



The more you know, the more you dare®

**HEC / Master in Management**

# **STATISTICS (AND BASIC ECONOMETRICS)**

**Course coordinator:**

**Gilles Fortin-Stoltz (stoltz@hec.fr)**

**Instructors:**

**Vasiliki Kostami (kostami@hec.fr)**

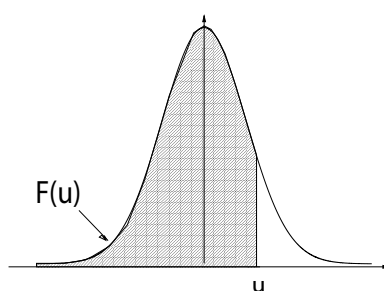
**Benjamin Petiau (benjamin.petiau@gmail.com)**

**Dinah Rosenberg (rosenberg@hec.fr)**

**Dan Xie (dan.xie@hec.edu)**

**Academic year 2022–2023**





	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

$u$	3.0	3.1	3.2	3.3	3.4
$F(u)$	0.99865	0.999032	0.999313	0.999517	0.999663
$u$	3.5	3.6	3.7	3.8	3.9
$F(u)$	0.999767	0.999841	0.999892	0.999928	0.999952
$u$	4.0	4.1	4.2	4.3	4.4
$F(u)$	0.999968	0.999979	0.999987	0.999991	0.999995
$u$	4.5	4.6	4.7	4.8	4.9
$F(u)$	0.999997	0.999998	0.999999	0.999999	1





## Contents

Standard normal table	3
Rules and evaluation	9
Chapter 1. Statistical thinking: sample the world!	13
1 Methodology: sampling!	14
2 Various sampling biases	16
3 More advanced exercises (quiz-like exercises)	20
Discounts to increase the number and amounts of orders	20
Car insurance company	20
Chapter 2. Confidence intervals: the basics	21
1 Notion of confidence intervals	21
2 Confidence intervals for proportions $p_0$	23
3 Confidence intervals for general population averages $\mu_0$	25
4 How nice these formulas are!	27
5 Conclusion: business insights and stories	28
6 Elementary exercises	30
7 Mathematical appendix	31
8 More advanced exercises (quiz-like exercises)	32
Discounts, continued	32
Car insurance company, continued	32
An advertisement featuring statistics (Lapeyre carpenter company)	33
TwitterAudit	34
Chapter 3. Confidence intervals: advanced notions	37
1 Corrections to perform in two specific cases	38
2 Simultaneous confidence intervals — the Bonferroni correction	39
3 Sample-size determination (survey planning)	40
4 Elementary exercises	42
5 More advanced exercises (quiz-like exercises)	43
Gender pay gap?	43
Sample size determination, in a different way though...	43

Budget planning for travelling costs (AirBnb and SNCF) . . . . .	44
The effect of touch, first version . . . . .	46
Success rate of a new dating method . . . . .	48
Car insurance company: planning . . . . .	48
Value of a stock . . . . .	49
Chapter 4. Hypothesis testing: methodology . . . . .	51
1 The Lady tasting tea . . . . .	52
2 Detecting cheaters . . . . .	55
3 Advertisement campaign for nicotine patches . . . . .	56
4 General methodology and notion of P-value . . . . .	59
5 Elementary exercises . . . . .	62
6 More advanced exercises (quiz-like exercises) . . . . .	63
Nicotine patches, continued . . . . .	63
Managing customers' dissatisfaction . . . . .	63
Chapter 5. One-sample tests . . . . .	65
1 Learning objectives . . . . .	65
2 A smooth start with proportions . . . . .	66
3 The case of general means . . . . .	68
4 Using a statistical software . . . . .	72
5 Some food for thought: one-sided versus two-sided tests . . . . .	74
6 Elementary exercises . . . . .	76
7 More advanced exercises (quiz-like exercises) . . . . .	78
A controversial governmental reform, short version . . . . .	78
Seizure of MegaUpload and side effects . . . . .	78
Public health surveillance . . . . .	80
"We look like our names" . . . . .	80
A controversial governmental reform, long version . . . . .	82
Walking many steps a day . . . . .	84
Chapter 6. Two-sample tests . . . . .	87
1 Paired data / for general means . . . . .	88
2 Independent data / for proportions . . . . .	92
3 Independent data / for general means . . . . .	96
4 Independent data / two mistakes to avoid! . . . . .	101
5 Elementary exercises . . . . .	103
6 More advanced exercises (quiz-like exercises) . . . . .	105
Separate or pooled marketing campaign? . . . . .	105
Comparing prices in two local supermarkets . . . . .	106
Gender pay gap, revisited . . . . .	108
Alcohol consumption during the POWs at HEC . . . . .	108
Cash in the wallet, by country . . . . .	110
Speedy self-assessment . . . . .	111
The effect of touch, re-worked . . . . .	112
Choosing between two gifts . . . . .	114
Lead levels in children's blood . . . . .	116

Chapter 7.	$\chi^2$ -tests	119
1	$\chi^2$ -test of goodness of fit . . . . .	120
2	$\chi^2$ -test of independence (homogeneity) . . . . .	125
3	Elementary exercises . . . . .	129
4	More advanced exercises (quiz-like exercises) . . . . .	132
	Reading habits per socio-professional category . . . . .	132
	The sinking of the Titanic . . . . .	134
	Hair color by gender . . . . .	134
	Call centers . . . . .	135
	“Junior entreprise” campaign at HEC Paris . . . . .	136
	Mendel’s experiment . . . . .	138
	Alcohol consumption at HEC Paris . . . . .	139
	M&M colors . . . . .	140
	Satisfaction survey . . . . .	142
	Births by month . . . . .	144
Chapter 8.	Simple linear regression	145
1	Basic notions, descriptive statistics, and a first significance test . . . . .	150
2	Inferential statistics: significance tests on the coefficients . . . . .	153
3	General summary: how to read SPSS outputs . . . . .	154
4	Prediction of new values, detection of outliers . . . . .	158
5	Mathematical appendix: formulas . . . . .	161
6	Elementary exercises . . . . .	162
7	More advanced exercises (quiz-like exercises) . . . . .	164
	The historical example of regression . . . . .	164
	An example of a spurious correlation . . . . .	166
	Some French politics: demonstrations . . . . .	170
	Prices of ski passes . . . . .	172
Chapter 9.	Multiple linear regression	173
1	What to read in a given, single output . . . . .	175
2	Model comparison, variable selection . . . . .	179
3	Regression problems extracted from past exams . . . . .	180
	Wage discrimination? . . . . .	181
	Modeling life expectancy . . . . .	195



## Rules and evaluation

### Tentative schedule (of the classes and of the required homework)

This course is in the flipped classroom format:

- at home and before a course session, you will read the course material and solve the elementary exercises of a given chapter (the corrections thereof are to be found on BlackBoard);
- during the session, we will discuss and solve some of the advanced exercises stated in the present handbook (all extracted from past exams).

Session	What we will do in class	Beforehand preparation
Jan. 26	Advanced exercises of Chapter 1	Chapter 1: read + solve elementary ex.
Feb. 2	Advanced exercises of Chapter 2	Chapter 2: read + solve elementary ex.
Feb. 9	Advanced exercises of Chapter 3	Chapter 3: read + solve elementary ex.
Feb. 16	Course recap on Chapter 4	Chapter 4: read + solve elementary ex.
Feb. 23	Advanced exercises of Chapter 5	Chapter 5: read + solve elementary ex.
Mar. 9	<b>Mid-term exam</b>	Review all material seen so far
Mar. 30	Advanced exercises of Chapter 6	Chapter 6: read + solve elementary ex.
Apr. 6	More exercises on Chapter 6	Chapter 6: read again
Apr. 13	Advanced exercises of Chapter 7	Chapter 7: read + solve elementary ex.
Apr. 20	Advanced exercises of Chapter 8	Chapter 8: read + solve elementary ex.
Apr. 27	One regression problem of a past exam	Chapter 9: read
May 11	Another regression problem	Review material of Chapters 8 and 9
After May 22	<b>Final exam</b>	Review all material seen so far

The above schedule was established based on the BDE campaign and on the academic calendar:

- No campaign (thus full academic work) till March 5, including a vacation week from February 25 to March 5; this corresponds to the first five sessions of this course;
- Pre-campaign from March 6 to March 26, which should not prevent students from obtaining an excellent performance in the mid-term for this course, scheduled on March 9;
- Campaign from March 27 to April 17, leaving a large amount of time after it to catch up and get prepared for the final exam (to be taken at earliest on May 22).

## Evaluation and grading rules

### One mid-term exam in class

Takes place on March 9, 2023, see the previous page

Set-up: all students of a given group will be physically present in a single classroom

Will last about 1h 25 minutes, no document allowed (but formulas will be recalled in the statement)

Calculator needed!

### One final exam consisting of two independent problems

Takes place in May 2023

Length: 1h30

Open book for printed documents (but no access to electronic documents: the Theia exam platform will be in full-screen mode)

With a strong recommendation: Write a two-page summary of the course (a “cheat sheet”)

Calculator needed as well!

### Evaluation structure

A grade between 0 and 100, converted into a A–F letter; the conversion is made based on quotas dictated by the school (main rules thereof are: best 10% to 20% grades get an A, best next batch of 10% to 20% grades get a B, best next batch of 20% to 40% grades get a C).

Determination of the grade over 100: denote by

- M the grade over 35 obtained at the mid-term exam,
- P1 grade over 35 obtained at problem #1 of the final exam,
- P2 grade over 30 for problem #2 of the final exam,

then the final grade is

$$\max\{M + P1 + P2, 2 * P1 + P2\}$$

I.e.: the mid-term exam can only increase the final grade and can be replaced by Problem 1 of the final exam. That is, the mid-term exam is meant to be an opportunity to test yourselves, not a way to punish students whose learning curve is slower than others.

Exception to this formula:

Students that were absent for no good reason at the mid-term will not benefit from the “max” in the formula and will get  $M + P1 + P2$  as a final grade, with  $M = 0$  due to their absence. I.e., they will be graded over 65, and not over 100. This is a harsh penalty, therefore, students should make sure in advance that their reasons for absence are considered valid—do so by contacting your instructor, see below.

### Mid-term exam absence policy

Since the mid-term exam can only increase the final grade, we will not offer any make-up mid-term exam to absentees.

Good reasons for absence are only: 1. sickness (in which case you must spontaneously provide a medical certificate within 48h) and 2. vital family or personal events (e.g.: death of a relative, your own graduation ceremony, etc.). In these cases, as indicated above, the final score will be given by  $2 * P1 + P2$ .

Internship interviews, sports competitions, actions for associations, representing HEC at some forum or event, etc., are not valid reasons of absence. In these cases, as indicated above, the final score will be given by  $0 + P1 + P2$ .

## Miscellaneous comments

### M1 students: waivers

A massive waiver campaign took place in June/July (150+ students participated to it!). We expect all students with bachelors in Business Administration, Economics, Engineering Sciences, etc., to have applied for and have been granted a waiver for this basic statistics course. If you have not heard from us, the waiver was granted. We only reached out to students whose waiver application was incomplete or denied.

Our basic assumption, if you are reading this handbook, will thus be that you have almost no mathematical background—e.g., that you did not further study mathematics after high school.

Given this assumption and given that we only have 18h for our crash course in statistics, we will not perform any introduction to probability theory and will just take the statistical formulas as given. We will not explain at length how they are derived, we will merely learn to apply them on data. A second and equally important aim will be to learn how to read/write a statistical report and take business actions accordingly while understanding that there is a variability / an uncertainty associated with each statistical study and that there are risks in each action or lack of action undertaken.

Note: It is too late now to apply for an official waiver. Waivers of attendance can be discussed (but you will have to take the mid-term and final exams). Please email your instructor and put the course coordinator in CC.

### Review sessions

Review sessions will be organized for students struggling with the statistical concepts and formulas. Vasiliki Kostami will be in charge for English groups, and Gilles Fortin-Stoltz for French groups. Details will be communicated in due time. We expect to hold 5 sessions.

### Further references

Should you feel the need of alternative explanations or additional exercises, you may use:

- *Statistics for Business and Economics*, by Newbold, Carlson and Thorne
- *Statistics for Business and Economics*, by McClave, Benson and Sincich

(both available at the HEC learning center).



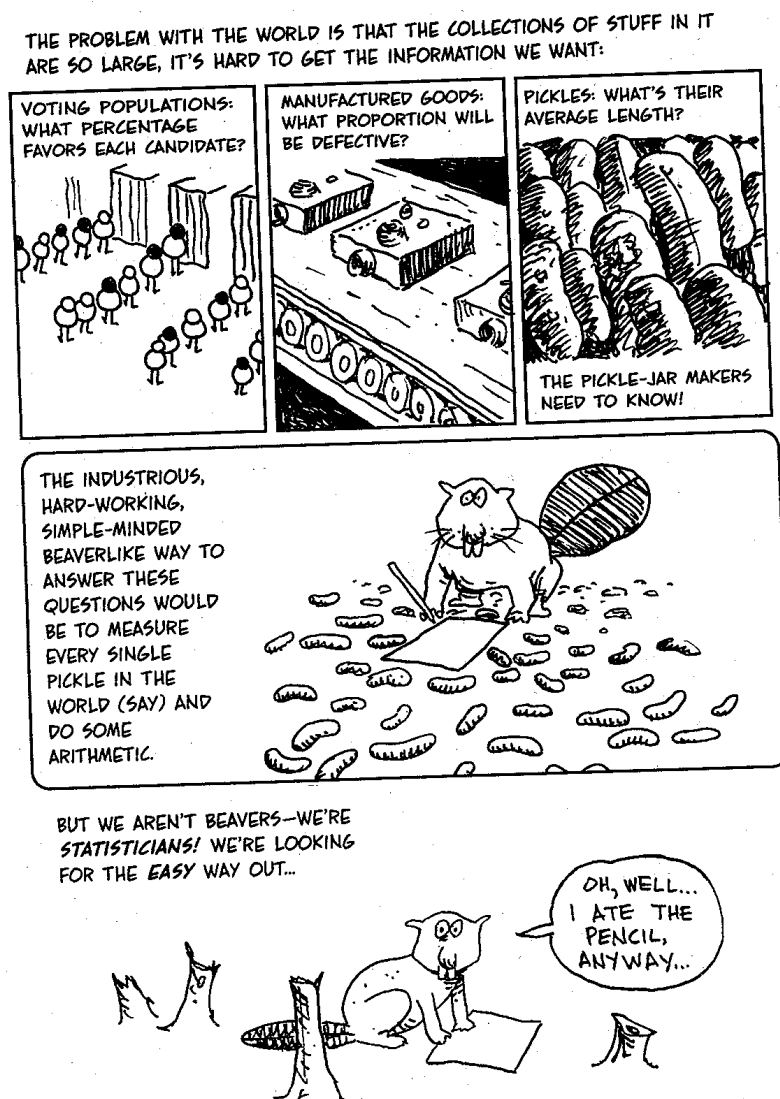


## Statistical thinking: sample the world!

*In God we trust, all others bring data.*

William Edwards Deming

(American engineer, statistician, professor, author, lecturer, and management consultant, 1900–93)



## 1. Methodology: sampling!

A population is the set of units (e.g., objects, people, events, etc.) which we are interested in studying. An experimental (or observational) unit is an object (e.g., thing, person, event, etc.) upon which we collect data.

The sample is the set of all experimental units.

We cannot, in general, collect data for all population units. The sample is a strict (and small) subset of the (possibly very large) population.

A variable is a characteristic or property of a unit (e.g., its size, its price, etc.).

We are interested in its average value on the population.

The latter cannot be computed because the population cannot be studied in its entirety.

We measure the variable of interest on the sample units.

It turns out that the sample average of this variable is most of the time close to the population average, as long as the sample was large enough.

(This result is called the “law of large numbers” in mathematics.)

### Extract the relevant statistical information from an exercise statement

This is typically the first question of any exercise statement! You need to identify:

- the population, the sample and the variable of interest;
- the parameter of interest (the population average related to the variable of interest);
- give names to the data available and interpret each data element;
- summarize the data (sample average and if applicable, sample standard deviation).

Beware! We will introduce some mathematical symbols to refer to some of the elements above.

### Situation 1: Binary data

#### Statement:

A survey on the popularity rating of President Emmanuel Macron is conducted on the phone between June 14 and June 24, 2017 (a few weeks after his election). Out of the respondents, 1,883 expressed an opinion on Emmanuel Macron: 1,205 had a positive opinion of his action, while 678 had a negative opinion.

Extract the relevant statistical information from this exercise statement.

#### Answer:

The statistical units at hand are French inhabitants (with a phone number and with an opinion about Emmanuel Macron). The population is made of all these inhabitants: this is the target of the survey, among whom we sample.

A sample of 1,883 respondents is created. (We discard the probably many respondents that picked up the phone but had no opinion or did not want to share it.)

The variable of interest (what we measure for each unit of the sample) is whether the respondent has a positive opinion (to be coded by 1) about Emmanuel Macron or not (to be coded by 0).



The corresponding parameter of interest is the average positive notoriety of Emmanuel Macron in the population: it is a proportion or frequency denoted by  $p_0$ . (Put differently, a fraction  $p_0$  of the population has a positive opinion about Emmanuel Macron.)

The data collected consists of numbers  $x_1, \dots, x_{1883}$  in  $\{0, 1\}$ : we denote  $x_j = 1$  when the  $j$ -th respondent declared a positive opinion and  $x_j = 0$  otherwise.

The data can be summarized as follows: in the sample, a fraction

$$\bar{x}_{1883} = \frac{1,205}{1,883} = 64\%$$

of the respondents declared a positive opinion about Emmanuel Macron.

**Remark 1.1** (Do not mix up  $\bar{x}$  and  $p_0$ !). Of course, this sample average  $\bar{x}_{1883} = 64\%$  is not exactly equal to the population average  $p_0$ , which is unknown (and this is why we conduct a survey!). But we are confident that the two quantities should not be too far away...

If you conduct the survey again, even during the same period of time, you will obtain a different sample average  $\bar{x}_n$ , with  $n$  the actual number of respondents, while the parameter of interest  $p_0$  does not change. This illustrates a certain variability in the statistics that you can compute. However, we will see in the next chapter why this variability is fortunately limited.

### Situation 1, continued: Categorical data

Categorical data corresponds to the case where there is a finite number of possible values for the data points  $x_j$ , which we denote by  $\{0, 1, \dots, k-1\}$ . The binary cases is the special case when  $k = 2$ .

This type of data will be omitted for now but studied extensively in Chapter 7.

### Situation 2: General quantitative data

#### Statement:

The typical French quick meal is (used to be?) a sandwich with buttered baguette bread, a slice of ham, and possibly a slice of emmental cheese. It is called a «jambon-beurre» (ham-butter). The price of this basic sandwich can be used to compare the levels of prices within France (as the Big Mac index helps comparing price levels around the world!). To determine the level of price for Paris, we sample 200 bakeries and check the corresponding prices. The average sample price equals 4.35 euros, with a standard deviation of 1.55 euros.



Extract the relevant statistical information from this exercise statement.

#### Answer:

The population is made of all Parisian bakeries.

It is, by the way, rather easy to sample within this population: get a list of all Parisian bakeries (thanks to the phonebook) and use Excel to sample at random in this file. Such a sample of 200 bakeries is created. Some cheap labor force is then sent on the field to enter in each bakery and read the price of a «jambon-beurre» sandwich.

As we all realized now, the variable of interest is the price of the sandwich (while the statistical units are the bakeries).

The parameter of interest, which we could call the «jambon-beurre» index of Paris, will be denoted by  $\mu_0$ . (The Greek letter for «m», where «m» stands for mean.) It is the average of sandwich prices over all Parisian bakeries (there are so many that you do not want to survey them all).

The data collected can be denoted by  $x_1, \dots, x_{200}$ , where each  $x_j$  denotes the price observed in the  $j$ -th bakery.

The sample data can be summarized as follows: its mean equals

$$\bar{x}_{200} = \frac{1}{200} \sum_{i=1}^{200} x_i = 4.35$$

while its standard deviation equals  $s_{x,200} = 1.55$ , where we recall that

$$s_{x,200} = \sqrt{\frac{1}{199} \sum_{i=1}^{200} (x_i - \bar{x}_{200})^2}.$$

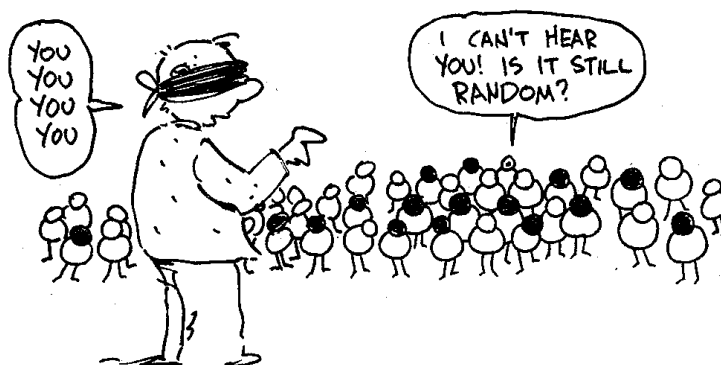
**Remark 1.2** (What is standard deviation?). If you have never seen standard deviation so far, do not panic. The exercise statements will always give its value and you do not need to understand in a fine way what it measures. Just take for granted that it measures the spread of the data (the larger the standard deviation, the more spread the data). Its general formula is the following: given data  $x_1, \dots, x_n$  with mean  $\bar{x}_n$ , its standard deviation equals

$$s_{x,n} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

## 2. Various sampling biases

A good way to collect data (with good theoretical guarantees) is to sample uniformly at random in the population.

NOT TO PROLONG THE MYSTERY, THE WAY TO GET STATISTICALLY DEPENDABLE RESULTS IS TO CHOOSE THE SAMPLE AT **random**.



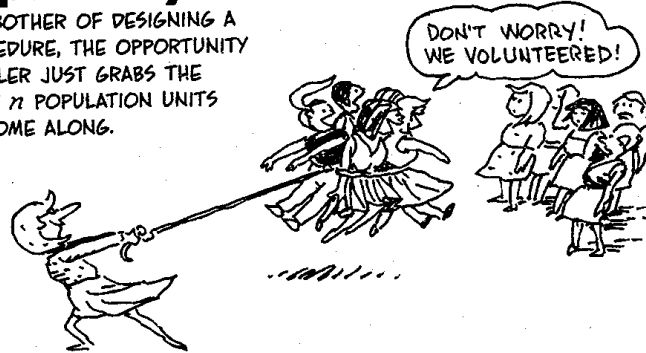
This is easy when you sample out of a customer database.

However, this may be difficult in many other cases due to several biases which we review below.

In practice, statisticians rather use quotas in their surveys to make sure that their sample represents well the population at hand (that it is a representative sample).

**Opportunity / Motivation bias:** Consists of getting the respondents come to you. The correct practice is to figure them out in advance.

A COMMONLY USED METHOD IS ESPECIALLY PRONE TO BIAS: IT'S CALLED AN **opportunity** SAMPLE. AVOIDING ALL THE BOTHER OF DESIGNING A PROCEDURE, THE OPPORTUNITY SAMPLER JUST GRABS THE FIRST  $n$  POPULATION UNITS TO COME ALONG.



A CLASSIC EXAMPLE IS SHERE HITE'S BOOK, *WOMEN AND LOVE*. 100,000 QUESTIONNAIRES WENT TO WOMEN'S ORGANIZATIONS (AN OPPORTUNITY SAMPLE), ONLY 4.5% WERE FILLED OUT AND RETURNED (RESPONSE BIAS). SO HER "RESULTS" WERE BASED ON A SAMPLE OF WOMEN WHO WERE HIGHLY MOTIVATED TO ANSWER THE SURVEY'S QUESTIONS, FOR WHATEVER REASON.



Another example features unhappy customers: these are more willing to participate in satisfaction surveys. This is why you need to figure out the respondents in advance (out of the customer database) and chase them till they answer; you may think of offering a small gift or reward to make sure that even the happy customers express themselves.

**Phone surveys:** Many young adults only have cell phones and there is no phonebook for cell phones!



**Internet surveys:** Same kind of issue, except that this time, you rather lose the elder ones.

**Endogamy bias:** Do not create a survey intended to students and have it circulate via your Face-Book friends! Otherwise, you will absolutely not sample uniformly all students and get a sample representative of the average students. On the contrary, your sample will be full of students that resemble you!

**Temporal biases:** If you are to study the customers of a store, do not just come when it is convenient for you (e.g., on a Saturday morning). You need to come regularly throughout the week and interview people in proportion of the total crowd in the store at this moment.

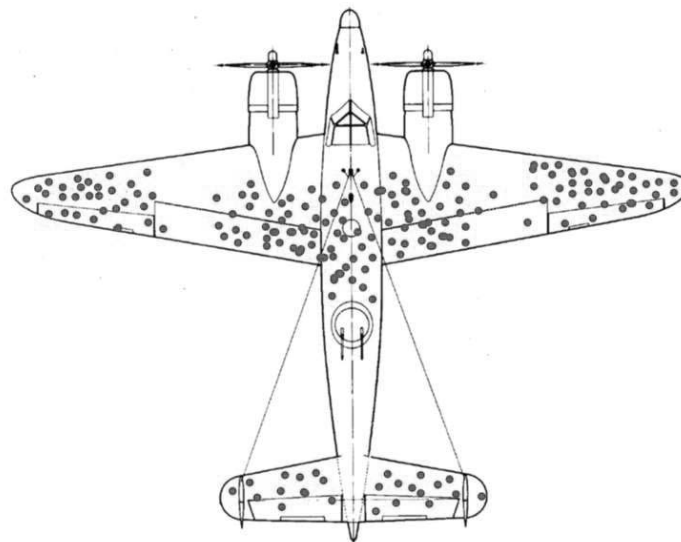
**And so on:** It is easy to list all possible mistakes one can make. It is much harder to conduct a study the right way (except for the case when the population at hand is formed by a customer database, in which sampling uniformly at random is easy). In practice, polling organizations use quotas (in terms of age, income, geography) to make sure that the samples gathered look like the general population (in the case when a preliminary census indicates how general population is distributed among various categories). Alternatively, they may weigh their respondents to correct for various biases. A good, though extreme, example of such weights is provided in the article on the next page. Another famous example of a biased sample is provided below.

