqrt{n}\frac{\widehat{\pi}-\pi_0}{s}$ could have also been acceptable. They would have yielded different values and p-values. In general several types of tests have various properties. We compare them according to their *power* (see next section).
Remark: To use the CLT for the estimator of the proportion we always need to check that

$$\begin{cases} n > 30 \\ n\pi(1-\pi) > 5 \quad under\ \mathsf{H}_0 \end{cases}$$

5.5 Power functions

When setting a test, we are in practice trying to minimize the probability that we are mistaken in our decision.

This is why we decide upon the test size α or probability of type I error. We hence control the error under the null.

But decreasing α means decreasing the rejection rate. Hence, by setting α too low, we might not reject H_0 even when it is not correct, i.e. the probability β of incorrectly accepting H_0 might be too high.

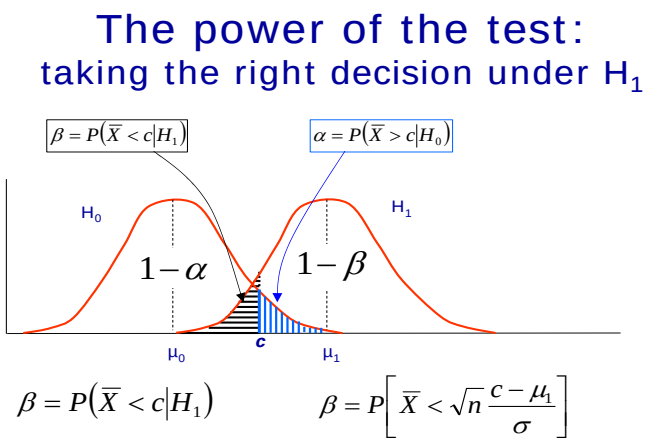
In the previous remark above, we saw that different test statistics are possible for the same hypothesis. One way to compare them is using their *power* $= 1-\beta$.

Definition 14 *The **power of a test** is the probability to correctly reject the null hypothesis when in fact the alternative hypothesis is true, i.e. it is*

$$1-\beta = \mathsf{P}\left(\textit{reject } \mathsf{H}_0 \mid \mathsf{H}_1 \textit{ is true}\right)$$

Assume that we reject H_0 if Z is greater than the critical value c, i.e. $\alpha = \mathsf{P}\left(Z > c \mid \mathsf{H}_0 \text{ is true}\right)$, then

$$\begin{aligned} \beta &= \mathsf{P}\left(Z < c \mid \mathsf{H}_1 \text{ is true}\right) \\ 1-\beta &= \mathsf{P}\left(Z > c \mid \mathsf{H}_1 \text{ is true}\right) \end{aligned}$$



The power of a test is the probability to correctly reject H_0 when it is false and H_1 is true. We use here the statistic $\overline{X}_n$.

Remark: $\alpha = 0.05$ leads often to an optimum generating low α and low β (Neyman-Pearson's approach)

Example 14 *The director of a chain of supermarkets wishes to open a new store. He/She is interested in the average income of the families living in the*

chosen area. A national survey made in 2001 estimated the average monthly income at 2148 Euros. The director wishes to find out if, in the area of interest, this parameter has the same value.

1. Fix the null and the alternative hypotheses.
   Denote by the r.v. $(I_i)_i$ *the monthly income of an average family* i *with*

$$I_i \sim iid\left(\mu, \sigma^2\right) \ \forall i$$

   then the problem can be translated as testing

$$\begin{array}{rcl} \mathsf{H}_0 & : & \mu = \mu_0 = 2148 \\ \mathsf{H}_1 & : & \mu \neq \mu_0 \end{array}$$

2. On a sample of 200 families living in the area, the average income in 2007 (taking into account the inflation rate) is equal to 2296 Euros; this average has a standard deviation of 125 Euros. According to this information, fix the decision rule and the critical threshold (error risk $= 1\%$).
   Working with the mean μ*, we consider the estimator* $\widehat{\mu} = \overline{I}_n$ *of* μ*. We take a sample of* n *=200 families, this is large enough to consider the normal approximation to the cdf of* $\overline{I}_n$*. Then, under* H_0*,*

$$Z_0 = \sqrt{n}\frac{\widehat{\mu} - \mu_0}{\sigma} \sim \mathcal{N}(0,1)$$

   The test we suggest at size $\alpha = 0.01$ *admits the following decision rule: Reject* H_0 *if*

$$z_0 \notin \left[\widetilde{z}_{\alpha/2}, \widetilde{z}_{1-\alpha/2}\right]$$

   with the standard normal quantiles $\widetilde{z}_{1-\alpha/2} = 2.58 = -\widetilde{z}_{\alpha/2}$.
3. Take your decision.
   In our sample, the realization $z_0 = \frac{2296-2148}{125} = 1.184 < 2.58$*. Hence, we accept the null hypothesis that the average income has not changed.*
   Note that the reported standard deviation is of the average income, hence it is $\sigma/\sqrt{n}$.
4. Supposing that, under the alternative hypothesis, the mean is 2400 Euros, and $\sigma/\sqrt{n} = 125$, compute the power $(1-\beta)$ of the test.
   Let $\mu_1 = \mathsf{E}_{\mathsf{H}_1}[I_i]$ *where* $\mathsf{H}_1 : \mu = \mu_1 = 2400$*. If the latter is true, then in fact*

$$Z_1 = \sqrt{n}\frac{\widehat{\mu} - \mu_1}{\sigma} \underset{\mathsf{H}_1}{\sim} \mathcal{N}(0,1)$$

but $Z_0 \underset{\mathsf{H}_1}{\not\sim} \mathcal{N}(0,1)$ *. Indeed*

$$Z_1 = \sqrt{n}\frac{\widehat{\mu} - \mu_0}{\sigma} + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma}$$

hence

$$Z_0 = Z_1 - \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} \underset{\mathsf{H}_1}{\sim} \mathcal{N}\left(-\sqrt{n}\frac{\mu_0 - \mu_1}{\sigma}, 1\right)$$

and we see that if the null hypothesis is not true, then Z_0 *does not follow a standard Normal. The power is given by*

$$1 - \beta = \mathsf{P}\left(|Z_0| > c \mid \mathsf{H}_1\right)$$

where $c = \widetilde{z}_{1-\alpha/2}$*. Let us compute* $1 - \beta$*:*

$$\begin{aligned}
1 - \beta &= \mathsf{P}\left(Z_0 < -c \mid \mathsf{H}_1\right) + \mathsf{P}\left(Z_0 > c \mid \mathsf{H}_1\right) \\
&= \mathsf{P}\left(Z_0 + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} < -c + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} \mid \mathsf{H}_1\right) \\
&\quad + \mathsf{P}\left(Z_0 + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} > c + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} \mid \mathsf{H}_1\right) \\
&= \mathsf{P}\left(Z_1 < -c + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} \mid \mathsf{H}_1\right) \\
&\quad + \mathsf{P}\left(Z_1 > c + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma} \mid \mathsf{H}_1\right) \\
&= \Phi\left(-c + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma}\right) + 1 - \Phi\left(c + \sqrt{n}\frac{\mu_0 - \mu_1}{\sigma}\right)
\end{aligned}$$

where Φ *is the cdf of the standard normal distribution (*$Z_1 \underset{\mathsf{H}_1}{\sim} \mathcal{N}(0,1)$*).*

Numerical application:

$$1 - \beta = \mathsf{NORMSDIST}(-4,596) + 1 - \mathsf{NORMSDIST}(0,564) = 28.6\%$$

The power of this test is 28.6% with a probability of type II error $\beta = 71.4\%$*.*

5.6 Tests on two variables

5.6.1 Chi-square test of independence of two r.v.

Suppose that the r.v. X takes values x_i and the r.v. Y takes values y_j with (*marginal*) probabilities $p_{i\cdot}$ and $p_{\cdot j}$ respectively. Let $p_{ij} = P(X = x_i \cap Y = y_j)$ denote the *joint* probabilities (recall chapter 2)

We want to test

$$\begin{aligned} \mathsf{H}_0 &: X \, and \, Y \, are \, independent \\ \mathsf{H}_1 &: X \, and \, Y \, are \, dependent \end{aligned}$$

under $\mathsf{H}_0 : P(X = x_i \cap Y = y_j) = P(X = x_i) P(Y = y_j)$, i.e.

$$p_{ij} = p_{i\cdot} p_{\cdot j} \tag{5.1}$$

The idea of the test is to check whether equality (5.1) holds. We saw in chapter 2 that we could compute the sum of square differences

$$d_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{q} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$$

where $n_{ij}^* = n p_{i\cdot} p_{\cdot j}$ and $n_{ij} = n p_{ij}$. Under H_0, d_0^2 should be random. Unfortunately, we do not know the true n_{ij}^* and estimate them as $\widehat{n}_{ij} = n \widehat{p}_{i\cdot} \widehat{p}_{\cdot j}$ with $\widehat{p}_{i\cdot} = n_{i\cdot}/n$ and $\widehat{p}_{\cdot j} = n_{\cdot j}/n$. This provides $(r-1)$ and $(q-1)$ independent estimates for the distributions of X and Y ($r-1$ because of the constraint $\sum_i \widehat{p}_{i\cdot} = 1$).

We then compute the test statistic

$$d^2 = \sum_{i=1}^{r} \sum_{j=1}^{q} \frac{\left(n_{ij} - \widehat{n}_{ij}\right)^2}{\widehat{n}_{ij}}$$

and we have, provided that $\widehat{n}_{ij} > 5$ for all i and j (otherwise we merge the categories, e.g. x_i and $x_{i'}$)

$$d^2 \sim \chi^2_{(r-1)(q-1)}.$$

The decision rule is therefore:

$$\textbf{Reject } \mathsf{H}_0 \textbf{ if } d^2 > (1-\alpha)\, th \textbf{ quantile of a } \chi^2_{(r-1)(q-1)}.$$

Example 15 *Table 1 presents the performances of three funds that are managed by the same agent. The performances have been measured over the last ten years and are computed relative to the market. The agent is offering to manage*

			Performance			
			Low	Medium	High	Total
Fund	A	Freq.	13	33	38	84
		%	15.5%	39.3%	45.2%	100.0%
	B	Freq.	38	102	40	180
		%	21.1%	56.7%	22.2%	100.0%
	C	Freq.	90	45	20	155
		%	58.1%	29.0%	12.9%	100.0%
Total		Freq.	141	180	98	419
		%	33.7%	43.0%	23.4%	100.0%

Table 1.Performance of three funds, recorded over time, relative to the market.

your firm's liquid assets. In order to diversify your portfolio and hence minimize your risks, you want to make sure that performances of the three funds are not related to each other. Can you confidently assume that the variables fund (F) and performance (P) are independent?

Answer Independence of the two variables would mean that the probability of each cell is the product of the probability of the row multiplied by the probability of the column:

$$p_{ij} = p_{i\cdot}p_{\cdot j}$$

We compute estimates of the theoretical (under independence) frequencies $\widehat{n}_{ij} = n_{i\cdot}n_{\cdot j}/n = np_{i\cdot}p_{\cdot j}$ as follows:

$\widehat{n}$	Low	$Medium$	$High$	$Total$
A	28	36	20	84
B	61	77	42	180
C	52	67	36	155
$Total$	141	180	98	419

which yields

$$d^2 = 83.78$$

Now, we compare d^2 to a $\chi^2_{(r-1)(q-1)}$ where $r = 3 = q$. Using Excel, the quantile of a χ^2_4 at $1-\alpha$, where $\alpha = 0.05$ is

$$\mathsf{CHIINV}\,(.05, 4) = 9.49$$

which implies that the null hypothesis of independence between the three funds is strongly rejected. The associated *p*-value is

$$\mathsf{CHIDIST}\,(83.78, 4) = 2.8 \times 10^{-17}.$$

Investing in the three funds will not minimize risks.

5.6.2 Test for equality in expectations between two samples

Often, one asks whether two random variables in different samples have the same expectations.

Consider for instance two samples A and B of respectively n_A and n_B observations and where the random variables X^A and X^B have empirical means $\overline{X}^A = \frac{1}{n_A}\sum_{i=1}^{n_A} X_i^A$, $\overline{X}^B$ with realizations $\overline{x}^A$, $\overline{x}^B$. Denote $\mathsf{E}\left[X^A\right] = \mu_A$, $\mathsf{Var}\left[X^A\right] = \sigma_A^2$ and $\mathsf{E}\left[X^B\right] = \mu_B$, $\mathsf{Var}\left[X^B\right] = \sigma_B^2$. The hypothesis we test is

$$\mathsf{H}_0 : \mu_A = \mu_B$$

against

$$\mathsf{H}_1 : \mu_A \neq \mu_B$$

Assume all r.v. are independent, so the CLT implies

$$\sqrt{n_A}\frac{\overline{X}^A - \mu_A}{S_A} \xrightarrow{d} \mathcal{N}(0,1); \quad \sqrt{n_B}\frac{\overline{X}^B - \mu_B}{S_B} \xrightarrow{d} \mathcal{N}(0,1)$$

where S_A and S_B are the usual estimators. Also

$$\begin{aligned}
\mathsf{Cov}\left(\overline{X}^A, \overline{X}^B\right) &= \mathsf{Cov}\left(\frac{1}{n_A}\sum_{i=1}^{n_A} X_i^A, \frac{1}{n_B}\sum_{j=1}^{n_B} X_j^B\right) \\
&= \frac{1}{n_A}\frac{1}{n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \mathsf{Cov}\left(X_i^A, X_j^B\right) \\
&= 0
\end{aligned}$$

as all $\mathsf{Cov}\left(X_i^A, X_j^B\right) = 0$.
Now define the r.v.

$$Z_\Delta = \overline{X}^A - \overline{X}^B$$

so

$$\mathsf{H}_0 : \mathsf{E}[Z_\Delta] = 0$$

and

$$\begin{aligned}
\mathsf{E}[Z_\Delta] &= \mu_A - \mu_B \underset{\mathsf{H}_0}{=} 0 \\
\mathsf{Var}[Z_\Delta] &= \mathsf{Var}\left[\overline{X}^A\right] - 2\mathsf{Cov}\left(\overline{X}^A, \overline{X}^B\right) + \mathsf{Var}\left[\overline{X}^B\right] \\
&= \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}
\end{aligned}$$

hence from the CLT, since both $\overline{X}^A$ and $\overline{X}^B$ are asymptotically normal,

$$Z_\Delta^* = \frac{Z_\Delta}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} \underset{\mathsf{H}_0}{\to} \mathcal{N}(0,1) \tag{5.2}$$

It follows that Z_Δ^* is a test statistic that is asymptotically $\mathcal{N}(0,1)$ distributed under H_0 and can be used for the test using standard critical values.

5.7 Test of Distributions

5.7.1 χ^2 Goodness of fit test

The χ^2 independence test allows testing whether a random variable follows a given law $\mathcal{F}$:

$$\mathsf{H}_0 : X \sim \mathcal{F} \qquad vs \qquad \mathsf{H}_1 : X \not\sim \mathcal{F}.$$

where the law $\mathcal{F}$ may need estimating some parameter θ.

Two cases arise, depending on whether $\mathcal{F}$ takes discrete or continuous values

1. For a discrete law, let p_i the probability that X takes value x_i, for $i = 1, ..., r$ under H_0 (i.e. the theoretical distribution). Define, for a sample of size n, the distance
$$d^2\left(\mathcal{F}_n, \mathcal{F}\right) = \sum_{i=1}^{r} \frac{\left(n_i - np_i\right)^2}{np_i}$$
where n_i is the number of observations that equal x_i.
2. If the distribution can take continuous values with density function f then one must group the possible observations into r categories (intervals) $k_i = [c_{i-1}, c_i]$ for some carefully chosen values c_i, with $i = 0, ..., r$. We then define the theoretical probabilities
$$p_i = \int_{\kappa_{i-1}}^{\kappa_i} f\left(u\right) du$$
Note that the intervals must be chosen so that $\sum_{i=1}^{r} p_i = 1$. The last interval may be $[c_{r-1}, +\infty[$ in which case $p_r = 1 - \sum_{i=1}^{r-1} p_i$.

Under H_0, $d^2\left(\mathcal{F}_n, \mathcal{F}\right)$ is a realization of random variable which is asymptotically χ^2_{r-1} distributed.

If establishing $\mathcal{F}$ needs the estimation of q parameters, then the asymptotic law is χ^2_{r-1-q}.

Note that as in the test of independence, the theoretical expected number of realizations within each interval, i.e. np_i, must be greater than 5. If it is not the case, then intervals must be merged to increase the probabilities.

5.8 Exercises

Exercise 1 *Study of the test $\mathsf{H}_0 : \mu = \mu_0$ and $\mathsf{H}_1 : \mu < \mu_0$*
You are in charge of purchases for your enterprise. Your main supplier of metallic wires has sold you a batch where the average breaking point is guaranteed to be 80 kg, with a standard deviation of 4 kg.

A sample of 64 wires has been selected, and it has been reported to you that the average breaking point is 79 kg.
With an error risk equal to 5%, are you going to accept the batch of wires, or are you going to send it back to the supplier if it does not respect the contractual specifications? You should proceed as follows:

1 *Fix the null and the alternative hypotheses.*
2 *Obtain the decision rule; then conclude*

Answer**.** *You should reject* H_0 *since* $\overline{x} < 79.18$ *and your decision should be:* with an error risk of 5%, send the batch back.

Exercise 2 (Test on a proportion) ***(+)*** *The pharmaceutical enterprise Beta has recently developed a new medicine against allergy symptoms. The scientific director assures that the medicine works on 90% of the cases, on a period of 8 hours. In order to verify this affirmation, the medicine has been tested over a sample of 200 people showing allergy symptoms. The medicine worked on 160 cases. What is your conclusion?*

Exercise 3 (Test on a proportion (2)) ***(+) Publishing Group AES***
The editorial group AES (Avenir et Société) is specialized in publishing scientific books and magazines. One of these magazines, Sciences du Futur, is only sold to members. The sales direction wishes to advertise in a segment of customers from medical professions by sending membership offers at promotional prices. For this reason, the direction is planning to acquire the list of members of the magazine CADUCOR.
CADUCOR declares that past experience shows that 8 to 12% of the doctors in the list answer positively to mail offers (memberships, books, objects, etc.). After a first evaluation, AES estimates that the list will be interesting if at least 10% answers.

1. *Define the population, the descriptive variable and the parameter object of the study.*
2. *Define the problem as a statistical test. Give the general formulation of the acceptance and rejection regions for the null hypothesis* H_0*. Interpret the two error types.*
3. *AES wishes to fix the probability of type I error at* α*=0.05. Define the rejection region if the sample size is 400.*
4. *A membership offer has been sent to 400 doctors; 58 have responded positively. Given this result, should AES buy CADUCOR?*

Chapter 6
Confidence Intervals

6.1 Motivations

We have focused so far on point estimation: we obtained one realization of the estimator.

$\hookrightarrow$ From a given estimation, what is the range of possible values?

Now, we wish to associate a measure or reliability to the numerical value yielded on the sample; this measure shall give information on the most likely values for the parameter θ.

$\hookrightarrow$ It is then natural to determine confidence intervals for a parameter θ.

- How do we build a confidence interval?
- How do we judge the optimality of a confidence interval?

More Generally: what is the accuracy of an estimation, i.e.

what is the order of magnitude of the difference (deviation) between the estimator $\widehat{\theta}$ and the parameter θ that we can expect?

$\hookrightarrow$ Confidence interval estimation

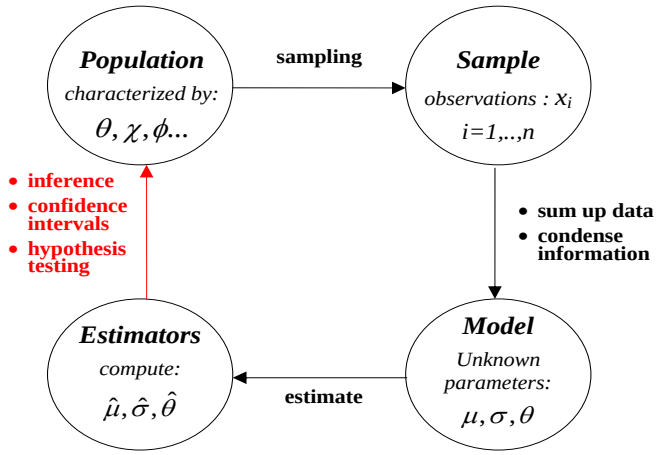
Example: *During the project phase of a new car, the engineers would like to know the height of men, so as to define the measures for the seats. A mean value has been calculated over a sample. What range of values should we take into account?*

$\hookrightarrow$ Hypothesis testing

Example: *After a new advertisement campaign, an increase in sales of 5% has been observed. The managers would like to find out if such an increase is only an artefact or whether it may be considered as an effect after the sales campaign.*

In both cases, the leap from the sample to the population never leads to certainty. For this reason, an evaluation of the reliability of results, in terms of probability, must always be associated to all conclusions. Such an evaluation will be based on probability laws.

We study estimation uncertainty using the distribution of the r.v. estimator, function of the $(X_i)_i$ as in the figure:



The term inference refers to what can be said about the distribution of the r.v. that generated the data. In particular the uncertainty surrounding estimation.

6.2 Definitions

Question: From an estimator $\widehat{\theta}$ of a parameter θ, can we find a set $\mathcal{C}\left(\widehat{\theta}\right)$ such that there is a 'good' chance that the true value $\theta = \theta_0 \in \mathcal{C}\left(\widehat{\theta}\right)$?

To define $\mathcal{C}\left(\widehat{\theta}\right)$, we set a **confidence level** p such that we wish to find $\mathcal{C}\left(\widehat{\theta}\right)$ that satisfies

$$\begin{aligned} \mathsf{P}\left(\theta \in \mathcal{C}\left(\widehat{\theta}\right)\right) &= p = 1-\alpha \\ \mathsf{P}\left(\theta \notin \mathcal{C}\left(\widehat{\theta}\right)\right) &= \alpha \text{ (RISK)} \end{aligned} \tag{6.1}$$

Examples: choosing $p = 95\%$ means taking a risk of $\alpha = 5\%$; or $p = 99\%$ with a risk of $\alpha = 1\%$.

For a confidence level that is chosen by the modeler, it is possible to derive a probability interval for θ to be in $\mathcal{C}\left(\widehat{\theta}\right)$ using the distribution of $\widehat{\theta}$.

Similarly, we can find an interval $\mathcal{C}^*\left(\widehat{\theta}\right)$ such that

$$\mathsf{P}\left(\widehat{\theta} \in \mathcal{C}^*\left(\theta\right)\right) = p = 1-\alpha \tag{6.2}$$

Expressions (6.1) and (6.2) are two different representations which are related to the definition of the estimator $\widehat{\theta}$.

$\hookrightarrow$ For scalar parameters and estimators, $\mathcal{C}$ **is an interval**, which we will write

$$\mathcal{C} = [L, U]$$

Definition 15 (Confidence Interval) *A confidence interval at probability p for a parameter θ is a probability interval $[L_p, U_p]$ at p around the parameter θ*

$$\mathsf{P}(\theta \in [L_p, U_p]) = p$$

The interval above is two-sided (or bilateral). It is also possible to construct intervals such that

$$\mathsf{P}(\theta > U_p) = p, \text{ or, } \mathsf{P}(\theta < L_p) = p$$

$\hookrightarrow$ To compute (L_p, U_p), we will most probably need assumptions concerning the distribution of $\widehat{\theta}$, and hence of $(X_1, ..., X_n)$. In particular, we will see that the **quantiles** of the distribution of $\widehat{\theta}$ will come into play.

Note that the pair (L_p, U_p) is function of the problem at hand and of the chosen estimation method. A realization (l_p, u_p) of (L_p, U_p) can be computed from the sample $(x_1, .., x_n)$ which is itself a draw from $(X_1, ..., X_n)$.

6.3 Methodology when the estimator is the sample mean

6.3.1 Confidence interval for μ where the parent distribution is $X \sim \mathcal{N}(\mu, \sigma^2)$, where σ is known

This case is of interest since we know from the CLT that the normal distribution arises naturally when the distribution of the parent r.v. is unknown.

Example 16 *Assume that you dispose of a sample of $n = 4$ observations $(X_1, ..., X_4)$ that you assume drawn independently from a $\mathcal{N}(\mu, 1)$. Give a confidence interval for μ at probability $p = 0.95$ for $\overline{x} = 2.10$.*

1. *Statistical problem: from the 4-sample $(X_i)_{i=1,...,4}$ with parent law $X \sim \mathcal{N}(\mu, 1)$ on $(\Omega, \mathcal{A}, \mathsf{P})$ we construct the estimator of μ as $\overline{X}_4$, the empirical mean, with distribution $\overline{X}_4 \sim \mathcal{N}\left(\mu, \sigma^2_{\overline{X}_4}\right)$ where*

$$\sigma^2_{\overline{X}_4} = \frac{1}{16} \sum_{i=1}^{4} \mathsf{Var}[X_i] = 1/4.$$

2. *Let* $Z = \left(\overline{X} - \mu\right) / \sigma_{\overline{X}}$*, then*

$$Z \sim \mathcal{N}(0, 1)$$

Z admits $\widetilde{z}_{\alpha/2} = -1.96$ *and* $\widetilde{z}_{1-\alpha/2} = +1.96$ *as quantiles at probabilities* 0.025 *and* 0.975 *respectively. Hence*

$$\mathsf{P}\left(Z \in [-1.96, 1.96]\right) = 0.95$$

3. *From the definition of* Z*:*

$$\begin{aligned}
\mathsf{P}\left(Z \in [-1.96, 1.96]\right) &= \mathsf{P}\left(\frac{\overline{X}_4 - \mu}{1/\sqrt{4}} \in [-1.96, 1.96]\right) \\
&= \mathsf{P}\left(\overline{X}_4 - \mu \in [-0.98, 0.98]\right) \\
&= \mathsf{P}\left(\mu \in \left[\overline{X}_4 - 0.98, \overline{X}_4 + 0.98\right]\right)
\end{aligned}$$

Therefore

$$\mathsf{P}\left(\mu \in \left[\overline{X}_4 - 0.98, \overline{X}_4 + 0.98\right]\right) = 95\%$$

We deduce a ***realization*** *of a 95% confidence interval (a random interval), namely* $[1.12, +3.08]$.

Remark: the range of this interval is large, hence it is not very informative. We could choose a larger risk, α, which would reduce the interval but also the probability. The sample size $n = 4$ needs be larger in order to get a reasonable precision.

In a symmetric approach for $\overline{X}$ or μ, $\mathsf{P}\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \in \left[\widetilde{z}_{\alpha/2}, \widetilde{z}_{1-\alpha/2}\right]\right) = 1 - \alpha$ implies either of

$$\begin{aligned}
\mathsf{P}\left(\mu \in \left[\overline{X} - \widetilde{z}_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} - \widetilde{z}_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) &= 1 - \alpha \text{ (Conf. Interval.)} \\
\mathsf{P}\left(\overline{X} \in \left[\mu + \widetilde{z}_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + \widetilde{z}_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) &= 1 - \alpha \text{ (Prob. Interval.)}
\end{aligned}$$

Here $\mathsf{P}\left(\overline{X} \in [\mu - 0.98, \mu + 0.98]\right) = 95\%$, but we need to know μ.

Note that for a one-sided interval at probability 0.95, we need the quantiles at α and/or $1 - \alpha$, i.e. ± 1.65. Then one-sided confidence intervals at 5% risk are

$$\begin{aligned}
\mathsf{P}\left(\mu > \overline{X}_n + 1.65\sigma/\sqrt{n}\right) &= \mathsf{P}\left(\mu \in \left(\overline{X}_n + 1.65\sigma/\sqrt{n}, +\infty\right)\right) = 0.05 \\
\mathsf{P}\left(\mu < \overline{X}_n - 1.65\sigma/\sqrt{n}\right) &= \mathsf{P}\left(\mu \in \left(+\infty, \overline{X}_n - 1.65\sigma/\sqrt{n}\right)\right) = 0.05
\end{aligned}$$

6.3.2 Confidence interval for μ where $X \sim \mathcal{N}(\mu, \sigma^2)$, σ unknown

We will now need to work with the Student T_n distribution. The expectation and variance of the Student distribution are

$$\begin{aligned} \mathsf{E}[T_n] &= 0 \text{ for } n > 1 \\ \mathsf{V}[T_n] &= \frac{n}{n-2} \text{ for } n > 2 \end{aligned}$$

and, $T_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$. In the case $n = 1$, the distribution T_1 is called a Cauchy distribution and it has the property that both its expectation and variance are infinite. More generally the T_n distribution arises as the ratio

$$\frac{U}{\sqrt{V/n}}$$

where $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_n$ such that U and V are independent.

To construct a confidence interval for μ in the case where $X \sim \mathcal{N}(\mu, \sigma^2)$, we proceed as when σ^2 is known, but replacing σ^2, when we need it, with an unbiased estimator:

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

It can be shown that

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Now, since

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$

hence

$$T = \frac{Z}{\sqrt{(n-1)\frac{S_n^2}{\sigma^2}/(n-1)}} = \frac{\overline{X} - \mu}{S_n/\sqrt{n}} \sim T_{n-1}. \tag{6.3}$$

We now work with T instead of Z and obtain (defining the quantiles of the

T_{n-1} distribution as $\widetilde{t}_{\alpha/2,n-1}$ and $\widetilde{t}_{1-\alpha/2,n-1}$):

$$\begin{aligned} & P\left(\widetilde{t}_{\alpha/2,n-1} < T < \widetilde{t}_{1-\alpha/2,n-1}\right) \\ = \; & P\left(\overline{X} - \frac{S_n}{\sqrt{n}}\widetilde{t}_{1-\alpha/2,n-1} < \mu < \overline{X} - \frac{S_n}{\sqrt{n}}\widetilde{t}_{\alpha/2,n-1}\right) \\ = \; & 1-\alpha \end{aligned}$$

which provides a confidence interval for μ. For large n ($n \geq 30$), $\widetilde{t}_{\alpha/2,n-1}$ can be approximated by $\widetilde{z}_{\alpha/2}$, the quantiles of the normal distribution.

6.3.3 Confidence interval for the expectation, μ, of the parent distribution when the latter is unknown

When the distribution of the data is unknown, the only solution is to work with large samples so that the CLT applies. In practice, it suffices that $n > 30$. But when estimating a proportion p, we also need to check that $np(1-p) > 5$, where p is the true value (which we approximate by its estimate).

If these requirements are satisfied then we use $\sqrt{n}\frac{\overline{X}-\mu}{\sigma} \xrightarrow{\text{d}} \mathcal{N}(0,1)$ as $n \to \infty$. If we do not know σ, we must estimate it via S_n.

6.3.4 Example: confidence interval for a proportion

Consider the following example

Example 17 *"Are you personally in favor of the European Union?" In a random phone survey (simple draws with even probabilities and a survey rate below 10% of the whole population), 171 respondents out of 300 answer yes, the rest says no.*

Question: How is it possible to estimate the true proportion for the whole population at a confidence level of $p = 95\%$?

To construct a confidence interval at a given level p, for an estimator $\widehat{\theta}$, we must first find the distribution of the r.v. $\widehat{\theta}$, a function of the sample $(X_1, ..., X_n)$. Here since respondents can only answer yes $(x = 1)$ or no $(x = 0)$ (binary response), the distribution is Bernoulli.

We assume the following:
1. *each individual is representative of the whole population, i.e. all respondents have identical distribution and the probabilities of their answers are that of the national electorate*
2. *the responses are independent across individuals*

So, we consider the sample $(X_1, ..., X_n)$ of $i.i.d.$ r.v. with parent $X \sim B(q)$ with probability mass function

$$f_X(x) = \mathsf{P}(X = x) = q^x (1-q)^{(1-x)}, \; x \in \{0, 1\}.$$

- **What we estimate:** *The proportion of voters in favor of a new constitution, i.e.* q.

Since we know that $q = E[X]$, we can consider the empirical mean as an estimator of q : $\widehat{Q} = \overline{X}_n$. We compute an estimate from the data:

$$\widehat{q} = 171/300 \simeq 0.57.$$

We know that $\widehat{Q}$ defined as above is unbiased and that its variance is

$$\mathsf{Var}\left(\widehat{Q}\right) = \mathsf{Var}\left(\overline{X}_n\right) = \frac{\mathsf{Var}(X)}{n} = \frac{q(1-q)}{n}$$

so the CLT implies

$$\frac{\widehat{Q} - q}{\sqrt{\mathsf{Var}\left[\widehat{Q}\right]}} = \sqrt{n}\frac{\widehat{Q} - q}{\sqrt{q(1-q)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Since we do not know q, and hence $\mathsf{Var}\left[\widehat{Q}\right]$, we use the LLN $\overline{X}_n \xrightarrow{p} \mathsf{E}[X]$, so that .

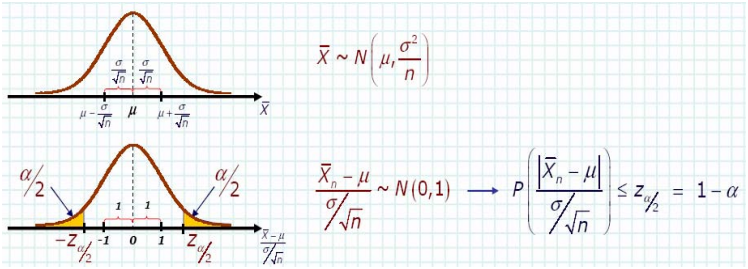
$$\widehat{Q}\left(1 - \widehat{Q}\right) \xrightarrow{p} q(1-q)$$

together with Slutsky's formula to ensure that

$$\sqrt{n}\frac{\widehat{Q} - q}{\sqrt{\widehat{Q}\left(1 - \widehat{Q}\right)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Remark: In the case of Bernoulli distributions, we can only use the CLT for $\mathcal{B}(q)$ if $n > 30$ and $nq(1-q) > 5$.

Looking at the sampling distribution of the mean from a different perspective



To construct a confidence interval, we transform the estimator $\widehat{Q} = \overline{X}_n$ to obtain a distribution that does not depend on any parameters. We then look at the *tail probability*, i.e. how likely it is for the realization of the r.v. to be away from its expectation.

Introducing the quantiles of order r of the standard normal distribution $\widetilde{z}_r$, for $r \in [0,1]$, we can construct a probability interval around $\widehat{\theta}$ or a confidence interval around θ.

1. For instance, for a symmetric two-sided interval, we know that $\mathsf{P}\left(Z \in \left[\widetilde{z}_{\alpha/2}, \widetilde{z}_{1-\alpha/2}\right]\right) = 1-\alpha$ if $Z \sim \mathcal{N}(0,1)$. Hence

$$\mathsf{P}\left(\sqrt{n}\frac{\widehat{Q}-q}{\sqrt{\widehat{Q}\left(1-\widehat{Q}\right)}} \in \left[\widetilde{z}_{\alpha/2}, \widetilde{z}_{1-\alpha/2}\right]\right) \underset{n\to\infty}{\to} 1-\alpha$$

Numerical application for a 5% risk, $\alpha = 0.05$: $\widetilde{z}_{1-\alpha/2} = 1.96 = -\widetilde{z}_{\alpha/2}$ and

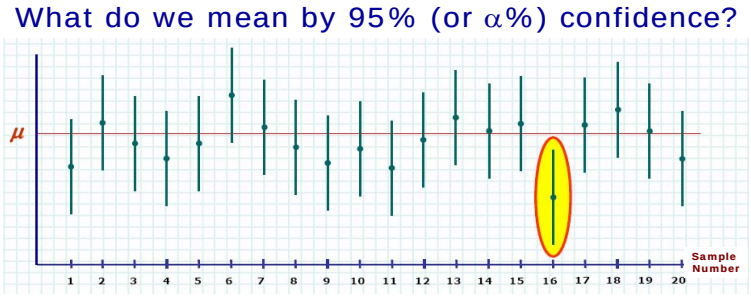
$$\mathsf{P}\left(q \in \left[\frac{51.4}{100}, \frac{62.6}{100}\right]\right) \approx 95\%$$

i.e. with a 95% level of confidence, a majority of the electorate is in favor of the European Union.

For a one-sided interval, we would consider $\widetilde{z}_\alpha$ or $\widetilde{z}_{1-\alpha}$, e.g.

$$\mathsf{P}\left(\sqrt{n}\frac{\widehat{Q}-q}{\sqrt{\widehat{Q}\left(1-\widehat{Q}\right)}} \in \left(-\infty, \widetilde{z}_{1-\alpha}\right]\right) \underset{n\to\infty}{\to} 1-\alpha$$

6.3.5 Generalities about confidence intervals



When we say that the confidence interval is for a probability $1-\alpha$, it means that if we were to construct 100 intervals based on 100 similar samples, α intervals would not contain the true parameter.

Definition 16 (Accuracy of the estimation) *We call* accuracy of the estimation *the value of* ***half the length of the confidence interval.***

$\hookrightarrow$ In the previous examples, the quantity

$$\nu = \frac{S_n}{\sqrt{n}}\widetilde{t}_{1-\alpha/2,n-1}$$

is the accuracy of the estimation of μ at the confidence level $1-\alpha$.

Definition 17 (Optimal sample) *The* optimal sample size *is the number of observations required to obtain a required accuracy ν at the $1-\alpha$ confidence level*

$\hookrightarrow$ In the examples above, let n denote the unknown optimal size of the sample. The problem is to find n such that, for a given ν,

$$\nu = \frac{S_n}{\sqrt{n}}\widetilde{t}_{1-\alpha/2,n-1}$$

Approximate solution: in the formula, S_n is replaced with its value for some n large enough and $\widetilde{t}_{\alpha/2,n-1}$ with $\widetilde{z}_{\alpha/2}$.