Summary (author unknown) of a blog note by Chris Dobbert, March 4, 2019

See <https://medium.com/the-innovation/the-missing-bullet-holes-and-abraham-wald-25e68d7a870f>

### The missing bullet holes and Abraham Wald

During WWII, the Navy tried to determine where they needed to armor their aircraft to ensure they came back home. They ran an analysis of where planes had been shot up, and came up with this.



Obviously the places that needed to be up-armored are the wingtips, the central body, and the elevators. That's where the planes were all getting shot up.

Abraham Wald, a statistician, disagreed. He thought they should better armor the nose area, engines, and mid-body. Which was crazy, of course. That's not where the planes were getting shot.

Except Mr. Wald realized what the others didn't. The planes were getting shot there too, but they weren't making it home. What the Navy thought it had done was analyze where aircraft were suffering the most damage. What they had actually done was analyze where aircraft could suffer the most damage without catastrophic failure. All of the places that weren't hit? Those planes had been shot there and crashed. They weren't looking at the whole sample set, only the survivors.

The Upshot  
THE 2016 RACE

## How One 19-Year-Old Illinois Man Is Distorting National Polling Averages

Nate Cohn @Nate\_Cohn OCT. 12, 2016

There is a 19-year-old black man in Illinois who has no idea of the role he is playing in this election. He is sure he is going to vote for Donald J. Trump. And he has been held up as proof by conservatives — including outlets like Breitbart News and The New York Post — that Mr. Trump is excelling among black voters. He has even played a modest role in shifting entire polling aggregates, like the Real Clear Politics average, toward Mr. Trump.

How? He's a panelist on the U.S.C./LAT poll, which has emerged as the biggest polling outlier of the presidential campaign. Despite falling behind by double digits in some national surveys, Mr. Trump has generally led in the U.S.C./LAT poll. He held the lead for a full month until Wednesday, when Hillary Clinton took a nominal lead.

Our Trump-supporting friend in Illinois is a surprisingly big part of the reason. In some polls, he's weighted as much as 30 times more than the average respondent, and as much as 300 times more than the least-weighted respondent. Alone, he has been enough to put Mr. Trump in double digits of support among black voters. He can improve Mr. Trump's margin by 1 point in the survey, even though he is one of around 3,000 panelists. He is also the reason Mrs. Clinton took the lead in the U.S.C./LAT poll for the first time in a month on Wednesday. The poll includes only the last seven days of respondents, and he hasn't taken the poll since Oct. 4. Mrs. Clinton surged once he was out of the sample for the first time in several weeks.

How has he made such a difference? And why has the poll been such an outlier? It's because the U.S.C./LAT poll made a number of unusual decisions in designing and weighting its survey. It's worth noting that this analysis is possible only because the poll is extremely and admirably transparent: it has published a data set and the documentation necessary to replicate the survey. Not all of the poll's choices were bound to help Mr. Trump. But some were, and it all combined with some very bad luck to produce one of the most persistent outliers in recent elections.

### Tiny Groups, Big Weights

Just about every survey is weighted — adjusted to match the demographic characteristics of the population, often by age, race, sex and education, among other variables. The U.S.C./LAT poll is no exception, but it makes two unusual decisions that combine to produce an odd result.

*It weights far very tiny groups, which results in big weights.*

A typical national survey usually weights to make sure it's representative across pretty broad categories, like the right number of men or the right number of people 18 to 29. The U.S.C./LAT poll weights for many tiny categories: like 18- to 21-year-old men, which U.S.C./LAT estimates make up around 3.3 percent of the adult citizen population. Weighting simply for 18- to 21-year-olds would be pretty bold for a political survey; 18- to 21-year-old men is really unusual.

On its own, there's nothing necessarily wrong with weighting for small categories like this. But it's risky: Filling up all of these tiny categories generally requires more weighting. A run of the U.S.C./LAT poll, for instance, might have only 15 or so 18- to 21-year-old men. But for those voters to make up 3.3 percent of the weighted sample, these 15 voters have to count as much as 86 people — an average weight of 5.7.

When you start considering the competing demands across multiple categories, it can quickly become necessary to give an astonishing amount of extra weight to particularly underrepresented voters — like 18- to 21-year-old black men. This wouldn't be a problem with broader categories, like those 18 to 29, and there aren't very many national polls that are weighting respondents up by more than eight or 10-fold. The extreme weights for the 19-year-old black Trump voter in Illinois are not normal.

### It weights by past vote.

The U.S.C./LAT poll does something else that's really unusual: it weights the sample according to how people said they voted in the 2012 election. Its weights are such that Obama voters represent 27 percent of the sample and Romney voters represent 25 percent, reflecting the split of 51 to 47 percent among actual voters in 2012. The rest include those who stayed home or who are newly eligible to vote. I'm not aware of any reputable public survey that weights self-reported past vote back to the actual reported results of an election. You can read more about the U.S.C./LAT "past vote" issue in this August article (link broken), but the big problem is that people don't report their past vote very accurately. They tend to over-report three things: voting, voting for the winner and voting for some other candidate. They underreport voting for the loser.

The same thing is true in the U.S.C./LAT poll. If the survey didn't include a past vote weight, the past vote of its respondents would be Obama 38, Romney 30. This is a lot like national surveys that were published around the same time as the U.S.C./LAT poll, like those from NBC/WSJ or the NYT/CBS News. By emphasizing past vote, they might significantly underweight those who claim to have voted for Mr. Obama and give much more weight to people who say they didn't vote.

### Two Key Factors

These two factors — an overweighted sample and the use of past vote — seem to explain the preponderance of the difference between the U.S.C./LAT poll and other surveys.

If the poll was weighted to a generic set of census categories like most surveys (four categories of age, five categories of education, gender and four categories of race and Hispanic origin), Mrs. Clinton would have led in every iteration of the survey except the period immediately after the Republican convention. The U.S.C./LAT poll weights for all of these demographic categories; it just weights to smaller groups.

About half of the difference is attributable to the small demographic categories that lead the 19-year-old black Trump voter in Illinois to get huge weights. The other half of the difference is because of the past vote weight. Of the two factors, it was probably inevitable that using "past vote" would create a problem. The potential biases of weighting by past vote are pretty well established.

But the costs of the U.S.C./LAT poll's extensive weighting were not so inevitable. Jill Darling, the survey director at the U.S.C. Center for Economic and Social Research, noted that they had decided not to "trim" the weights (that's when a poll prevents one person from being weighted up by more than some amount, like five or 10) because the sample would otherwise underrepresent African-American and young voters.

This makes sense. Gallup got itself into trouble for this reason in 2012: it trimmed its weights, and nonwhite voters were underrepresented. In general, the choice in "trimming" weights is between bias and variance in the results of the poll. If you trim the weights, your sample will be biased — it might not include enough of the voters who tend to be underrepresented. If you don't trim the weights, a few heavily weighted respondents could have the power to sway the survey. The poll might be a little noisier, and the margin of error higher (note that the margin of error on the U.S.C./LAT poll for black voters surges every time the heavily weighted young black voter enters the survey).

But the U.S.C./LAT poll is a panel — which means it recontacts the same voters over and over — and so it wound up with the worst of both worlds. If the U.S.C./LAT poll were a normal poll, the 19-year-old from Illinois might have been in the poll only once. Most of the time, the heavily weighted young black voters would lean toward Mrs. Clinton — ensuring that the poll both had the appropriate number of black voters, and a relatively representative result. But the U.S.C./LAT poll had terrible luck: The single most overweighted person in the survey was unrepresentative of his demographic group. The people running the poll basically got stuck at the extreme of the added variance.

By design, the U.S.C./LAT poll is stuck with the respondents it has. If it had a slightly too Republican sample from the start — and it seems it did, regardless of weighting — there was little it could do about it.

### 3. More advanced exercises (quiz-like exercises)

**Advanced exercise 1.1** (Discounts to increase the number and amounts of orders). Suppose that Nozama.com is a leading web site primarily selling books and, on second thoughts, all types of products. Their customer database contains hundreds of thousands of customers. On average, 13% of them place at least one order per given trimester. When multiple orders were placed by the same customer, we combine them by summing their amounts. The per-customer average amount of such (aggregated) orders is 70 euros.

Out of all customers, 1,000 are picked at random and get a discount coupon on their purchase for the upcoming trimester: 5% extra discount on everything or almost<sup>1</sup>. The results of this policy are summarized in the table below.

	Sample size	1,000
	Number of customers with at least one order	170
	Per-customer average amount of orders (before discount is applied)	73 €
	Standard deviation of these amounts	8 €

Extract the relevant statistical information from this exercise statement. Beware of the interpretations for the parameters of interest: think twice about the verb tenses to use!

**Advanced exercise 1.2** (Car insurance company). An HEC alumni has a brilliant idea: create a car-insurance company, operating in France, dedicated<sup>2</sup> to students, with insurance policies and premiums tailored to their needs and financial resources. To get a first rough idea of the viability of this idea, he has a survey conducted on French students about their car accidents in the past year, and whether they have been held responsible for them or not, and if so, how much it costs in total. (Insurance companies need only to pay for the expenses in this case.) Out of the 3,000 students interviewed, 1,472 only had an insurance in place in their names; out of them 256 report an accident for which they were held responsible, with an average amount of damages of 1,865 euros (and a standard deviation of 524 euros).

1. Explain how to conduct the survey, how to select the 3,000 students, and which biases to avoid.
2. Extract the relevant statistical information from this exercise statement. Beware of the interpretations for the parameters of interest: think twice about the verb tenses to use!

**Advanced exercise 1.3.** Your instructor may pick virtually any exercise of this textbook and extract the relevant statistical information from its statement. Good candidates are: “TwitterAudit” (page 34), “Value of a stock” (page 49), and “Walking many steps a day” (page 84).

<sup>1</sup>French culture tip: The prices for books are not freely set. By law, they are determined by the publisher and apply to all stores. Maximal discounts on book prices are of 5%. Major online bookshops typically offer(ed) a systematic 5% discount on book prices so that no further discount may be offered on these products.

<sup>2</sup>French culture tip: There are many companies dedicated to given groups of people, e.g., MAIF for the teachers and professors. Their business model is that similar people share similar risks, which may be lower in case of well-behaving, sensible people as we all professors are — aren't we? And by the way, do you think that it is a good business idea to have young adults like students share their risks?



## Confidence intervals: the basics

The following is of course a very ironic statement!

*The inaccuracy of any statistic is compensated by the precision of the decimals.*

Georges Elgozy (French<sup>1</sup> economist, 1909–89)

It merely indicates that sample averages are not to be used directly; they need to be rounded off and be given with a margin of uncertainty.

The newspaper article on the next page illustrates both facts. (We will see in a minute how the margins of error indicated therein were obtained.)

### 1. Notion of confidence intervals

In all statistical problems we encountered in the previous chapter, we had a parameter of interest (either  $p_0$ , a proportion of the population, or  $\mu_0$ , the population mean of a variable) and a sample average ( $\bar{x}_n$ , where  $n$  is the sample size; either a sample proportion or a sample mean).

We say that  $\bar{x}_n$  estimates  $p_0$  or  $\mu_0$ : the two quantities should be close (by the so-called “law of large numbers”). However, giving just one number  $\bar{x}_n$  is not informative enough.

Instead we will rather construct an interval of plausible values for  $p_0$  or  $\mu_0$ , which is called a confidence interval. The name comes from the fact that we are confident that the (unknown) parameter of interest  $p_0$  or  $\mu_0$  lies in this interval.

In the sequel we provide the formulas for these confidence intervals.

---

<sup>1</sup>Original quote: “Dans toute statistique, l’inexactitude du nombre est compensée par la précision des décimales.”

## POLITICS 2016 ELECTION

# Hillary Clinton Holds Double-Digit Lead Over Donald Trump, Poll Finds

Melissa Chan @melissalchan | June 26, 2016

## Clinton leads Trump by 51% to 39% in a new national poll

Democratic presidential candidate Hillary Clinton holds a double-digit lead over presumptive GOP nominee Donald Trump, according to a new national poll released Sunday.

The former Secretary of State leads Trump by 51% to 39% among registered voters nationwide, a new Washington *Post*-ABC News poll shows. The poll also found that 56% of the public at large say Trump stands against their beliefs and 64% say he does not have the necessary credentials to be president. Fifty-six percent feel strongly that he is unqualified.

The poll, which randomly sampled 1,001 adults between June 20 and June 23, also found that half of Americans are anxious about Clinton possibly leading the country should she win in November. (The margin of sampling error for overall results is plus or minus 3.5 percentage points.) On Sunday, Clinton's opponent Bernie Sanders told CNN "the vast majority of my voters will vote for her" if she works to "embrace" some of his policy positions. The Vermont Senator, who has not officially suspended his campaign, has said he will vote for Clinton.



John Locher—AP

Democratic presidential candidate Hillary Clinton speaks at a rally in Los Angeles on June 6, 2016.

Trump by just 5 points. Forty-six percent of registered voters support Clinton, while 41% support Trump, according to this poll, conducted from June 19 to June 23 among 1,000 registered voters. It has a margin of error of plus-minus 3.1 percentage points.

The Democratic National Convention will be held in Philadelphia from July 25 to July 28. The GOP convention will take place in Cleveland from July 18 to July 21.

In another national poll by NBC News and the *Wall Street Journal*, Clinton is ahead of

## 2. Confidence intervals for proportions $p_0$

We start with confidence intervals for a population proportion  $p_0$  and will try to provide smarter statements than on the picture below.



### 2.1. Plain formulas: with “high confidence”

We can either offer plausible bounds for  $p_0$

- from above and from below;
- just from below;
- just from above.

It all depends on the strategic considerations at hand and on the conclusions we want to draw. What do we want to write?

- With high confidence, the population proportion lies between [...] and [...].
- With high confidence, the population proportion is larger than [...].
- With high confidence, the population proportion is smaller than [...].

In the exercises, to decide which kind of statement we should write, we will determine the underlying objectives of the study (and in particular, who funds it and why).

The corresponding formulas are the following ones, where  $n$  denotes the sample size and  $\bar{x}_n$  denotes the sample average:

$$\left[ \bar{x}_n \pm 1.96 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \right] = \left[ \bar{x}_n - 1.96 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}}, \bar{x}_n + 1.96 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \right];$$

$$\left[ \bar{x}_n - 1.65 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}}, 100\% \right];$$

and

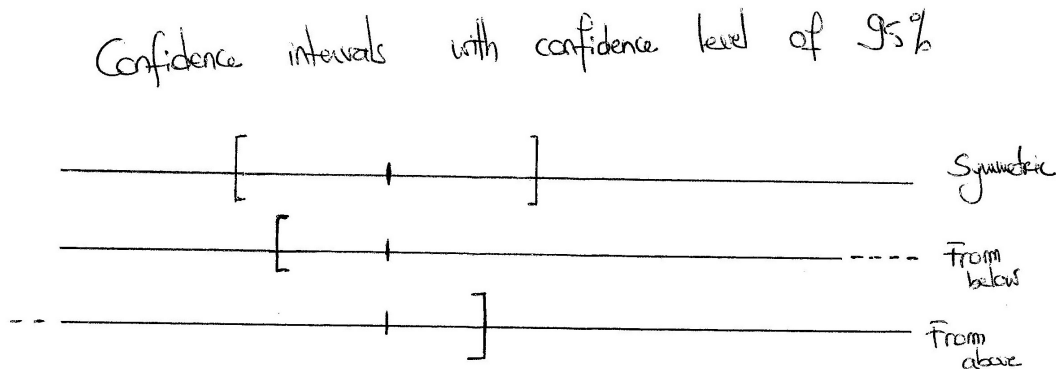
$$\left[ 0\%, \bar{x}_n + 1.65 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \right].$$

The first interval is called the symmetric or two-sided confidence interval; the last two intervals are called one-sided confidence intervals.

### 2.2. Illustration

We illustrate the formulas on the picture below: the one-sided intervals are shorter than the symmetric interval on one side but provide no information on the other side. They should be applied if (and

only if) one is only interested in bounding the parameter of interest from one side.



Let us find again the margins claimed in the news article!

For instance, the first margin: The total proportion  $p_0$  of American voters that were anxious, as of June 2016, about Hillary Clinton leading the country should she win lied, with high confidence, in the interval

$$50\% \pm 1.96\sqrt{\frac{0.5 \times 0.5}{1,001}} = 50\% \pm 3.1\% = [46.9\%, 53.1\%].$$

(As you can see, I found a margin of 3.1% instead of the 3.5% claimed in the article...)

Alternatively, more political (try to attribute them either to Democrats or Republicans!) statements could have been:

With high confidence, the total proportion  $p_0$  of American voters that were anxious about Hillary Clinton leading the country was larger than

$$50\% - 1.65\sqrt{\frac{0.5 \times 0.5}{1,001}} = 50\% - 2.7\% = 47.3\%,$$

which was the sign of a vast rejection feeling against this candidate. Put in terms of intervals, the confidence interval is  $[47.3\%, 100\%]$ .

With high confidence, the total proportion  $p_0$  of American voters that were anxious about Hillary Clinton leading the country was smaller than

$$50\% + 1.65\sqrt{\frac{0.5 \times 0.5}{1,001}} = 50\% + 2.7\% = 52.7\%,$$

which was a good score, given Donald Trump's such score and given the overall distrust feeling against politics in the USA. Put in terms of intervals, the confidence interval equals here  $[0\%, 52.7\%]$ .

Note again what we wrote earlier: the one-sided intervals are shorter than the symmetric interval on one side but provide no information (uninformative bounds of 0% or 100%) on the other side.

### 2.3. What does “high confidence” mean?

The above formulas actually correspond to a confidence level of 95%.

What does this mean? It means that 95% of the times when we apply these formulas, we obtain an interval that indeed contains  $p_0$ . This thus also means that 5% of the times, we are wrong! But we

accept this risk<sup>2</sup> of failure, as otherwise, the only way of being 100% safe and always be correct is to claim that  $p_0$  lies between 0% and 100%...

Statistical formulas for confidence intervals give you, in some sense, a trade-off between relative safety (it works 95% of the times) and precision (the intervals are narrower than the stupid 0% – 100% interval).

The only bugging point is that you never know whether you are right or wrong! The only way would be to actually compute  $p_0$  by exhaustively surveying the population, which is too costly and too time-consuming in general...

**What happens with larger confidence levels.**

However, in some applications, you may want to be more risk-averse and increase the confidence level; the side effect will be an increase of the width of the confidence intervals. (The larger they are, the safer they are.) Put differently, you may want to set the trade-off level between safety and precision towards more safety. This is performed by replacing the numbers 1.96 and 1.65 in the plain formulas above by larger numbers to be read in statistical tables.

We omit the corresponding details and will only work with 95%-confidence intervals during this course.

### 3. Confidence intervals for general population averages $\mu_0$

The formulas will look very similar except that we will need to take into account not only the sample mean  $\bar{x}_n$  but also the sample standard deviation  $s_{x,n}$ .

(It is still unimportant to deeply understand what the standard deviation means: just remember that it measures some spread in the sample data.)

#### 3.1. Plain formulas: with 95% confidence level

$$\begin{aligned} \left[ \bar{x}_n \pm 1.96 \frac{s_{x,n}}{\sqrt{n}} \right] &= \left[ \bar{x}_n - 1.96 \frac{s_{x,n}}{\sqrt{n}}, \bar{x}_n + 1.96 \frac{s_{x,n}}{\sqrt{n}} \right]; \\ &\left[ \bar{x}_n - 1.65 \frac{s_{x,n}}{\sqrt{n}}, +\infty \right]; \\ \text{and} \quad &\left[ -\infty, \bar{x}_n + 1.65 \frac{s_{x,n}}{\sqrt{n}} \right]. \end{aligned}$$

The  $-\infty$  and  $+\infty$  bounds just indicate that we have no control on the  $\mu_0$  population average on this side of the one-sided interval.

---

<sup>2</sup>Note that we could say, instead of our intervals being given with high confidence, that they are associated with a small risk; but this would be pessimistic terminology, so that we encourage you to rather focus on the high confidence than on the small risk.

### 3.2. Example

Compute a confidence interval on the «jambon-beurre» index of Paris (see page 15).

A neutral statement would be:

With high confidence, the  $\mu_0$  «jambon-beurre» index of Paris lies in the interval

$$4.35 \pm 1.96 \frac{1.55}{\sqrt{200}} = 4.35 \pm 0.22 = [4.13, 4.57] .$$

A consumers' association would rather compute a lower bound on the index and complain that the cost of living in Paris is high: it would issue a statement like “the «jambon-beurre» index of Paris is larger than [some value],” hoping that the value that came out of the analysis would be considered by everyone too large of a value. With the data we consider, it would write, for instance:

With high confidence, the  $\mu_0$  «jambon-beurre» index of Paris is larger than

$$4.35 - 1.65 \frac{1.55}{\sqrt{200}} = 4.35 - 0.19 = 4.16 ,$$

which is much higher than the indices computed in major<sup>3</sup> cities of France (Marseille, Lyon, Toulouse, Nice, Nantes, Strasbourg, Montpellier, Bordeaux, Rennes, etc.).

On the contrary, the federation of Parisian bakers (an association grouping all owners of Parisian bakeries) would try to convey that, given the same data, the index is not so high after all; it would (by construction) compute a larger value but comment that it is not prohibitively large. For instance, it would write:

With high confidence, the  $\mu_0$  «jambon-beurre» index of Paris is smaller than

$$4.35 + 1.65 \frac{1.55}{\sqrt{200}} = 4.35 + 0.19 = 4.54 ,$$

which is very comparable to the Parisian beer index; in all cities of France, the beer and the «jambon-beurre» indices are close, so that Parisian prices for «jambon-beurre» sandwiches are just aligned with the overall price level. There is nothing to complain about.

### 3.3. Interpretation

Beware of the interpretation:

We should write “With 95% confidence, the average price of sandwiches over Paris lies somewhere between 4.13 and 4.57 euros”.

But it is incorrect to write “95% of bakeries have sandwich prices lying somewhere between 4.13 and 4.57 euros”.

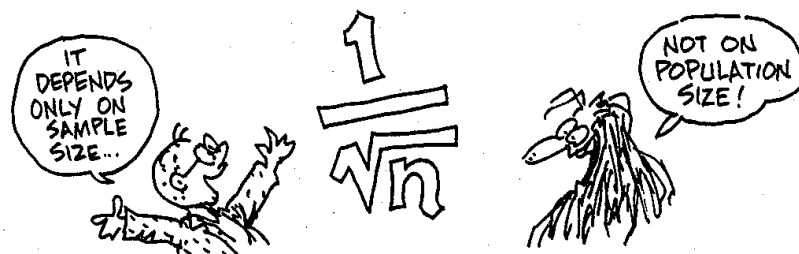
We only issue a statement on the average price, not on individual prices.

---

<sup>3</sup>French culture: these are the major 10 cities of France!

#### 4. How nice these formulas are!

OK, the above formulas (for  $p_0$  and  $\mu_0$ ) might look scary. But they carry a fantastic news about the precision of our estimates:



The (half-)width of these two-sided intervals are indeed given by

$$1.96 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \quad \text{and} \quad 1.96 \frac{s_{x,n}}{\sqrt{n}}$$

and the size of the population plays no role therein!

It is a common mistake to think that an accurate survey consists of sampling a given fraction of the population (e.g., 10% of the population).

But no: accuracy depends rather on the absolute size  $n$  of the sample; not on its relative size (relative to the population).

This is why political polls can make statements<sup>4</sup> about the hundreds of thousands of American voters simply by interviewing 1,000 such voters drawn at random.

Isn't this a miracle that will enlighten your day?

Can you imagine how much money can be saved thanks to this fact?

Note also that accuracy depends on the standard deviations: the larger the standard deviation  $s_{x,n}$ , the poorer the precision (the larger the error margin).

<sup>4</sup>Mathematical remark: the margin of error when surveying  $n$  individuals is upper bounded by the maximum of  $\pm 1.96 \sqrt{x(1-x)/n}$  as  $x$  varies between 0% and 100%. This maximum can be computed and is smaller than  $\pm 1/\sqrt{n}$ , which approximatively equals  $\pm 3\%$  when  $n = 1,000$ . Keep this order of magnitude in mind!

## 5. Conclusion: business insights and stories

### Business insight 1/2.

Remember the words of Georges Elgozy: “The inaccuracy of any statistic is compensated by the precision of the decimals”.

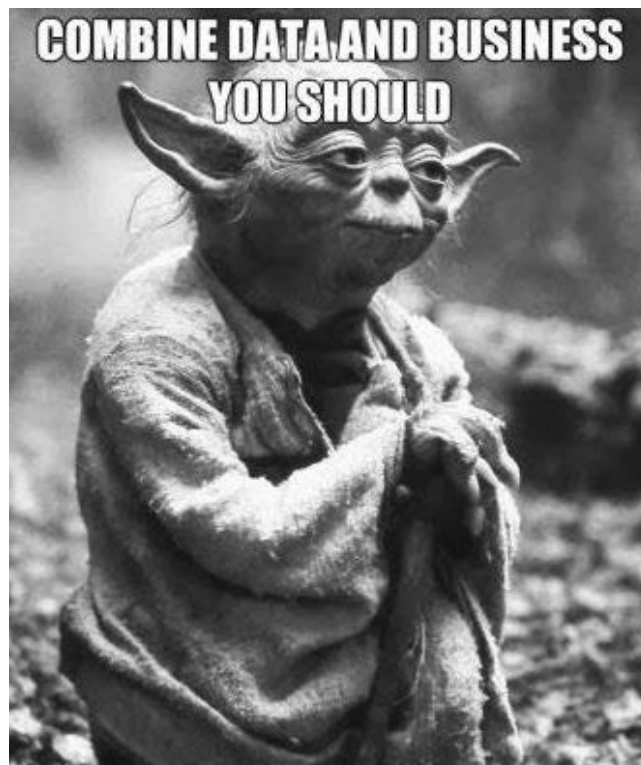
This means that whenever you conduct a survey, collect data and compute sample averages, you should go one step further: not only report the sample average but also provide an error margin (in case of a two-sided interval) or “correct” the value into the lower or upper bound of a one-sided interval; we will refer to these “corrected values” as under-estimates and over-estimates of the population proportion or of the population average. Note that even in the symmetric case, you need to adequately round off the numbers (average value, error margin) you provide.