6.3.6 Exercises

Exercise 4 *A machine produces bolts whose weight is normally distributed with mean $\mu = 63$ gr. and variance $\sigma^2 = 0.8$. Having drawn 8 bolts at random, what is the interval of values that, with a 95% probability, will comprise their mean weight? (*answer: $\mathsf{P}\left(\overline{X}_n \in [62.38, 63.62]\right) = 0.95$*).*

Exercise 5 *A machine produces bolts whose weight is normally distributed with mean μ gr. and variance $\sigma^2 = 0.8$. Having drawn 8 bolts at random, their mean weight $\overline{x} = 62.6$ gr. What is the interval of values that, with a 95% probability, comprises the population mean? (*answer: $[61.98, 63.22]$*).*

Exercise 6 *The owner of a restaurant wishes to expand in a new neighborhood. He wishes, first of all, to estimate the average monthly expenditure for restaurant dinners of the population of the neighborhood. We may suppose that the monthly expenditures of the 20,000 families of the population follow a normal distribution with unknown expectation and variance. He selects a random sample of 20 families: the mean observed in the sample is equal to 234 Euro/month and the estimated variance $s^2 = 25$.*

1. *Give a confidence interval for the population mean with a confidence level $1 - \alpha = 0.95$, then 0.99.*
2. *What would the confidence interval become if the estimates were computed over a sample of 250 families?*
3. *Instead, consider a sample of 20 families with estimate $s^2 = 80$, define a confidence interval for the population mean.*

Exercise 7 *By measuring reaction times, a psychologist has estimated a standard deviation at 0.05 sec. What is the optimal size of the sample to have a confidence level (a) equal to 0.95 and (b) equal to 0.99, knowing that the accuracy must be lower than 0.01?*
Answer: $n = \left(\widetilde{z}_{1-\alpha/2}\sigma/\nu\right)^2$, *i.e. 96.04 ($n = 97$) and 166.4 ($n = 167$) at 95% and 99% respectively.*

Exercise 8 ***Optimal Sample for a proportion***
An airline company wishes to establish the number of business seats on a new line. On a test flight, 20 passengers have chosen a business ticket, over 118 passengers.

1. *How many business seats should be created by the company with a confidence level of 95%?*
   *(*answer: $[10.21\%, 23.79\%]$*)*
2. *Do you think that the sample has the optimal size for a precision of 0.05?*
   *(*answer: optimal sample size under the worse case scenario of proportion $p = 0.5$: $n = 384 > 118$*)*

PART III

Linear Regression

Chapter 7
Least Squares

7.1 Introduction

- Whenever two quantitative variables are available, we may be interested in studying the dependence relationship between these two variables by means of a mathematical model
- The objective of the mathematical model is to study the behavior of the logically dependent variable as a function of the explanatory variable
- Exact deterministic versus approximate (more realistic) probabilistic models
- Understanding the past
- Forecasting the future

7.2 Statistical Interpolation

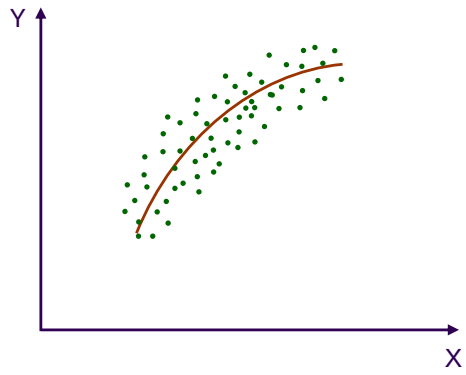
Let us give a series of n pairs of numbers (x_i, y_i) corresponding to the observation of two numerical variables on n statistical units. When there are many points, most likely they will form a **cloud of points** on the plan (scatter diagram).

In such cases, mathematical interpolation is replaced by statistical interpolation, where the constraint of touching all points is abandoned in favor of a more realistic condition: to pass amongst the given points.

While in mathematical interpolation there is a unique solution related to the system of equations, in statistical interpolation there exist infinite interpolating curves that actually pass amongst the given points.

It is necessary to establish some conditions to be satisfied by the interpolating function so as to uniquely define the problem.

In statistical interpolation there is no fixed relation between the number of parameters and the number of points. It is enough that the number of points is larger than the number of parameters.



- Polynomial Function

As the observed and the interpolated values are different, they need to be clearly distinguished as:

y_i: the observed y-values for a given value x_i
$\widehat{y}_i$: the computed y-values for a given value x_i

The theory of statistical interpolation is quite simple when the interpolating function is a complete k-th degree polynomial function, represented by the expression:

$$\widehat{y} = b_0 + b_1 x + b_2 x^2 + ... + b_k x^k$$

where $b_0, b_1, ..., b_k$ are constant values representing the curve parameters.

Functions comprising all first degree (power = 1) parameters are defined to be linear in the parameters. Functions whose parameters have exponents other than 1 are called nonlinear.

7.3 Ordinary Least Squares Method

This method dates back to Gauss, 1795 and Legendre, 1805.

- Define the interpolating function:

$$\begin{aligned}\widehat{y} &= b_0 + b_1 x + b_2 x^2 + ... + b_k x^k \\ &= \phi\left(x; b_0, b_1, ..., b_k\right)\end{aligned}$$

- The Least Squares condition determines the unknown parameters so as to minimize the sum of squares of deviations between interpolated and observed values: we choose $b_0, b_1, ..., b_k$ to find the minimum of

$$S = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

i.e. we look for

$$\min_{b_0, b_1, ..., b_k} \sum_{i=1}^{n} (y_i - \phi(x_i; b_0, b_1, ..., b_k))^2$$

The quantity to be minimized, S, is a function of $k+1$ unknown parameters. A necessary (but not sufficient) condition to achieve the solution is that the $k+1$ partial derivatives of S with respect to its parameters are zero.

7.3.1 Fitting a Straight Line (the simple regression)

- The Linear Interpolating function is

$$\begin{aligned} \widehat{y}_i &= \phi(x_i; b_0, b_1) \\ &= b_0 + b_1 x_i \end{aligned}$$

we need to choose (b_0, b_1) to obtain the **maximum fit**. This is done in the Ordinary Least Squares methodology but choosing (b_0, b_1) that minimize S, where

$$S = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- The first order conditions are defined by setting the partial derivatives to zero

$$\frac{\partial S}{\partial b_0} = 0; \quad \frac{\partial S}{\partial b_1} = 0$$

with here

$$\begin{cases} \frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \\ \frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i \end{cases}$$

- The second-order conditions are

$$\begin{aligned}\frac{\partial^2 S}{\partial b_0^2} &= 2n > 0,\\ \frac{\partial^2 S}{\partial b_1^2} &= 2\sum_{i=1}^{n} x_i^2 > 0\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 S}{\partial b_0^2}\frac{\partial^2 S}{\partial b_1^2} - \left(\frac{\partial^2 S}{\partial b_0 b_1}\right)^2 &= 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2\\ &= 4n\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 > 0\end{aligned}$$

so the Hessian

$$\begin{bmatrix} \frac{\partial^2 S}{\partial b_0^2} & \frac{\partial^2 S}{\partial b_0 b_1} \\ \frac{\partial^2 S}{\partial b_0 b_1} & \frac{\partial^2 S}{\partial b_1^2} \end{bmatrix} = 2\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

is positive definite and the first-order conditions define a minimum.

- We solve the system

$$\begin{cases} -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) = 0 \\ -2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) x_i = 0 \end{cases}$$

$$\Rightarrow \begin{cases} n\overline{y} - nb_0 - b_1 n\overline{x} = 0 \\ \sum_{i=1}^{n} y_i x_i - b_0 n\overline{x} - b_1\sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\Rightarrow \begin{cases} b_0 = \overline{y} - b_1\overline{x} \\ \sum_{i=1}^{n} y_i x_i - \left(\overline{y} - b_1\overline{x}\right) n\overline{x} - b_1\sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\Rightarrow \begin{cases} b_0 = \overline{y} - b_1\overline{x} \\ \sum_{i=1}^{n} y_i x_i - n\overline{x}\overline{y} = b_1\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \end{cases}$$

so

$$\begin{aligned} b_0 &= \overline{y} - b_1\overline{x}\\ b_1 &= \frac{\sum_{i=1}^{n} y_i x_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\end{aligned}$$

We can go further noticing that

$$\frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

The final result is therefore

$$b_0 = \overline{y} - b_1\overline{x} = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

and our model of linear regression gives us:

$$\widehat{y_i} = \left(\overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}\right) + \frac{s_{xy}}{s_x^2} x_i$$

which we can also write

$$\widehat{y_i} = \overline{y} + \frac{s_{xy}}{s_x^2}(x_i - \overline{x})$$

- In conclusion: the OLS (ordinary least-squares) method of estimation yields the following
  - ► Slope for the Estimated Regression Equation:

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

  - ► y-intercept for the Estimated Regression Equation:

$$b_0 = \overline{y} - b_1\overline{x} = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}$$

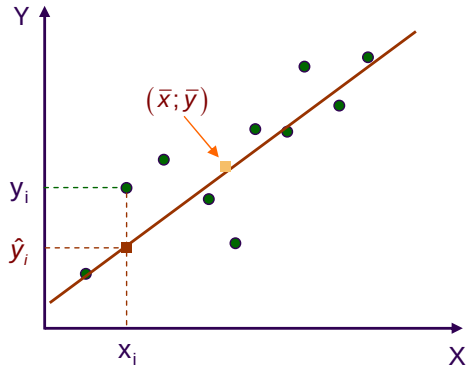
    where:
    x_i = value of explanatory variable for i-th observation
    y_i = value of dependent variable for i-th observation
    $\overline{x}$ = mean value for explanatory variable
    $\overline{y}$ = mean value for dependent variable
    n = total number of observations.

7.3.1.1 Interpretation of the OLS line

- X : Explanatory variable
  Y : Dependent variable, or variable to explain
- $b_0 = \overline{y} - b_1\overline{x}$: It is the **intercept** on the y-axis. It can be interpreted as the value of Y when $X = 0$ (if this is meaningful for the specific application at hand).
  - ► this also implies:

$$\overline{y} = b_0 + b_1\overline{x}$$

    The point with coordinates $(\overline{x}, \overline{y})$ is a point of the regression line. The regression line goes through the barycenter of the cloud of points if b_0 is different from 0.
- $b_1 = \frac{s_{xy}}{s_x^2}$: it is the slope (regression coefficient) of the straight line as it is a function of the angle between the regression line and the x-axis. It expresses the inclination (positive, negative or null) of the regression line.
  - ► Statistically speaking, it is interpreted as the expected (mean) change of the Y-variable when there is a unitary change of the X-variable.



7.3.2 Residuals

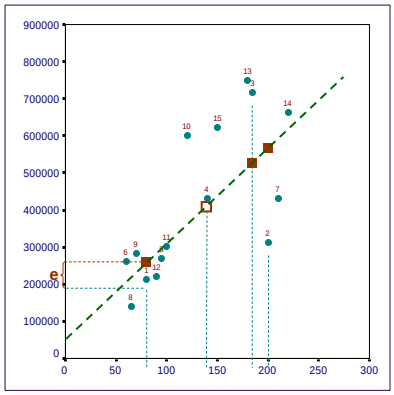
- Model predictions are not perfect, we denote the difference between observed values of y_i and our model predictions as **residuals**
- The residuals are the part of y_i that cannot be explained by our model

$$e_i = y_i - \widehat{y_i}$$

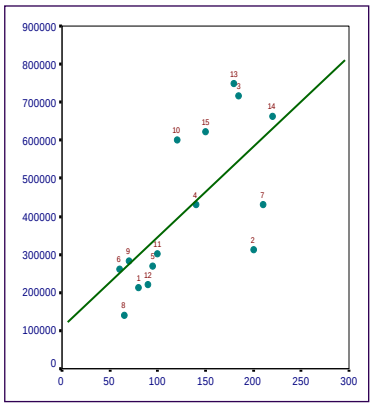
- e_i is the modeling error, it is equal to

$$e_i = y_i - b_0 - b_1 x_i$$

Flat	Sq.m. (X)	Price in € (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	80	212,000	282,107	-70,107
2	200	313,000	593,227	-280,227
3	185	717,000	554,337	162,663
4	140	431,000	437,667	-6,667
5	95	270,000	320,997	-50,997
6	60	261,000	230,254	30,746
7	210	431,000	619,154	-188,154
8	65	140,000	243,217	-103,217
9	70	282,000	256,181	25,819
10	120	600,000	385,814	214,186
11	100	303,000	333,961	-30,961
12	90	220,000	308,034	-88,034
13	180	749,000	541,374	207,626
14	220	663,000	645,081	17,919
15	150	623,000	463,594	159,406
	1,965	6,215,000	6,215,000	0



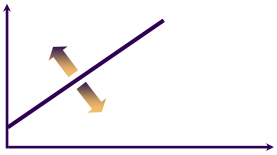
Residuals



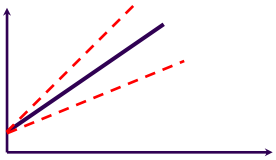
In order to make the prediction more accurate, we can narrow the field of application of the model by removing some anomalous observations (outliers, i.e. showing "important" values for the residuals), but we need to justify the removal by means of extra-statistical considerations.

Outliers' effect

1. The *intercept* may change



2. The *slope* may change



Note: when we introduce the *statistical regression model* we will make a difference between the *model errors* and *the residuals.*

7.3.3 Application to the "Flats in Naples" example

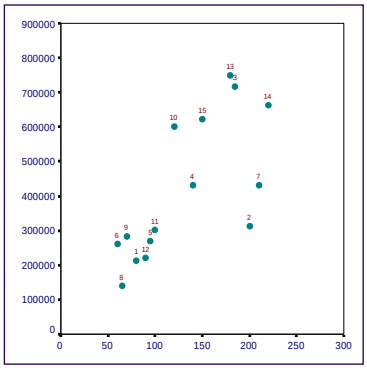
X : Explanatory variable (Surface of a Flat in Naples, Italy)

Y : Dependent variable (Price of a Flat in Naples, Italy)

X ⟶ Explanatory variable (Surface of a Flat in Naples, Italy)

Y ⟶ Dependent variable (Price of a Flat in Naples, Italy)

Flat	sq.m. (X)	Price in € (Y)	X-M(X)	Y-M(Y)	[X-M(X)]*[Y-M(Y)]
1	80	212000	-51,0	-202333,3	10 319 000
2	200	313000	69,0	-101333,3	-6 992 000
3	185	717000	54,0	302666,7	16 344 000
4	140	431000	9,0	16666,7	150 000
5	95	270000	-36,0	-144333,3	5 196 000
6	60	261000	-71,0	-153333,3	10 886 667
7	210	431000	79,0	16666,7	1 316 667
8	65	140000	-66,0	-274333,3	18 106 000
9	70	282000	-61,0	-132333,3	8 072 333
10	120	600000	-11,0	185666,7	-2 042 333
11	100	303000	-31,0	-111333,3	3 451 333
12	90	220000	-41,0	-194333,3	7 967 667
13	180	749000	49,0	334666,7	16 398 667
14	220	663000	89,0	248666,7	22 131 333
15	150	623000	19,0	208666,7	3 964 667
	1 965	*6 215 000*	*0,0*	*0,0*	*115 270 000*



with statistics: $\overline{y} = 414333$, $\overline{x} = 131$, $s_x = 54.44$, $s_y = 197061$ and $s_{xy} = 7,684,667$. Hence

$$b_1 = 2,593$$
$$b_0 = 74,694$$

so the model is

$$\widehat{y}_i = 74,694 + 2,593x_i$$

e.g. we predict a flat of 100 square meters should cost $74,694 + 2,593 \times 100 = 333\,994$ euros.

7.3.3.1 Interpretation and Extrapolation

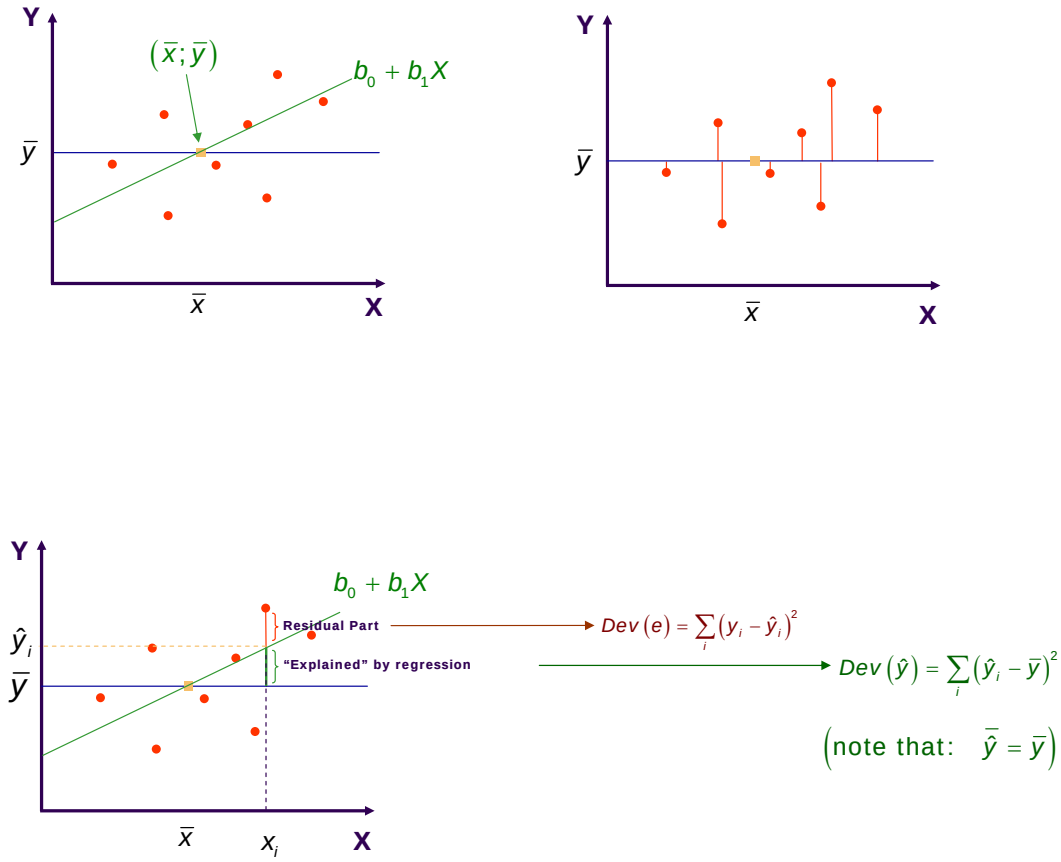
What is the "expected" price for a flat of 160 sq. m.?

$$\begin{aligned}\widehat{y} &= 74693.88 + 2592.67 \times 160 \\ &= 489520.7\end{aligned}$$

What is the **expected** price for a flat of **400** sq. m.?

Extrapolation might be irrelevant!

7.3.4 Goodness of Fit: the R^2 index



Definition 18 *The linear determination index (of Goodness of Fit coefficient) is defined as*

$$R^2 = \frac{\sum_i \left(\widehat{y}_i - \overline{\widehat{y}}\right)^2}{\sum_i (y_i - \overline{y})^2} = \frac{\sum_i (\widehat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}$$

We used in the previous definition that

$$\overline{\widehat{y}} = \overline{y} \tag{7.1}$$

which follows from the expression we showed before

$$\widehat{y}_i = \overline{y} + \frac{s_{xy}}{s_x^2}(x_i - \overline{x})$$

so

$$\begin{aligned}\overline{\widehat{y}} &= \overline{y} + \frac{s_{xy}}{s_x^2}(\overline{x} - \overline{x}) \\ &= \overline{y}\end{aligned}$$

and in particular

$$\overline{e} = \overline{y} - \overline{\widehat{y}} = 0 \tag{7.2}$$

Importantly, the R^2 coefficient is always between zero and one,

$$0 \leq R^2 \leq 1$$

and the closer to unity it is, the better the fit of the regression.

Proof. We show that

$$\sum_i (y_i - \overline{y})^2 = \sum_i (\widehat{y}_i - \overline{y})^2 + \sum_i (e_i - \overline{e})^2$$

which implies that

$$1 = \frac{\sum_i (\widehat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} + \frac{\sum_i (e_i - \overline{e})^2}{\sum_i (y_i - \overline{y})^2}$$

where $\frac{\sum_i (e_i - \overline{e})^2}{\sum_i (y_i - \overline{y})^2} \geq 0$. So $\frac{\sum_i (\widehat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} \in [0, 1]$.

First notice that

$$\begin{aligned}\sum_i (e_i - \overline{e})^2 &= \sum_i e_i^2 \\ &= \sum_i (y_i - \widehat{y}_i)^2\end{aligned}$$

and

$$\begin{aligned}\sum_i (y_i - \overline{y})^2 &= \sum_i (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\ &= \sum_i (y_i - \widehat{y}_i)^2 + \sum_i (\widehat{y}_i - \overline{y})^2 + 2\sum_i (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \\ &= \sum_i (e_i - \overline{e})^2 + \sum_i (\widehat{y}_i - \overline{y})^2 + 2\sum_i e_i (\widehat{y}_i - \overline{y}) \\ &= \sum_i (e_i - \overline{e})^2 + \sum_i (\widehat{y}_i - \overline{y})^2 + 2b_1 \sum_i e_i (x_i - \overline{x}) \\ &= \sum_i (e_i - \overline{e})^2 + \sum_i (\widehat{y}_i - \overline{y})^2 + 2b_1 \left[\sum_i e_i x_i - \overline{x} \sum_i e_i\right]\end{aligned}$$

where $\sum_i e_i = 0$ and the first-order conditions we used when differentiating to obtain b_1 where that

$$\sum_i (y_i - b_0 - b_1 x_i) x_i = \sum_i e_i x_i = 0$$

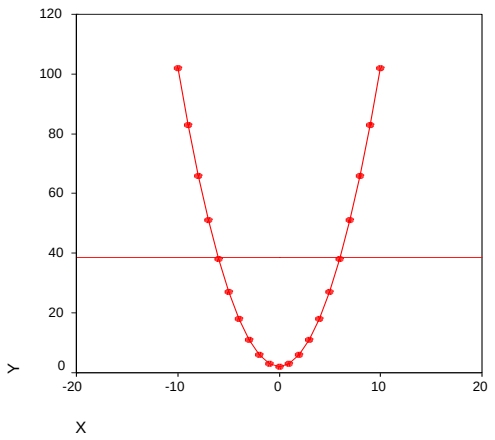
it follows that

$$\sum_i (y_i - \overline{y})^2 = \sum_i (e_i - \overline{e})^2 + \sum_i (\widehat{y}_i - \overline{y})^2$$

■

- Interpretation
  - ► R^2 measures the strength of the linear relationship between X and Y.
  - ► The closer the points to a straight line, the better the R^2.
  - ► R^2 means that X and Y are *linearly independent* but it does NOT mean X and Y are independent in general: consider the following nonlinear Model:

$$y = ax^2 + bx$$



- Application to the "Flat in Naples" example

Flat	Sq.m. (X)	Price in € (Y)	$\hat{Y}$	$(Y - \bar{Y})^2$	$(\hat{Y} - \bar{Y})^2$	$(Y - \hat{Y})^2$
1	80	212,000	282,107	40,938,777,778	17,483,727,210	4,915,031,754
2	200	313,000	593,227	10,268,444,444	32,003,085,447	78,527,393,139
3	185	717,000	554,337	91,607,111,111	19,601,133,620	26,459,127,322
4	140	431,000	437,667	277,777,778	544,475,934	44,453,442
5	95	270,000	320,997	20,832,111,111	8,711,614,942	2,600,724,704
6	60	261,000	230,254	23,511,111,111	33,885,224,478	945,320,444
7	210	431,000	619,154	277,777,778	41,951,534,609	35,401,954,462
8	65	140,000	243,217	75,258,777,778	29,280,705,777	10,653,805,643
9	70	282,000	256,181	17,512,111,111	25,012,283,333	666,640,808
10	120	600,000	385,814	34,472,111,111	813,352,938	45,875,646,835
11	100	303,000	333,961	12,395,111,111	6,459,770,030	958,561,153
12	90	220,000	308,034	37,765,444,444	11,299,556,109	7,749,978,661
13	180	749,000	541,374	112,001,777,778	16,139,342,188	43,108,537,570
14	220	663,000	645,081	61,835,111,111	53,244,368,793	321,099,637
15	150	623,000	463,594	43,541,777,778	2,426,614,965	25,410,267,386
	1,965	6,215,000	6,215,000	582,495,333,333	298,856,790,373	283,638,542,960

$\hat{y} = 74.693,88 + 2.592,67x$

Interpretations:
51.3% of the overall variability of flats' price is explained by its linear depending on the flats' surface.

The regression model is able to account for 51.3% of the price's variability (i.e. there is 48.7% of residual variability unexplained by the model).

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 = 582495333333$$

$$Dev(\hat{Y}) = \sum_i (\hat{y}_i - \bar{y})^2 = 298856790373$$

$$Dev(e) = \sum_i (y_i - \hat{y})^2 = 283638542960$$

$$R^2 = \frac{Dev(\hat{Y})}{Dev(Y)} = \frac{298856790373}{582495333333} = 0.513$$

7.3.5 Links between Regression and Correlation (*)

1. Regression coefficients with Y acting as the **dependent** variable:

$$b_0^{(y)} = \overline{y} - b_1^{(y)}\overline{x}$$
$$b_1^{(y)} = \frac{s_{xy}}{s_x^2}$$

2. Regression coefficients with X acting as the **dependent** variable:

$$b_0^{(x)} = \overline{x} - b_1^{(x)}\overline{y}$$
$$b_1^{(x)} = \frac{s_{xy}}{s_y^2}$$

- Interdependence (both variables play the same role in the relationship):

$$b_1^{(x)} \times b_1^{(y)} = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

where using the linear correlation coefficient $r_{xy} = sign\left(b_1^{(\cdot)}\right)\sqrt{\frac{s_{xy}^2}{s_x^2 s_y^2}}$:

$$\frac{s_{xy}^2}{s_x^2 s_y^2} = r^2 \leq 1$$

so when computing

$$y_i - \overline{y} = b_1^{(y)}\left(x_i - \overline{x}\right)$$
$$x_i - \overline{x} = b_1^{(x)}\left(y_i - \overline{y}\right)$$

the fact that the models are not perfect means information is lost since we do not have

$$y_i - \overline{y} \neq b_1^{(y)}\left(b_1^{(x)}\left(y_i - \overline{y}\right)\right) = b_1^{(y)} b_1^{(x)}\left(y_i - \overline{y}\right)$$

this is where the term regression comes from.

7.3.6 Limitations of this Model

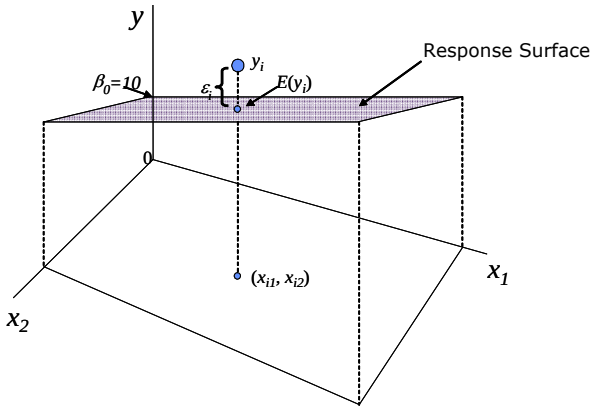
- Simple linear regression, as presented in this chapter, is a mathematical model stating a deterministic relationship between the dependent and the explanatory variables.
- The stochastic nature of relationships between real phenomena demands for a statistical model that comprises a random part (stochastic model). This will be the purpose of the second part of the book.
- Simple linear regression studies the effect of one explanatory variable on one dependent variable
  - ► a 2-dimensional space with a feasible and intuitive graphical representation.
- Most often, several explanatory variables influence and explain the behavior of the dependent variable

► higher dimensional space with unfeasible graphical representations.

7.4 Multiple linear regression

7.4.1 Introduction

- Multiple Linear Regression (MLR) is an extension of Simple Linear Regression that considers a set of $k > 1$ explanatory X-variables as predictors of the single dependent (response) y-variable.
- In case of $k = 2$ predictors:



7.4.1.1 MLR Model in Matrix Form for $k = 2$

Assume we now have two explanatory variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

which we observe over a sample of n observations. We write the model using vectors

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{1,1} \\ x_{1,2} \\ x_{1,3} \\ \vdots \\ x_{1,n-1} \\ x_{1,n} \end{pmatrix} + \beta_1 \begin{pmatrix} x_{2,1} \\ x_{2,2} \\ x_{2,3} \\ \vdots \\ x_{2,n-1} \\ x_{2,n} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_{n-1} \\ e_n \end{pmatrix}$$

and then using a matrix:

$$\mathbf{y} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_{n-1} \\ e_n \end{pmatrix}$$

which can be condensed into

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- For a multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e$$

  β_j should be interpreted as the expected change in y when a unitary change is observed in x_j, while all other predictors are kept constant. This statement is not very clear when x_j and

the other predictors are related. Think for instance of the limiting case where $x_{j'} = x_j$ for $j' \neq j$.

7.4.2 Estimation of the regression coefficients: Ordinary Least Squares again

- Notations: for $i = 1, ..., n$
  - y_i : observation of the dependent variable
  - $x_1, ..., x_k$: predictors, regressors
  - $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \widehat{\beta}_2 x_{2,i} + ... + \widehat{\beta}_k x_{k,i}$: predicted value for y_i
  - e_i : model error
  - $\widehat{e}_i = y_i - \widehat{y}_i$: residual. This is equal to

$$\begin{aligned} \widehat{e}_i &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + e_i - \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \widehat{\beta}_2 x_{2,i} + ... + \widehat{\beta}_k x_{k,i} \\ &= e_i + \left(\beta_0 - \widehat{\beta}_0\right) + \left(\beta_1 - \widehat{\beta}_1\right) x_{1,i} + \left(\beta_2 - \widehat{\beta}_2\right) x_{2,i} + ... + \left(\beta_k - \widehat{\beta}_k\right) x_{k,i} \end{aligned}$$

- **Errors** versus **residuals**: when we introduce the statistical regression model later in the course we will assume that there exist 'true' values of the

parameters $\beta_0, ..., \beta_k$ so the **error** is

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

and the **residual** is the "estimated" error

$$\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- Least Squares Method (OLS): we choose the estimates $\widehat{\beta}_0, \widehat{\beta}_1, ...\widehat{\beta}_k$ of $\beta_0, ..., \beta_k$ by minimizing the in-sample sum of squared errors

$$\begin{aligned} & \min_{\beta_0,...,\beta_k} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_k x_k\right)^2 \\ \left(\widehat{\beta}_0, \widehat{\beta}_1, ...\widehat{\beta}_k\right) \quad = \quad & \underset{(\beta_0,...,\beta_k)}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_k x_k\right)^2 \end{aligned}$$

- Using matrix notation (and differentiation) it can be shown that this amounts to choosing

$$\begin{aligned} \widehat{\boldsymbol{\beta}} \quad &= \quad \underset{\boldsymbol{\beta}}{\text{argmin}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \\ &= \quad \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} \end{aligned}$$

with corresponding estimate of the variance of the errors:

$$s_e^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \widehat{e}_i^2$$

7.5 Exercises

7.5.1 (+) Simple Linear Regression

- Dataset: “Flats in Paris”

1. CENSIER, bas de R. Mouffetard, pied-à-terre, 28m^2, tt confort. Visite vendredi, samedi, dim. 130.000 € à discuter. Facilités	2. CONTRESCARPE, imm. Ancien, pierre de taille, beau duplex caractère, 50m^2, poutres, refait neuf, 280.000 €
3. R. St-Simon, en pleine verdure, calme, plein soleil, Superbe appt 4p., 106m^2, cuis. aménagée, s. de bains moderne, chff. cent. Parfait état. Px 650.000 à discuter. Agence s'abstenir. Direct. Propriétaire.	4. RAPP 7P., 196m^2 standing, 9 fenêtres plein soleil, 800.000 €.
5. R. St André-des-Arts, beau liv + chbre, imm. XVIII^e siècle, 55m^2, 268.000 €.	6. 5^e PRES QUAIS, 7 pces, 190m^2 caractère, standing, 790.000 €
7. GOBELINS, Beau 5p., 110m^2, gd cft, soleil, 500.000 €	8. GOBELINS, et. élevé, calme, asc., 2 pièces, 60m^2, 320.000 €
9. CENSIER, très grand studio + entrée 48m^2, tt cft, ensoleillé, calme, bel imm., 250.000 €	10. PANTHEON, 7^e étage, ascenseur, grand studio 35m^2 + terrasse. Vue. 250.000 €.
11. RUE MADAME, 3P. + Serv., 86m^2, 350.000 €.	12. RUE DE SEINE, 3P., tt cft, 65m^2, calme, soleil, 300.000 €.
13. PANTHEON, bel imm., verdure, magnifique studio 32m^2, caractère, 155.000 €.	14. SEVRES BAB, 1er ét., 2P., gde cuis., bns, 52m^2, état neuf, 245.000 €.
15. MONTPARNASSE, Part. vend atelier d'artiste 40m^2, duplex, vue imprenable, tout confort, Prix 200.000 €.	16. RUE D'ASSAS, imm. gd standing, bel appart 260m^2, triple récept. + 5 ch., tt cft (travaux) 2 park., 2 ch. Serv., Prix 1.500.000 € à déb.
17. BD St-GERMAIN, 4P., 70m^2, à amén., 4^e ét., 325.000 €.	18. ILE St-LOUIS, Lux. appt., 117m^2, en duplex, gde récept., gde chambre, 2 sdb, Terras., parf. et., décor tr. bon goût, 950.000 €.
19. JUSSIEU, Charme, gd 3 pces, 90m^2, 378.000 €.	20. QUARTIER LATIN, 30m^2 à aménager, prix 78.000 €.
21. MONTPARNASSE, Imm. p.d.t., 4-5 P., 105m^2, bon état, 375.000 €.	22. RUE MAZARINE, 4^e ét., sans ascens., 52m^2 à rénover. Prix total 200.000 €.
23. CENSIER, Bel imm., 4P. 80m^2, tt cft, petits travaux, 270.000 €.	24. ASSAS LUXEMBOURG, 3P. 60m^2 s/arbres, imm. caractère, 295.000 €.
25. SUR JARDINS OBSERVATOIRE, 140m^2, grand charme, 990.000 €.	26. RUE DE SAVOIE, 4^e ét., Studio 20m^2, dche, 85.000 €. crédit possible.
27. PRES LUXEMBOURG, Bel imm., pierre de taille, Appartement 100m^2, salon, sal. à manger, 2 chbres, office, cuis., bains, chf. cent., asc., prix : 495.000 €.	28. Mo GOBELINS, studio, cuis., s. de bains, 28m^2, calme. Prix 85.000 €.

	Localisation	Surface	Prix (en milliers d'euros)
1	censier	28	130
2	contrescarpe	50	280
3	rue saint-simon	106	650
4	rapp	196	800
5	saint-andré des arts	55	268
6	5-ième, près quais	190	790
7	gobelins	110	500
8	gobelins	60	320
9	censier	48	250
10	panthéon	35	250
11	rue madame	86	350
12	rue de seine	65	300
13	panthéon	32	155
14	sèvres-babylone	52	245
15	montparnasse	40	200
16	rue d'assas	260	1500
17	saint-germain	70	325
18	ile saint-louis	117	950
19	jussieu	90	378
20	quartier-latin	30	78
21	montparnasse	105	375
22	rue mazarine	52	200
23	censier	80	270
24	assas luxembourg	60	295
25	jardins de l'observatoire	140	990
26	rue de savoie	20	85
27	près luxembourg	100	495
28	gobelins	28	85

- **Questions:**
  (a) Study the linear relationship between Price (Y) and Surface (X) of flats in Paris.
  (b) Give an interpretation of the regression coefficients.
  (c) Identify the presence of eventual outliers and explain their anomalous

behavior.

(d) Compute the expected price for a flat of 150 sq. m.

7.5.2 (+) Multiple Regression

- Dataset: “Sales”.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	**Y**
SEM.	Total Market Sales (MM)	Disc. to Wholesalers (M)	Price	Research Budget (M)	Investm. (M)	Advert. (M)	Sales Expenses (M)	Total Adv. of the Sector (M)	Total Sales (M)
1	398	138	56	12	50	77	229	98	5540
2	369	118	59	9	17	89	177	225	5439
3	268	129	57	29	89	51	166	263	4290
4	484	111	58	13	107	40	258	321	5502
5	394	146	59	13	143	52	209	407	4872
6	332	140	60	11	61	21	180	247	4708
7	336	136	60	25	-30	40	213	328	4627
8	383	104	60	21	-45	32	201	298	4110
9	285	105	63	8	-28	12	176	218	4123
10	277	135	62	11	76	68	175	410	4842
11	456	128	65	22	144	52	253	93	5741
12	355	131	65	24	113	77	208	307	5094
13	364	120	64	14	128	96	195	107	5383
14	320	147	66	15	10	48	154	305	4888
15	311	143	67	22	-25	27	181	60	4033
16	362	145	67	23	117	73	220	239	4942
17	408	131	66	13	120	62	235	141	5313
:	:	:	:	:	:	:	:	:	:

which contains data over $n = 38$ semesters about

- y: dependent variable is Sales
- $x_1, ..., x_8$ 8 explanatory variables:

- **Question:** assume your experts predict for the next semester the following scenario

$$\begin{aligned} PRICE &= 83 \quad RESEARCH = 30 \\ INVESTMENT &= 50 \quad ADVERTIZING = 90 \\ EXPENSES &= 300 \quad TOTAL\ ADVERTIZING = 200 \\ MARKET &= 500 \quad DISCOUNT = 100 \end{aligned}$$

What is the predicted value for Total Sales?

- Simple Correlation Coefficients

Correlation Matrix

	Correlation								
	Marché total	Remises aux grossistes	Prix	Budget de recherche	Investissements	Publicité	Frais de ventes	Total publicité de la branche	Ventes
Marché total	1.000	-.069	.549	.164	.144	.200	.903	-.020	.721
Remises aux grossistes	-.069	1.000	.022	.010	-.093	-.120	-.050	-.146	-.084
Prix	.549	.022	1.000	.455	-.058	.255	.625	-.181	.287
Budget de recherche	.164	.010	.455	1.000	.157	.105	.364	-.128	.084
Investissements	.144	-.093	-.058	.157	1.000	.241	.216	-.123	.453
Publicité	.200	-.120	.255	.105	.241	1.000	.134	-.195	.568
Frais de ventes	.903	-.050	.625	.364	.216	.134	1.000	-.022	.637
Total publicité de la branche	-.020	-.146	-.181	-.128	-.123	-.195	-.022	1.000	-.096
Ventes	.721	-.084	.287	.084	.453	.568	.637	-.096	1.000

- First Steps of the Analysis
  (a) What would the best choice be for a simple linear regression model?
  (b) Write down the equation for the multiple regression model with all explanatory variables included.
  (c) Run this model with XLStat and compare its R^2 to the one obtained with the best simple regression model previously chosen. Give a comment.
  (d) Give an interpretation of the estimated regression coefficients.
  (e) Use the final model to predict the Sales for the scenario given for next semester
- Limitations?
  (a) How confident are you about the precision of the model?
  (b) How confident are you that each variable that you used actually matters?
  (c) Do you think you could use any explanatory variable?
  (d) How could you discriminate between the variables that matter and those that do not?
  (e) Give an interpretation of the regression coefficients, if you can.
- We will revisit this example in chapter 8.