In any case you should never believe that the sample average  $\bar{x}_n$  equals the population average  $\mu_0$  or the population proportion  $p_0$ . The former is only a rough approximation (an estimation) of the former and a statistical treatment, namely, the computation of confidence intervals, is necessary.

If you follow the above guidelines (easy, right?), everybody will rightfully believe that you are a God of statistics!

### Business insight 2/2.

The accuracy of confidence intervals only depends on the sample size and not on the fraction of the population that was sampled. It is a costly mistake to think that, e.g., 5% of the population always has to be surveyed. No, accuracy only depends on the sample size  $n$  and not at all on the population size! You can save so much money with this in mind...





**Stories: how journalists announce election results.**

French journalists typically never report error margins for polls; they rather insist on these polls being a snapshot of the opinion and on the fact that the general opinion can vary in time.

This is the same for election days. Voting desks close at latest at 8:00pm and forecasts can then be announced. Polling organizations (IFOP, CSA and IPSOS are the main, private, French such companies) of course always indicate error margins for their forecasts but journalists never (never!) reproduce them, they just give the raw forecasts, which roughly correspond to the sample averages computed.

With one notable exception, though: April 21, 2002 at 8:00pm, the day of the first round of the 2002 French presidential elections. France 2 (second channel of French TV) announced the following scores for the candidates ranking second and third: Jean-Marie Le Pen (extreme right),  $17\% \pm 0.5\%$ , and Lionel Jospin (the then socialist Prime Minister),  $16.5\% \pm 0.5\%$ . The two confidence intervals were overlapping... Journalists wanted to (at the same time!) announce that they believed that Jean-Marie Le Pen scored second and would be present in the second round, but that it was not a clear-cut situation. Final scores, revealed the next day, were equal to 16.86% and 16.18% (and in particular, the order was kept). That night was a trauma for many French inhabitants.

(Note: why polling organizations failed to predict Trump's election is a different story, that we might comment on in class.)

More recently, for 2017 presidential elections in France, below is what the two major TV channels reported at 8:00pm (I was lucky enough to follow both announcements on two contiguous TV screens) — my wild guess is that polling organizations reported overlapping confidence intervals with respective centers 23.7% and 21.7%, with different interpretations and presentations made by the two sets of journalists. Final scores were 24.01% and 21.30%.



TF1 channel at 8:00pm

vs.



France 2 channel at 8:00pm

In 2022, nothing noticeable took place when announcing forecasts at 8pm...

## 6. Elementary exercises

**Elementary exercise 2.1.** A customer manager has a portfolio of 1,536 customers, out of whom she surveys 100 customers. Out of them, 78 are happy about how she deals with them, thus leading to a sample satisfaction rate of  $\bar{x}_{100} = 78\%$ . She wants to use these numbers for her annual review meeting with her manager. She wants to declare that her satisfaction rate  $p_0$  among her entire portfolio is (with high confidence) at least [...]%. Which number should she use?

**Elementary exercise 2.2.** Users of the Jouy-en-Josas post office are currently pissed off by the (long) waiting time before being served on Saturday mornings. They want to file a complaint but need numbers. One of them being a famous statistician decides to proceed as follows: for 3 consecutive Saturdays, he goes to the post office and picks at random 17 users, whom he follows in the post office to see how long they wait before being served. He thus gets 51 waiting times, with average value  $\bar{x}_{51} = 16$  minutes and standard deviation  $s_{x,51} = 5$  minutes. What can be said about the average waiting time  $\mu_0$  on Saturdays? As the complaint should be as neutral and as factual as possible for better efficiency, provide a neutral statement about the actual value  $\mu_0$  of the waiting time.

**Elementary exercise 2.3.** A survey on the popularity rating  $p_0$  of President Emmanuel Macron was conducted on the phone between September 8 and 10, 2017. Out of the respondents, 1,002 expressed an opinion on Emmanuel Macron: 285 had a positive opinion of his action, while 717 had a negative opinion, thus resulting on a sample popularity rating of  $\bar{x}_{1002} \approx 28.4\%$ . A polling organization wants to comment on the low popularity rating  $p_0$  achieved by President Macron only a few months after his election. However, they want to do so in an honest and fully informative way. Therefore, they will tell the general public that “their study proves that his popularity rating  $p_0$  is already lower than [...]% (with high confidence)”. Which number should they use?

## 7. Mathematical appendix

You should probably skip it! We are just giving you an idea of what you would have seen in a mathematics curriculum when speaking of confidence intervals (and we are of course in a business school, not in a mathematics department!).

The core result is the so-called “central limit theorem”. It states that given random variables  $X_1, X_2, \dots$  independent and identically distributed according to a distribution with expectation  $\mu_0$  and standard deviation  $\sigma_0$ , we have the convergence in distribution

$$\sqrt{\frac{n}{\sigma_0^2}} (\bar{X}_n - \mu_0) \rightarrow \mathcal{N}(0, 1),$$

where  $\bar{X}_n$  is the average of  $X_1, X_2, \dots, X_n$  and  $\mathcal{N}(0, 1)$  is the standard Gaussian distribution.

Denoting by

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2$$

the estimator of the variance  $\sigma_0^2$ , we know, by the “law of large numbers”, that we have the convergence in probability

$$\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma_0^2.$$

Slutzky’s lemma then yields that we have the final convergence in distribution

$$\sqrt{\frac{n}{\hat{\sigma}_n^2}} (\bar{X}_n - \mu_0) \rightarrow \mathcal{N}(0, 1).$$

This implies, by definition of the convergence in distribution, that for all real numbers  $z$ ,

$$\mathbb{P} \left( \sqrt{\frac{n}{\hat{\sigma}_n^2}} (\bar{X}_n - \mu_0) \in [-z, z] \right) \rightarrow \mathbb{P}(N \in [-z, z]),$$

where  $N$  is a random variable with distribution  $\mathcal{N}(0, 1)$ .

The events

$$\frac{n}{\hat{\sigma}_n^2} (\bar{X}_n - \mu_0) \in [-z, z] \quad \text{and} \quad \mu_0 \in \left[ \bar{X}_n \pm z \frac{\sqrt{\hat{\sigma}_n^2}}{\sqrt{n}} \right]$$

are equal, and choosing  $z = 1.96$  ensures that  $\mathbb{P}(N \in [-z, z]) = 95\%$ .

These lead to the claimed (asymptotic) two-sided confidence interval for  $\mu_0$  of confidence level 95%: the interval estimate

$$\left[ \bar{X}_n \pm 1.96 \frac{\sqrt{\hat{\sigma}_n^2}}{\sqrt{n}} \right]$$

containing the true mean  $\mu_0$  with probability converging to 95%, and whose realized value on the data equals indeed

$$\left[ \bar{x}_n \pm 1.96 \frac{s_{x,n}}{\sqrt{n}} \right],$$

for which we say that we are confident at level 95% that it contains  $\mu_0$ .

## 8. More advanced exercises (quiz-like exercises)

The two exercises below are the continuations of the statements of page 20.

**Advanced exercise 2.1** (Discounts, continued). The question is to determine whether the discount policy, if offered to all customers, would be profitable or not. Without the discount, the average gross margin rate on an order is 40%; that is, without the discount, Nozama.com gets an average 3.64 euros of gross margin per customer per trimester. (How did we get this number?) What happens with the discount? More orders are placed, with larger amounts, but the margin is smaller. Which phenomenon takes the lead? (I.e., what is the elasticity of the phenomenon at hand?) Answer the following questions to know!

1. All relevant statistical information was already extracted in the previous chapter but it might be good to recall it.
2. Compute a confidence interval on the order rate with the discount. Beware! you first need to think about its shape (two-sided or one-sided? and in the latter case, with a bound from above or below?).
3. Same question for the amount of order (before application of the discount).
4. Conclude, by exploiting the two<sup>5</sup> confidence intervals constructed.

**Advanced exercise 2.2** (Car insurance company, continued). Solve this exercise three times, with the following mindsets:

- the entrepreneur’s mindset — this is the dream of his life, he will live it, no matter what happens, he just needs to get a better idea of how much money to leverage to satisfy the Basel prudential rules;
- the banker’s mindset — bankers try to assess business plans in a rigorous and rather neutral way, they do not want to lend money and not get it back, while simultaneously aiming at some profits;
- his best friend’s mindset — this friend feels that creating this insurance company dedicated to students is a disaster waiting to happen.

Now, here are the questions!

1. Recall all relevant statistical information.
2. Compute confidence intervals on the two parameters of interest.
3. Deduce a confidence<sup>6</sup> interval on the average expected expenses per insured student.

**Other advanced exercises:** The next pages feature exercises extracted from past quiz statements

---

<sup>5</sup>As you will read in the next chapter there is a slight loss in confidence when simultaneously exploiting two confidence intervals to make a statement: a 90% only confidence level is achieved. More details to come in the next chapter.

<sup>6</sup>Same remark as in the footnote above.

## An advertisement featuring statistics (6 points)

The aim of the advertisement below is to show that stairs are so important in your daily life (it was designed for a French carpenter company named Lapeyre). The text says: “On average you will walk up and down your stairs 89,019 times: choose it carefully!” Some footnote indicated that this number had been provided by a survey conducted by BVA Group (a French polling organization) in July 2017. We did not get the corresponding raw data and made up some plausible data instead.



Suppose that about a thousand people were interviewed: we of course only keep the 534 of them that have stairs in their homes. Each of them was asked to indicate how many times they walk it up and down per day, as well as for how long they think they will keep their current stairs before the next renovation work. Answers were:

- an average number of 12.60 walks up and down (with a corresponding standard deviation of 2.41);
- an average period of 19.34 years before the next renovation (corresponding standard deviation: 4.35 years).

We will consider below that 1 year is made of 365.25 days.

- ☐ Which type of confidence interval (symmetric, underestimate, overestimate) should you pick, and why?

Write a nice and enjoyable sentence indicating

- ☐ the parameter of interest at hand and  
☐ a confidence interval on its value (please provide the details of your calculation):

- ☐ Provide the numerical value of the confidence interval on the second parameter (no need for a nice sentence, just the number[s] with details of your calculation):

- ☐ ☐ Conclude by filling the gaps in the sentence below: the first gap (the confidence level) would not be part of the advertisement; for the second gap, write all needed words.

With confidence and on average, you will walk up and down your stairs times.  
technical statement

## TwitterAudit (7 points)

TwitterAudit defines its methodology as follows (see <https://www.twitteraudit.com>):

“Each audit takes a sample of up to 5,000 [...] Twitter followers for a user and calculates a score for each follower. This score is based on number of tweets, date of the last tweet, and ratio of followers to friends. We use these scores to determine whether any given user is real or fake. Of course, this scoring method is not perfect but it is a good way to tell if someone with lots of followers is likely to have increased their follower count by inorganic, fraudulent, or dishonest means.”



Screenshot of  
<https://www.twitteraudit.com/realdonaldtrump>  
 (webpage accessed on Wednesday October 4, 2017, at 9.20am)

In what follows we will assume that TwitterAudit always uses 5,000 followers, and that its classification as real or fake is accurate. Consider some random Twitter user, say, Donald Trump: see the associated picture above. The picture reports the sample proportion of real users.

- ☐ Define in detail the population considered here; in particular, provide a population count.
  
- ☐ Indicate the parameter of interest.
  
- ☐ Spell out the available sample data and summarize it.  
*Beware, the sample proportion actually equals 59.8% (how do we know that it is not just 59%?).*
  
- ☐ Should we compute a symmetric confidence interval, an underestimate, or an overestimate? Explain.

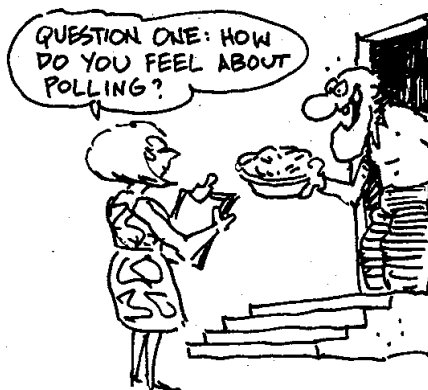
- 
- Statistics / © Gilles Fortin-Stoltz, HEC Paris, 2022–23. All rights reserved. 35





## Confidence intervals: advanced notions

OUR METHOD IS TO TAKE A **SAMPLE**... A RELATIVELY SMALL SUBSET OF THE TOTAL POPULATION, THE WAY POLLSTERS DO AT ELECTION TIME.



AN OBVIOUS QUESTION IS: HOW BIG A SAMPLE DO WE HAVE TO TAKE TO GET MEANINGFUL RESULTS?



AND THE ANSWER, WHICH YOU SHOULD INSCRIBE IN YOUR BRAIN FOREVERMORE, WILL TURN OUT TO BE: IF  $n$  IS THE NUMBER OF ITEMS IN THE SAMPLE, THEN EVERYTHING IS GOVERNED BY

$$\frac{1}{\sqrt{n}}.$$

GOVERNED BY  $\frac{1}{\sqrt{n}}$ ? DIDN'T EVEN KNOW IT WAS ON THE BALLOT!



In this chapter, we will review more advanced topics as far as confidence intervals are concerned:

- Corrections to perform in two specific cases (small sample size or small population size relative to the sample size);
- What happens when several confidence intervals are used simultaneously (the Bonferroni correction);
- How large a sample should we take to get meaningful results (survey planning).

And then, we will provide exercises, as usual.

## 1. Corrections to perform in two specific cases

The formulas indicated in the previous chapter hold under two assumptions:

- The sample size is large enough, say,  $n \geq 30$  or  $n \geq 50$ ;
- The sample size  $n$  is small relative to the population size  $N$ , say,  $n/N \leq 5\%$ .

We now indicate how to relax each of these assumptions.

### 1.1. Small sample size (and statistical softwares)

Because of the underlying mathematics, which involve what are called convergences, the formulas in the previous chapter are to be used when the sample size  $n$  is large enough, say,  $n \geq 30$  or  $n \geq 50$ .

When  $n$  is smaller than 30, it is usually unsafe to build confidence intervals. If one really has to, then the  $z_{1-\alpha/2}$  and  $z_{1-\alpha}$  quantiles of the Gaussian distribution are to be replaced by quantiles of Student's distribution with  $n - 1$  degrees of freedom, denoted by  $t_{n-1,1-\alpha/2}$  and  $t_{n-1,1-\alpha}$ .

Actually, most of the statistical softwares resort to the  $t_{n-1,1-\alpha/2}$  and  $t_{n-1,1-\alpha}$  quantities, irrespective of a small or large  $n$ . This is not an issue as these quantities are larger than the  $z_{1-\alpha/2}$  and  $z_{1-\alpha}$  quantities, hence the confidence intervals are larger and safer.

And for  $n \geq 30$ , the two sets of quantities are actually extremely close.

This all was mostly meant for your general culture; we will almost only deal with the case of large samples in the quizzes and in the exam.

### 1.2. Correction for the ratio of sample size to population size

Again because of the underlying mathematics, the formulas in the previous chapter only hold when the population size  $N$  is large relative to the sample size  $n$ . More precisely, we would like to only consider situations where sampling units with or without replacement are the same (i.e., where it is unlikely to sample twice the same unit if picking at random). This is not the case anymore when the ratio  $n/N$  of sample size  $n$  to population size  $N$  is large, say, larger than 5%.

In the latter case, we slightly decrease the width of the confidence intervals by a multiplicative factor  $\sqrt{(N-n)/(N-1)}$  on the error margins. For instance, the symmetric 95%-confidence intervals on a proportion  $p_0$  and on a mean  $\mu_0$  then become

$$\left[ \bar{x}_n \pm 1.96 \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \sqrt{\frac{N-n}{N-1}} \right] \quad \text{and} \quad \left[ \bar{x}_n \pm 1.96 \frac{s_{x,n}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right].$$

This correction too was mostly meant for your general culture.

## 2. Simultaneous confidence intervals — the Bonferroni correction

The rule is the following:

Any statement drawn from the simultaneous consideration of a confidence interval of confidence level  $1 - \alpha$  and another one of confidence level  $1 - \beta$  only holds with confidence  $1 - (\alpha + \beta)$ .

Put differently: the error risks  $\alpha$  and  $\beta$  need to be added up. (This should seem intuitive, no?)

For instance, any statement drawn from the simultaneous consideration of two confidence intervals of same confidence level 95% only holds with confidence level 90%.

But what do we mean by “any statement drawn from the simultaneous consideration of two confidence intervals”?

Let us illustrate it with examples. You got some first examples in the exercises on page 32, for the computation of confidence intervals on the gross margin per customer per trimester and on the average expected expenses per insured student in the year to come. We were then “multiplying” confidence intervals.

All arithmetical operations are possible; here are examples with subtraction, addition, and division. The basic rule is to determine the two most extreme (minimal and maximal) values that can be taken, given the two confidence intervals at hand (i.e., when considering all pairs of points of the two intervals).

**Examples.** Consider a toy case in which two proportions  $p_0$  and  $q_0$  are to be estimated:  $p_0$  is the proportion of Parisian men eating bread at least once a day, and  $q_0$  is the same proportion for women. Suppose that after some survey, we computed the 95%-confidence-level intervals for these quantities, with respective values:

$$[30\%, 45\%] \quad \text{for } p_0 \quad \text{and} \quad [20\%, 25\%] \quad \text{for } q_0.$$

Suppose that we are interested in the excess proportion  $p_0 - q_0$  of men eating bread with respect to women. Given the intervals above for  $p_0$  and  $q_0$ , the minimal excess is given by  $30\% - 25\% = 5\%$  and the maximal one by  $45\% - 20\% = 25\%$ . That is, we suggest the confidence interval  $[5\%, 25\%]$  for  $p_0 - q_0$ ; it is associated with a confidence level of 90% only.

Assuming that the Parisian inhabitants are half men and half women, the global proportion of Parisians eating bread at least once a day is given by  $(p_0 + q_0)/2$ . The minimal value for this quantity given these intervals equals  $(30\% + 20\%)/2 = 25\%$  and the maximal value is  $(45\% + 25\%)/2 = 35\%$ . The global proportion of Parisians eating bread is thus estimated by the interval  $[25\%, 35\%]$ , which is associated with a confidence level of 90% only.

Finally, consider the excess ratio  $p_0/q_0$  of men eating bread at least once a day with respect to women. Given the intervals for  $p_0$  and  $q_0$ , the minimal value for this ratio equals  $30\%/25\% = 1.2$  and the maximal value is  $45\%/20\% = 2.25$ . Therefore, we get the confidence interval  $[1.2, 2.25]$  for the excess ratio  $p_0/q_0$  at hand; it holds with a confidence level of 90% only.

### 3. Sample-size determination (survey planning)

So far, we have been collecting data and then analyzing it, computing confidence intervals as functions of these data (and of the confidence level  $1 - \alpha$ ). When considering two-sided intervals, we had some precision  $\varepsilon$  (some error margin), given by the half-width of the interval: for  $1 - \alpha = 95\%$ ,

$$\varepsilon = 1.96 \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \quad \text{and} \quad \varepsilon = 1.96 \frac{s_{x,n}}{\sqrt{n}}.$$

We now want to reverse the process: to set a target margin  $\varepsilon$  (and a confidence level of  $1 - \alpha = 95\%$ ) and determine the sample size  $n$ .

The issue is that we have too many unknowns:  $n$  is our main unknown, but  $s_{x,n}$  and  $\bar{x}_n(1 - \bar{x}_n)$  are not known either.

We solve this by a two-step procedure:

1. We first sample about 30 units and compute  $s_{x,30}$ .
2. We hope<sup>1</sup> that  $s_{x,30}$  will be close to  $s_{x,n}$  and solve the equation  $1.96 s_{x,30}/\sqrt{n} \leq \varepsilon$ , where now, the only unknown is  $n$ . The solution is

$$n \geq \left( \frac{1.96 s_{x,30}}{\varepsilon} \right)^2.$$

3. We sample  $n - 30$  additional units.

The above procedure is for general averages; but it can be adapted in a straightforward manner to the case of proportions: just solve  $1.96 \sqrt{\bar{x}_{30}(1 - \bar{x}_{30})/n} \leq \varepsilon$ , that is, sample  $n - 30$  additional units, where

$$n \geq \frac{1.96^2}{\varepsilon^2} \bar{x}_{30}(1 - \bar{x}_{30}).$$

Of course, the 1.96 value above for the two formulas for  $n$  can be rounded upwards to 2.

#### Alternative formulation.

You may consider difficult to recall and implement the above procedure. It can be equivalently stated as follows. Call

$$\gamma = 1.96 \frac{s_{x,30}}{\sqrt{30}} \quad \text{or} \quad \gamma = 1.96 \sqrt{\frac{\bar{x}_{30}(1 - \bar{x}_{30})}{30}}$$

the error margin guaranteed with the first sample with 30 units. We want to decrease this error to  $\varepsilon$ , that is, to reduce it by a factor of  $\gamma/\varepsilon$ . With the same hopes<sup>2</sup> as above, we should thus sample in total  $30(\gamma/\varepsilon)^2$  units, including the initial 30 ones.

Of course, if a different number  $m$  of units were first sampled, then we adapt the formulas above by replacing 30 by  $m$ .

<sup>1</sup>And by the “law of large numbers”, this should actually be the case!

<sup>2</sup>That  $s_{x,30}$  is close to  $s_{x,n}$  and/or that  $\bar{x}_{30}$  is close to  $\bar{x}_n$

**A note for proportions.**

We already mentioned (on page 27) that

$$1.96\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \leq \sqrt{\frac{1}{n}},$$

irrespectively of the value of  $\bar{x}_n$ . This is a reasonable bound when  $\bar{x}_n$  is close to 50% and is a crude bound when  $\bar{x}_n$  is close to the extreme. To guarantee a precision  $\varepsilon$  ex-ante (without first sampling some units) we can take  $n = 1/\varepsilon^2$ . This is a conservative value and we recommend to rather use the two-step procedure described above.

**Example: accurately determine the customers' satisfaction.**

During your gap year you are recruited by a retail company to, among others, assess customers' satisfaction. You have no idea of how happy or unhappy the customers currently are, yet your manager wants an accurate estimate, with an error margin of at most 3% ideally. She is willing to provide some money to give incentives to customers to participate to your survey — say, a 5 euros coupon. The total budget she has in mind is only 1,500 euros because, you know, these are difficult times in a difficult economy.

Is this enough? Should you negotiate with her (more budget and/or a different precision)? You need to act quickly. She will be unhappy if you spend the whole budget and then realize that you cannot reach the desired precision (you should have told her in advance!). She will be similarly disappointed if you spend the whole budget unnecessarily (money is precious, what do you think, young man?).

Your plan: call 100 customers right now, this can be done in 2 hours, assess the situation and report to her at the lunch break. She will appreciate your dedication and how quickly you can get some first results.

**Scenario 1:** Out of the 100 customers, 32 are unhappy with the products sold. Wow! that was unexpected. Something is going wrong somewhere. Anyway, you compute the needed sample size to get the  $\pm 3\%$  precision:

$$n \geq \frac{1.96^2}{\varepsilon^2} \bar{x}_{100}(1 - \bar{x}_{100}) = \frac{1.96^2}{0.03^2} 0.32(1 - 0.32) \approx 929.$$

The alternative calculation notes that the current error margin equals

$$1.96\sqrt{\frac{\bar{x}_{100}(1 - \bar{x}_{100})}{100}} = 1.96\sqrt{\frac{0.32(1 - 0.32)}{100}} = 0.09142933 \approx 9.15\%$$

and that we want to reduce it to  $\pm 3\%$ , that is, to reduce it by a factor of  $9.15/3 = 3.05$ . The total number of customers to be interviewed (including the 100 customers already interviewed) is about  $100 \times 3.05^2 \approx 931$ . The small difference with respect to the first calculation is only due to rounding issues in the intermediate calculations.

In both cases — even by reducing a bit the amount on the coupon, there is no way you can interview that many people! But given the high dissatisfaction rate, even a rougher precision would be enough. This, at least, is what you will explain to and negotiate with your manager!

**Scenario 2:** Out of the 100 customers, 5 only are unhappy with the products sold. Similar calculations directly suggest a sample size of

$$n \geq \frac{1.96^2}{\varepsilon^2} \bar{x}_{100}(1 - \bar{x}_{100}) = \frac{1.96^2}{0.03^2} 0.05(1 - 0.05) \approx 203$$

or note that the current error margin equals

$$1.96\sqrt{\frac{\bar{x}_{100}(1 - \bar{x}_{100})}{100}} = 1.96\sqrt{\frac{0.05(1 - 0.05)}{100}} \approx 4.3\%$$

and thus that in total,  $100 \times (4.3/3)^2 \approx 206$  customers should be interviewed.

You can save some money! Only about 1,030 euros will be needed out of the budget. Isn't this thrilling news for your manager that will make her day?

## 4. Elementary exercises

**Elementary exercise 3.1.** A study about shopping in physical stores (that is, shopping in the old-fashioned way) shows that when customers are served with a smile and are offered a candy, they are more likely to buy a new product. More precisely, the study showed that with confidence 95%, at least 40% of the customers would buy the new product with the nice treatment, while, with confidence 95%, at most 25% of them would buy it in a neutral setting. Quantify the impact of a smile and a candy on the purchase rate of a new product; associate your quantification with a confidence level.

**Elementary exercise 3.2.** Can you solve the previous exercise under the following two statements? “With confidence 95%, at most 50% of the customers would buy the new product with the nice treatment” and “with confidence 95%, at least 15% of them would buy it in a neutral setting”.

Same question with these other two statements: “with confidence 95%, at most 50% of the customers would buy the new product with the nice treatment” and “with confidence 95%, at most 25% of them would buy it in a neutral setting”.