Chapter 8
The Statistical Regression Model

This chapter provides an overview of the analysis of the regression model. It provides a formal analysis of the linear regression model. This allows to evaluate the uncertainty surrounding the predicted values, the parameters and also to test whether the variables that we propose as *explanatory* are at all relevant.To understand and use it, you need elements from all previous chapters. This chapter is very important as it gives the conditions for which the regression provides estimators that are consistent and unbiased.

In addition we show below how to construct a confidence interval for the parameters of the regression and how to test whether a variable has significant explanatory power (i.e. whether we can use it in the regression).

8.1 Statistical Model and Assumptions

- Each observed value y_i is considered as a realization of a random variable Y_i defined by :

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

  where x_i is itself considered to be the realization of X_i and ε_i is a random variable
- The question is once we observe x_i, what can we say about the distribution of Y_i.
  - ► This means we are attempting to model the conditional distribution of Y_i given that $X_i = x_i$.
  - ► This is in fact an assumption about the **joint distribution** of (Y_i, X_i) which is assumed to be bivariate normal. We summarize it here: if

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \right)$$

    then the conditional distribution of Y given X is also Gaussian:

$$Y|X \sim \mathcal{N}\left(\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}\left(X - \mu_X\right), \sigma_X^2\left(1 - \rho_{XY}^2\right) \right)$$

Define

$$\begin{aligned}\alpha &= \mu_Y - \beta\mu_X \\ \beta &= \rho_{XY}\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2}\end{aligned}$$

then

$$\mathsf{E}[Y|X] = \alpha + \beta X$$

and define $\varepsilon = Y - \mathsf{E}[Y|X]$, then, letting $\sigma_\varepsilon^2 = \sigma_X^2(1-\rho_{XY}^2)$

$$\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$$

so since $Y = \mathsf{E}[Y|X] + \varepsilon$

$$Y = \alpha + \beta X + \varepsilon$$

In fact, we can also extend the previous analysis to considering several "explanatory variables" in the form of a vector $X = (X_1, ..., X_k)$. If (X, Y) is Gaussian, then so is $Y|X$ and so on.

- The fact that the conditional expectation is a linear function of X_i in $\mathsf{E}[Y_i|X_i] = \alpha + \beta X_i$ is a direct consequence of the assumption that the variables are jointly Gaussian.
  - ► Yet we can in more general settings assume directly that the conditional expectation is linear (we will have to see whether this makes sense as an assumption, but think of Taylor expansions, if the function $\mathsf{E}[Y_i|X_i] = f(X_i)$ is non linear, we can develop it as $\mathsf{E}[Y_i|X_i] = f(0) + f'(0)X_i + \frac{1}{2}f''(0)X_i^2 + ...$ and consider only the linear terms)

Remark 5 *The assumption* $\mathsf{E}[Y_i|X_i] = \alpha + \beta X_i$ *then directly implies that the error* ε_i *and the so-called* regressor X_i *are uncorrelated:*

$$\begin{aligned}\mathsf{Cov}(X_i, \varepsilon_i) &= \mathsf{Cov}(X_i, Y_i - \mathsf{E}[Y_i|X_i]) \\ &= \mathsf{E}[X_i(Y_i - \mathsf{E}[Y_i|X_i])] \\ &\quad -\mathsf{E}[X_i]\mathsf{E}[Y_i - \mathsf{E}[Y_i|X_i]] \\ &= \mathsf{E}[X_iY_i] - \mathsf{E}[X_i\mathsf{E}[Y_i|X_i]] \\ &\quad -\mathsf{E}[X_i]\mathsf{E}[Y_i] + \mathsf{E}[X_i]\mathsf{E}[\mathsf{E}[Y_i|X_i]]\end{aligned}$$

The law of iterated expectations tells us that

$$\mathsf{E}[\mathsf{E}[Y_i|X_i]] = \mathsf{E}[Y_i]$$

and also, since $\mathsf{E}[X_iY_i|X_i] = X_i\mathsf{E}[Y_i|X_i]$:

$$\begin{aligned}\mathsf{E}[X_i\mathsf{E}[Y_i|X_i]] &= \mathsf{E}[\mathsf{E}[X_iY_i|X_i]] \\ &= \mathsf{E}[X_iY_i]\end{aligned}$$

so all the terms in the decomposition of the covariance cancell out:

$$\begin{aligned}\mathsf{Cov}(X_i,\varepsilon_i) &= \mathsf{E}[X_iY_i] - \mathsf{E}[X_iY_i] \\ &\quad -\mathsf{E}[X_i]\mathsf{E}[Y_i] + \mathsf{E}[X_i]\mathsf{E}[Y_i] \\ &= 0\end{aligned}$$

so by construction X_i *and* ε_i *are not correlated.*

- Generic assumptions about the error term ε_i:
  - ▶ The error ε_i is a random variable with a zero mean:
  $$\mathsf{E}(\varepsilon_i) = 0$$
  - ▶ The variance of ε_i, denoted by σ_ε^2 is the same for all values of the independent variable: for all i
  $$\mathsf{Var}(\varepsilon_i) = \sigma_\varepsilon^2$$
  - ▶ The random variables ε_i are uncorrelated with each other: for $i \neq j$
  $$\mathsf{Cov}(\varepsilon_i,\varepsilon_j) = 0$$
  Remark: The errors ε_i with these assumptions are called "white noise".
  - ▶ In practice, a stronger condition is also assumed: Gaussian white noise, i.e. the ε_i are i.i.d.
  $$\mathcal{N}\left(0,\sigma_\varepsilon^2\right)$$
- Taking into account the assumptions $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma_\varepsilon^2)$, as well as the non random nature of the X-variable, we have :
$$\begin{aligned}\mathsf{E}(Y_i|X_i = x_i) &= \mathsf{E}(\alpha + \beta x_i + \varepsilon_i|X_i = x_i) \\ &= \alpha + \beta x_i + E(\varepsilon_i|X_i = x_i)\end{aligned}$$
if we make the additional assumption that X_i and ε_i are **independent** (so

what we do not model, ε_i, is independent of the explanatory variable)[1]

$$\mathsf{E}\left(\varepsilon_i | X_i = x_i\right) = \mathsf{E}\left(\varepsilon_i\right) = 0$$

then

$$\mathsf{E}\left(Y_i | X_i = x_i\right) = \alpha + \beta x_i$$

- Error Variance Estimation

$$\widehat{\Sigma}_{\varepsilon}^2 = S_{\varepsilon}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 \tag{8.1}$$

where $\widehat{\varepsilon}_i$ are the residuals. The denominator is $n-2$, i.e. we lose two degrees of freedom, because two parameters are estimated α and β prior to estimating σ_{ε}^2.

Example 18 *Computing the Beta.*
An asset's β with respect to a portfolio or market m is measured via linear regression. Denoting returns r_t and r_t^m, the beta is given by

$$r_t = \alpha + \beta r_t^m + \varepsilon_t$$

The β coefficient describes the way the asset correlates with the market:

- *$\beta = 0$ means the asset and the market are not correlated*
- *$\beta > 0$ means the asset generally follows the market (both bulls and bears)*
- *$\beta < 0$ means the asset has a behavior opposite that of the market (a good hedge or safe haven)*

Within the CAPM (Capital Asset Pricing Model), the β between an asset and a portfolio measures the part of the asset's volatility that cannot be offset by diversifying using the portfolio's assets. It is the part of the asset's return that correlates with the portfolio's components.
Figures 8.1 and 8.2 present a scatterplots of AIG's and the Gold Bullion's daily returns against the S&P500 over 2004-2008.

A question is what is the associated β, i.e. does it make sense to draw a line

[1] In the bivariate Gaussian model, the regression errors are always independent of the regressors as they are not correlated.

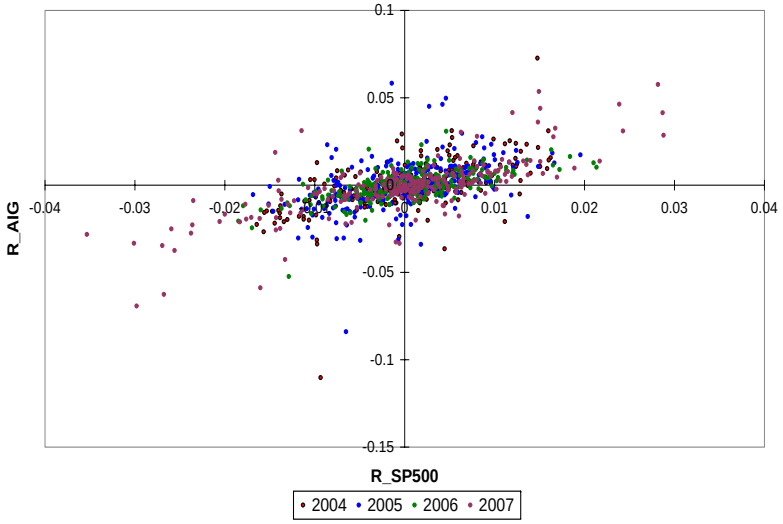


Figure 8.1Scatterplot of daily returns: AIG vs S&P500

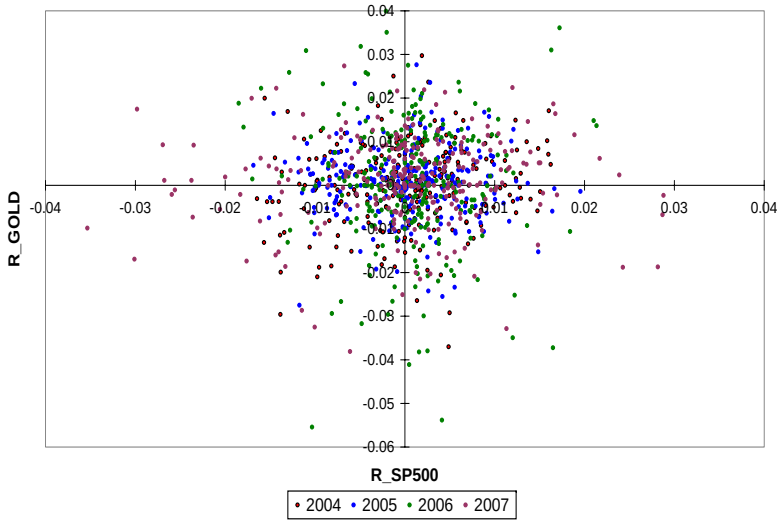


Figure 8.2Scatterplot of daily returns: Gold Bullion vs S&P500

through the clouds? Consider for instance the following correlations

Correlation (2004-2008)	*S&P 500*	*AIG*	*GOLD*
S&P 500	*1*	*0.548*	*-0.043*
AIG	*0.548*	*1*	*-0.082*
GOLD	*-0.043*	*-0.082*	*1*

Correlations measure the strength of the link but do not allow to quantify variations:

- *assume that when S&P500 grows by 10%, AIG* always *grow by 20%, then the correlation is* $\rho = 1$*, but this is also the case if AIG always grow by 30%.*
- *by contrast, the* β *is then either 2 or 3: the regression is more informative.*

The β *represents market risk, systematic risk (a good manager always perform better then her* β*). It is a combination of correlation* ρ_{xm} *and relative volatility:*

$$\beta = \frac{\sigma_x}{\sigma_m}\rho_{xm}$$

Assume the asset x *has a low volatility (*σ_x*) but a high correlation, and asset* z *has a high volatility and a low correlation. Comparing the betas show which is more risky.*

8.2 Estimation of the parameters of a regression

8.2.1 Method of moments

From the assumption that $\mathsf{E}[\varepsilon_i]$ and $\mathsf{Cov}[X_i, \varepsilon_i] = 0$ we can directly estimate the parameters using the *method of moments* which simply replaces expectations by the empirical moments to carry out the estimation:
$\mathsf{E}[\varepsilon_i] = \mathsf{E}[Y_i - \alpha - \beta X_i] = 0$ leads to choosing $\left(\widehat{\alpha}, \widehat{\beta}\right)$ such that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{\alpha} - \widehat{\beta}X_i\right) &= 0 \\ \therefore \quad \overline{Y} &= \widehat{\alpha} + \widehat{\beta}\overline{X} \end{aligned}$$

and $\mathsf{Cov}\left[X_i, \varepsilon_i\right] = \mathsf{Cov}\left[X_i, Y_i - \alpha - \beta X_i\right] = 0$ leads to

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y} - \left(\widehat{\alpha} + \widehat{\beta}X_i - \widehat{\alpha} - \widehat{\beta}\overline{X}\right)\right) &= 0 \\ \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) &= \widehat{\beta}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\end{aligned}$$

i.e.

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

which shows that the Least-Squares estimators imply that the errors and regressors must be uncorrelated for the estimators to be consistent. Remember also that this is only valid if the conditional expectation is linear in the regressor, in practice we often need to assume Gaussianity to ensure that the latter assumption holds.

8.2.2 Bias

What about the bias of $\left(\widehat{A}, \widehat{B}\right)$? We saw that

$$\widehat{B} - \beta = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

so

$$\mathsf{E}\left[\widehat{B} - \beta\right] = \mathsf{E}\left[\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right] = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\mathsf{E}\left[\varepsilon_i\right]}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = 0$$

and

$$\begin{aligned}\mathsf{E}\left[\widehat{A}\right] &= \mathsf{E}\left[\overline{Y} - \beta\overline{x} + \left(\beta - \widehat{B}\right)\overline{x}\right] = \mathsf{E}\left[\overline{Y} - \beta\overline{x}\right] - \mathsf{E}\left[\left(\widehat{B} - \beta\right)\right]\overline{x} \\ &= \alpha + 0 = \alpha\end{aligned}$$

the **Estimator of** (α, β) **is unbiased**.

8.2.3 Variance

Now, consider the variance of this estimator so that we can then perform tests and construct confidence intervals about the estimators.

The variance of $\widehat{B}$ satisfies:

$$\begin{aligned}\mathsf{Var}\left[\widehat{B}\right] &= \mathsf{Var}\left[\frac{\sum_{i=1}^{n}(x_i-\overline{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] = \frac{1}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\,\mathsf{Var}\left[\varepsilon_i\right] \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2}\sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\end{aligned}$$

and it can be shown that the variance of $\widehat{A}$ is given by it follows that

$$\mathsf{Var}\left[\widehat{A}\right] = \frac{\sigma_\varepsilon^2}{n}\frac{\sum_{j=1}^{n}x_j^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}$$

These formulas are used by softwares to when they report the standard error of the estimators, next to the estimated coefficients. We will see below how they are used to construct confidence intervals and to compute the statistics for the tests that there is no impact of the X variable onto the Y variable (i.e. that X is irrelevant and should not be used to explain the variations in Y).

- From what we saw previously since the ε_i are normally distributed r.v. and $\widehat{A}$ and $\widehat{B}$ are linear combination of them, both $\widehat{A}$ and $\widehat{B}$ are normally distributed:

$$\begin{aligned}\widehat{A} &\sim \mathcal{N}\left(\alpha, \frac{\sigma_\varepsilon^2}{n}\frac{\sum_{j=1}^{n}x_j^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}\right) \\ \widehat{B} &\sim \mathcal{N}\left(\beta, \frac{\sigma_\varepsilon^2}{\sum_{j=1}^{n}(x_j-\overline{x})^2}\right)\end{aligned}$$

  it follows that

$$\begin{aligned}\sigma_\varepsilon^{-1}\sqrt{\frac{\sum_{i=1}^{n}(x_j-\overline{x})^2}{\frac{1}{n}\sum_{j=1}^{n}x_j^2}}\left(\widehat{A}-\alpha\right) &\sim \mathcal{N}(0,1) \\ \sigma_\varepsilon^{-1}\sqrt{\sum_{i=1}^{n}(x_j-\overline{x})^2}\left(\widehat{B}-\beta\right) &\sim \mathcal{N}(0,1)\,.\end{aligned}$$

  And these are the test statistics that can be used for tests on the parameters.

Example 19 *The beta in Finance*
The equations, estimated over 1306 daily observations, are (standard errors of

the parameters in parentheses)

$$AIG \quad : \quad \widehat{r}^{\,AIG} = \underset{(0.00116)}{-\,0.003} + \underset{(0.0877)}{2.07}\, r^{\,SP500}$$

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-0.003	0.001	-2.190	0.029	-0.005	0.000
R_SP500	2.075	0.088	23.648	< 0.0001	1.903	2.248

and

$$GOLD \quad : \quad \widehat{r}^{\,GOLD} = \underset{(0.00036)}{0.001} - \underset{(0.0273)}{0.043}\, r^{\,SP500}$$

Model parameters:

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	0.001	0.000	1.524	0.128	0.000	0.001
R_SP500	-0.043	0.027	-1.557	0.120	-0.096	0.011

It appears that AIG is more risky than gold, in fact the negative β for Gold shows that this is seen as a hedging asset.

The software reports the estimated standard errors of the parameters together with tests for their significance and confidencen intervals. We will see below what these mean and how they are constructed but first we show the link between the variance of the parameters and the confidence interval on the predicted values.

8.3 Inference on the parameters of a regression

8.3.1 Confidence intervals around the estimated coefficients

To construct a confidence interval for α, β, we first need an estimator of $\sigma_\varepsilon^2 = \mathsf{Var}\left[\varepsilon_i\right]$. We can e.g. use the MLE $\widehat{\Sigma}^2$ defined previously. Alternatively, we do not observe the errors $\varepsilon_i = Y_i - \alpha - \beta x_i$ (as we do not know β) but we observe the residuals $\widehat{\varepsilon}_i = Y_i - \widehat{A} - \widehat{B}x_i$ and we know that the mean $n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i = 0$, hence we can use the estimator:

$$S^2 = \frac{1}{n-2}\sum_{i=1}^n \widehat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \widehat{A} - \widehat{B}x_i\right)^2 \xrightarrow{p} \sigma_\varepsilon^2.$$

(we divide by $n-2$ as we lose two degrees of freedom from the estimation of $\widehat{A}$ and $\widehat{B}$).

Using the true (unknown) values of (α, β), we define the r.v. T_α and T_β as

$$\begin{aligned} T_\alpha &= S^{-1}\sqrt{\frac{\sum_{i=1}^n (x_j - \overline{x})^2}{\frac{1}{n}\sum_{j=1}^n x_j^2}}\left(\widehat{A} - \alpha\right) \\ T_\beta &= S^{-1}\sqrt{\sum_{i=1}^n (x_j - \overline{x})^2}\left(\widehat{B} - \beta\right) \end{aligned}$$

The distributions of T_α and T_β are Student (T) with $n-2$ degrees of freedom.