**Elementary exercise 3.3.** Suppose that we statistics instructors would like to determine the average time that students devoted to do their homework for last week's session. Suppose that there are 7 groups for the statistics course, that each of the instructor asks 5 students at random in each of her/his groups, and that we obtain a sample average homework time of 62 minutes over these 35 students. We want to determine the population average homework time up to a  $\pm 1$  minute margin. We assume that the population is composed of hundreds of students: each year we only have about 300 students but their working behaviors should be representative of the ones of students of past and future years. How many more students should we interview if the sample standard deviation over 35 students equals 4 minutes? And if it happens to equal 15 minutes? Comment on the sample sizes obtained, if needed.

## 5. More advanced exercises (quiz-like exercises)

**Advanced exercise 3.1** (Gender pay gap?). Data used here are extracted from a recent survey by the French public statistics office called INSEE<sup>3</sup>. Variables reported for each respondent include: age, gender, region of residence, socio-professional category, level of study, etc., and of course, monthly net salary.

Let us consider a gender pay gap, in the French socio-professional category number 37 (company executives) and in the Ile-de-France region (that is, Paris and its suburbs). There's a heavy context and previous evidence about gender wage discrimination taking place in France. Available data were processed with a statistical software, which produced the summary below.

**Group Statistics**

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Monthly net salary	Men	179	3431,46	3895,437	291,159
	Women	147	2434,90	1282,947	105,816

1. Do these data show a significant difference between men's and women's monthly net salaries in this category and in this region? (Answer with the techniques seen so far but a better method to answer this question will be studied on page 92.)
2. Does the answer to the question above, by itself, prove or disprove the existence of a gender pay gap in this category and in this region?
3. How many respondents in each category would you need to estimate the average monthly net salaries up to a margin of  $\pm 100$  euros?

**Advanced exercise 3.2** (Sample size determination, in a different way though...).

You want to conduct a street survey in a pedestrian street (where is the street depicted here?). The shopkeepers' association asks you to interview 2,000 customers at random on a given Saturday. How many volunteers should you recruit for this purpose?

You first assess the situation: you go to this street two or three weeks before the given day and you start interviewing people as follows. You select your targets at random, ask them if they are up to being interviewed, and if so, you conduct the interview. The interview is short, about 5 minutes. But you are mostly interested in the participation rate to determine the overall (wo)man hours needed to get the 2,000 respondents.



In your test, you asked 100 customers for participation and 15 accepted.

How many customers in total do you think will need to be interviewed before you get the 2,000 desired respondents? (Hint: Try to think about it from scratch. The solution is not in the formulas of the survey-planning section above.)

Assuming one needs 1 minute on average to establish a contact and get a yes/no answer on the participation to the survey, how many (wo)man hours are needed to fully conduct the survey and get the desired 2,000 interviews?

<sup>3</sup>INSEE means in French "Institut National de la Statistique et des Etudes Economiques", that is, National Institute for Statistics and Economic Studies.

## Exercise 2 — Budget planning for traveling costs — 10 points

*This exercise is based on a statistical experiment that I am currently conducting.* Assume that I do not live in a neighboring area of HEC Paris, but rather in some farther away place to the West, in France's countryside. I come to HEC Paris 2 days / 1 night a week. When planning my monthly budget, I need to take into account weekly traveling costs (one fast-train trip and housing costs for one night). It turns out that train prices and accomodation prices (through AirBnB usually) are quite volatile and are difficult to predict. This is why, as a trained statistician, I collected data for 30 weeks. To study the data collected, I of course implicitly assume that my sample of 30 weeks (a bit more than a semester) is representative of the semesters to come (i.e., that the various favorable or unfavorable price situations I met will take place in a similar fashion in the upcoming years). The data set looks like that (the lines below are only an excerpt of the data set):

Date	Train	Housing
Feb. 7-8	44	53.10
Feb. 14-15	25	31.83
...	...	...
Sept. 19-20	46.30	34.97
Sept. 26-27	60.20	65.00
Oct. 3+4	54	29.35



The data set can be summarized as follows:

- Sample average price for the train trips = 39.25 euros
- Sample standard deviation of these prices = 18.92 euros
- Sample average housing costs = 41.54 euros
- Sample standard deviation of these prices = 14.24 euros

### Study of the prices of the train trips

- ☐ Indicate the parameter of interest out of the four following statements:
- 1A. the individual prices of the past train trips
  - 1B. the average price of the past train trips
  - 1C. the individual prices of the (past and) future train trips
  - 1D. the average price of the (past and) future train trips

Assume that I am conducting this study because on second thoughts, I am worried that I moved so far away from my workplace and I need to reassured as far as the traveling costs are concerned.

- ☐ What is the *best* shape for a confidence interval on the parameter of interest?
- 2A. a symmetric interval (= with high confidence, the [parameter of interest] lies between [...] and [...] euros)
  - 2B. an overestimate (= with high confidence, the [parameter of interest] is at most [...] euros)
  - 2C. an underestimate (= with high confidence, the [parameter of interest] is at least [...] euros)

We now want to compute the number(s) to put in the conclusion stated right above.

- ☐ Spell out the calculation you will type on your calculator (i.e., which formula with which numbers):



- ☐ Provide your final numerical value(s), as read on your calculator (no need for rounding yet):

Provide your final numerical value(s), after rounding to integer value(s), i.e., without cents:

### Study of the housing costs

We are now interested in housing costs and proceed similarly to obtain a confidence interval on the parameter of interest corresponding to housing costs.

- ☐ Spell out the calculation you will type on your calculator (i.e., which formula with which numbers):

- ☐ Provide your final numerical value(s), as read on your calculator (no need for rounding yet):

Provide your final numerical value(s), after rounding to integer value(s), i.e., without cents:

We are now putting all results together.

- ☐ Provide a concluding sentence on the total costs for one week (please provide all necessary adjectives, think of the verb tense, etc.):

- ☐ Quantify the confidence level guaranteed for the conclusion stated above:

### A more precise picture

- ☐ ☐ How many weeks will I have to wait before my error margin on the parameter of interest for train trips will be (of the order of)  $\pm 5$  euros?

**Exercise 1 — The effect of touch (10 points)**

It is well documented, e.g., in marketing studies (Jacob Hornik, “Tactile stimulation and consumer response”, *Journal of Consumer Research*, 1992) that light tactile contacts influence human beings in a subtle way towards the requests of the contact-maker. For instance, if a seller touches you lightly, you should be more inclined to buy a product.



We want to illustrate this fact by performing the following experiment. We consider two similar stores (e.g., two Hollister stores) and ask the sellers of the first store to avoid any physical contact with the customers, while the ones of the second store are asked to lightly touch the customers' arm. We are interested in the corresponding purchase rates, which we denote by  $p_0$  (without any contact) and  $q_0$  (with a light contact), respectively. Data collected are that 12 out of the 120 customers served without a contact purchased an item, while 23 out of the 120 served with such a contact did so.

We want to quantify the impact  $q_0 - p_0$  of a light contact by exhibiting a confidence interval for it.

**Symmetric interval on  $p_0$** 

We exhibit first a symmetric confidence interval on  $p_0$ .

- ☐ Spell out the calculation you will type on your calculator (i.e., which formula with which numbers):
  
- ☐ Provide your final numerical value(s),
  - as read on your calculator (no need for rounding yet):
  
  - after rounding the error margin to a X.X% format:
  
- ☐ If 6,000 customers are served without any contact every week, how many purchases will be made each week, based on the previous result? Fill the following sentence by including all necessary numbers and words to avoid any ambiguity:  
  

*With high confidence, the store will get every week**purchases*
  
- ☐ ☐ How many customers should have been considered to get an estimation of  $p_0$  at a  $\pm 2\%$  margin? Provide calculation details for your answer.

**Symmetric intervals on  $q_0$  and  $q_0 - p_0$** 

- ☐ Provide the final numerical value of the symmetric confidence interval for  $q_0$ , rounded into a X.X% format (do not write the calculation details):
- ☐ Same question for  $q_0 - p_0$  (with some calculation details or with a picture):
- ☐ What is the confidence level of the interval calculated in the previous question? \_\_%

**Shape of confidence intervals**

Let us consider an academic researcher and a shopkeeper. In which shape of a confidence interval on the difference  $q_0 - p_0$  (symmetric interval, underestimate, overestimate) would they be most interested? If your answer is not “symmetric”, then explain which respective shapes for the confidence intervals on  $p_0$  and  $q_0$  should have been considered to that end.

- ☐ Academic researcher:
- ☐ Shopkeeper:

**Advanced exercise 3.3** (Success rate of a new dating method).

You are targeting the online dating<sup>4</sup> market. Your method consists of finding matches via a complicated machine-learning algorithm and your users have to honor each match proposal and meet in flesh (which is checked by their cell phones exchanging some information via Bluetooth — hence the need of physical closeness). In short, the users do not get to choose their dates, your algorithm does it for them! If they do not meet within a week after a match, then both get expelled from your system. And they really do not want this, because registering to your system was costly. They accepted the principle because it is fun and adds some spice to the meeting, because they feel coached, because they save a lot of time not browsing profiles, and because you promised them a high success rate.



But, well, it is getting time to quantify this success rate, and how much higher it is with respect to other, more traditional, online dating applications.

You hire an independent polling organization. They interview 200 of your customers and 200 of another major application, and ask them how long they needed to get a serious relationship. (They defined a serious relationship as: lasting more than 1 month; including meetings of each other's friends; and of course, several sexual intercours). Data were the following ones: 46 days on average (with a standard deviation of 23 days) for your method, and 78 days on average (standard deviation: 18 days) for the traditional application. We assume that all 400 customers got a serious relationship at some point (possibly after waiting a long time).

By which guaranteed percentage is your method more effective than the traditional methods?

**Advanced exercise 3.4** (Car insurance company: planning). This exercise is a continuation of the corresponding exercises on pages 20 and 32.

Data was that out of 1,472 students with a car insurance interviewed, 256 had reported an accident for which they were held responsible, with an average amount of damages of 1,865 euros (and a standard deviation of 524 euros). We had already computed a symmetric confidence interval on the average expenses generated by students held responsible for an accident, namely,  $1,865 \pm 65$  euros.

1. How many additional students held responsible for an accident should we interview to reduce the margin of error to  $\pm 15$  euros?
2. In total, how many additional students with a car insurance (with or without an accident) should we interview to guarantee with high confidence that we will get the sample required in the previous question?

---

<sup>4</sup>French culture tip, for those interested in getting French soulmates: #1 operator for longer-term relationships would be Meetic (a French company! actually also world leader, I think, when you consider all its subsidiaries). They display advertisements though every media channel, and in particular, on TV and on the radio.

**Advanced exercise 3.5** (Value of a stock<sup>5</sup>). Consider a company producing high-tech products. It has a large stock of spare parts, which is the subject of a permanent inventory managed by a central information system, based on entry vouchers (deliveries from suppliers) and exit slips issued by production. Alice is an auditor who has been tasked with auditing the actual value of the stock of spare parts. The diversity of the items in stock led Alice to consider two categories:

- items of small value (less than 10 euros), for which there are many (1,532) references;
- items of significant unit cost (10 euros or more), for which there are only 180 references.

Auditing a given reference consists of two operations:

- checking the price of the corresponding article (i.e., check no input error occurred when the price was entered in the information system);
- recounting the available units; the latter operation takes some time.

From these two pieces of information, one can determine the *actual* value of the existing stock for this reference.

Five trained employees are in charge of recounting items, and they manage to check 50 items per hour (collectively). In view of the entries found in the general ledger account (reproduced in Table 3.1) and the time available, Alice decides:

- to check all the references in Category 2 (items over 10 euros);
- to conduct random checks on a number of references in Category 1 (items less than 10 euros).

To determine how many references should be randomly checked, Alice performs a first poll of 50 randomly selected references in the catalog, and determines the average values of these 50 references: in the general ledger account (before the audit) and their actual values (from the audit). The results are reproduced in Table 3.2.

Unit cost	Number of references	Total value
< 10 euros	1,532	3,366,495
≥ 10 euros	180	2,625,380

Table 3.1: Stock value according to the general ledger account.

Variable	Mean	Standard deviation
From ledger account	2,315.83	777.35
From audit	2,304.10	753.74
Difference	−11.73	110.32

Table 3.2: Results from a poll on 50 randomly selected references.

The following questions are aimed at understanding Alice's approach, and determining the number of additional references needed to achieve a certain accuracy.

<sup>5</sup>Warning: this is a difficult exercise, that was a third of a past exam statement; this should give you a first hint at what the exam will look like.

1. Explain Alice's approach; why does she audit all the references in Category 2, but uses a random method to audit the other category?

We first consider the information provided by the *actual* value of 50 sampled references (among 1532 references in total).

2. Extract the relevant statistical information corresponding to this sample.
3. Deduce from the result of the first poll a 95% confidence interval for the total (actual) value of the stock of articles of small value; indicate its accuracy.
4. It is assumed that the required accuracy was  $\pm 1\%$  of the total current value of these items: check that the accuracy obtained in Question 3 is not sufficient, and determine the sample size that would ensure the desired accuracy. What do you think of this size?

We are therefore looking for a smarter and more accurate approach. To this end, we now review the information provided by the *differences* between the ledger-account and actual (audit) values of the 50 references checked.

5. Extract the corresponding relevant statistical information.
6. Compute a 95% confidence interval of the difference between the ledger-account and the actual (audit) values.
7. Deduce a confidence interval for the actual total stock value; indicate accuracy.
8. Determine the sample size that would achieve the accuracy desired in Question 4. This time, what do you think of this size?
9. Take a moment of reflection: fundamentally, why is this second approach more effective?

## Hypothesis testing: methodology

In this chapter we will learn how to answer questions!

To do so, we consider two statements (examples will be given below):

- There is a starting point referred to as the null hypothesis  $H_0$ , from where we will move only if there is strong evidence that we must.
- There is therefore an alternative statement  $H_1$  which we will consider only if we had to discard  $H_0$ .

There is a strong asymmetry between the two hypotheses, they do not play similar roles:

- We will stick to  $H_0$  unless  $H_0$  is severely contradicted by the data (slight contradictions or doubts raised are not enough, we need a severe contradiction).
- If and only if such a severe contradiction is shown, we will discard  $H_0$  and turn to  $H_1$ .

The associated vocabulary is enlightening:

- We never accept  $H_0$ , we merely say that we fail to reject  $H_0$  (in case the data do not contradict  $H_0$  in a severe way). In this case, either  $H_0$  is true or  $H_0$  is incorrect but the data are not conclusive enough (e.g., because of a small sample size).
- In case of a severe contradiction, we reject  $H_0$  in favor of  $H_1$ . This means that we turn to  $H_1$  because we have no other option, but you should not write that we accept  $H_1$ .

A good example is provided by trials in courts:

- The starting point  $H_0$  is that the defendant is innocent; public prosecution must prove that he/she is guilty.
- If there is any possibility that the defendant could actually be innocent, he should not be convicted. Doubts are not enough to convict someone, strong facts are needed.
- If and only if strong facts and convincing evidence are raised, may the starting point  $H_0$  of innocence be rejected for the alternative statement  $H_1$  that the defendant is guilty.

### Guiding rules for the choices of $H_0$ and $H_1$

Let us try to generalize this principle and state some guiding rules for the choices of  $H_0$  and  $H_1$ . They are typically given by one of the following pairs. We will explain in detail, when solving exercises, which pair to pick and why; this requires practice and is one of the main difficulties when implementing statistical hypothesis testing.

Pair 1

- $H_0$ : the contrary of what is to be proved
- $H_1$ : what is to be proved

Pair 2

- $H_0$ : a reasonable<sup>1</sup> viewpoint
- $H_1$ : some statement that would require further thoughts or (costly) actions

Pair 3

- $H_0$ : a statement that you want to challenge
- $H_1$ : your own, personal, alternative view

Pair 4

- $H_0$ : a prudent action or consideration
- $H_1$ : a risky action or consideration (typically more profitable than the prudent action or consideration but costly in the case it was picked wrongly)

### Wording of statistical conclusions

Our conclusions will be of one of the following forms (always data-based, and either not rejecting  $H_0$  or rejecting  $H_0$ , but never accepting  $H_0$ ):

- The data collected fail to show that [ $H_0$  is not true] or cannot rule out that [ $H_0$  is true]
- The data collected show that [ $H_0$  is not true]

## 1. The Lady tasting tea

See the Wikipedia article about this famous (true) story: Fisher, one of the inventors of statistics, was having tea with his colleagues and a colleague's wife claimed that pouring milk before or after hot tea had an effect on the taste. The gentlemen constructed an experiment: prepare 8 cups of tea with milk, 4 of each category, and have her taste them in a random order.

The hypotheses these gentlemen had in mind were:

- $H_0$  [reasonable point of view / contrary of what is to be proved]: She is fantasizing.
- $H_1$  [what is to be proved]: She has the superpower<sup>2</sup> of distinguishing the tastes of these two milk-and-tea preparations.

Data collected are binary: for each cup, whether she succeeded (1) or failed (0) to correctly determine the order of tea and milk.

If  $H_0$  was correct, then she would have been guessing at random. (Unimportant) calculations show that she would only have had a  $1/70 \approx 1.4\%$  chance to get all answers right.

It turns out that she got all answers right.

The methodology here is to declare that

- this was not a matter of good or back luck (though it could have been, of course);
- this shows that  $H_0$  has to be rejected in favor of  $H_1$ .

---

<sup>1</sup>Reasonable = what your manager or the general public thinks, i.e., a common opinion

<sup>2</sup>Isn't that a wonderful talent anyone would like to have? Superman, please be quiet, you would also wish you had this power.



Indeed, such perfect 1, 1, 1, 1, 1, 1, 1, 1 data severely contradict the hypothesis  $H_0$  of random guessing.

Of course, there is always a risk to be incorrect. She could just have been super-lucky. The next examples will highlight that there are actually two risks to be taken into account.

## Lady tasting tea

From Wikipedia, the free encyclopedia

In the design of experiments in statistics, the **lady tasting tea** is a randomized experiment devised by Ronald Fisher and reported in his book *The Design of Experiments* (1935).<sup>[1]</sup> The experiment is the original exposition of Fisher's notion of a null hypothesis, which is "never proved or established, but is possibly disproved, in the course of experimentation".<sup>[2][3]</sup>

The lady in question claimed to be able to tell whether the tea or the milk was added first to a cup. Fisher proposed to give her eight cups, four of each variety, in random order. One could then ask what the probability was for her getting the specific number of cups she identified correct, but just by chance.

Fisher's description is less than 10 pages in length and is notable for its simplicity and completeness regarding terminology, calculations and design of the experiment.<sup>[4]</sup> The example is loosely based on an event in Fisher's life. The lady in question was Muriel Bristol, and the test used was Fisher's exact test.

### The experiment

The experiment provided the Lady with 8 randomly ordered cups of tea—4 prepared by first adding milk, 4 prepared by first adding the tea. She was to select the 4 cups prepared by one method. This offered the Lady the advantage of judging cups by comparison. She was fully informed of the experimental method.

The null hypothesis was that the Lady had no ability to distinguish the teas. In Fisher's approach, there is no alternative hypothesis;<sup>[2]</sup> this is instead a feature of the Neyman–Pearson approach.

The test statistic was a simple count of the number of successes in selecting the 4 cups. The null hypothesis distribution was computed by the number of permutations. The number of selected permutations equalled the number of unselected permutations. Using a combination formula, with  $n = 8$  total cups and  $k = 4$  cups chosen, there are  $\frac{8!}{4!(8-4)!} = 70$  possible combinations.



The experiment asked whether a taster could tell if the milk was added before the brewed tea, when preparing a cup of tea



Ronald Fisher in 1913

**Tea-Tasting Distribution**

Success count	Permutations of selection	Number of permutations
0	oooo	$1 \times 1 = 1$
1	ooox, ooxo, oxoo, xooo	$4 \times 4 = 16$
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	$6 \times 6 = 36$
3	xxxx, xoxx, xxox, xxxo	$4 \times 4 = 16$
4	xxxx	$1 \times 1 = 1$
<b>Total</b>		70

The critical region was the single case of 4 successes of 4 possible based on a conventional probability criterion ( $< 5\%$ ;  $1 \text{ of } 70 \approx 1.4\%$ ).

If and only if the Lady properly categorized all 8 cups was Fisher willing to reject the null hypothesis – effectively acknowledging the Lady's ability at a 1.4% significance level (but without quantifying her ability). Fisher later discussed the benefits of more trials and repeated tests.

David Salsburg reports that a colleague of Fisher, H. Fairfield Smith, revealed that in the test, the woman got all eight cups correct.<sup>[5][6]</sup> The chance of someone who just guesses getting all correct, assuming she guesses that four had the tea put in first and four the milk, would be only 1 in 70 (the combinations of 8 taken 4 at a time).

In popular science, Salsburg published a book entitled *The Lady Tasting Tea*,<sup>[5]</sup> which describes Fisher's experiment and ideas on randomization. Deb Basu wrote that “the famous case of the ‘lady tasting tea’” was “one of the two supporting pillars ... of the randomization analysis of experimental data.”<sup>[7]</sup>

## See also

- Hypergeometric distribution
- Permutation test
- Random assignment
- Randomization test

## References

1. Fisher 1971, II. The Principles of Experimentation, Illustrated by a Psycho-physical Experiment.
2. Fisher 1971, Chapter II. The Principles of Experimentation, Illustrated by a Psycho-physical Experiment, Section 8. The Null Hypothesis.
3. OED quote: **1935** R. A. Fisher, *The Design of Experiments* ii. 19, "We may speak of this hypothesis as the 'null hypothesis', and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation."
4. Fisher, Sir Ronald A. (1956) [*The Design of Experiments* (1935)]. "Mathematics of a Lady Tasting Tea". In James Roy Newman. *The World of Mathematics, volume 3*. Courier Dover Publications. ISBN 978-0-486-41151-4.
5. Salsburg (2002)
6. Box, Joan Fisher (1978). *R.A. Fisher, The Life of a Scientist*. New York: Wiley. p. 134. ISBN 0-471-09300-9.
7. Basu (1980a, p. 575; 1980b)

## 2. Detecting cheaters

Assume a professor asks the students to draw 200 times a fair coin and to write on a sheet of paper the results obtained, heads (H) and tails (T). Some students do their homework while some others cheat and try to write H's and T's at random. The professor has a trick to determine (and punish) the cheaters: he checks the presence or the absence of 6 consecutive heads or 6 consecutive tails; the absence of any such sequence reveals a cheater. Why and how?

The professor knows that there is a probability 97% that out of 200 random draws, there is such a sequence of 6H or 6T. Students usually ignore it and would not dare writing 6 consecutive H's and T's when making up the results (because to them, such sequences seem unlikely: they think the H's and T's should somehow compensate each other at all times).

He has two hypotheses in mind:

- $H_0$  [prudent point of view]: Think that the student is honest.
- $H_1$  [risky statement]: Declare him/her a cheater.

Why? Because you should only punish a student if you have enough evidence; otherwise, he/she is entitled to complain to the Dean. Therefore, it is prudent to think that the student is honest and did his/her homework; the risky statement is to accuse him/her of cheating.

There are two possible errors, one corresponding to a more harmful situation than the other.

- Type I error — Rejecting  $H_0$  while it was correct, i.e., punishing an honest student: this is extremely unfair, chances are high that the student complains and/or gives a bad evaluation to the teacher. This error is to be minimized as much as possible!
- Type II error — Failing to reject  $H_0$  while it was incorrect, i.e., not punishing a student that made up his/her sequence of 200 H's and T's. The associated danger is not as harmful and consists of loosing one's authority. This error should be as low as possible, but people sometimes cheat in smart ways and you cannot catch them all. This error is what it is...

Here, the professor's decision rule is based on the presence or absence of a sequence of 6 consecutive H's or T's. If such a sequence is not present for a given student, which had only a 3% chance to happen, we decide that this is not due to bad luck, but that it is the sign that  $H_0$  is incorrect, i.e., that the student cheated. The student's data severely contradict  $H_0$  and we reject it.

Therefore,

- Type I error equals 3% by construction (on average, 97% of honest students will get the desired sequence).
- Type II error cannot be quantified (all students who know the trick will make up a sequence with the desired subpattern; the other ones will be caught; but we ignore the respective proportions of the two groups of aware and ignorant students).

Of course, there is a way of never incorrectly punishing an honest student: never punish anyone. But that is too loose. We have to accept a small type I error to be able to occasionally reject  $H_0$ . We cannot stick to  $H_0$  at all times, otherwise we would never learn from data!

**Elementary exercise 4.1 (Discount).** Consider a discount or a new commercial policy: between "The discount is profitable" and "The discount is not profitable", which one should be  $H_0$  and which one should be  $H_1$ ? Spell out (in words) what the two errors correspond to (use, e.g., the words "invisible shortfall" and "tangible losses").

### 3. Advertisement campaign for nicotine patches

A typical exercise in our quizzes or exams would be the following.

#### Statement

A pharmaceutical company designed a new nicotine patch and would like to communicate on its high success rate: a 60% efficiency rate in quitting smoking for at least six consecutive months. To that end, it recruits, somewhat at random, 100 volunteers.

1. Extract all relevant statistical information.
2. Define the relevant hypotheses  $H_0$  and  $H_1$  and conduct a test.
3. It turns out that 64 of the volunteers reported a smoking break of at least six months. Take a business decision: what should the company do?



#### Answers

1. Extract all relevant statistical information.

The targeted population (from a statistical and a marketing viewpoint) is formed by smokers motivated for quitting smoking and ready to consider nicotine patches to this end. A sample of 100 such smokers is formed (hopefully at random). We get as data points whether the  $j$ -th smoker quitted smoking ( $x_j = 1$ ) or not ( $x_j = 0$ ) for at least six months. The objective of our study is to determine  $p_0$ , the actual proportion of the population (that is, of the future market) for whom the patch would be effective. It corresponds to the efficiency rate the company would like to communicate on. In the sample results (only revealed later in the exercise statement), we read an average efficiency of  $\bar{x}_{100} = 64\%$ .

Do these data prove or not whether this rate  $p_0$  is higher than 60%? I.e., is the 64% average efficiency reported in the sample significantly larger than 60%?

2. Define the relevant hypotheses  $H_0$  and  $H_1$  and conduct a test.

First, keep in mind that the hypotheses must never be set depending on the data collected; they must be stated before any data collection. This is why the statement only revealed the sample data after asking for the hypotheses.

The hypotheses are as follows.

- $H_0$  [prudent viewpoint]: The desired efficiency is not achieved, i.e.,  $p_0 \leq 60\%$ .
- $H_1$  [what we want to prove]: The desired efficiency is achieved, i.e.,  $p_0 > 60\%$ .

We tend to stick to  $H_0$  unless it is severely contradicted by the data. Therefore, when we want to prove something, we need to take it as  $H_1$ , take its contrary as  $H_0$ , and hope that  $H_0$  will be rejected. Is this the case here?

A first simplification is to consider the “limit” hypothesis  $H_0 : p_0 = 60\%$  versus  $H_1 : p_0 > 60\%$ .

Now, if  $H_0 : p_0 = 60\%$  is true, the following quantity, called a test statistic, should take its values on a normal curve (as indicated by a sophisticated result of probability theory called the central limit theorem); it is sometimes called the  $z$ -score:

$$t_{100} = \frac{\sqrt{100}}{\sqrt{0.60(1-0.60)}} (\bar{x}_{100} - 0.60).$$

Its value on our sample equals

$$t_{100} = \frac{\sqrt{100}}{\sqrt{0.60(1-0.60)}}(0.64 - 0.60) = 0.82.$$

Was this a likely value?

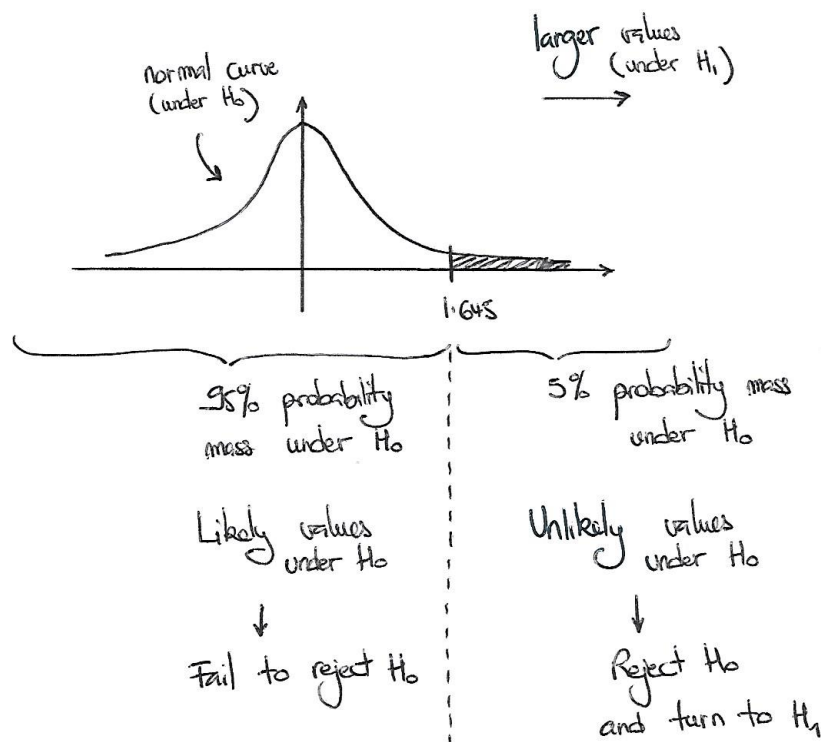
To answer this question, we first need to see what happens to the  $t_{100}$  quantity under  $H_1$ : as  $p_0$  is larger under  $H_1$  than under  $H_0$ , and as  $\bar{x}_{100}$  is close to  $p_0$  (by the law of large numbers), the  $t_{100}$  test statistic tends to take larger values under  $H_1$  than under  $H_0$ .

Under  $H_0$ , the values of the  $t_{100}$  test statistic were distributed according to a normal curve, i.e., were typically around 0.

A reasonable decision rule is therefore the following, and it involves a threshold  $r$  to be determined:

- If  $t_{100}$  is larger than  $r$ , then reject  $H_0$  and turn to  $H_1$ ; namely, we consider that unlikely values are not due to bad luck but to  $H_0$  being incorrect.
- If  $t_{100}$  is smaller than  $r$ , then do not reject  $H_0$ .

It only remains to set  $r$  so that the type I error is controlled: the probability of incorrectly rejecting  $H_0$  is the probability mass under the normal curve after the  $x$ -axis point  $r$ . See the table page 61 and the picture below: the appropriate value of  $r$  is 1.645.



All in all, with the data of our sample ( $t_{100} = 0.82$ ), we fail to reject  $H_0$ . We do not have enough evidence to reject the statement that the efficiency rate is smaller than 60% and need to stick with this claim.

3. Take a business decision: what should the company do?

What it should not do is to advertise a 60% efficiency rate straight away: there is not enough evidence to back up this claim.

But simply writing this is not enough, you need to reach a business conclusion, i.e., suggest an action; not just reword the statistical conclusion.

Here, you can, e.g., decide that the lack of statistical evidence is likely due to the sample size and suggest recruiting more volunteers (despite its cost). You can also ask the R&D department to work on a better product (but that is probably even more expensive and time consuming).

Also, you may run out of time and need to act now, in which case none of the two business conclusions above is suitable. In this case, you could check whether you can guarantee a 50% or 55% efficiency rate for the nicotine patches with the available data. (See an exercise page 63.)

#### **Additional remarks.**

A dishonest pharmaceutical company could have thought of picking the following hypotheses:

- $H'_0$ : The efficiency rate  $p_0$  is larger than 65%, i.e.,  $p_0 \geq 65\%$ .
- $H'_1$ : The efficiency rate  $p_0$  is smaller than 65%, i.e.,  $p_0 < 65\%$ .

Show that with the collected data, we would have failed to reject  $H'_0$ . (See an exercise page 63.)

This however does not mean that  $H'_0$  is true and that  $p_0$  is indeed larger than 65%; it merely shows that the data cannot lead to the conclusion that  $H'_0$  is incorrect.

When we stick to  $H'_0$ , we do not know in general whether this is because  $H'_0$  is true or the data are inconclusive (e.g., because of a small sample size).

Observe that with the same data, we failed to reject both

$$H_0 : p_0 \leq 60\% \quad \text{and} \quad H'_0 : p_0 \geq 65\%.$$

And we know that these two hypotheses cannot be both true at the same time!

This is in strong contrast with all the cases where we reject the null hypothesis and turn to the alternative hypothesis. Then we know that up to a reasonable risk of error,  $H_0$  is indeed incorrect.

In conclusion, hypothesis testing can only exclude scenarios/cases (and thus, can only lead to negative statements/conclusions):

- either the available data cannot show that  $H_0$  is incorrect (which does not mean that  $H_0$  is true);
- or they show that up to a reasonable risk of error,  $H_0$  is indeed incorrect.

#### **Final quote.**

Remember Fisher's words (see the document on the Lady-tasting-tea problem page 53): the hypothesis  $H_0$  is "never proved or established, but is possibly disproved, in the course of experimentation".

## 4. General methodology and notion of P-value

In the next chapters, we will see which test statistics to use in which case, and their behaviors under  $H_0$  and  $H_1$  (steps 3 and 4 of the methodology). We will explain after the box what P-values are (step 5 of the methodology).

### *Statistical hypothesis testing — general methodology*

1. Preliminary step: extract all relevant statistical information from the sample with size  $n$ .
2. Pick the hypotheses  $H_0$  (e.g., prudent or reasonable viewpoint) and  $H_1$  (e.g., the risky statement or what is to be proved); do not base your statements of these hypotheses on the data available, but on the context only.
3. Consider a test statistic  $t_n$ , with known behavior under  $H_0$  (e.g., normal-curve behavior). We will refer to this behavior as the expected behavior under  $H_0$ .
4. Take into account how  $t_n$  changes under  $H_1$  (does it take larger or smaller values than under  $H_0$ ?). We will refer to this behavior as the expected behavior under  $H_1$ .
5. Determine the sets of likely and unlikely values for  $t_n$ ; the set of unlikely values is usually referred to as the rejection region.  
Even better: compute the P-value.
6. Statistical conclusion: failure to reject or rejection of  $H_0$ , together with a comment on the P-value if applicable.
7. Business decision: what actions do you recommend to take, given the statistical conclusion?

### **An important methodological note.**

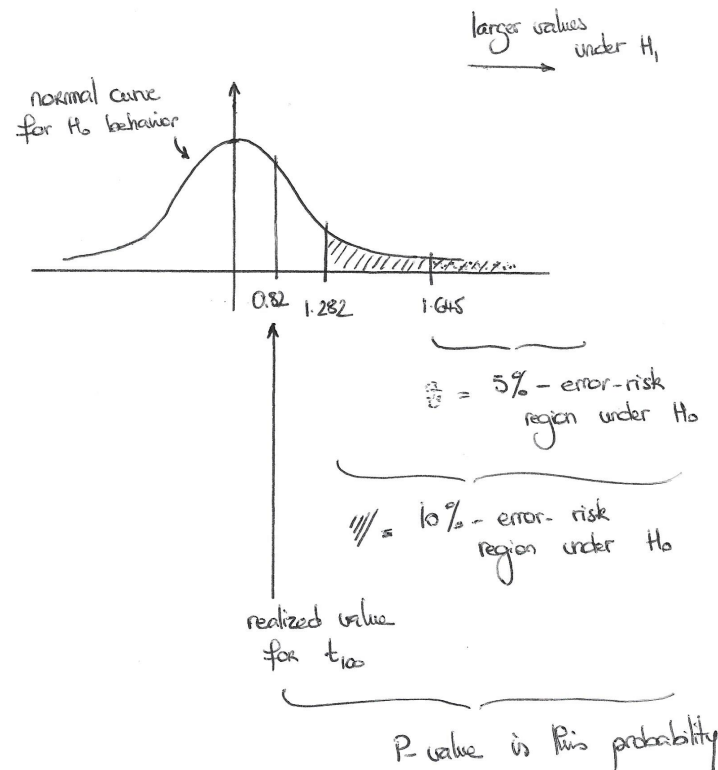
The hypotheses  $H_0$  and  $H_1$  need to be decided in advance, before you collect the data. It is a methodological mistake to set them in view of the data. We will see why in the next chapters, when dealing with two-sided tests.

### **P-values: calculation.**

So far, we set rejection regions corresponding to a given type I error and checked whether the test statistic  $t_n$  lies or not in this region in order to reject or fail to reject  $H_0$ . Our statistical conclusion was binary, thus not so informative.

A more informative way to proceed is to report a P-value: the probability mass of the rejection region with boundary set by the numerical value of the test statistic  $t_n$ . See the figure on the next page.

Application on the example with nicotine patches:



The P-value in this example is given by the probability mass under the normal curve beyond 0.82.

The table of page 61 (see also a larger version of it on the last page of this textbook) shows that the probability mass under the normal curve before 0.82 equals  $0.7939 = 79.39\%$ .

Therefore, the probability beyond 0.82, which is our P-value here, equals  $100\% - 79.39\% = 20.61\%$ .

### P-values: interpretation and decision rule.

The P-value is to be interpreted as the credibility level of  $H_0$  given the data and in view of the alternative  $H_1$ . Indeed, it is the probability to get a value of the test statistic as contradictory or more contradictory to  $H_0$  if we performed the experiment again. If the P-value is already small, the probability of a stronger contradiction is small only because the current value is already unlikely; this is the sign that  $H_0$  is not credible and should be rejected. If, on the contrary, the P-value is large, it would be easy to get more contradictory values of the test statistic, which shows that the current value does not contradict too much  $H_0$ . We set a conventional threshold of 5%.

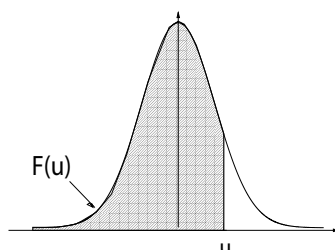
All in all, the summary is:

- A small P-value (i.e., typically smaller than 5%) indicates that  $H_0$  is not credible and should be rejected.
- A larger P-value (i.e., typically larger than 5%) means that  $H_0$  is credible (or at least, not too implausible) and prevents us from rejecting  $H_0$ .

In the nicotine patch example, the P-value computed (equal to 20.61%) was larger than 5%, therefore we fail to reject  $H_0$ .



This is of the course the same conclusion that we reached earlier, on page 57, except that we know now that this was not a borderline case: the P-value 20.61% is significantly larger than 5%, there is no ambiguity in not rejecting  $H_0$  given these data.



	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

**Reminder:** what type of conclusion to write (taking 5% as a reference threshold).

In the quizzes and exams, the wording of your conclusions will have to match the P-value:

- P-value larger than 5%: The data collected fail to show that [ $H_0$  is not true] or cannot exclude that [ $H_0$  is true]
- P-value below 5%: The data collected show that [ $H_0$  is not true]
- P-value below 1%: The data collected strongly show that [ $H_0$  is not true]

Where of course you should replace the technical statement [ $H_0$  is true] that only a few people may understand by a sentence in plain English conveying what is being tested and the conclusion.

The 5% threshold looks like a hard threshold, but it is not. Actually, when the P-value is only slightly larger than 5% (e.g., equals 6% or 7%), we could well write: “The data collected suggest that [ $H_0$  is not true].”

## 5. Elementary exercises

A first elementary exercise is stated on page 55, please solve it first!

**Elementary exercise 4.2.** Sometimes, companies changes the names of some of their products, e.g., Raider  $\rightarrow$  Twix as Wikipedia indicates:

Twix is a chocolate bar made by Mars, Inc., consisting of biscuit applied with other confectionery toppings and coatings (most frequently caramel and milk chocolate). [...] Twix was called Raider in mainland Europe for many years before its name was changed in 1991 [...] to match the international brand name. The name Twix is a portmanteau of twin biscuits, or “twin bix.”

Suppose that a product manager is thinking about changing the name of the product she is in charge of. She would only do it if at least 50% of the customers would prefer the new name. Which hypotheses  $H_0$  and  $H_1$  should she choose?

**Elementary exercise 4.3.** The economics department of a bank is looking for early signs of an economic crisis to issue a public signal, if necessary (i.e., it will not study the available data for internal purposes only). To do so, it monitors closely the expected delinquency rate of the mortgages undertaken by its customers. In a steady state of the economy (i.e., outside crises), the delinquency rate equals  $p_{\text{ref}} = 8.5\%$ . The bank considers as a crisis signal a significant increase of the delinquency rate. It studies a random number of credit files to determine for each of them whether a delinquency is likely to occur; this analysis is carried over by human experts, so only a small number of randomly chosen files is inspected. The bank denotes by  $p_0$  the current delinquency rate and tests  $H_0 : p_0 = p_{\text{ref}}$  (no deviation in the delinquency rate, no signal of a crisis) versus  $H_1 : p_0 > p_{\text{ref}}$  (an increase in the delinquency rate, which is the early signal of a crisis).

1. Explain the choice of  $H_0$  and  $H_1$ .
2. The collected data are as follows: out of 50 manually inspected files, 7 are extremely likely to default on the mortgage in the upcoming months. The bank considers that these 7 files already defaulted. Will it then reject or fail to reject  $H_0$ ?

In case  $H_0$  is rejected, it identified an early signal of a crisis; in case  $H_0$  failed to be rejected, no such signal was identified.

## 6. More advanced exercises (quiz-like exercises)

**Advanced exercise 4.1** (Nicotine patches, continued). There are several calculations you need to perform on this example:

- first, test whether the 64% efficiency rate on the sample of size 100 shows that the population efficiency rate  $p_0$  is larger than 55%;
- second, test whether the 64% efficiency rate on the sample of size 100 is not significantly smaller than 65%, i.e., that the population efficiency rate  $p_0$  may still be equal to or larger than 65%.

In both cases, the best would be to provide P-values and conclude from them.

**Advanced exercise 4.2** (Managing customers' dissatisfaction). In the company you were hired by, the typical customers' dissatisfaction rate was equal to 10%: boy, the phone was ringing all the time at the customer service center, they were overwhelmed with emails, etc. Fortunately, every complaint was recorded and this is how they could check that each year, about 10% of the customers were complaining about the products or service sold. You suspected that this all was in part because customers were not guided in their choices when buying your products or services. By hiring sales advisers you think you decreased drastically the dissatisfaction rate in less than two months (and that the volumes of sales boomed accordingly). But as a trained statistician, you want to fact-check your impressions. You cannot afford to wait one or two years to get the new dissatisfaction rate, so you conduct a statistical survey on recently served customers.

You sample 500 of these customers and ask them whether they are satisfied with the products or service bought or not, and whether they intend to complain to your customer service center. 32 intend to complain, 41 have no opinion yet, and 427 are satisfied.

Is your intuition confirmed? (Provide a P-value and then state both a statistical and a business conclusion.)



## One-sample tests (Testing equality to a reference value)

It has been a while since you did not get a quote on statistics. Here comes one!

*Statistics are like a bikini<sup>1</sup>. What they reveal is suggestive. What they conceal is vital.*

Arthur Koestler (Hungarian writer and journalist, 1905–1983)

### 1. Learning objectives

With the exercises at the end of the previous chapter you actually learned how to test whether a population proportion  $p_0$  was equal to or not (larger or smaller than) a reference value  $p_{\text{ref}}$ . We only considered one-sided alternative hypotheses, of the form  $H_1 : p_0 < p_{\text{ref}}$  and  $H_1 : p_0 > p_{\text{ref}}$ .

We will study in this chapter

- how to consider two-sided alternatives, i.e., test  $H_0 : p_0 = p_{\text{ref}}$  versus  $H_1 : p_0 \neq p_{\text{ref}}$ ;
- the case of general population means  $\mu_0$ , for which the methodology will be very similar, but the formula for the test statistics  $t_n$  will be slightly modified.

We will also start learning how to read outputs from a statistical software: the latter operates all the calculations for you. However, you have to read, interpret and convey the message of the P-value thus obtained.

---

<sup>1</sup>Or: a (boxer) brief; both genders can participate to the citation!

## 2. A smooth start with proportions

**Test 5.1.** *One-sample test for a proportion*

(Testing the equality of a population proportion  $p_0$  to a reference value  $p_{\text{ref}}$ )

**Data:**  $x_1, \dots, x_n$  with values either 0 or 1, and where  $n \geq 30$

**Parameter(s) of interest:** population proportion  $p_0$

**Hypothesis  $H_0$ :**  $p_0 = p_{\text{ref}}$

**Test statistic:**

$$t_n = \sqrt{n} \frac{\bar{x}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})}}$$

**Behavior under  $H_0$ :** normal curve

**Behavior under  $H_1$ :**

- if  $H_1$  includes  $p_0 > p_{\text{ref}}$ , then  $t_n$  tends to take larger values;
- if  $H_1$  includes  $p_0 < p_{\text{ref}}$ , then  $t_n$  tends to take smaller values.

**One-sided tests (deviations from one side only).** They correspond to the case when we test the pairs

$$\begin{cases} H_0 : p_0 \leq p_{\text{ref}} \\ H_1 : p_0 > p_{\text{ref}} \end{cases} \quad \text{or} \quad \begin{cases} H_0 : p_0 \geq p_{\text{ref}} \\ H_1 : p_0 < p_{\text{ref}} \end{cases}$$

In these cases, we only expect deviations on one side (e.g., when granting a discount, you expect that the order rate remains the same or increases, you do not expect it to decrease).

To perform these tests, it suffices to consider the limit cases (the values in  $H_0$  that are the closest to the ones in  $H_1$ ):

$$\begin{cases} H_0 : p_0 = p_{\text{ref}} \\ H_1 : p_0 > p_{\text{ref}} \end{cases} \quad \text{or} \quad \begin{cases} H_0 : p_0 = p_{\text{ref}} \\ H_1 : p_0 < p_{\text{ref}} \end{cases}$$

Mathematical details of this reduction are omitted but we hope you find it natural!

You can then apply the test principle stated above.

**Two-sided tests (deviations from both sides allowed).** We mean hypotheses of the form

$$\begin{cases} H_0 : p_0 = p_{\text{ref}} \\ H_1 : p_0 \neq p_{\text{ref}} \end{cases}$$

corresponding, for instance, to the following story.

A tech company produces smartwatches and is thinking of offering them in rose gold. It already tried several other pink or rose gold products, like smartphones, and they were very popular; typically, 20% of the customers bought them. Just to be on the safe side, the company conducts a small survey to see whether they indeed should also have 20% of their watches in rose gold within the first series of production.

What are the hypotheses that the company should consider? (Remember: hypotheses have to be decided before data are collected.)

Call  $p_0$  the proportion of customers that will buy a watch in rose gold rather than in any other color. The reasonable or prudent hypothesis  $H_0 : p_0 = 20\%$  is to consider that this rate is equal to the rate observed for the (several) other, similar products. We have no clue as in which direction a deviation from this reference value  $p_{\text{ref}} = 20\%$  would go: would rose gold watches sell more or less than other pink / rose gold products? Therefore, we take an uninformed alternative hypothesis  $H_1 : p_0 \neq 20\%$ .

The company marketing division recruits 200 customers and has them choose which color they would buy: white, black, grey, or rose gold. 52 customers indicate a preference for rose gold. What do you think: should the 20% production proportion be revised?

Collected data are  $x_1, \dots, x_{200}$  where  $x_j = 1$  if the  $j$ -th recruited customer bought a rose gold watch, and  $x_j = 0$  otherwise. The sample proportion is  $\bar{x}_{200} = 52/200 = 26\%$ . Is it significantly different from the reference value 20%?

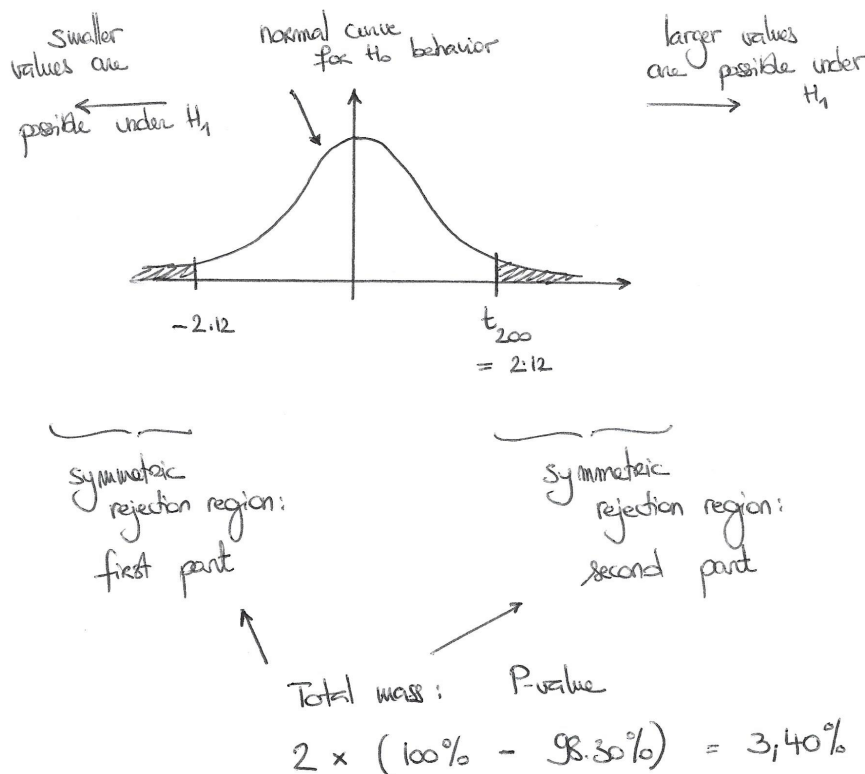
The test statistic equals

$$t_{200} = \sqrt{200} \frac{\bar{x}_{200} - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})}} = \sqrt{200} \frac{0.26 - 0.20}{\sqrt{0.20(1 - 0.20)}} \approx 2.12.$$

Under  $H_0$  we expected a normal-curve behavior, that is, values rather centered around 0.

Under  $H_1$ , larger or smaller values are expected: we reject  $H_0$  above a given threshold  $r$  or below  $-r$ . The rejection region is formed by two symmetric components.

All in all, we obtain the following picture, where we compute the P-value: 3.40%, which is below 5%.



The statistical conclusion is that we reject  $H_0$ . We suspect that the typical 20% rate will not apply here: in view of the data, we suspect that the demand for rose gold will be higher than 20%.

How higher? And does it matter? It might be OK to quickly run out of the most popular color, because it conveys the idea that only privileged people get these smartphones and thus creates a desire among potential consumers. The business conclusion is that further (qualitative and quantitative) marketing studies are needed. You do not want to build a several millions-of-dollars business policy on a survey with 200 respondents, do you?

### 3. The case of general means

**Test 5.2.** *One-sample test for a population mean*  
*(Testing the equality of a population mean  $\mu_0$  to a reference value  $\mu_{\text{ref}}$ )*

**Data:**  $x_1, \dots, x_n$  with general quantitative values and where  $n \geq 30$

**Parameter(s) of interest:** population mean  $\mu_0$

**Hypothesis  $H_0$ :**  $\mu_0 = \mu_{\text{ref}}$

**Test statistic:**

$$t_n = \sqrt{n} \frac{\bar{x}_n - \mu_{\text{ref}}}{s_{x,n}}$$

**Behavior under  $H_0$ :** normal curve

**Behavior under  $H_1$ :**

- if  $H_1$  includes  $\mu_0 > \mu_{\text{ref}}$  then  $t_n$  tends to take larger values;
- if  $H_1$  includes  $\mu_0 < \mu_{\text{ref}}$  then  $t_n$  tends to take smaller values.

The same comments as in the case of proportions apply (namely: the reduction to a limit case for one-sided pairs of hypotheses; and the two symmetric components of the rejection region for a two-sided alternative hypothesis  $H_1$ ).

Let us consider an example, in which we will discuss various decision-makers, each thinking of different pairs of hypotheses. It was a long time since we last talked about French politics, let us therefore revisit the topic!

**Example: a debate on the salary evolutions of French civil servants.** Or lack of evolutions, according to some! Let us first review how the pay of French civil servants is determined.

The pay of a civil servant is composed of:

- a base pay known as “*traitement*”;
- possible overtime pay;
- possible bonuses, which depend on the particular job assignment and possibly of the individual worker.