Using for instance a 95% confidence interval, and defining the 2.5% and 97.5% quantiles of the T distribution with $n-2$ degrees of freedom as $\widetilde{t}_{2.5/100,n-2}$ and $\widetilde{t}_{97.5/100,n-2}$ respectively, we can construct a confidence interval using

$$\begin{aligned} \mathsf{P}\left(T_\alpha \in \left[\widetilde{t}_{2.5/100,n-2}, \widetilde{t}_{97.5/100,n-2}\right]\right) &= 0.95 \\ \mathsf{P}\left(T_\beta \in \left[\widetilde{t}_{2.5/100,n-2}, \widetilde{t}_{97.5/100,n-2}\right]\right) &= 0.95 \end{aligned}$$

Therefore

$$\begin{aligned} \mathsf{P}\left(S^{-1}\sqrt{\frac{\sum_{i=1}^n (x_j - \overline{x})^2}{\frac{1}{n}\sum_{j=1}^n x_j^2}}\left(\widehat{A} - \alpha\right) \in \left[\widetilde{t}_{2.5/100,n-2}, \widetilde{t}_{97.5/100,n-2}\right]\right) &= 0.95 \\ \mathsf{P}\left(S^{-1}\sqrt{\sum_{i=1}^n (x_j - \overline{x})^2}\left(\widehat{B} - \beta\right) \in \left[\widetilde{t}_{2.5/100,n-2}, \widetilde{t}_{97.5/100,n-2}\right]\right) &= 0.95 \end{aligned}$$

which implies that

$$\begin{aligned} & \mathsf{P}\left(\alpha \in \left[\widehat{A} - \frac{1}{S}\sqrt{\frac{\frac{1}{n}\sum_{j=1}^n x_j^2}{\sum_{i=1}^n (x_j - \overline{x})^2}}\widetilde{t}_{0.975,n-2}, \widehat{A} - \frac{1}{S}\sqrt{\frac{\frac{1}{n}\sum_{j=1}^n x_j^2}{\sum_{i=1}^n (x_j - \overline{x})^2}}\widetilde{t}_{0.025,n-2}\right]\right) \\ = \; & 0.95 \\ & \mathsf{P}\left(\beta \in \left[\widehat{B} - \frac{1}{S}\sqrt{\frac{1}{\sum_{i=1}^n (x_j - \overline{x})^2}}\widetilde{t}_{0.975,n-2}, \widehat{B} - \frac{1}{S}\sqrt{\frac{1}{\sum_{i=1}^n (x_j - \overline{x})^2}}\widetilde{t}_{0.025,n-2}\right]\right) \\ = \; & 0.95 \end{aligned}$$

These constitute confidence intervals that can be computed in practice.

Remark 6 *for large samples (i.e. large number of degrees of freedom) the Student distribution becomes a standard Normal.*

Remark 7 *The same results can be derived when considering the model with stochastic regressors* $Y_i = \beta X_i + \varepsilon_i$ *with* $\mathsf{Cov}[X_i, \varepsilon_i] = 0$, *instead of* $Y_i = \beta x_i + \varepsilon_i$. *But then,* $\mathsf{Var}[\varepsilon_i] = \mathsf{Var}[Y_i]$ *no longer holds true.*

Remark 8 *When the* ε_i *are not normally distributed, it still holds that under the other conditions on* ε_i *that*

$$\sigma_\varepsilon^{-1} \sqrt{\frac{\sum_{i=1}^n (x_j - \overline{x})^2}{\frac{1}{n}\sum_{j=1}^n x_j^2}} \left(\widehat{A} - \alpha\right) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\sigma_\varepsilon^{-1} \sqrt{\sum_{i=1}^n (x_j - \overline{x})^2} \left(\widehat{B} - \beta\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Example 20 *The beta*
Given the large samples, we can use normal distributions instead of the Student. Hence 95% confidence intervals for the α *and* β *parameters are*

$$\begin{aligned}
\alpha^{AIG} &: [-0.003 - 1.96 \times 0.00116, -0.003 + 1.96 \times 0.00116] \\
&= \left[-5.27 \times 10^{-3}, -7.26 \times 10^{-4}\right] \\
\beta^{AIG} &: [2.07 - 1.96 \times 0.0877, 2.07 + 1.96 \times 0.0877] \\
&= [1.90, 2.24] \\
\alpha^{GOLD} &: [0.001 - 1.96 \times 0.00036, 0.001 + 1.96 \times 0.00036] \\
&= \left[2.94 \times 10^{-4}, 1.71 \times 10^{-3}\right] \\
\beta^{GOLD} &: [-0.043 - 1.96 \times 0.0273, -0.043 + 1.96 \times 0.0273] \\
&= \left[-9.65 \times 10^{-2}, 1.05 \times 10^{-2}\right]
\end{aligned}$$

We see that the confidence interval for α^{AIG} *does not contain zero, which implies that* α *has at least 95% chance of being negative. By contrast, we cannot be certain at 95% that* β^{GOLD} *is nonzero. AIG's beta is quite large, between 1.9 and 2.25.*

8.3.2 Test on the significance of a regressor (the relevance of the explanatory variable)

In the regression model, with variables X_i and Y_i satisfying

$$\mathsf{E}\left[Y_i | X_i = x_i\right] = \alpha + \beta x_i$$

where we estimated α, β using maximum likelihood. We now test whether the link between X and Y truly exists, i.e.

$$\begin{aligned} \mathsf{H}_0 &: \beta = 0 \\ \mathsf{H}_1 &: \beta \neq 0 \end{aligned}$$

For this we recall that $\sigma_\varepsilon^{-1}\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\left(\widehat{B} - \beta\right) \sim \mathcal{N}(0,1)$ which implies that under H_0 :

$$\frac{\widehat{B}}{\sqrt{Var\left(\widehat{B}\right)}} = \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\frac{\widehat{B}}{\sigma_\varepsilon} \sim \mathcal{N}(0,1)$$

We estimate σ_ε^2 as before so that, using Slutsky's formula, as $n \to \infty$,

$$T_\beta = \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\frac{\widehat{B}}{S} \xrightarrow[\mathsf{H}_0]{\mathsf{d}} \mathcal{N}(0,1)$$

and in finite samples $\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\frac{\widehat{B}}{S}$ follows a T distribution with $n-2$ degrees of freedom.

We base our decision on the following rule:

Reject H_0 if the realization t_β of T_β satisfies:

$$|t_\beta| = \sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{S^2}}\left|\widehat{\beta}\right| > \widetilde{t}_{97.5/100,n-2}$$

where $\widetilde{t}_{97.5/100,n-2}$ is the 97.5% quantile of the T distribution with $n-2$ degrees of freedom and $\widehat{\beta}$ is the realization of the estimator $\widehat{B}$, i.e. the estimate of β.

Even when the errors are not normally distributed (but if the other conditions holds) we can use the Central Limit theorem and Slutsky's formula to show

that

$$T_\beta = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \frac{\widehat{B}}{S} \xrightarrow[\mathsf{H}_0]{\mathrm{d}} \mathcal{N}(0,1)$$

so the decision rule becomes:
Reject H_0 if the realization t_β of T_β satisfies:

$$\sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{S^2}} \left|\widehat{\beta}\right| > \widetilde{z}_{0.975}.$$

where $\widetilde{z}_{0.975}$ is the 97.5% quantile of the standard normal (i.e. 1.96, or approximately 2).

More generally for the regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i$$

or in the matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the least-squares estimator of $\boldsymbol{\beta} = {}^t(\beta_0, \beta_1, ..., \beta_k)$ is

$$\widehat{\boldsymbol{\beta}} = \left({}^t\mathbf{X}\mathbf{X}\right)^{-1} {}^t\mathbf{X}\mathbf{Y}$$

then, if the errors are $i.i.d.$ with variance σ_ε^2 and they are uncorrelated to the regressors X_{ji}

$$\left({}^t\mathbf{X}\mathbf{X}\right)^{1/2} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_r\right)$$

i.e. for large samples

$$\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \mathcal{N}\left(\mathbf{0}, \sigma_\varepsilon^2 \left({}^t\mathbf{X}\mathbf{X}\right)^{-1}\right)$$

and we can test for individual β_j to be zero, i.e. for the regressor to be relevant. Softwares automatically report the t statistic for

$$\mathsf{H}_0 : \beta_j = 0 \text{ versus } \mathsf{H}_1 : \beta_j \neq 0$$

for all $j = 1, ..., k$. The distribution we need to use (when the errors are normally distributed) to compute the critical values is T_{n-k-1}, the Student distribution with $n - k - 1$ degrees of freedom. In large samples, we can use the standard normal distribution instead.

Example 21 *The beta*

The t-stats are as follows

$$\begin{aligned} t_{\alpha^{AIG}} &= \frac{-0.003}{0.00116} = -2.59 \\ t_{\beta^{AIG}} &= \frac{2.07}{0.0877} = 23.6 \\ t_{\alpha^{GOLD}} &= \frac{0.001}{0.00036} = 2.78 \\ t_{\beta^{GOLD}} &= \frac{-0.043}{0.0273} = -1.58 \end{aligned}$$

which shows at a significance level of 95% we can reject the hypotheses that α^{AIG}, β^{AIG} and α^{GOLD} are zero, but we cannot reject the hypothesis that $\beta^{GOLD} = 0$. There is not enough evidence to show that GOLD and stock markets are not disconnected. Gold can therefore be used for diversification, but probably not for hedging against stock market fluctuations.

8.3.3 Logistic Regression and Probit Model (*)

Firms show considerable interest in conditional regression models for binary variables, with

$$\mathsf{P}(Y_i = 1|X = x) = g(\beta' x).$$

The most well model of this type is the logistic regression where

$$\mathsf{P}(Y_i = 1|X = x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)}$$

Notice in this case, no direct attempt to model the joint distribution of X, Y is made, instead we have just decided that our interest focuses on the parameters which index the conditional probability.

More generally, we can think of models for a binary dependent variables; the most common are the Probit and Logit models.

Example 22 *We can imagine a bank giving a loan to a client i, the client's ability to repay the loan is an unobservable variable Y_i^*. Say Y_i^* is normally distributed across individuals*

$$Y_i^* \sim \mathsf{N}\left(\mu, \sigma_\varepsilon^2\right).$$

If a client ability to repay is less than the loan ℓ_i, the client fails to repay. The problem is that we cannot observe Y_i^*: we only observe whether the loan is repaid That is we observe Y_i such that

$$Y_i = \begin{cases} 1 & \textit{if loan is repaid} \\ 0 & \textit{otherwise} \end{cases}$$

Given this setup, the question of interest is: $\text{what is the probability that client } i \text{ repays the loan?}$ *It is merely*

$$P\left(Y_i = 1\right) = P\left(Y_i^* \geq \ell_i\right)$$

and hence the observations are generated by the following rule:

$$Y_i = \begin{cases} 1 & \textit{if } Y_i^* \geq \ell_i \\ 0 & \textit{otherwise} \end{cases}$$

In this example y_i^* is called a $\text{latent or index variable}$

8.3.4 Formulating a probability model

Now our model is

$$y_i^* = X_i\beta + \varepsilon_i$$

but we only observe y_i which takes values 0 or 1. (X_i is a row vector $X_i = [x_{i1}...x_{ik}]$ of characteristics of the client, such as age, gender, previous history with the bank...). We would like to transform $X\beta$ into a *probability*. That is, we need a function F such that

$$P\left(y_i = 1\right) = F\left(X_i\beta\right)$$

A natural choice for F is a distribution function or cumulative density as they lie between 0 and 1. The identity

$$P\left(y_i = 1\right) = X_i\beta$$

does not yield the type of functions that we want.

Choosing F to be the standard Normal leads to an attractive possibility; this is called the *Probit* model:

$$P\left(y_i = 1\right) = \Phi\left(X_i\beta\right) = \int_{-\infty}^{X_j\beta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

The standard normal transformation constrains the probability to lie between zero and one.

Choosing F to be the logistic distribution yields another attractive possibility, the *logit* model

$$P(y_i = 1) = \Lambda(X_i\beta) = \frac{\exp X_i\beta}{1 + \exp X_i\beta}.$$

We are not constrained to restrict our choice to these two functions, but these are often used as they have **behavioral** interpretations. (we denote $\partial\Phi = \phi$ and $\partial\Lambda = \lambda$)

8.3.5 The Probit

Assume that we observe y that takes on one of two values 0 and 1. Define a latent variable y^* such that

$$y_i^* = X_i\beta + \epsilon_i$$

We do not observe y^* but rather y, which takes values 0 or 1 according to the following rule

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

We also assume that $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$.

Hence y_i^* is distributed Normally, conditional on X, but y_i is not. It is straightforward to show that the previous rule generates a probit:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0) = P(X_i\beta + \epsilon_i > 0) \\ &= P(\epsilon_i > -X_i\beta) = P\left(\frac{\epsilon_i}{\sigma} > -\frac{X_i\beta}{\sigma}\right) \end{aligned}$$

and

$$\frac{\epsilon_i}{\sigma} \sim \mathsf{N}(0, 1)$$

so by symmetry

$$\begin{aligned} P(y_i = 1) &= P\left(\frac{\epsilon_i}{\sigma} > -\frac{X_i\beta}{\sigma}\right) = P\left(\frac{\epsilon_i}{\sigma} < \frac{X_i\beta}{\sigma}\right) \\ &= \Phi\left(\frac{X_i\beta}{\sigma}\right) \end{aligned}$$

Another important aspect is that the coefficients β and σ always appear together: they are not separately identified; only their ratio matters. It is therefore convenient to normalize $\sigma = 1$. The log-Likelihood is globally concave but cannot be solved analytically, yet it is easy to maximize it using standard algorithms

How can me move from β to the impact of a variable on y? Using

$$\frac{\partial \mathsf{E}[y]}{\partial x_j} = \frac{\partial \mathsf{P}[y=1|X]}{\partial x_j} = \phi(X\beta)\beta_j$$

with the standard normal density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Hence *the derivative of the probability with respect to x varies with the level of X,* i.e. *all the variables in the model.* Hence, the probit necessitates more information than just reporting the coefficients. It can be useful to compute the derivatives of the model at the mean values.

8.3.6 The Logit

The logit is similar to the probit, with now

$$P(y_i = 1) = \Lambda(X_i\beta) = \frac{\exp X_i\beta}{1 + \exp X_i\beta}$$

The latent variable interpretation is the same, but with ϵ_i which now follows what is called *a logistic distribution.* The main difference between the Normal and Logistic distributions is that the latter has more weight in the tails.

The derivative of the probability with respect to one element is now given by

$$\begin{aligned} \frac{\partial \mathsf{E}[y]}{\partial x_j} &= \frac{\exp X\beta}{(1+\exp X\beta)^2}\beta_j \\ &= q(1-q)\beta_j \end{aligned}$$

where $q = \frac{\exp X\beta}{1+\exp X\beta}$

8.3.7 Estimation

Estimation in the logit and probit models is done via means of so-called nonlinear least-squares. The principle is simple, we know that when we try to estimate a proportion, say if $Y_i \sim Bernoulli(\pi)$, then the estimator is the empirical mean

$$\widehat{\pi} = \overline{Y}_n.$$

This amounts in fact to estimating a regression model

$$Y_i = \pi + u_i$$

where

$$\begin{aligned} u_i &= 1 - \pi \text{ with probability } \pi, \text{ so } Y_i = 1 \\ u_i &= -\pi \text{ with probability } 1 - \pi, \text{ so } Y_i = 0 \end{aligned}$$

Hence the least squares estimator in this case is

$$\begin{aligned} \widehat{\pi} &= argmin_{\pi \in (0,1)} \sum_{i=1}^{n} (Y_i - \pi)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} Y_i. \end{aligned}$$

For the logit and probit models, the principle is the same:

$$Y_i = g(X_i\beta) + u_i$$

where $g(X_i\beta) = \Phi(X_i\beta)$ or $\Lambda(X_i\beta)$ and

$$\begin{aligned} u_i &= 1 - g(X_i\beta) \text{ with probability } g(X_i\beta), \text{ so } Y_i = 1 \\ u_i &= -g(X_i\beta) \text{ with probability } 1 - g(X_i\beta), \text{ so } Y_i = 0 \end{aligned}$$

Hence the *nonlinear least-squares* estimator is given by

$$\widehat{\beta} = argmin \sum_{i=1}^{n} (Y_i - g(X_i\beta))^2$$

and we can solve it numerically.

This is easily extended to general models where there are several explanatory variables X_i and we can compute the R^2 as usual

$$R^2 = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}.$$

The usual t-tests are also possible.

8.4 Exercises

8.4.1 (+) Revisiting the multiple linear regression model page 121

Consider the exercise page 121. Answer the following questions that constitute further steps to the analysis:

1. Test the significance of the different regression coefficients by looking at the confidence intervals and the p-values. Shall we retain all explanatory variables in the model?
2. Improve the initial model by means of a Backward Stepwise procedure: remove the least significant explanatory variable and run again the model; repeat the two steps till all regression coefficients are significant.
3. Give an interpretation of the regression coefficients in the finally selected model.
4. Use the final model to predict the *Sales* for the scenario given for next semester.

8.4.2 (+) Some more Apartment regressions

Ayant l'intention d'acheter un appartement dans le centre de Paris, vous avez noté figure 8.3 parmi les 9 premiers arrondissements les annonces du Figaro du 14 Mai 2009 (en milliers d'euros).

1. ► Quel est le prix moyen par mètre carré dans ces arrondissements ?
   ► Proposez un modèle régression permettant de prévoir le prix de vente affiché en fonction de la surface.
   ► Votre tante possède un studio de 19 m^2 dans le V°, quel prix lui recommandez-vous ?