The “*traitement*” is for most civil servants determined by multiplying an index by the value of the index point in euros. The value of the index point is set by the executive and is raised regularly to compensate for inflation. The index depends on the body (“*corps*”), rank and seniority in rank (“*échelon*”).

Source: Wikipedia

Around 2010, when France’s major right-wing party was leading the country (called UMP at that time, now LR), with Nicolas Sarkozy being the president of the Republic and François Fillon being the Prime Minister, it was decided to freeze the value of the index point of civil servants. It remained frozen till Spring 2016. Even before 2010, under right-wing governments, it was raised by multiplicative factors that were smaller than the inflation.

So, you would think that civil servants were being paid less (in constant euros) year after year, wouldn’t you? But civil service unions and the government were disagreeing on this point, and it turns out that both are right in some sense!

Consider that we are in early 2012 (a few months before the presidential elections) and let us start with the unions’ viewpoint. Since common sense is not enough to convince public opinion (common sense is not so common!) the unions decide to resort to facts and numbers. Everybody loves facts and numbers; they can convince the journalists better than any mathematical argument about the index point being frozen or reevaluated slower than the inflation rate.

Now, France has no human resources department for its civil servants. There is no easy way to access the current salary data. A phone survey will be needed! The result of the survey will be compared, e.g., to the data collected in the last census for which the average monthly salary of civil servants was calculated over the whole population. Suppose<sup>2</sup> that such censuses take place every 8 years and that the last of them was in 2006. It showed an average monthly salary of 2,245 in constant euros (i.e., in euros of 2012: adjusted for inflation).

But before we have the survey conducted, we need to state the hypotheses the unions have in mind (remember: hypotheses have always to be set in advance).

### 1. Relevant statistical information and hypotheses $H_0$ , $H_1$

The population studied is formed by all civil servants in France. The parameter of interest is  $\mu_0$ , their average monthly salary. The reference parameter is  $\mu_{\text{ref}} = 2,245$ , the average monthly salary calculated with the last exhaustive census in 2006 and converted into constant euros.

The unions want to prove that the average salary decreased: this statement will be their  $H_1$  hypothesis. The contrary of this statement, that salaries remained the same (or increased), is taken as  $H_0$ . We hope that the future data will severely contradict  $H_0$  and that we can reject it.

Mathematically, this corresponds to

$$\begin{cases} H_0 : \mu_0 \geq \mu_{\text{ref}} & \text{(same or larger average salary)} \\ H_1 : \mu_0 < \mu_{\text{ref}} & \text{(smaller average salary),} \end{cases}$$

where  $\mu_{\text{ref}} = 2,245$ ; with limit case

$$\begin{cases} H_0 : \mu_0 = \mu_{\text{ref}} \\ H_1 : \mu_0 < \mu_{\text{ref}} \end{cases}$$

<sup>2</sup>Actually, this last exhaustive census took place in 1999 and since 2004, census is organized by surveying at random 8% of the population every year, at least in cities of more than 10,000 inhabitants. In the good old times, censuses were indeed taking place every 8 years.

So, unions have 1,000 French adults interviewed about their salaries, out of which 345 are civil servants. They declare an average monthly salary of 2,193 euros (with associated standard deviation of 573 euros).

Now, we may at last answer the question we had in mind: are these data a smoking gun against governmental lies?

2. We first compute a P-value.

Collected data are  $x_1, \dots, x_{345}$  taking positive values, where  $x_j$  is the monthly salary declared by the  $j$ -th respondent. Their mean and standard deviation equal  $\bar{x}_{345} = 2,193$  and  $s_{x,345} = 573$ .

Is 2,193 a sample mean significantly smaller than  $\mu_{\text{ref}} = 2,245$ ?

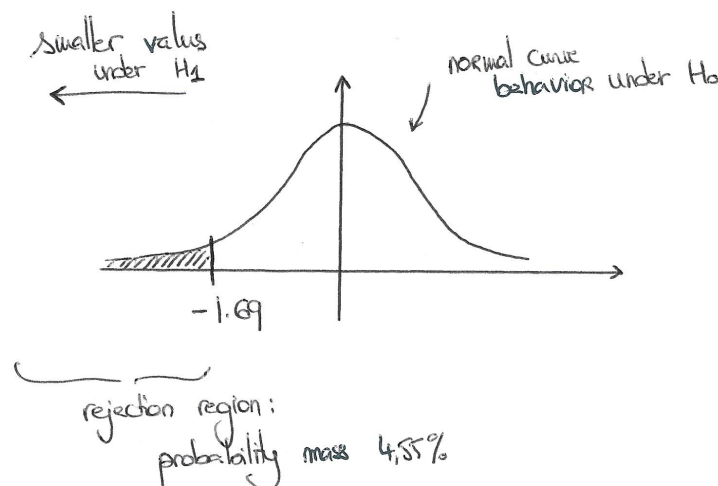
The test statistic equals

$$t_{345} = \sqrt{345} \left( \frac{\bar{x}_{345} - \mu_{\text{ref}}}{s_{x,345}} \right) = \sqrt{345} \left( \frac{2,193 - 2,245}{573} \right) \approx -1.69.$$

Under  $H_0$  we expected a normal-curve behavior, that is, values rather centered around 0.

Under  $H_1$ , smaller values were expected to be taken: we reject  $H_0$  below a given threshold  $r$ .

All in all, we obtain the following picture. By now, you should be able to calculate the P-value in a routinely manner: it equals 4.55%, which is below 5%.



3. What are the conclusions that the unions should reach?

The statistical conclusion is that we reject  $H_0$  in favor of  $H_1$ : these data prove that the average salary decreased in the considered period (2006–2012).

The business action that the unions should take is, e.g., to discuss with journalists and send them their study. They do have a smoking gun!

Now, let us re-do all the calculations from someone else's viewpoint. Consider a neutral newspaper (like Le Monde). They do not want to have any prior. To them, the most reasonable statement to start with is that salaries remained constant. They will depart from this neutral and reasonable statement only if data prove that it is necessary. In this case, they are ready to believe that the average salary increased or decreased: they are ready to follow the government's argument or the unions'. These neutral journalists do not exclude any possibility in advance.

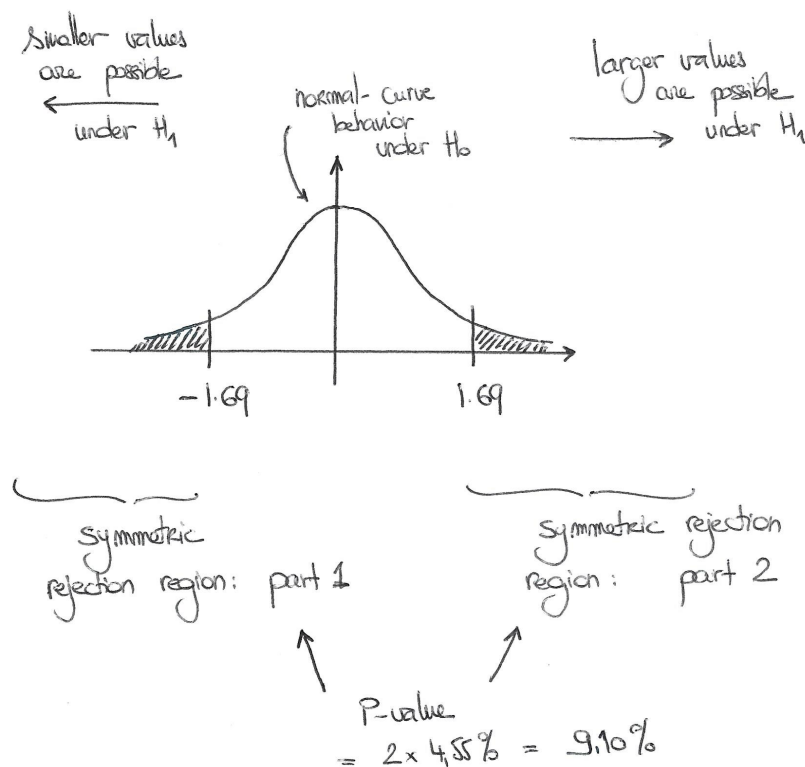
4. Let us answer all questions again from this new viewpoint.

The hypotheses are changed into

$$\begin{cases} H_0 : \mu_0 = \mu_{\text{ref}} & (\text{constant average salary}) \\ H_1 : \mu_0 \neq \mu_{\text{ref}} & (\text{smaller or larger average salary}), \end{cases}$$

where  $\mu_{\text{ref}} = 2,245$ . The test statistic still equals  $t_{345} \approx -1.69$  and under  $H_0$  we still expect a normal-curve behavior.

But now, under  $H_1$ , smaller or larger values were expected to be taken: we reject  $H_0$  according to a symmetric rejection region. The picture below shows that the P-value equals 9.10%.



The P-value is now larger than 5%: the neutral journalists do not have enough evidence to reject  $H_0$ . Of course, they feel that there is some tendency but evidence is, in some sense, weak.

Their action would be to write a prudent statement like: "Unions come with interesting preliminary data suggesting a decrease in salary, but further and independent investigations are needed as these data, per se, do not prove the claimed decrease." (Note the different uses of the two verbs "suggest" and "prove".)

**What will marketing professors tell you?** In marketing classes you will typically only perform two-sided tests. But you realize from the example above that if you have a prior and resort to a one-sided test, then the P-value may be divided by a factor of 2. Sometimes, you then go below the 5% bar; sometimes not.

That is why marketing professors usually use the following categories to state their statistical conclusions when dealing with a two-sided test:

- P-value below 5% = reject  $H_0$ ;

- P-value between 5% and 10% and the test is two-sided = grey zone;
- P-value above 10% = fail to reject  $H_0$ .

The rationale behind these categories is that in the grey zone, conclusions would differ depending on your prior beliefs, while the latter would not affect the former in the other two categories.

**Back to unions versus government!** We had promised that we would explain why both unions and the government were right.

The unions' argument was already explained in theory and illustrated in practice (on data that we invented): the unions are interested in the average salary, where the average is computed over all civil servants.

The government used to say in 2012 that 83% of the civil servants had had an increase of their salaries in the past 6 years. True: because they grew older and got some (rather automatic) raises due to seniority, which were larger than the difference between inflation and the small increases in the index point. Those people who indeed were civil servants in 2006 and 2012 were earning more in 2012.

To make it clear, the unions complained about the fact that a 30-year-old teacher with 6 years of seniority in 2012 would earn less than a 2006 teacher with the same profile. The government noted with satisfaction that the 2006 teacher with the same profile earned more in 2012 when she/he is 36-years old and has 12 years of seniority.

Journalists could typically not explain both viewpoints clearly enough and were satisfied by just pointing out that government and unions had dissenting viewpoints.

## 4. Using a statistical software

Fortunately, in the real world, you do not need to perform all these computations by yourself: statistical softwares will take care of it. This is, by the way, also the case for confidence intervals.

We will use SPSS for simplicity, because it is clickable and because of its popularity in the business environment. There are also other available choices, the outputs of all statistical softwares are similar anyway. These other choices include Microsoft Excel (with the XLStat package) or the open-source software R, preferred by professional statisticians due to the multiple available packages that allow for recent methods to be used.

**SPSS outputs.** Consider the following SPSS outputs corresponding to the two tests conducted in the previous sections.

As far as the test for a mean is concerned:

**T-Test****One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
Salary	345	2193,01	572,956	30,847

**One-Sample Test**

	Test Value = 2245					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Salary	-1,685	344	,093	-51,990	-112,66	8,68

These tables report the following quantities (where  $n$  denotes the sample size):

N	Mean	Std. Deviation	Std. Error. Mean
$n$	$\bar{x}_n$	$s_{x,n}$	$s_{x,n}/\sqrt{n}$

t	df	Sig. (2-tailed)	Mean Diff.	95% confidence interval
$\sqrt{n} \left( \frac{\bar{x}_n - \mu_{\text{ref}}}{s_{x,n}} \right)$	$n - 1$	P-value	$\bar{x}_n - \mu_{\text{ref}}$	$\bar{x}_n - \mu_{\text{ref}} \pm 1.96 s_{x,n}/\sqrt{n}$

The P-value for the two-sided test as indicated by SPSS equals 9.30%; we had computed 9.10% and the small difference is only due to two rounding errors. First, we rounded off the value of the test statistic to  $-1.69$  instead of the  $-1.685$  that SPSS uses. Second, we computed the P-value with the normal curve while SPSS resorts to a distribution called Student's t-distribution with 344 degrees of freedom (sample size minus 1). Student's t-distributions are very close, but not exactly equal, to the normal distribution when the degrees of freedom are large enough, say, larger than 30. A similar remark holds for the confidence interval formula above in the table: SPSS does not use 1.96, but the Student's quantile of level 97.5%, which is very close to 1.96 whenever  $n$  is large.

We had already drawn your attention to these issues on page 38.

For the test for a proportion:

**Binomial Test**

		Category	N	Observed Prop.	Test Prop.	Sig. (G-tailed)
Color	Group 1	Rose gold	52	,26	,2	,034
	Group 2	Other color	148	,74		
	Total		200	1.0		

SPSS indicates a P-value ("Sig.") equal to 3.40% for a two-sided test (see "2-tailed"). We are actually cheating a bit: for pedagogical reasons, we edited the SPSS output above to match our calculations. This is because SPSS, as surprising as it seems, does not feature the test we studied in Section 2, it only features a more complicated test called Fisher's exact test. The original SPSS output performed this more efficient test, in a one-sided way, and thus the P-value differed.

**R outputs.** Just for fun, let us see what the outputs of the more advanced software R look like:

```
> prop.test(52,200,0.2,correct = FALSE)

1-sample proportions test without continuity correction

data:  52 out of 200, null probability 0.2
X-squared = 4.5, df = 1, p-value = 0.03389
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.2041383 0.3249075
sample estimates:
      p 
0.26
```

```
> salaries = read.table("Salary.txt")
> t.test(salaries, alternative="less", mu=2245)
```

```
One Sample t-test

data:  salaries
t = -1.6854, df = 344, p-value = 0.04641
alternative hypothesis: true mean is less than 2245
95 percent confidence interval:
 -Inf 2243.885
sample estimates:
mean of x 
2193.01
```

This time, we can choose the type of test (two-sided or one-sided). R gets 3.389%  $\approx$  3.4% as a P-value for the proportion test (as we obtained) and 4.641%  $\approx$  4.6% for the one-sided test about the mean (we had obtained 4.55% but the same rounding-off remarks apply as for SPSS).

## 5. Some food for thought: one-sided versus two-sided tests

Andy Field (University of Sussex, UK) is the author of the textbook *Discovering Statistics Using IBM SPSS Statistics*—DSUS in short, possibly followed by a number indicating the edition, e.g., DSUS4 is the fourth edition of the textbook.

When moving from DSUS3 to DSUS4, he wrote a long blog post about one-sided tests and their caveats (their advantage is clear: they lead to smaller P-values and thus to more frequent rejections of  $H_0$ ). Please read this post! You will realize that even experienced statisticians may have a hard time setting their hypotheses right...

Source: article retrieved on August 2nd, 2021 from <https://www.discoveringstatistics.com/2012/07/21/one-tailed-tests/>

One-tailed tests are problematic for three reasons:

1. As the question I was sent illustrates, when scientists see interesting and unexpected findings their natural instinct is to want to explain them. Therefore, one-tailed tests are dangerous because like a nice piece of chocolate cake when you're on a diet, they waft the smell of temptation under your nose. You know you shouldn't eat the cake, but it smells so nice, and looks so tasty that you shovel it down your throat. Many a scientist's throat has a one-tailed effect in the opposite direction to that predicted wedged in it, turning their face red (with embarrassment).
2. One-tailed tests are appropriate only if a result in the opposite direction to the expected direction would result in exactly the same action as a non-significant result (Lombardi & Hurlbert, 2009; Ruxton & Neuhaeuser, 2010). This can happen, for example, if a result in the opposite direction would be theoretically meaningless or impossible to explain even if you wanted to (Kimmel, 1957). Another situation would be if, for example, you're testing a new drug to treat depression. You predict it will be better than existing drugs. If it is not better than existing drugs (non-significant p) you would not approve the drug; however it was significantly worse than existing drugs (significant p but in the opposite direction) you would also not approve the drug. In both situations, the drug is not approved.
3. One-tailed tests encourage cheating. If you do a two-tailed test and find that your p is .06, then you would conclude that your results were not significant (because .06 is bigger than the critical value of .05). Had you done this test one tailed however, the p you would get would be half of the two tailed value (.03). This one-tailed value would be significant at the conventional level. Therefore, if a scientist finds a two-tailed p that is just non-significant, they might be tempted to pretend that they'd always intended to do a one-tailed test, half the p value to make it significant and report that significant value. Partly this problem exists because of journal's obsessions with p-values, which therefore rewards significance. This reward might be enough of a temptation for some people to half their p-value just to get a significant effect. This practice is cheating (for reasons explained in one of the Jane Superbrain boxes in Chapter 2 of my SPSS/SAS/R books). Of course, I'd never suggest that scientists would half their p-values just so that they become significant, but it is interesting that two recent surveys of practice in ecology journals concluded that "all uses of one-tailed tests in the journals surveyed seemed invalid." (Lombardi & Hurlbert, 2009), and that only 1 in 17 papers using one-tailed tests were justified in doing so (Ruxton & Neuhaeuser, 2010).

For these reasons, DSUS4 is going to discourage the use of one-tailed tests unless there's a very good reason to use one (e.g., 2 above).

PS Thanks to Shane Lindsay who, a while back now, sent me the Lombardi and Ruxton papers.

## References

- Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, 54(4), 351-353. doi: 10.1037/h0046737
- Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecology*, 34(4), 447-468. doi: 10.1111/j.1442-9993.2009.01946.xISTEX
- Ruxton, G. D., & Neuhaeuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, 1(2), 114-117. doi: 10.1111/j.2041-210X.2010.00014.x

## Discovering Statistics: The Blog



Professor Andy Field  
 Blog of Professor Andy Field.  
 Author of *An adventure in statistics: the reality enigma*;  
 'Discovering Statistics Using SPSS/SAS' (amongst other things).

SATURDAY, JULY 21, 2012

### One-Tailed Tests

I've been thinking about writing a blog on one-tailed tests for a while. The reason is that one of the changes I'm making in my re-write of DSUS4 is to alter the way I talk about one-tailed tests. You might wonder why I would want to alter something like that – surely if it was good enough for the third edition then it's good enough for the fourth? Textbook writing is quite an interesting process because when I wrote the first edition, I was very much younger, and to some extent the content was driven by what I saw in other textbooks. Even as the book has evolved over certain editions, the publishers will get feedback from lecturers who use the book, I get emails from people who use the book, and so, again, content gets driven a bit by what people who use the book want and expect to see. People expect to learn about one-tailed tests in an introductory statistics book and I haven't wanted to disappoint them. However, as you get older you also get more confident about having an opinion on things. So, although I have happily entertained one-tailed tests in the past, in more recent years I have felt that they are one of the worse aspects of hypothesis testing that should probably be discouraged.

Yesterday I got the following question landing in my inbox, which was the perfect motivator to write this blog and explain why I'm trying to deal with one-tailed tests very differently in the new edition of DSUS4:

"I need some advice and thought you may be able to help. I have a one-tailed hypothesis, ego depletion will increase response times on a Stroop task. The data is parametric and I am using a related T-Test. Before depletion the Stroop performance mean is 70.66 (12.36). After depletion the Stroop performance mean is 61.95 (10.36). The t-test is,  $t(138) = 2.07$ ,  $p = .02$  (one-tailed). Although the t-test comes out significant, it goes against what I have hypothesised. That Stroop performance decreased rather than increased after depletion. So it goes in the other direction. How do I acknowledge this in a report? I have done this so far. Is it correct? Although the graph suggests there was a decrease in Stroop performance times after ego-depletion. Before ego-depletion ( $M=70.66$ ,  $SD=12.36$ ) after ego-depletion ( $M=61.95$ ,  $SD=10.36$ ), a t-test showed there was a significance between Stroop performance phase one and two  $t(138) = 10.94$ ,  $p < .001$  (one-tailed)."

This question illustrates perfectly the confusion people have about one-tailed tests. The author quite rightly wants to acknowledge that the effect was in the opposite direction, but quite wrongly still wants to report the effect... and why not, effects in the opposite direction and interesting and intriguing and any good scientists want to explain interesting findings.

The trouble is that my answer to the question of what to do when you get a significant one-tailed p-value but the effect is in the opposite direction to what you predicted is (and I quote my re-written chapter 2 here): "If you do a one-tailed test and the results turn out to be in the opposite direction to what you predicted you must ignore them, resist all temptation to interpret them, and accept (no matter how much it pains you) the null hypothesis. If you don't do this, then you have done a two-tailed test using a different level of significance from the one you set out to use"

[Quoting some edited highlights of the new section I wrote on one-tailed tests]:

## 6. Elementary exercises

The following elementary exercises all deal with the same data set, which will be presented and studied in detail in Chapter 6. It was gathered over the years by surveying a small number (some dozens) of students of some Dutch business school (ABS, Arnhem Business School) every year. The version that we consider contains measurements (height, age, etc.) of several hundreds of students. ABS students come from various countries and we are interested in relating their heights to the heights of Dutch people. As Wikipedia<sup>3</sup> indicates, the average heights of Dutch adults equal 181 and 169 centimeters, for men and women, respectively.

**Elementary exercise 5.1.** We first deal with the average height of male ABS students, which we denote by  $\mu_0^\sigma$ . We want to relate it to the average height of male Dutch adults, which we refer to as  $\mu_{\text{ref}}^\sigma = 181$  centimeters. We take an agnostic viewpoint and do not assume any preliminary thoughts or observations on the heights of the (international) ABS students (as opposed to the second question of the next exercise).

1. Explain why one should choose the hypotheses  $H_0 : \mu_0^\sigma = 181$  versus  $H_1 : \mu_0^\sigma \neq 181$ .

The considered data set reports the heights of 253 male ABS students, with sample average height 181.24 centimeters (and associated sample standard deviation of 8.11 centimeters).

2. Based on the described data set, conduct the test and compute the associated P-value for the hypotheses stated above. Show that we fail to reject  $H_0$ .

**Elementary exercise 5.2.** We repeat the same exercise with  $\mu_0^\circ$ , the average height of female ABS students. The reference value is  $\mu_{\text{ref}}^\circ = 169$  centimeters in this case. On the sample of 223 female ABS students, the average height was 166.03 centimeters (with an associated standard deviation of 6.70 centimeters).

1. In case of an agnostic viewpoint: mimic the arguments and calculations of the first exercise.

Suppose now that the statistics professor had already had the feeling, before collecting any data, that Dutch women were particularly tall compared to South-European women. ABS students coming from all over Europe, their average height must be close to the average height of European women, thus it should certainly be smaller than the average height of Dutch women.

2. With this mindset before collecting data, what would have the hypotheses been, and which P-value would have been obtained? Compare to the agnostic situation.

---

<sup>3</sup>URL: [https://en.wikipedia.org/wiki/List\\_of\\_average\\_human\\_height\\_worldwide](https://en.wikipedia.org/wiki/List_of_average_human_height_worldwide), retrieved on September 26, 2017



**Elementary exercise 5.3.** In this final exercise, we consider the following outputs generated by the statistical software SPSS.

### Test (for men)

#### One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Height (in cm)	253	181,24	8,108	,510

#### One-Sample Test

	Test Value = 181					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Height (in cm)	,473	252	,637	,241	-,76	1,25

### Test (for women)

#### One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Height (in cm)	223	166,03	6,700	,449

#### One-Sample Test

	Test Value = 169					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Height (in cm)	-6,616	222	,000	-2,969	-3,85	-2,08

1. Find again all numerical values provided in the statements of or in your answers to the previous two exercises.
2. Try to recompute all other numerical values present in these outputs. (Hint: "standard error" equals the standard deviation divided by the root of the sample size.)

## 7. More advanced exercises (quiz-like exercises)

### Advanced exercise 5.1 (A controversial governmental reform, short version).

Assume that you are the leader of a government that wants to put in place a highly controversial<sup>4</sup> reform. Recent history shows that as long as not more than 15% of the total population is ready to actively fight it (with street demonstrations or via social networks), the reform can be safely adopted. Otherwise, bad things can happen, such as: the government might loose the next elections; the government might need to repeal the reform shortly after it is voted by the Parliament; etc. Therefore, the government wants to be sure that the 15% threshold will not be reached.



1. Start extracting the relevant statistical information. State the hypotheses to be tested.

The government mandates a polling organization, which conducts a survey over 1,000 adults living in France. Among them, 980 express an opinion: 131 are ready to actively fight the reform, while the 849 other ones will not (though some of them are also against the reform).

2. What should the government do? Back up your answer with figures, namely, a P-value.

### Advanced exercise 5.2 (Seizure of MegaUpload and side effects).

The seizure of MegaUpload, a popular filesharing website with 150 million registered users, occurred on January 19, 2012 following a US indictment accusing MegaUpload of harboring millions of copyrighted files (source: Wikipedia). Legal websites offering replay or on-demand streaming, like the ones of the major TV channels, were of course happy with the news. A few weeks later, they wanted to find out whether their dreams of getting more visits and more users (thus, ultimately, more advertisement revenue!) became true or not. Data points are given by the percentage of global Internet users visiting a web site at a given time, as reported by the website [www.alexa.com](http://www.alexa.com). We assume that the latter website measures the audience every 30 minutes.