8.4.3 (+) BioPro Marketing

Grâce à votre expérience chez « Bio-net », vous êtes nouvellement embauché(e) chez « Bio-Pro », le leader des fabricants de détergent professionnel. Le marché des produits commercialisés par Bio-Pro répond à des règles proches du marché de Bio-net de sorte que vous espérez bien exploiter votre précédente expérience. Vous souhaitez établir des prévisions de ventes pour le « B05 », l'un des produits phares de Bio-Pro. En accord avec le directeur commercial et lui relatant votre précédente mission, vous définissez les variables suivantes pour expliquer la demande Y (exprimée par 100 000 unités) de « B05 » :
X_1 : le prix de vente (en euros d'une unité pour le mois considéré)
X_2 : la moyenne des prix de vente (en euros) sur le mois pour des produits concurrents,
X_3 : le budget de la publicité (en K€) pour promouvoir le « B05 » pendant le

	Localisation	Surface	Prix
1	Réaumur-Sébastopol	131	1150
2	Bourse	55	450
3	Marais Archives	81	850
4	Turenne	100	1180
5	Luxembourg	152	1575
6	Soufflot	145	1650
7	Cardinal Lemoine	59	495
8	Notre-Dame	56	670
9	Gay-Lussac	80	843
10	St-Medard	53	448
11	Arenes de Lutèce	78	650
12	Rue du Four	95	1350
13	ND des Champs	50	389
14	ND des Champs	98	850
15	St-Germain des Prés	97	1150
16	Vavin	98	850
17	Rue de Seine	30	390
18	Suffren	160	1470
19	Bourdonnais	120	1250
20	Rue de Sèvres	68	650
21	Vaneau	68	775
22	Babylone	69	833
23	Bourdonnais	64	725
24	Raspail-Varenne	373	3500
25	Raspail-Varenne	101	836
26	Vaneau	91	1060
27	St-Germain	65	645
28	Tour Eiffel	132	1160
29	Verneuil	56	525
30	Madeleine	51	800
31	Madeleine	67	950
32	Berri-Artois	200	1650
33	Monceau	141	1100
34	Lisbonne	118	948
35	Liège	132	1030
36	Villiers	20	185
37	Fb St-Honoré	50	511
38	Fb St-Honoré	83	865
39	Fb St-Honoré	121	1367
40	Montaigne	178	2310
41	Montaigne	195	2380
42	Villiers	104	870
43	Madeleine	191	2200
44	Trudaine	158	895
45	Drouot	145	1300
46	Bonne nouvelle	105	795

Figure 8.3Appartements en vente dans Le Figaro du 14 mai 2009

	Y	X1	X2	X3	X4
1	7.38	5.9	5.8	55.0	-0.1
2	8.51	5.7	6.1	67.5	0.4
3	9.52	5.6	6.6	72.5	1.0
4	7.50	5.6	5.6	55.0	0.0
5	9.33	5.5	5.9	70.0	0.4
6	8.28	5.5	5.8	65.0	0.3
7	8.75	5.5	5.7	67.5	0.2
8	7.87	5.8	5.9	52.5	0.1
9	7.10	5.8	5.6	52.5	-0.2
10	8.00	5.9	6.1	60.0	0.2
11	7.89	6.0	6.3	65.0	0.3
12	8.15	6.0	6.1	62.5	0.1
13	9.10	5.6	6.3	70.0	0.7
14	8.56	5.7	6.4	69.0	0.7
15	8.90	5.7	6.3	68.0	0.6
16	8.87	5.8	6.3	68.0	0.5
17	9.26	5.6	6.4	71.0	0.8
18	9.00	5.8	6.6	70.0	0.8
19	8.75	5.6	6.3	68.0	0.7
20	7.95	5.8	5.7	65.0	-0.1
21	7.65	5.8	5.7	62.5	-0.1
22	7.27	5.7	5.6	60.0	-0.1
23	8.00	5.6	6.0	65.0	0.4
24	8.50	5.4	5.6	70.0	0.2
25	8.75	5.5	6.3	68.0	0.8
26	9.21	5.6	6.5	68.0	0.9
27	8.27	5.6	5.6	65.0	0.0
28	7.67	5.7	5.7	57.0	0.0
29	7.93	5.8	5.9	58.0	0.1
30	9.26	5.6	6.5	68.0	0.9

Figure 8.4Données de ventes du produit “B05”.

mois
$X_4 : X_2 - X_1$
Le service commercial vous communique les données suivantes sur les derniers 30 mois (voir figure 8.4)

- Dans une première étape, vous vous interrogez sur l’effet de variable X_3 (budget de publicité) et décidez d’étudier la régression :

$$Y = a_0 + a_3 X_3 + \epsilon$$

Les résultats, avec Excel, de cette régression figurent dans le tableau suivant :

(tâchez de le reproduire)

SUMMARY OUTPUT

Regression Statistics

Multiple R	0,87037
R Square	0,75754
Adjusted R Square	0,74888
Standard Error	0,33884
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10,04437	10,04437	87,48303	0,00000
Residual	28	3,21482	0,11482		
Total	29	13,25919			

	Coeff.	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1,75869	0,70983	2,47760	0,01953	0,30466	3,21271
X3	0,10252	0,01096	9,35324	0,00000	0,08006	0,12497

(a) A partir de ces résultats, donner les estimations $\widehat{a}_0$ et $\widehat{a}_3$ des coefficients a_0 et a_3, et l'équation de la droite de régression.
(b) Quelle est la signification du coefficient de régression estimé $\widehat{a}_3$?
(c) D'après le tableau ci-dessus, quelle est la part de variance empirique des ventes restituée par ce modèle ? (i.e. quel est le coefficient d'ajustement linéaire ?) Expliquez en une phrase l'intérêt de cet indicateur.
(d) Le résultat de la régression de la variable Y sur la variable X_4 donne les résultats suivants. Si vous deviez choisir entre la régression simple étudiée en 1. et celle-ci, laquelle choisiriez-vous ? Justifiez clairement et

rapidement votre réponse.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0,88041
R Square	0,77513
Adjusted R Square	0,76710
Standard Error	0,32632
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10,27759	10,27759	96,51606	0,00000
Residual	28	2,98160	0,10649		
Total	29	13,25919			

	Coeff.	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	7,81419	0,08235	94,89291	0,00000	7,64551	7,98287
X4	0,26179	0,02665	9,82426	0,00000	0,20720	0,31637

- Proposez un modèle de régression multivariée pour les ventes de B5.

Chapter 9
Answers to exercises

9.1 Chapter 4

1. Exercise 4.3.2 page 75
   - ▶ On suppose que la variable aléatoire X_i indique pour chaque individu i son nombre d'enfants. On note par X sa loi parente. La v.a. prend des valeurs entières non-négatives. On fait l'hypothèse ici que le nombre maximum d'enfants dans la population concernée des "jeunes managers" est de 4.
     On dispose de deux types d'échantillons:
     - $*$ le premier est de grande taille (1000) mais malheureusement nous n'observons pas les réponses de tous les individus, seulement des 20% qui ont répondu. Ces derniers n'ont pas exactement les mêmes caractéristiques que l'ensemble de la population, car les "jeunes managers" répondent différemment selon leur nombre d'enfants (la variable qui nous intéresse ici). On nomme ce problème un **effet de sélection.** On fait l'hypothèse que les 200 qui ont répondu sont issus d'une sous-population qui n'est pas celle qui nous intéresse directement, mais toutefois cet échantillon est de grande dimension $m = 200$.
     - $*$ le second échantillon est issu de la population totale (celle qui nous intéresse) mais il est plus petit ($n = 25$).

     La population totale comprend $N = 50\,000$ individus. Dans un sondage réel, on ne connaît pas la répartition des individus dans la population, mais afin d'évaluer les problèmes d'estimation, on suppose ici qu'on connaît la population totale ainsi que la sous-population des managers prêts à nous répondre.
   - ▶ Le nombre moyen d'enfants dans la population totale est l'espérance de X. Si on définit $\mathbf{1}_{\{x_i=j\}}$ la variable indicatrice qui, pour tout j, prend la

valeur 1 si $x_i = j$ et 0 si $x_i \neq j$, l'espérance est

$$\begin{aligned}
\mu_X &= \mathsf{E}[X] = \frac{1}{N}\sum_{i=1}^{N} x_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(0 \times \mathbf{1}_{\{x_i=0\}} + 1 \times \mathbf{1}_{\{x_i=1\}} + 2 \times \mathbf{1}_{\{x_i=2\}} + 3 \times \mathbf{1}_{\{x_i=3\}} + 4 \times \mathbf{1}_{\{x_i=4\}}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{4} j \times \mathbf{1}_{\{x_i=j\}} = \sum_{j=0}^{4} j \sum_{i=1}^{N} \frac{\mathbf{1}_{\{x_i=j\}}}{N}
\end{aligned}$$

où on reconnaît $\sum_{i=1}^{N} \frac{\mathbf{1}_{\{x_i=j\}}}{N} = f_j$ la fréquence d'observation de la valeur j. Et donc:

$$\begin{aligned}
\mu_X &= \sum_{j=0}^{4} j \times f_j \\
&= 0 \times 0.40 + 1 \times 0.24 + 2 \times 0.20 + 3 \times 0.12 + 4 \times 0.04 \\
&= 1.16
\end{aligned}$$

- De même, dans la sous-population des individus prêts à répondre: on note la v.a. Y_i (de loi parent Y) car les individus sont issus d'une autre population (celles des gens prêts à répondre à un sondage) de taille 10 000 indiquée sur la colonne gauche du tableau:

$$\begin{aligned}
\mu_Y &= \mathsf{E}[Y] = \frac{1}{10000}\sum_{i=1}^{10000} y_i = ... \\
&= 0 \times 0.62 + 1 \times 0.21 + 2 \times 0.12 + 3 \times 0.04 + 4 \times 0.01 \\
&= 0.61
\end{aligned}$$

- On note à présent les estimateurs "moyenne empirique" par $\overline{X}_n$ sur l'échantillon issu de la population totale de taille $n = 25$ et $\overline{Y}_m$ sur l'échantillon de taille $m = 200$ issu de la sous-population. Par linéarité

de l'espérance:

$$\begin{aligned}\mathsf{E}\left[\overline{X}_n\right] &= \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\left[X\right] \\ &= \mathsf{E}\left[X\right] = \mu_X\end{aligned}$$

et de même

$$\mathsf{E}\left[\overline{Y}_n\right] = \mu_Y.$$

Le paramètre qu'on cherche à estimer est μ_X, il en suit que $\overline{X}_n$ est un estimateur sans biais, mais que $\overline{Y}_n$ est biaisé puisque $\mathsf{E}\left[\overline{Y}_n\right] \neq \mu_X$.

► A présent, pour calculer le risque quadratique des estimateurs de μ_X,

$$\begin{aligned}\mathsf{E}\left[\left(\overline{X}_n - \mu_X\right)^2\right] &= \left(\mathsf{E}\left[\overline{X}_n - \mu_X\right]\right)^2 + \mathsf{V}\left[\overline{X}_n\right] \\ \mathsf{E}\left[\left(\overline{Y}_m - \mu_X\right)^2\right] &= \left(\mathsf{E}\left[\overline{Y}_m - \mu_X\right]\right)^2 + \mathsf{V}\left[\overline{Y}_m\right]\end{aligned}$$

et si on fait l'hypothèse selon laquelle les individus des échantillons sont choisi de manière que les v.a. soient indépendantes

$$\begin{aligned}\mathsf{E}\left[\left(\overline{X}_n - \mu_X\right)^2\right] &= \left(\mathsf{E}\left[X - \mu_X\right]\right)^2 + \frac{\mathsf{V}\left[X\right]}{n} \\ \mathsf{E}\left[\left(\overline{Y}_m - \mu_X\right)^2\right] &= \left(\mathsf{E}\left[Y - \mu_X\right]\right)^2 + \frac{\mathsf{V}\left[Y\right]}{m}\end{aligned}$$

Il nous faut calculer la variance de X et celle de Y :

$$\begin{aligned}\mathsf{V}\left[X\right] &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu_X\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{4}\left(j - \mu_X\right)^2 \times \mathbf{1}_{\{x_i=j\}} \\ &= \sum_{j=0}^{4}\left(j - \mu_x\right)^2 \sum_{i=1}^{N}\frac{\mathbf{1}_{\{x_i=j\}}}{N} = \sum_{j=0}^{4}\left(j - \mu_X\right)^2 \times f_j \\ &= (0 - 1.16)^2 \times 0.40 + (1 - 1.16)^2 \times 0.24 + (2 - 1.16)^2 \times 0.20 \\ &\quad + (3 - 1.16)^2 \times 0.12 + (4 - 1.16)^2 \times 0.04 \\ &= 1.4144\end{aligned}$$

et de la même manière

$$\begin{aligned}\mathsf{V}[Y] &= (0-0.61)^2 \times 0.62 + (1-0.61)^2 \times 0.21 + (2-0.61)^2 \times 0.12 \\ &\quad + (3-0.61)^2 \times 0.04 + (4-0.61)^2 \times 0.01 \\ &= 0.8379\end{aligned}$$

ainsi

$$\begin{aligned}\mathsf{E}\left[\left(\overline{X}_n - \mu\right)^2\right] &= (1.16-1.16)^2 + \frac{1.4144}{25} \\ &= 0 + 0.057 = 0.057 \\ \mathsf{E}\left[\left(\overline{Y}_m - \mu\right)^2\right] &= (0.61-1.16)^2 + \frac{0.8379}{200} \\ &= 0.303 + 0.004 = 0.307\end{aligned}$$

le risque quadratique associé à l'estimateur sur le petit échantillon est plus faible car le biais d'estimation sur le grand échantillon est très important, et ce malgré une variance plus faible.

	# enfants	Population Totale	Sous-population
	0	0.4	0.62
	1	0.24	0.21
	2	0.2	0.12
	3	0.12	0.04
	4	0.04	0.01
Population	moyenne	1.16	0.61
	variance	1.41	0.84
Estimateur	biais	0	-0.55
	variance	0.057	0.004
	MSE	0.057	0.307

9.2 Chapter 5

1. Exercice page 91.
   Denote p the proportion of cases where the medecine works over 8 hours. The hypothesis that we want to test is

$$\mathsf{H}_0 : p = 0.9,$$

versus

$$\mathsf{H}_1 : p \neq 0.9.$$

For this, we estimate p as the empirical proportion $\widehat{p}$ over a sample of $n = 200$ independent individuals. We know from the Bernoulli distribution, that if we define the r.v. X_i that takes value 1 if the treatment is succesful over 8h on individual i, and $X_i = 0$ otherwise, then $\mathsf{P}(X_i = 1) = p$ with $\mathsf{E}[X_i] = p$, $\mathsf{Var}[X_i] = p(1-p)$. Since $\widehat{p} = n^{-1}\sum_{i=1}^{n} x_i$ is the realization of $\widehat{P} = n^{-1}\sum_{i=1}^{n} X_i$, then the Central Limit Theorem implies that

$$\sqrt{n}\frac{\widehat{P} - p}{\sqrt{p(1-p)}} \xrightarrow{L} \mathcal{N}(0,1),$$

which, using Slutsky's theorem, we can replace with

$$\sqrt{n}\frac{\widehat{P} - p}{\sqrt{\widehat{P}\left(1-\widehat{P}\right)}} \xrightarrow{L} \mathcal{N}(0,1),$$

and under H_0 :

$$\sqrt{n}\frac{\widehat{P} - 0.9}{\sqrt{\widehat{P}\left(1-\widehat{P}\right)}} \xrightarrow[\mathsf{H}_0]{L} \mathcal{N}(0,1).$$

We perform a test at size $\alpha = 5\%$ so, using $\widehat{p} = 160/200 = 0.8$, the test statistic is

$$z = \sqrt{200}\frac{0.8 - 0.9}{\sqrt{0.8(1-0.8)}} = -3.54$$

and, since $n > 30$ and $200 \times .9 \times (1 - 0.9) = 18 > 5$, we can use the CLT so the critical value is 1.96 for a two-sided test. Then

$$|z| > 1.96$$

so we reject H_0 and conclude that the medecine does not work as claimed by the firm.

Notice that we could have also used

$$\sqrt{n}\frac{\widehat{P} - 0.9}{\sqrt{P(1-P)}} \xrightarrow[\mathsf{H}_0]{L} \mathcal{N}(0,1)$$

with test statistic

$$\sqrt{200}\frac{0.8 - 0.9}{\sqrt{0.9(1-0.9)}} = -4.71$$

which also leads to rejecting H_0.

2. Exercise page 91: **Publishing Group AES**

   (a) The population of interest is the doctors from the Population of individuals that receive CADUCOR, the random variable is X_i which takes value 1 if the individual i responds positively to a mail offer and 0 otherwise. Hence X_i follows a Bernoulli distribution with parameter π that is the object of the study.

   (b) The hypothesis we want to test is

   $$\mathsf{H}_0 : \pi < 0.1 \quad vs \quad \mathsf{H}_1 : \pi \geq 0.1$$

   but it could also be

   $$\mathsf{H}_0 : \pi \geq 0.1 \quad vs \quad \mathsf{H}_1 : \pi < 0.1$$

   which one is preferable? Remember that we never accept a hypothesis but only fail to reject it. This means that when we do not reject we are not sure that it is true, simply we do not have enough evidence to claim that it is false. By contrast, when we reject, we have enough evidence to be quite sure (at α) that it is not true. The degree of confidence in our conclusion is higher when we reject since we control our probability of being wrong (this is the size of the test, the probability of type I error). By contrast, when we do not reject, we may in fact be doing a Type II error, whose probability we do not know (since we don't know we are wrong...).
   As a consequence, we only want to buy the CADUCOR list if we are confident that is will be useful, i.e. if we can reject the claim that it will be useless! So the problem is

   $$\mathsf{H}_0 : \pi < 0.1 \quad vs \quad \mathsf{H}_1 : \pi \geq 0.1$$

   i.e. H_0 : the list is useless, H_1 : the list is useful.
   We use the test statistic (see previous exercises)

   $$z = \sqrt{n} \frac{\widehat{\pi} - 0.1}{\sqrt{0.1\,(1 - 0.1)}}$$

   Sufficiently positive values of z lead us to rejecting H_0. Hence the critical region will be $\mathcal{C} = [c, \infty)$ for some $c > 0$; we reject H_0 if $z \in \mathcal{C}$.

   (c) Fixing $\alpha = 0.05$, and given that the sample size is large $n = 400 > 30$ and if $n\widehat{\pi}\,(1 - \widehat{\pi}) > 5$, we use the Central Limit Theorem to use as critical value the 95% percentile of the standard normal distribution (z is the realization of Z which is approximately $\mathcal{N}\,(0, 1)$), i.e. $c = 1.645$.

Hence

$$z \in \mathcal{C} \Leftrightarrow \widehat{\pi} \geq 0.1 + c\sqrt{\frac{0.1\,(1-0.1)}{n}} \Leftrightarrow \widehat{\pi} \geq 0.125$$

(d) The estimated proportion is $\widehat{\pi} = 58/400 = 0.145$ so we reject H_0 and conclude that AES should buy the CADUCOR list.