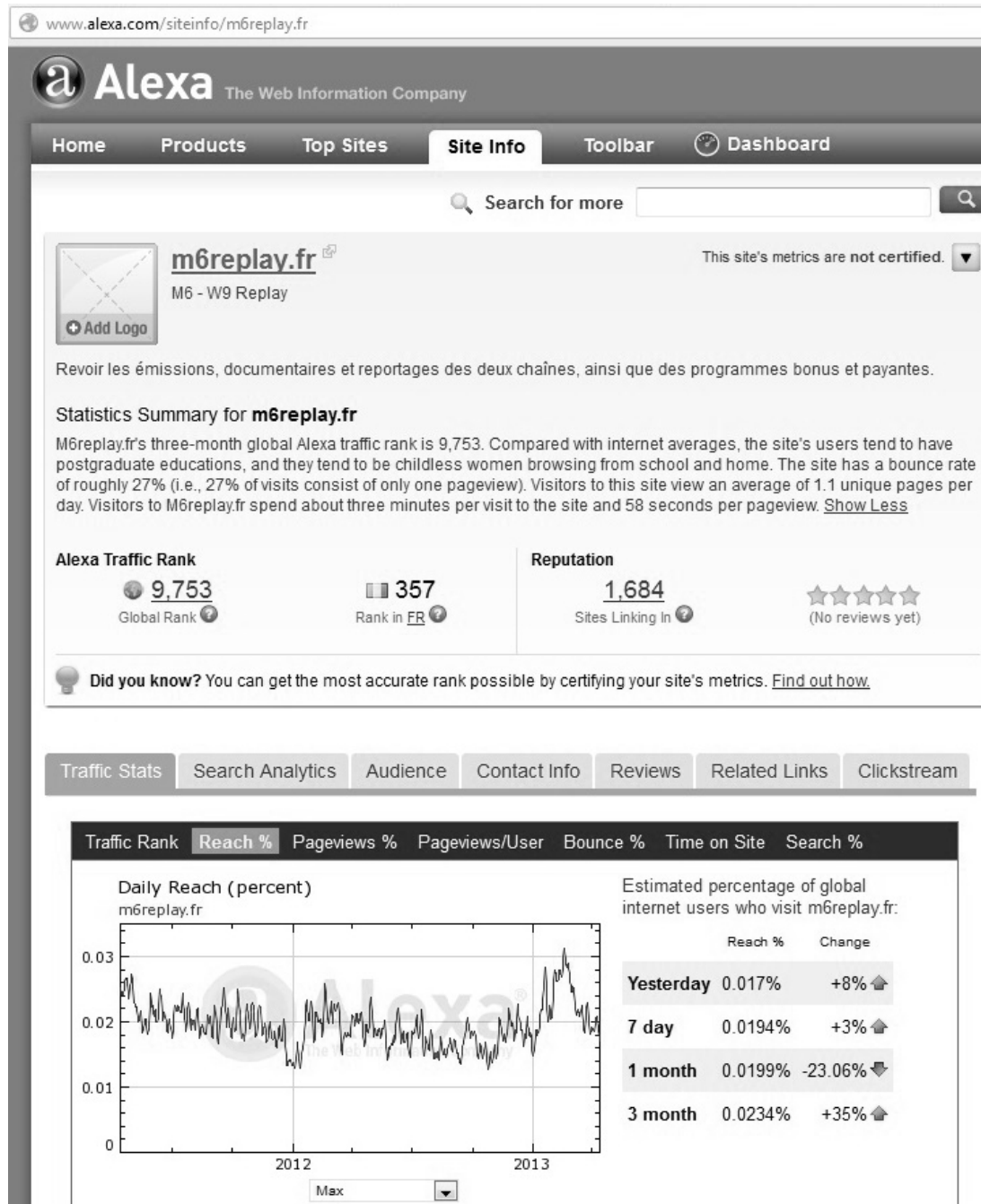
We now consider the replay service associated with the sixth national French channel, M6: it used to be [m6replay.fr](http://m6replay.fr) as in the picture (and is now [www.6play.fr](http://www.6play.fr)). Its visit rate used to be of 0.021% in the years 2010 and 2011. It wonders whether its visit rate changed after the seizure of MegaUpload.

1. Start extracting the relevant statistical information. State the hypotheses to be tested.

A graphical display of the visit rates measured in 2012 is provided in the picture below. We focus on the first 40 days after the seizure of MegaUpload, that is, we conduct a study between January 20, 2012 and February 29, 2012. For the sake of time, we only read one measurement a day on [www.alexa.com](http://www.alexa.com), but at a different time (picked at random). The obtained data are summarized under the picture.

2. What should be concluded? Which actions should the replay service undertake?

<sup>4</sup>Recent examples in France: special working contract for the first job, called the CPE (“contrat première embauche”) in 2006, which resulted in mass demonstrations of high-school and university students, as shown in the picture; marriage and adoption equality for all couples, in 2012–13, which resulted in mass demonstrations of traditional families; labor code reform in Spring 2016, which was opposed to by unions and anarchists and against which much smaller but more violent demonstrations and actions took place.



Sample	n = 40 measurements
Mean	0.023%
Standard deviation	0.005 %

**Advanced exercise 5.3 (Public health surveillance).**

This exercise is based on true stories. The prevalence rate of pediatric asthma is about 9.7%, as measured by a thorough, national, extensive survey. In this survey, children having experienced at least one asthma attack in the past six months had been classified as asthmatics. The same methodology is used in Parisian elementary schools by parents' associations. The parents indeed suspect that the Parisian air quality is low and that it has a detrimental effect on their children's health. But public health surveillance is typically too global and too national in France; it is difficult to access precise figures for local prevalence rates of the diseases. You need the help of the press. But to alert the press you need figures. What a chicken-egg problem!

1. Start extracting the relevant statistical information. State the hypotheses to be tested.

Therefore, a parents' association collects some preliminary figures. It selects at random 5 different Parisian schools, and in these schools, further sub-selects 2 classes, each of which with about 30 pupils. All parents of these selected pupils are literally begged to fill in the asthma questionnaires. All in all, the association gets exactly 300 responses: out of the 300 pupils, 37 report an asthma attack in the past six months.

2. What can be concluded from these data? What is the next step for the parents' association?
3. Would the situation have been different with 38 asthmatic children instead of 37?

**Other advanced exercises:** The next pages feature exercises extracted from past quiz statements

## Exercise 1 — “We look like our names” — 4 points / 9 minutes

This exercise is based on the article “We look like our names: The manifestation of name stereotypes in facial appearance” (co-authored by an HEC Paris professor of marketing, Anne-Laure Sellier).

Question was whether people guess the name of a person based on her/his face, and actually, whether they do so better or worse than at random. If so, it would mean that we think that some faces look rather like this or that name (hopefully but not necessarily, the true name), rather than some other one.

A typical experiment performed is reproduced on the right. We denote by  $p_0$  the proportion of people in the same country (here, Israel) that would correctly guess the name based on the face. Guessing at random would result in a correct answer rate of  $p_{\text{ref}} = 25\%$ .

Try to determine, from among the offered list of names, which is the true given name of the person in the picture.



1. Jacob
2. Dan
3. Josef
4. Nathaniel

- ☐ State your hypotheses, in words and in equations.  
Briefly explain why you picked these hypotheses, in one sentence.

When the experiment was performed on 67 volunteers, 26 of them, that is,  $26/67 \approx 38.8\%$ , found out the correct name, Dan. Work out the test of your hypotheses, by drawing a picture summarizing

- ☐ the expected behaviors of your test statistic under  $H_0$  and  $H_1$ ;  
☐ the numerical value of your test statistic on the data and the associated P-value.

- ☐ Write a statistical conclusion (only; no business conclusion required). Beware, it must be most informative and formulated in plain words (do not use the words “reject” or “ $H_0$ ”).

## Exercise 2 — A controversial governmental reform (10 points)

Assume that you are the leader of a government that wants to put in place a highly controversial reform (e.g., on pensions) and wonders whether there will be massive actions against the reform. Sociologists have it that unless a fraction  $p_{\text{ref}} = 30\%$  of the population is strongly against the reform, not much will happen; and otherwise, some massive actions (massive strikes or demonstrations) may take place. The question is of course whether the fraction  $p_0$  of the population strongly against the reform under review is larger or smaller than 30%.



We will first consider two pairs of hypotheses and test each of these pairs; only then we will indicate which pair a given government should choose.

After figuring out its hypotheses, the government mandates a polling organization, which conducts a survey over 1,000 adults living in France. Among them, 979 express an opinion: 275 are strongly against the reform under review, while the 704 other ones are not (they have no strong opinion or are even indifferent).

### First case — Testing $H_0 : p_0 \geq 30\%$ against $H_1 : p_0 < 30\%$

Work out the test of the hypotheses  $H_0 : p_0 \geq 30\%$  against  $H_1 : p_0 < 30\%$

- ☐ by drawing a picture summarizing the expected behaviors of your test statistic under  $H_0$  and  $H_1$ ,
- ☐ by computing the numerical value of your test statistic (please spell out the calculation that you typed),
- ☐ by providing the associated P-value.

Write a conclusion consistent with the hypotheses and the P-value obtained, by picking the beginning and the middle of the sentence:

- |             |   |
|-------------|---|
|             | A. The data collected cannot exclude that |
| [Beginning] | B. The data collected suggest that        |
|             | C. The data collected show that           |

☐

- |          |                  |
|----------|------------------|
|          | 1. more than 30% |
| [Middle] | 2. less than 30% |

of the population is strongly against the reform under review.

**Second case — Testing  $H_0 : p_0 \leq 30\%$  against  $H_1 : p_0 > 30\%$** 

Same questions based on the hypotheses  $H_0 : p_0 \leq 30\%$  against  $H_1 : p_0 > 30\%$ .

- ☐ Draw a picture summarizing the expected behaviors of your test statistic under  $H_0$  and  $H_1$ ,
- ☐ Provide the P-value associated with the data collected.

Write a conclusion consistent with the hypotheses and the P-value obtained, by using the same coding as above:

- ☐ Letter: \_ Number: \_

**Picking the hypotheses**

A government can be ideological (it would try to implement its reforms by all means) or cautious (risk-averse). Which pair of hypotheses would be chosen by which profile? Circle the correct profile in each sentence:

- ☐  $H_0 : p_0 \geq 30\%$  against  $H_1 : p_0 < 30\%$  is for cautious / ideological governments
- $H_0 : p_0 \leq 30\%$  against  $H_1 : p_0 > 30\%$  is for cautious / ideological governments
- ☐ Provide a brief justification for your choices.

**SPSS output**

Consider the following fake SPSS output (assuming SPSS can run the kind of tests computed above, which surprisingly, it cannot in its default configuration).

- ☐ What number should be written in the empty cell, titled **Sig. (2-tailed)**?

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
Strongly against	979	,28	,450	,014

One-Sample Test					
Test Value = 0.3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference
					Lower Upper
Strongly against	-1,304	978		-,019	-,05 ,01

## Exercise 1 — Walking many steps a day — 10 points

There is a long story behind the trendy 10,000-steps-a-day recommendation issued in the recent years by fitness websites and magazines to experience health benefits. This story has strong links with the creation of pedometers: devices recording the number of steps taken. Nowadays, your smartphone can act as a pedometer via a suitable application.

Suppose that we want to offer a new such application; its distinguishing point would be that not only it would report the numbers of steps made so far but it would also be able to indicate by a green / orange / red color code whether the 10,000-steps-a-day target is reached or not. More precisely, assuming that the pace observed so far is maintained, it would be able to tell whether we are confident that the aim would be reached in the long term, with three possible outcomes:

- we are certain that it will be reached;
- we are certain that it will not be reached;
- we do not know yet.



### Design of the underlying test

- ☐ Indicate the parameter of interest  $\mu_0$  out of the four following statements:
  - 1A. the individual daily numbers of steps made so far
  - 1B. the average daily number of steps made so far
  - 1C. the individual daily numbers of steps (made so far and) to be made in the upcoming months
  - 1D. the average daily number of steps (made so far and) to be made in the upcoming months
- ☐ What pair of hypotheses should we consider based on our aim for a color code?
  - 2A.  $H_0 : \mu_0 \geq 10,000$  vs.  $H_1 : \mu_0 < 10,000$
  - 2B.  $H_0 : \mu_0 \neq 10,000$  vs.  $H_1 : \mu_0 = 10,000$
  - 2C.  $H_0 : \mu_0 = 10,000$  vs.  $H_1 : \mu_0 \neq 10,000$
- ☐ Provide a brief justification of your choice, based on our aim for a color code.

### First data set

A first user monitors his numbers of steps for 49 days and obtains a sample average number of steps equal to 10,532 steps, with a standard deviation in these data points of 3,154 steps. Work out the test of the hypotheses

- ☐ by drawing a picture summarizing the expected behaviors of your test statistic under  $H_0$  and  $H_1$ ,
- by computing the numerical value of your test statistic (please spell out the calculation that you typed),
- by providing the associated P-value.



Based on the same data set, SPSS provides the following output.

One-Sample Statistics						
	N	Mean	Std. Deviation	Std. Error Mean		
Number of steps	49	10532	3154	450,578401		

One-Sample Test						
Test Value = 10000						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Number of steps	1,181	48	,244	532	-373,95605	1437,94114

☐ Explain where to check your two numerical answers above and provide additional comments if needed.

☐ All in all, which color code should this user see?

### Second data set

We consider a second user: she monitored her numbers of steps for 115 days and obtained a sample average number of steps equal to 10,452 steps, with a standard deviation in these data points of 2,356 steps.

☐ ☐ Provide the P-value associated with this data set, as well as the color code that the user should see. (Indicate some of your intermediary calculations.)

### Third data set

A third user obtains a sample average number of steps equal to 9,759 steps, during 62 days.

☐ Based solely on this information, do we already know the color code? How many colors are ruled out already? Explain.



## Two-sample tests (Comparing two populations)

We will learn how to compare two series of data, corresponding to two population parameters  $p_1$  and  $p_2$  (for proportions) or  $\mu_1$  and  $\mu_2$  (for general means). A distinction has to be made between paired and independent data.

Paired data correspond to two measurements made on the same units coming from a single population. Independent data correspond to a single measurement on units coming from two different populations.

The tables below illustrate visually this distinction.

ID	Variable 1	Variable 2		ID	Group	Variable
1	16	12		1	0	12
2	18	16		2	1	16
3	30	29		3	0	29
4	29	32	vs.	4	0	32
5	18	17		5	1	17
6	22	25		6	1	25
7	46	42		7	0	42
8	32	38		8	0	38
9	33	34		9	1	34
...	...	...		...	...	...

Table 6.1: Left table: paired data (two measurements on each sample unit). Right table: independent data (two sub-samples corresponding to two distinct sub-populations, with a single measurement).

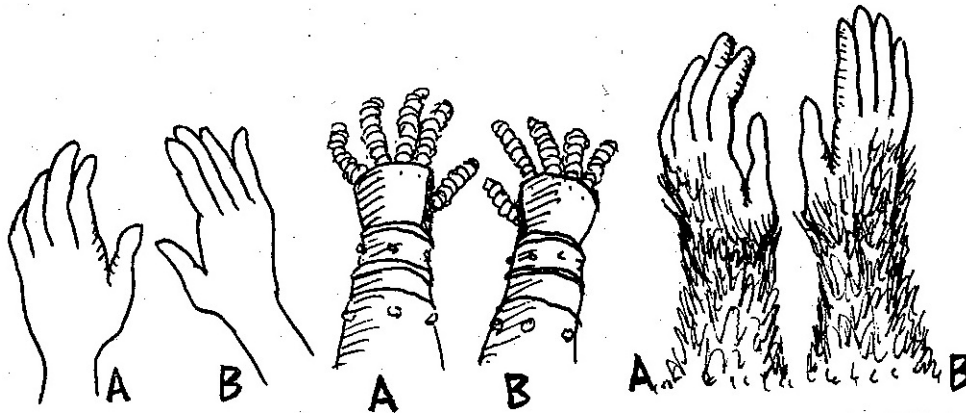
## 1. Paired data / for general means

What follows only covers general quantitative data; it does not apply to proportions.

The same sample units are considered in this case and measurements are made as follows:

- of the same quantity but at two time points; e.g., salary in 2010 and in 2016;
- of the same quantity but in two different sets of conditions; e.g., amount ordered without and with a discount;
- of the same quantity but on each element of a matched pair; e.g., heights of fathers and sons (as Galton did, we will study this historical example in a subsequent chapter).

An example could be testing whether a given moisturizing hand cream is effective; since the hand-hydration levels vary by people (see figure below), the fairest way to determine the cream's effect is to use it on one hand and compare the obtained hydration with the other, control, hand.



In the case of paired data the two series of data are not independent: they tend to take similar values. What we do is

- compute the differences between the two values reported for each sample unit;
- perform a significance test with respect to the reference value 0 based on these data.


Formally, we denote by  $y_1, \dots, y_n$  and  $z_1, \dots, z_n$  the two data series, compute the series of their differences  $x_j = y_j - z_j$ , and perform a one-sample test (see previous chapter) to test  $H_0 : \mu_0 = 0$  against an alternative hypothesis.

The  $\mu_0$  parameter corresponds to the average difference between the two quantities of interest over all members of the population (it is of course unknown).

**Example: prices of books.** The data set reproduced on the next page compares the prices of textbooks mandatorily required<sup>1</sup> in some randomly chosen UCLA classes: the price at the UCLA bookstore (variable: `uclaNew`) and the one on `amazon.com` (variable: `amazNew`). The difference between the two prices (`diff = uclaNew - amazNew`) is computed. A more thorough presentation of the data is given on the next page as well.

Suppose that the question was the following: are prices comparable or is `amazon.com` cheaper? You can see that the question is not “are prices comparable or is one of the sellers cheaper than the other?”

<sup>1</sup>In case several textbooks are required, only the most expensive one was considered.



	ID	deptAbbr	course	ibsn	more	uclaNew	amazNew	diff
1	1	Am-Ind	C170	978-0803272620	Y	27,67	27,95	-,28
2	2	Anthro	9	978-0030119194	Y	40,59	31,14	9,45
3	3	Anthro	135T	978-0300080643	Y	31,68	32,00	-,32
4	4	Anthro	191HB	978-0226206813	Y	16,00	11,52	4,48
5	5	Art-His	M102K	978-0892365999	Y	18,95	14,21	4,74
6	6	Art-His	118E	978-0394723693	Y	14,95	10,17	4,78
7	7	Asia-Am	187B	978-0822338437	Y	24,70	20,06	4,64
8	8	Asia-Am	191E	978-0816646135	N	19,50	16,66	2,84
9	9	Ch-Engr	C125	978-0195123401	N	123,84	106,25	17,59
10	10	Chicano	M145B	978-0896086265	Y	17,00	13,26	3,74
11	11	Chin	174	978-0791420621	Y	31,63	29,95	1,68
12	12	Com-Sci	180	978-0321295354	N	116,00	88,09	27,91
13	13	Comm-Lit	290	978-0393329254	Y	27,67	18,45	9,22
14	14	Comm-St	10	978-0195181234	Y	24,70	16,47	8,23
15	15	Comptng	20A	978-0470509487	N	126,67	97,38	29,29
16	16	E&S-Sci	7	978-0521711128	N	53,90	47,51	6,39
17	17	E&S-Sci	101	978-0393927634	N	89,73	74,93	14,80
18	18	Econ	101	978-0324421620	N	171,00	132,77	38,23
19	19	El-Engr	2	978-0131497269	N	152,00	121,50	30,50
20	20	El-Engr	176	978-0471287704	N	124,80	108,00	16,80
21	21	Engl	M104C	978-0582276024	Y	16,00	11,67	4,33
22	22	Engl	151	978-0199561339	Y	25,95	18,94	7,01
23	23	Engl	180	978-0374515362	Y	18,00	12,24	5,76
24	24	Engl	M260A	978-0804747295	Y	21,73	21,95	-,22
25	25	Engr	183EW	978-0132306416	N	40,59	31,14	9,45
26	26	ESL	38A	978-1577665304	Y	28,95	28,95	,00
27	27	Frnch	120	978-2020525718	Y	19,95	16,15	3,80
28	28	Geog	232	978-1405102667	Y	49,45	37,75	11,70
29	29	German	57	978-0205668946	N	41,09	31,78	9,31
30	30	Greek	130	978-1598561692	Y	50,95	32,97	17,98

textbooks

*Textbook data for UCLA Bookstore and Amazon***Description**

A random sample was taken of nearly 10% of UCLA courses. The most expensive textbook for each course was identified, and its new price at the UCLA Bookstore and on Amazon.com were recorded.

**Usage**

```
data(textbooks)
```

**Format**

A data frame with 73 observations on the following 7 variables.

deptAbbr Course department (abbreviated).

course Course number.

ibsn Book ISBN.

uclaNew New price at the UCLA Bookstore.

amazNew New price on Amazon.com.

more Whether additional books were required for the course (Y means "yes, additional books were required").

diff The UCLA Bookstore price minus the Amazon.com price for each book.

**Details**

The sample represents only courses where textbooks were listed online through UCLA Bookstore's website. The most expensive textbook was selected based on the UCLA Bookstore price, which may insert bias into the data; for this reason, it may be beneficial to analyze only the data where more is "N".

**Source**

This data was collected by David M Diez on April 24th.

(which would lead to a two-sided test) but that the question is one-sided. There seems to be some intuition or some folklore knowledge that amazon.com could be cheaper, but no one would believe that the UCLA bookstore is cheaper than amazon.com.

We denote by  $\mu_0$  the average difference in prices (computed as UCLA price minus Amazon price) of the mandatory textbooks over all UCLA courses. We want to test  $H_0 : \mu_0 = 0$  (comparable prices) versus  $H_1 : \mu_0 > 0$  (Amazon is cheaper, or, put differently, the UCLA bookstore is more expensive).

We summarize data below, with the help of SPSS.

Descriptive Statistics

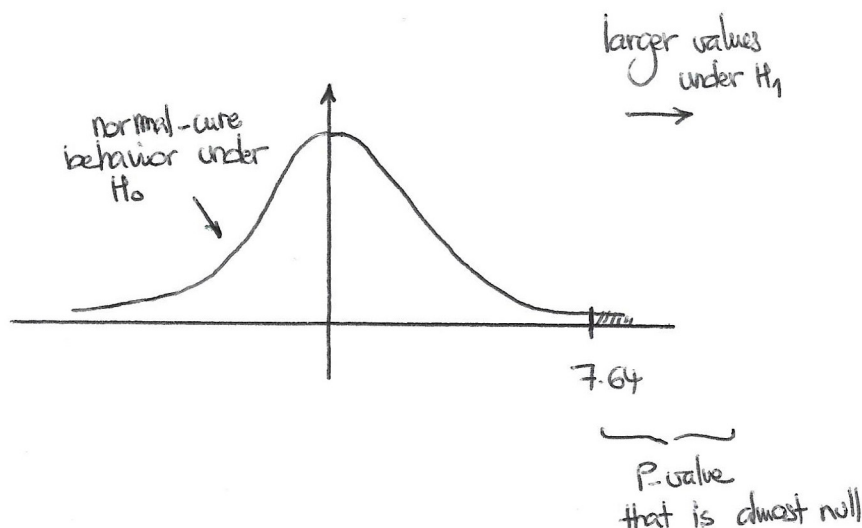
	N	Minimum	Maximum	Mean	Std. Deviation
uclaNew	73	10,50	214,50	72,2219	59,65913
amazNew	73	8,60	176,00	59,4603	48,99557
diff	73	-9,53	66,00	12,7616	14,25530
Valid N (listwise)	73				

Formally, we have 73 data elements  $x_1, \dots, x_{73}$ , where  $x_j$  denotes the difference in prices for the  $j$ -th book. (Only the first 30 of them were shown on the data screenshot.) The sample average and standard deviation equal  $\bar{x}_{73} \approx 12.76$  and  $s_{x,73} = 14.26$ , respectively. The question is to determine whether the sample average  $\bar{x}_{73} = 12.76$  is significantly larger than 0.

The test statistic equals

$$t_{73} = \sqrt{73} \left( \frac{\bar{x}_{73} - 0}{s_{x,73}} \right) = \sqrt{73} \left( \frac{12.76 - 0}{14.26} \right) \approx 7.64.$$

Under  $H_0$  we expected a normal-curve behavior that is, values rather centered around 0. Under  $H_1$ , larger values were expected to be taken: we reject  $H_0$  above a given threshold  $r$ .



All in all, we draw a graph and read a P-value that is very, very small. In any case, we strongly reject  $H_0$  and conclude

- on the statistical side, that there is a significant difference in prices in favor of Amazon (which is cheaper);

- on the business side, that we will never buy our books at the UCLA bookstore, our tuition fees are already high enough, we do not want to pay even more to UCLA!

The values above (the value of the  $t_{73}$  test statistics and the almost null P-value) can be found again in the following SPSS output.

#### One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
diff	73	12,7616	14,25530	1,66846

#### One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
diff	7,649	72	,000	12,76164	9,4356	16,0877

Note that SPSS performs a two-sided test. In this case, the P-value indicated by SPSS should be divided by two to correspond to our one-sided test. Of course, here, all these P-values (be the test one-sided or two-sided) are almost null...

## 2. Independent data / for proportions

We now turn to independent data, that is, data collected from two different populations, by drawing separately a sample from each of the populations. We deal first with the case of proportions.

The proportions  $p_1$  and  $p_2$  of a certain event or feature in the two populations are to be compared. The null hypothesis is that they are equal,  $H_0 : p_1 = p_2$  and alternative hypotheses are of the form

$$H_1 : p_1 \neq p_2, \quad H_1 : p_1 < p_2, \quad H_1 : p_1 > p_2.$$

Of course, a given pair of hypotheses  $H_0 : p_1 = p_2$  versus  $H_1 : p_1 > p_2$  can appear as the limit case of another pair of hypotheses, namely in this case,  $H_0 : p_1 \leq p_2$  versus  $H_1 : p_1 > p_2$ .

We collect data  $x_1, \dots, x_n$  on a sample extracted from the first population, and data  $y_1, \dots, y_m$  on a sample from the second population. These data elements take values 0 and 1. The sample proportions of the 1 are denoted by  $\bar{x}_n$  and  $\bar{y}_m$ , and we are interested in comparing the population proportions of the 1, which we denoted by  $p_1$  and  $p_2$ .

It is natural to compare  $\bar{x}_n$  and  $\bar{y}_m$ . Our test statistic will actually be proportional to  $\bar{x}_n - \bar{y}_m$ . But we need to normalize this quantity: by the sample size and by some standard deviation.