9.3 Chapter 6

1. Exercise page 101. **Restaurant Expansion.**

(a) The population we consider is the 20,000 families in the neighborhood, and the r.v. X_i gives the monthly expenditure of family i in restaurants. We observe a sample of $n = 20$ families whose expenses are independent. We assume that X_i is normally distributed so letting $\mu = \mathsf{E}[X_i]$ and $\sigma^2 = \mathsf{Var}[X_i]$,

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

and

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$$

We have seen that $(n-1)\,S_n^2/\sigma^2$ follows a χ^2_{n-1} distribution so

$$\frac{\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{(n-1)\,\frac{S^2}{\sigma^2}/\,(n-1)}} \sim T_{n-1}$$

where T_{n-1} denotes a Student distribution with $n-1$ degrees of freedom. Hence

$$\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} \sim T_{n-1}$$

The denote $\widetilde{t}_{q,n-1}$ the $100 \times q$ percentile of the T_{n-1} distribution.

$$\mathsf{P}\left\{\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} \in \left[\widetilde{t}_{q,n-1}, \widetilde{t}_{1-q,n-1}\right]\right\} = 2q$$

so setting $q = \alpha/2$ we obtain

$$\begin{aligned}
\mathsf{P}\left\{\frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} \in \left[\widetilde{t}_{\alpha/2,n-1}, \widetilde{t}_{1-\alpha/2,n-1}\right]\right\} &= 1-\alpha \\
\mathsf{P}\left\{\overline{X}_n - \mu \in \left[\widetilde{t}_{\alpha/2,n-1}\sqrt{S^2/n}, \widetilde{t}_{1-\alpha/2,n-1}\sqrt{S^2/n}\right]\right\} &= 1-\alpha \\
\mathsf{P}\left\{\mu - \overline{X}_n \in \left[-\widetilde{t}_{1-\alpha/2,n-1}\sqrt{S^2/n}, -\widetilde{t}_{\alpha/2,n-1}\sqrt{S^2/n},\right]\right\} &= 1-\alpha \\
\mathsf{P}\left\{\mu \in \left[\overline{X}_n - \widetilde{t}_{1-\alpha/2,n-1}\sqrt{S^2/n}, \overline{X}_n - \widetilde{t}_{\alpha/2,n-1}\sqrt{S^2/n},\right]\right\} &= 1-\alpha
\end{aligned}$$

By symmetry of the Student distribution, $\widetilde{t}_{1-\alpha/2,n-1} = -\widetilde{t}_{\alpha/2,n-1}$ so the confidence interval at probability $1-\alpha$ is

$$\left[\overline{X}_n - \widetilde{t}_{1-\alpha/2,n-1}\sqrt{S^2/n}, \overline{X}_n + \widetilde{t}_{1-\alpha/2,n-1}\sqrt{S^2/n},\right]$$

with $\widetilde{t}_{1-0.05/2,19} = 2.09$ and $\widetilde{t}_{1-0.01/2,19} = 2.86$. A realization of the 95% and 99% confidence intervals is

$$\mathsf{P}\left\{\mu \in \left[234 - \widetilde{t}_{1-\alpha/2,n-1}\sqrt{25/20}, 234 + \widetilde{t}_{1-\alpha/2,n-1}\sqrt{25/20},\right]\right\} = 1-\alpha$$

so the intervals are, at 95%

$$\begin{aligned}
at\ 95\% \quad &: \quad \left[234 - 2.09\sqrt{25/20}, 234 + 2.09\sqrt{25/20},\right] = [231.7, 236.3] \\
at\ 99\% \quad &: \quad \left[234 - 2.86\sqrt{25/20}, 234 + 2.86\sqrt{25/20},\right] = [230.8, 237.2]
\end{aligned}$$

(b) If the estimates were computed over a sample of 250 families, then we should use

$$\widetilde{t}_{1-0.05/2,250} = 1.97;\ \widetilde{t}_{1-0.01/2,250} = 2.60$$

so the intervals are much tighter:

$$\begin{aligned}
at\ 95\% \quad &: \quad \left[234 - 1.97\sqrt{25/250}, 234 + 1.97\sqrt{25/250},\right] = [233.4, 234.6] \\
at\ 99\% \quad &: \quad \left[234 - 2.60\sqrt{25/250}, 234 + 2.60\sqrt{25/250},\right] = [233.2, 234.8]
\end{aligned}$$

(c) Finally, with a larger estimated variance

$$at\ 95\% \quad : \quad \left[234 - 2.09\sqrt{80/20}, 234 + 2.09\sqrt{80/20},\right] = [229.8, 238.2]$$
$$at\ 99\% \quad : \quad \left[234 - 2.86\sqrt{80/20}, 234 + 2.86\sqrt{80/20},\right] = [228.3, 239.7]$$

the uncertainty is not too much larger.

9.4 Chapter 7 page 119

1. Dataset: "Flats in Paris"

(a) Linear relationship between Price (Y) and Surface (X) of flats in Paris. The estimated equation is

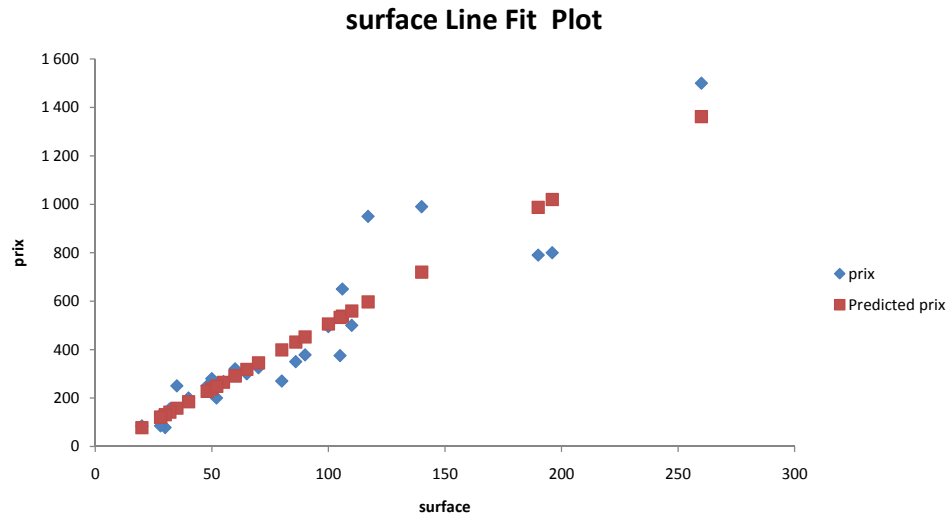
$$Y = 29.47 + 5.35X$$

here the units are not clear.

(b) Give an interpretation of the regression coefficients.
the slope coefficient of 5.35 indicates that differ by 1 unit (we assume it is square meter) will have a predicted price differential of 5.35. The fact that the intercept (the constant) is nonzero at 29.47 does not imply that a flat with zero surface has a price of 29.47. Instead, it simply means that the price is not just proportional to the price, this really affects only the very small flats. For instance, consider a flat of $20m^2$, then the predicted price is 136.47; for $100m^2$, the price is 564.47. If we were only to compute the average price per surface (for instance by estimating a regression without intercept), we find 5.11. This implies that a flat of $20m^2$ would cost 102.2 and 511 for $100m^2$. The estimated intercept implies a predicted price that is respectively 33% and 10% higher for the 20 and 100 square meter flats.

(c) Identify the presence of eventual outliers and explain their anomalous behavior.
The following graph shows the price and predicted price for the data. We see that the fit is very good for small flats (until $120m^2$) but that it is not so good for larger flats, except for the largest in the sample, rue d'Assas. Two flats (île Saint-Louis and Jardins de l'Observatoire) are more expensive than predicted and two (Avenue Rapp and Quais du 5ème) are cheaper.
There are two interpretations to this: either the four "outliers" are somewhat specific in that they differ from the rest of the sample

(depending on their own characteristics) or it is actually the Rue d'Assas flat that misleadingly leads us to believe in a linear relation that holds both for small and large flats. It may well be the case that for large flats (say above $140m^2$) the surface is not so much a good predictor as something else (location, number of bedrooms...) and that by trying to fit the model to all data points, we are making a mistake.



(d) Compute the expected price for a flat of 150 sq. m.
The expected price is $29.47 + 5.35 \times 150 \approx 832$, but as mentioned above, we should be careful here.

2. Dataset: "Sales". What is the predicted value for Sales using the available scenario?

► First Steps of the Analysis

I. What would the best choice be for a simple linear regression model? We look at the estimated correlation between sales and the other variables. The highest correlation is with total market sales. We have to be careful here, for it may be the case that our firm is the dominant one on the market and hence sales and market sales are essentially the same variable (so the correlation is of no practical use). In fact, if we look at the data, say the first observation, $SALES = 5540$ and $MARKET = 398$. To make sense of these values, it must hold that the units are not the same, so expressing *MARKET* in the same units as *SALES* implies that MARKET is $39\ 800$ or $398\ 000$ or even higher. No matter the actual value, we see that our sales represent at

most 14% of the market. Hence the correlation we find is useful.
The estimated equation is then

$$\begin{aligned} Sales &= 2960 + 5.27 \times MARKET \\ R^2 &= 0.52; \quad R^2_{adj} = 0.51 \end{aligned}$$

II. Write down the equation for the multiple regression model with all explanatory variables included.
The equation is

$$\begin{aligned} Sales = {} & b_0 + b_1\, MARKET + b_2\, DISCOUNT - b_3\, PRICE \\ & - b_4\, RESEARCH + b_5\, INVESTMENT + b_6\, AVERTIZING \\ & + b_7\, EXPENSES - b_8\, TOTAL\ ADVERTIZING \end{aligned}$$

III. Run this model with XLSTAT and compare its R^2 to the one obtained with the best simple regression model previously chosen. Give a comment.
The estimated multiple regression is

$$\begin{aligned} Sales = {} & 3130 + 4.42\, MARKET + 1.68\, DISCOUNT - 13.5\, PRICE \\ & - 3.41\, RESEARCH + 1.92\, INVESTMENT + 8.55\, AVERTIZING \\ & + 1.50\, EXPENSES - 0.02\, TOTAL\ ADVERTIZING \end{aligned}$$

with an $R^2 = 0.81,\ R^2_{adj} = 0.75$. A priori the model seems to fit a lot better since the adjusted R^2 is much higher.

IV. Give an interpretation of the estimated regression coefficients.
It is interesting to compare the signs of the multiple regression coefficient estimates with the linear correlations computed in the table. We see that when all the factors are taken into account (in the regression) the coefficient of Price on sales is negative, whereas the correlation is positive. This is one of the interesting features of multiple regression: it allows to see the specific impact of prices, everything else being constant. Simple linear correlations cannot achieve this.
Yet, the multiple regression model estimated above is not altogether satisfactory in the sense that one sign is still odd: what to make of the negative impact of Research? It is a priori counterintuitive, unless research is very badly done. One possibility is to remove this variable from the regression since it seems that its effect is not clearly estimated. We will see in later chapters how we can assess how confident we are in the estimated value. The new estimated

equation is

$$\begin{aligned} Sales &= 3140 + 4.76\, MARKET + 1.69\, DISCOUNT - 14.8\, PRICE \\ &\quad +1.88\, INVESTMENT + 8.51\, AVERTIZING \\ &\quad +0.95\, EXPENSES - 0.01\, TOTAL\ ADVERTIZING \\ R^2 &= 0.80, \quad R^2_{adj} = 0.76 \end{aligned}$$

Later chapters will show us how to evaluate this model further. One way is to see that the adjusted R^2 has increased, so it seems that our model is performing better!
To go further: we see that the data constitutes *time series* so we could also try to use the dynamic effect. For instance, we could use the price in the previous semester as an explanatory variable!

V. Use the final model to predict Sales for the scenario given for next semester
Using the model without research expenses, the predicted value for Sales next semester is

$$Sales = 5600$$

► Limitations?

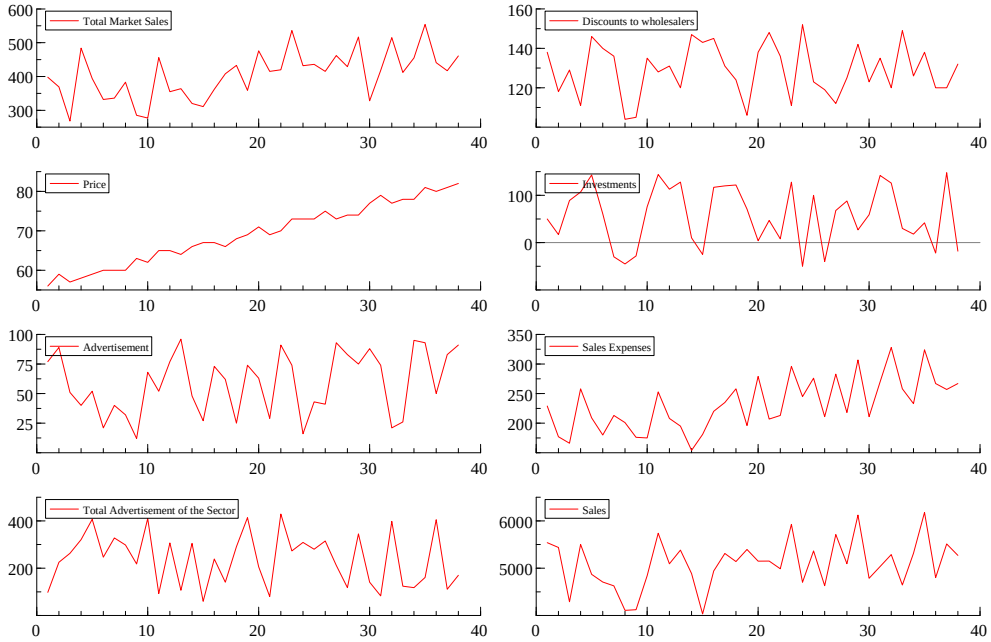
I. How confident are you about the precision of the model?
Above, we have rounded lots of the estimates but still, given that we only have 38 semesters at our disposal, it might be preferable only to use two significant digits.

II. How confident are you that each variable that you used actually matters?
Not much, some coefficients are very small, but this could be related to the fact the all variables are not expressed in the same units...
It is difficult overall here to see how to assess the validity of the estimated impacts. When the sign does not make sense, then it seems reasonable to remove the variable.

III. Do you think you could use any explanatory variable?

- We discussed this above, we have to be careful e.g. between sales and total market sales. But there are also some other problems. For instance, Total Market Sales and Sales Expenses are highly correlated (0.90) so it seems that we cannot disentangle their respective effects: we should probably use only one of the two.
- Given the time series nature of the data, we could also use *lags*

of the variables, e.g. sales in the previous period. If we plot the data as below:



we see that *Price* increase continuously. We have to be careful as none of the other variables do. So how can a continuously increasing variable be a linear combination of variables that do not grow? In fact**, it may be better to use not *Price* but the increase in *Price* from one semester to the other** (in units or in percents).

- Also, there are issues with some variables: for instance Sales Expenses is a variable that is **not exogenous** since it appears more as a consequence of Sales than as a cause (although this is debatable, we need to know more about what is measured here). Hence, we ought probably to remove it from the set of explanatory variables...

IV. How could you discriminate between the variables that matter and those that do not?

- We could for instance remove each variable in turn and see whether the adjsuted R^2 increases.
- We can also look at removing variables which are too highly correlated between themselves.

- Of course, we must always think of the signs of the estimated coefficients.
- We could also transform all explanatory variables by removing their sample means and dividing them by their standard deviation. Then all variables become comparable and so are their estimated coefficients. We can then try to remove the variables that seem to matter less (with coefficients that are closest to zero).

V. Give an interpretation of the regression coefficients, if you can.
The regression coefficients measure the impact on sales of increasing each explanatory variable by one unit, while keeping all the others constant. As mentioned above, variables which are highly correlated will vary together, so the model may not make sense: when we assume that Total Market Sales increase, we may not be justfied in assuming that sales expenses do not vary. Also, can we think of changing advertizing (and hence sales) while keeping sales expenses constant? If sales expenses are an automatic consequence of sales, the answer is no.

9.5 Chapter 8

9.5.1 Revisiting the multiple linear regression model

Consider the exercise page 121. Answer the following questions that constitute further steps to the analysis:

1. Test the significance of the different regression coefficients by looking at the confidence intervals and the p-values. Shall we retain all explanatory variables in the model?
   We only consider here the model without *Research* which anyway is not significant (t-stat: 0.519, with associated p-value: 0.608). The only variables which are significant at the 5% level, are *Total Market Size, Investment and Adverizing.* The other are not significant, with p-values above 0.5. Only *Price* is marginally insignificant,with a p-value of 0.11. Yet, this could change once we remove irrelevant variables (as it should sharpen the estimation, i.e. reduce the standard errors of the estimators, so the t-stats should increase, and hence p-values decrease, for the retained variables.)
2. Improve the initial model by means of a Backward Stepwise procedure: remove the least significant explanatory variable and run again the model; repeat the two steps till all regression coefficients are significant.

We remove *Expenses*, then *Total Advertizing*, then *Discounts*. *Price* remains non-significant even then, with a p-value of 0.12. We therefore remove it, but this reduces the adjusted R^2 so it may be interesting to keep it since we have only few observations.

3. Give an interpretation of the regression coefficients in the finally selected model
   The finally selected model is (standard errors in parentheses):

$$\begin{aligned}\text{Sales}_t &= \underset{(391)}{3205} - \underset{(7.06)}{11.03}\ \text{Price}_t + \underset{(0.711)}{5.087}\ \text{Total Market Sales}_t \\ &\quad + \underset{(0.687)}{2.039}\ \text{Investments}_t + \underset{(1.69)}{7.932}\ \text{Advertisement}_t\end{aligned}$$

   with an $R^2_{adj} = 0.78$, or

$$\begin{aligned}\text{Sales} &= \underset{(245)}{2722} + \underset{(0.598)}{4.457}\ \text{Total Market Sales}_t + \underset{(0.681)}{2.302}\ \text{Investments}_t \\ &\quad + \underset{(1.65)}{7.183}\ \text{Advertisement}_t\end{aligned}$$

   with an $R^2_{adj} = 0.77$.
   The coefficients are the marginal impact of increase each of the variable by one unit keeping the other constant. Here, we see that the explanatory variables are not too highly correlated, so the *ceteris paribus* assumption makes sense (careful: price and total market size are somewhat correlated, at 0.5, so this is why it might be better to leave price out of the model).

4. Use the final model to predict the *Sales* for the scenario given for next semester.
   Model 1 (with price):

$$\begin{aligned}Sales_t &= 3205 - 11.03 \times 83 + 5.087 \times 500 + 2.039 \times 50 + 7.932 \times 90 \\ &= 5650\end{aligned}$$

   and model 2 without price:

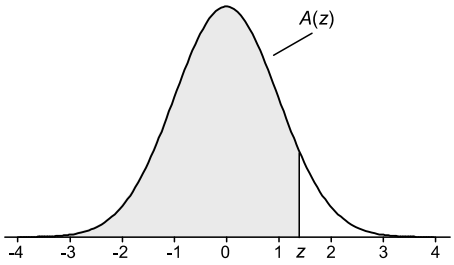
$$\begin{aligned}Sales_t &= 2722 + 4.457 \times 500 + 2.302 \times 50 + 7.183 \times 90 \\ &= 5710\end{aligned}$$

   both models predict higher sales than the model that includes all variables.

Chapter 10

Statistical Tables

Cumulative Standardized Normal Distribution

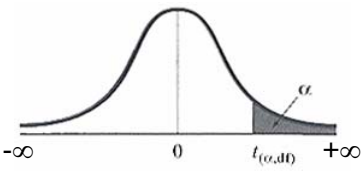


$A(z)$ is the integral of the standardized normal distribution from $-\infty$ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	$A(z)$	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail

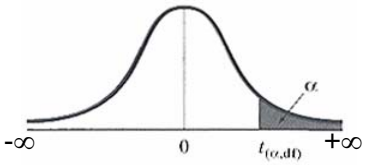
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

Student's T_n distribution



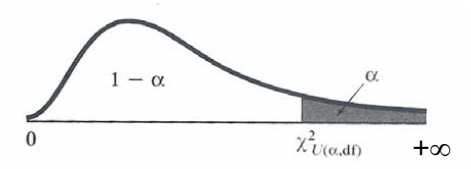
n degrees of freedom	Area on the right tail					
	0.25	0.1	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778

Student's T_n distribution



n degrees of freedom	Area on the right tail					
	0.25	0.1	0.05	0.025	0.01	0.005
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.6772	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.6772	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.6771	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.6771	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.6771	1.2905	1.6611	1.9853	2.3662	2.6286
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
101	0.6769	1.2900	1.6601	1.9837	2.3638	2.6254
102	0.6769	1.2899	1.6599	1.9835	2.3635	2.6249
103	0.6769	1.2898	1.6598	1.9833	2.3631	2.6244
104	0.6769	1.2897	1.6596	1.9830	2.3627	2.6239
105	0.6768	1.2897	1.6595	1.9828	2.3624	2.6235

Table for the distribution χ^2_n



n degrees of freedom	Area on the right tail										
	0.995	**0.990**	**0.975**	**0.950**	**0.900**	**0.750**	**0.250**	**0.100**	**0.050**	**0.025**	**0.010**
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635
2			0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.070	12.833	15.086
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.920	24.725
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932
22	8.643	9.542	10.982	12.338	14.041	17.240	26.039	30.813	33.924	36.781	40.289
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.195	46.963
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278
29	13.121	14.256	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.588
30	13.787	14.953	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892