In the case of one-sample tests, this standard deviation was given by  $\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})}$ , but there, the reference value  $p_{\text{ref}}$  was known under  $H_0$ . This is not the case here. Assuming that  $H_0$  is true and thus that  $p_1 = p_2$ , we perform a pooled estimation of the proportions  $p_1 = p_2$  by mixing the two data sets. We get the estimate

$$\bar{x}\bar{y}_{n+m} = \frac{x_1 + \dots + x_n + y_1 + \dots + y_m}{n + m} = \frac{n \bar{x}_n + m \bar{y}_m}{n + m}.$$

As far as the normalization by the sample size is concerned, we replace the  $\sqrt{n} = 1/\sqrt{1/n}$  factor for one-sample tests by a

$$\frac{\sqrt{1}}{\sqrt{1/n + 1/m}}$$

factor. The test statistic equals

$$t_{n,m} = \frac{1}{\sqrt{1/n + 1/m}} \left( \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\bar{x}\bar{y}_{n+m}(1 - \bar{x}\bar{y}_{n+m})}} \right)$$

and we get the test principle stated at the top of the next page.

**Example: small gifts to encourage impulse purchases.**

In France, most of the supermarkets organize wine fairs in the Fall for a given period of approximatively two weeks. Consider a supermarket chain, and two different stores of this chain under the same manager. In the first morning of the first day of the wine fair, she does the following. In supermarket A, she offers a (fancy-looking but cheap!) bottle opener to all customers spending more than 100 euros on wine, while in supermarket B, there is no offer (no small gift, no discount, nothing). Then, at noon, she reviews the results and has to decide whether she should generalize the bottle-opener gift to both places or whether the gift has no impact. The figures are the following: in supermarket A, out of 130 customers who bought wine, 26 spent more than 100 euros; while in supermarket B, 15 out of 96 did so. An excerpt of the corresponding data is provided on the next page.



**Test 6.1. Two-sample test for proportions***(Testing the equality of two population proportions  $p_1$  and  $p_2$ )***Data:**  $x_1, \dots, x_n$  and  $y_1, \dots, y_m$  with values either 0 or 1, and where  $n, m \geq 30$ **Parameter(s) of interest:** population proportions  $p_1$  and  $p_2$ **Hypothesis  $H_0$ :**  $p_1 = p_2$ **Test statistic:**

$$t_{n,m} = \frac{1}{\sqrt{1/n + 1/m}} \left( \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\bar{x}_{n+m}(1 - \bar{x}_{n+m})}} \right) \quad \text{where} \quad \bar{x}_{n+m} = \frac{n\bar{x}_n + m\bar{y}_m}{n + m}$$

**Behavior under  $H_0$ :** normal curve**Behavior under  $H_1$ :**

- if  $H_1$  includes  $p_1 > p_2$ , then  $t_{n,m}$  tends to take larger values;
- if  $H_1$  includes  $p_1 < p_2$ , then  $t_{n,m}$  tends to take smaller values.

What should the manager do?

We first need to decide on our hypotheses (without looking at the data). The prudent viewpoint is to think that [ $H_0$  : ] the gift has no impact (because gifts cost you something) and to only depart from this hypothesis if data show that [ $H_1$  : ] the gifts are indeed effective. The latter statement is the risky but possibly profitable hypothesis.

We denote by  $p_1$  and  $p_2$  the proportions of customers that would/will spend over 100 euros with and without getting a bottle opener. The sample collected in supermarket A corresponds to the  $p_1$  proportion, while the one of supermarket B corresponds to  $p_2$ . The hypotheses are

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 > p_2.$$

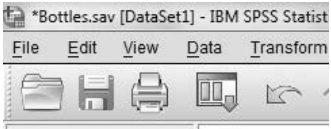
Data collected in the first supermarket are  $x_1, \dots, x_{130}$ , where  $x_j = 1$  if the  $j$ -th customer spent over 100 euros (and got a bottle opener) and  $x_j = 0$  otherwise; a fraction

$$\bar{x}_{130} = \frac{26}{130} = 0.2 = 20\%$$

of these customers indeed spent over 100 euros. In the second supermarket, data consist of  $y_1, \dots, y_96$  with the same convention; a fraction

$$\bar{y}_{96} = \frac{15}{96} \approx 0.156 = 15.6\%$$

of these customers spent over 100 euros. Our question can be rephrased as: is the sample average of  $\bar{x}_{130} = 20\%$  significantly larger than the sample average of  $\bar{y}_{96} = 15.6\%$ ?



	Supermarket	Over100
127	A	N
128	A	N
129	A	Y
130	A	N
131	B	N
132	B	Y
133	B	N
134	B	N

The pooled sample proportion equals

$$\bar{xy}_{130+96} = \frac{26 + 15}{130 + 96} = \frac{41}{226} \approx 0.181 = 18.1\%.$$

The test statistic equals

$$t_{130,96} = \frac{1}{\sqrt{1/130 + 1/96}} \left( \frac{\bar{x}_{130} - \bar{y}_{96}}{\sqrt{\bar{xy}_{130+96}(1 - \bar{xy}_{130+96})}} \right) = \frac{1}{\sqrt{1/130 + 1/96}} \left( \frac{0.2 - 0.156}{\sqrt{0.181(1 - 0.181)}} \right) \approx 0.85.$$

A normal-curve behavior was expected under  $H_0$  while larger values were expected under  $H_1$ . The top figure on the right page computes the P-value associated with our data:  $19.77\% \approx 20\%$ .

This P-value is much larger than 5%, we fail to reject  $H_0$  and we must conclude that these data fail to show that offering a bottle opener has an impact on the amounts of wine purchased (at least when compared to a threshold of 100 euros).

This was the statistical conclusion. On the business side, the manager should not further waste money in bottle openers, and should either think of a more effective way of encouraging purchases (a free wine bottle of the customer's choice?) or stick with the uncreative, current, lack-of-commercial policy of supermarket B.

Now, let us manipulate the data and perform the comparison-of-proportions test with SPSS. We get the following output.

**Supermarket \* Purchasesd over 100 euros Crosstabulation**

Count		Purchasesd over 100 euros		Total
		No	Yes	
Supermarket	A	104	26	130
	B	81	15	96
Total		185	41	226

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,712 <sup>a</sup>	1	,399		
Continuity Correction <sup>b</sup>	,448	1	,503		
Likelihood Ratio	,720	1	,396		
Fisher's Exact Test				,486	,253
N of Valid Cases	226				

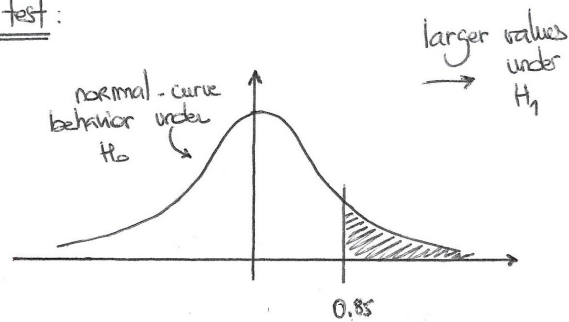
a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 17,42.

b. Computed only for a 2x2 table

You can see that SPSS performs several tests, usually two-sided ones. What we just did basically corresponds to the first test, except that SPSS uses as a test statistic the square of our  $t_{130,96}$  statistic (for reasons that would be too long to explain). It obtains a value of  $0.85 \times 0.85 \approx 0.72$  (well, up to the usual small rounding errors). The P-value indicated is  $39.9\% \approx 40\%$ , the double of what we obtained. This was expected as our test was one-sided and SPSS does it in a two-sided way. See the second graph to visualize how SPSS proceeds.

Note also here that all other tests (which we do not study in this course) fail to reject  $H_0$ .

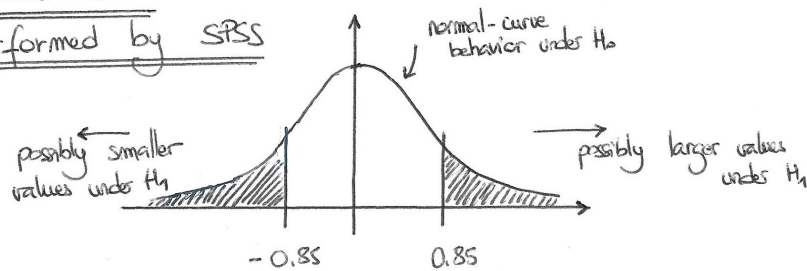
Our one-sided test:



$$P\text{-value} = 100\% - 80.23\% = 19.77\%$$

↑  
as read in the table       $\approx 20\%$

The double-sided test performed by SPSS



P-value,  
first part

P-value,  
second part

$$\approx 2 \times 20\% = 40\% \text{ in total}$$

### 3. Independent data / for general means

We still deal with independent data (i.e., two series of data collected on separate samples extracted from two different populations) but now, with general quantitative data. We are interested in comparing two population means  $\mu_1$  and  $\mu_2$ .

The null hypothesis is still the equality hypothesis,  $H_0 : \mu_1 = \mu_2$ . Alternative hypotheses can be

$$H_1 : \mu_1 \neq \mu_2, \quad H_1 : \mu_1 < \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We collect data  $x_1, \dots, x_n$  on a sample extracted from the first population, and data  $y_1, \dots, y_m$  on a sample from the second population.

To test our hypotheses, it is again natural to compare  $\bar{x}_n$  and  $\bar{y}_m$ : if the gap is large then we will tend to think that  $H_0$  is incorrect. The question is: what is large in this context? Some standard deviations need to be considered. Our test statistics will be of a similar form as before,

$$t_{n,m} = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{s_{x,n}^2/n + s_{y,m}^2/m}},$$

where the quantities  $s_{x,n}$  and  $s_{y,m}$  are standard deviations. They

- are either computed separately with the usual formula on each data series;
- or are computed in a pooled way,  $s_{x,n} = s_{y,m} = s_{xy,n+m}$ .

A first restriction is that in all that follows, the data should be normally distributed—an assumption that we should check first to be fully rigorous (but this initial check is beyond the scope of this course).

To determine how to proceed, we need to test whether the population standard deviations  $\sigma_1$  and  $\sigma_2$  (or variances  $\sigma_1^2$  and  $\sigma_2^2$ ) can be considered equal (an hypothesis  $H'_0$ ) or are different (an hypothesis  $H'_1$ ).

The  $H_0$  behaviors of the  $t_{n,m}$  statistic are (slightly) different depending on whether the variance pre-test rejected or failed to reject  $H'_0$  and whether we resorted to separate or pooled variance estimation. The matter is complicated, as you can see, and we do not want to further dig into these mathematical details.

**What you only need to know:** how to read software outputs when it comes to comparing means. You will not have to perform the calculations on your own! Isn't this fantastic?

See next page for details.

All the statistical softwares first work out the equality-of-variance pre-test (called Levene's test). They report the P-value for the hypotheses  $H_0'$  : of equal variances, versus  $H_1'$  : of non-equal variances. Depending on the P-value read here, the equality-of-mean test is performed in one way or another. What line to read is indicated in the first column.

The following are stylized outputs.

	Equality of variances		Equality of means		
	F	Sig.	t	Sig. (2-tailed)	...
Equal variances	0.404	0.526	− 0.155	0.877	...
Non-equal variances			− 0.154	0.878	...

In this first example, the P-value for equality of variances is 52.6% and we fail to reject the equality hypothesis  $H_0'$ . We thus consider, for the second part of the study, that variances should be considered equal. We should read the first line when it comes to the test of equality of the means. It is always performed in a two-sided way, i.e., with  $H_1 : \mu_1 \neq \mu_2$ . The corresponding P-value equals 87.7%. We fail to reject the hypothesis  $H_0 : \mu_1 = \mu_2$ .

Note: The columns F and t simply report the values taken by the test statistics on which the tests are based. You do not need to take them into account. Just focus on the P-values, denoted by “Sig.” under SPSS (for significance).

Another stylized output is reproduced below.

	Equality of variances		Equality of means		
	F	Sig.	t	Sig. (2-tailed)	...
Equal variances	4.931	0.027	2.034	0.043	...
Non-equal variances			2.009	0.045	...

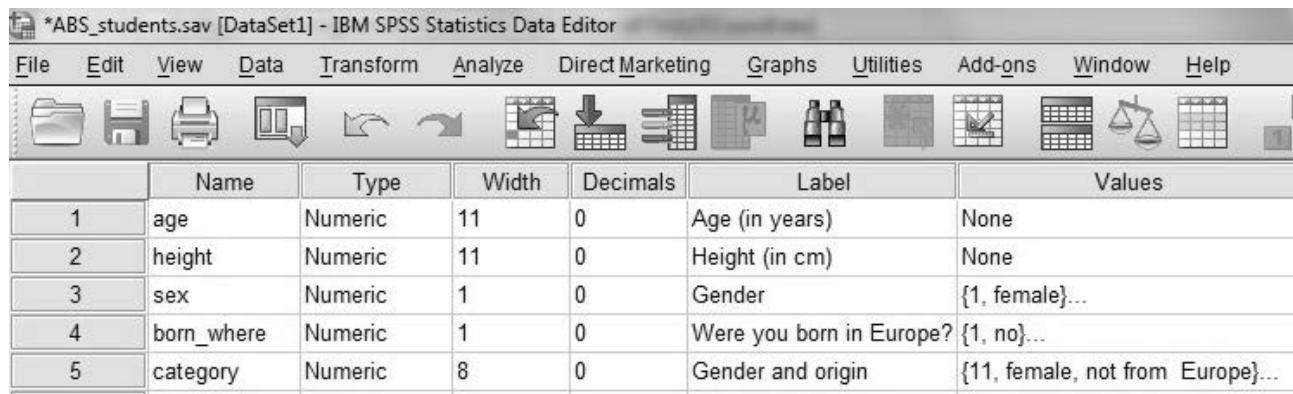
The hypothesis  $H_0'$  of equal variances is associated with a P-value of 2.7% and must be rejected. We thus read the second line when it comes to the test for the equality of the means. The P-value read therein is 4.5%, which is smaller than 5% and which therefore indicates that the hypothesis  $H_0$  of equal populations means should be rejected. Again, the alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  was two-sided here.

A final note: the two P-values for the equality of the means (in the equal variances or non-equal variances contexts) are usually close to each other.

**Example: a study on Arnhem BS students.** The nice data discussed below is extracted from [http://helpdeskspssabs.femplaza.nl/analysis/Independent\\_samples\\_t\\_test.htm](http://helpdeskspssabs.femplaza.nl/analysis/Independent_samples_t_test.htm) and was already hinted at on page 76.

Arnhem Business School professors write: “For a number of years we have asked foundation year students to fill in a simple questionnaire with some questions about who they are.”

The variables collected are presented below, and an excerpt of the data is provided on the next page, together with some descriptive statistics.



	Name	Type	Width	Decimals	Label	Values
1	age	Numeric	11	0	Age (in years)	None
2	height	Numeric	11	0	Height (in cm)	None
3	sex	Numeric	1	0	Gender	{1, female}...
4	born_where	Numeric	1	0	Were you born in Europe?	{1, no}...
5	category	Numeric	8	0	Gender and origin	{11, female, not from Europe}...

The coding for the sex (Gender) variable is: 1 for female, 2 for male. The coding for the born\_where (in Europe?) variable is: 1 if born outside Europe, 2 if undecided, and 3 if born in Europe. The category variable is a mere concatenation of these two variables, which results in values like 11 (a woman born outside Europe) or 23 (a man born in Europe). Why did they create the “undecided” category when it came to the birth place?

**Were you born in Europe? Yes / No**

What could be simpler than a question like this? But have a look at some of the answers:

Europe? ..... (Yes or No) *I don't know, Russia.*

Were you born in Europe? *Ukraine* ..... (Yes or No)

Indeed, Russia is a country that partly lies in Europe and partly in Asia. So neither yes nor no would be appropriate. The problem with the Ukraine is a little more complicated. Geographically it counts as part of Europe, but politically definitely not.

This shows once again that pretesting your questionnaire is a wise thing to do. Of course in our foundation year we keep this question to show our students how even seemingly simple things may go wrong. And how they want to solve this problem. What would you do?

We have created a separate category “undecided” and specified it to be a missing value in SPSS.

What could we test?

- Are male and female Arnhem BS students of the same average height?
- Are male and female Arnhem BS students of the same average age?
- Are Europe-born and outside-of-Europe-born Arnhem BS students of the same average age?

\*ABS\_students.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs

	age	height	sex	born_where	category
1	21	150	1	1	11
2	17	155	1	1	11
3	25	160	1	1	11
4	19	163	1	3	13
5	23	163	1	1	11
6	23	163	1	1	11
7	21	164	1	3	13
8	19	165	1	3	13
9	19	167	1	3	13
10	17	168	1	3	13
11	18	168	1	3	13
12	21	168	1	3	13
13	21	169	2	3	23
14	25	169	1	1	11
15	19	170	1	3	13
16	21	170	2	3	23
17	19	172	1	3	13
18	20	174	2	2	.
19	19	175	2	3	23
20	21	175	2	1	21
21	19	176	2	1	21

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	224	46,7	46,7	46,7
	male	256	53,3	53,3	100,0
	Total	480	100,0	100,0	

Were you born in Europe?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	217	45,2	45,4	45,4
	yes	261	54,4	54,6	100,0
	Total	478	99,6	100,0	
Missing	undecided	2	,4		
Total		480	100,0		

Gender and origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female, not from Europe	116	24,2	24,3	24,3
	female, from Europe	107	22,3	22,4	46,7
	male, not from Europe	101	21,0	21,1	67,8
	male, from Europe	154	32,1	32,2	100,0
	Total	478	99,6	100,0	
Missing	System	2	,4		
Total		480	100,0		

Thanks to the following outputs, we may answer the three questions.

### T-Test #1

Group Statistics					
	Gender	N	Mean	Std. Deviation	Std. Error Mean
Height (in cm)	female	223	166,03	6,700	,449
	male	253	181,24	8,108	,510

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
Height (in cm)	Equal variances assumed	4,931	,027	-22,131	474	,000	-15,210	,687	-16,560 -13,859
	Equal variances not assumed			-22,396	472,080	,000	-15,210	,679	-16,544 -13,875

### T-Test #2

Group Statistics					
	Gender	N	Mean	Std. Deviation	Std. Error Mean
Age (in years)	female	224	19,83	2,063	,138
	male	256	19,86	1,965	,123

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
Age (in years)	Equal variances assumed	,404	,526	-,155	478	,877	-,028	,184	-,390 ,333
	Equal variances not assumed			-,154	462,613	,878	-,028	,185	-,391 ,334

### T-Test #3

Group Statistics					
	Were you born in Europe?	N	Mean	Std. Deviation	Std. Error Mean
Age (in years)	no	217	20,06	2,143	,145
	yes	261	19,69	1,877	,116

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
Age (in years)	Equal variances assumed	3,868	,050	2,034	476	,043	,374	,184	,013 ,736
	Equal variances not assumed			2,009	433,052	,045	,374	,186	,008 ,740

The first test answers the question: “Are male and female Arnhem BS students of the same average height?” The pre-test indicates that the variances in heights are different between men and women, we thus read the second line when it comes to the test for equality of means. The P-value is almost null: we strongly reject the hypothesis of equal average heights. Now, looking at the sample means, we see which group is taller on average: men.

The second test answers the question: “Are male and female Arnhem BS students of the same average age?” The pre-test indicates that the variances in ages are not significantly different between men and women, we thus read the first line when it comes to the test for equality of means. The P-value read therein equals 87.7%: we fail to reject the hypothesis of equal average ages.



The third test answers the question: “Are Europe-born and outside-of-Europe-born Arnhem BS students of the same average age?” The pre-test gives a 5% P-value to the hypothesis that the variances in ages are equal between Europe-born and outside-of-Europe-born students. This is a borderline value! We thus read any line (first or second) of our choice when it comes to the test for equality of means. Anyway, the P-values reported are almost identical and slightly below 5% (namely, 4.3% and 4.5%). We reject the hypothesis of equal average ages. Now, looking at the sample means, we see which group is older on average: the students born out of Europe. (Which, by the way, is easy to explain, no?)

**Statistical significance versus effect size.** So, we proved that students born out of Europe were older than students born in Europe—the difference between the average ages is statistically significant. We exhibited an effect. But how large is this average difference in ages, what is the effect size? Not so large: the confidence interval for this parameter is 0.008 to 0.74 year (let us pick the largest interval), that is, 3 days to 9 months. The minimum guaranteed value for the difference is... 3 days, which is short! In a nutshell, we exhibited an effect but the effect size is small: the difference in average ages does exist but is small.

**Conclusion:** A good practice is to provide P-values smaller than 5% together with a quantification of the effect size.

## 4. Independent data / two mistakes to avoid!

In this section we highlight two common methodological mistakes.

**The layman’s mistake.** One may think that a good way to proceed is to take the first-sample average  $\bar{x}_n$  as a reference value  $\mu_{\text{ref}}$  and apply the one-sample test of  $H_0 : \mu_2 = \bar{x}_n$  based on the data  $y_1, \dots, y_m$ .

The mistake is that doing so, one neglects the randomness associated with the first sample.

Indeed, both samples values  $\bar{x}_n$  and  $\bar{y}_m$  can deviate (slightly) from the population values  $\mu_1$  and  $\mu_2$ , and the incorrect procedure described above neglects the deviations from  $\mu_1$ .

**A more subtle mistake.** It is described in the newspaper article reproduced on the next page (extracted from *The Guardian*).

It basically corresponds to the following situation. A reference value  $\mu_{\text{ref}}$  is fixed. People want to conclude that the population means  $\mu_1$  and  $\mu_2$  are significantly different as soon as

- the test of  $H_0 : \mu_1 = \mu_{\text{ref}}$  (against some one-sided or two-sided  $H_1$ ) based on  $x_1, \dots, x_n$  rejects  $H_0$ ;
- the test of  $H'_0 : \mu_2 = \mu_{\text{ref}}$  (against some one-sided or two-sided  $H'_1$ ) based on  $y_1, \dots, y_m$  fails to reject  $H'_0$ .

This is incorrect because failing to reject does not mean that  $H'_0$  is true! Of course, if we could prove that  $H_0 : \mu_1 = \mu_{\text{ref}}$  is not true while  $H'_0 : \mu_2 = \mu_{\text{ref}}$  is true, then there would have been a significant discrepancy.

But we cannot prove that  $H'_0 : \mu_2 = \mu_{\text{ref}}$  is true! We can only say that the data fail to reject it...

# The statistical error that just keeps on coming

Ben Goldacre

The same statistical errors - namely, ignoring the "difference in differences" - are appearing throughout the most prestigious journals in neuroscience

Friday 9 September 2011 20:59 BST

**W**e all like to laugh at quacks when they misuse basic statistics.

But what if academics, en masse, deploy errors that are equally foolish? This week Sander Nieuwenhuis and colleagues publish a mighty torpedo in the journal *Nature Neuroscience*.

They've identified one direct, stark statistical error so widespread it appears in about half of all the published papers surveyed from the academic neuroscience research literature.

To understand the scale of this problem, first we have to understand the error. This is difficult, and it will take 400 words of pain. At the end, you will understand an important aspect of statistics better than half the professional university academics currently publishing in the field of neuroscience.

Let's say you're working on nerve cells, measuring their firing frequency. When you drop a chemical on them, they seem to fire more slowly. You've got some normal mice and some mutant mice. You want to see if their cells are differently affected by the chemical. So you measure the firing rate before and after applying the chemical, first in the mutant mice, then in the normal mice.

When you drop the chemical on the mutant mice nerve cells, their firing rate drops, by 30%, say. With the number of mice you have this difference is statistically significant, and so unlikely to be due to chance. That's a useful finding, which you can maybe publish. When you drop the chemical on the normal mice nerve cells, there is a bit of a drop, but not as much - let's say 15%, which doesn't reach statistical significance.

But here's the catch. You can say there is a statistically significant effect for your chemical reducing the firing rate in the mutant cells. And you can say there is no such statistically significant effect in the normal cells. But you can't say mutant and normal cells respond to the chemical differently: to say that, you would have to do a third statistical test, specifically comparing the "difference in differences", the difference between the chemical-induced change in firing rate for the normal cells against the chemical-induced change in the mutant cells.

Now, looking at the figures I've given you here (for our made up experiment) it's very likely that this "difference in differences" would not be statistically significant, because the responses to the chemical only differ from each other by 15%, and we saw earlier that a drop of 15% on its own wasn't enough to achieve statistical significance.

But in just this situation, academics in neuroscience papers routinely claim to have found a difference in response, in every field imaginable, with all kinds of stimuli and interventions: comparing younger versus older participants; in patients against normal volunteers; between different brain areas; and so on.

How often? Nieuwenhuis looked at 513 papers published in five prestigious neuroscience journals over two years. In half the 157 studies where this error could have been made, it was. They broadened their search to 120 cellular and molecular articles in *Nature Neuroscience*, during 2009 and 2010: they found 25 studies committing this fallacy, and not one single paper analysed differences in effect sizes correctly.

These errors are appearing throughout the most prestigious journals for the field of neuroscience. How can we explain that? Analysing data correctly, to identify a "difference in differences", is a little tricky, so thinking generously, we might suggest that researchers worry it's too longwinded for a paper, or too difficult for readers. Alternatively, less generously, we might decide it's too tricky for the researchers themselves.

But the darkest thought of all is this: analysing a "difference in differences" properly is much less likely to give you a statistically significant result, and so it's much less likely to produce the kind of positive finding you need to look good on your CV, get claps at conferences, and feel good in your belly. Seriously: I hope this is all just incompetence.

This article was amended on 13 September 2011 to make clear that the Nieuwenhuis study looked specifically at neuroscience papers, not psychology research.

## 5. Elementary exercises

**Elementary exercise 6.1.** [Independent data, general means] Left-handed people are said to have shorter reaction times than right-handed people (and this is important in some sports!)—due to different, more efficient brain connections. Now, daily-life objects are designed for the vast majority of the population that is right-handed, so that, even with a shorter reaction time, left-handed people are slowed down by manipulating these objects designed for right-handed people. We perform an experiment on picking up the phone: we measure reaction times (RT) of 10 right-handed (R) and 10 left-handed (L) people, in centiseconds [cs]. Data and their statistical treatment under SPSS are reproduced below. (Note: this is simulated, not real, data.)

RT	L/R	RT	L/R
69.4	R	97.1	L
52.6	R	93.6	L
89.3	R	95.6	L
86.3	R	106.7	L
70.8	R	108.8	L
86.3	R	86.5	L
108.3	R	85.9	L
81.7	R	100.3	L
91.8	R	86.2	L
77.3	R	88.8	L

Group Statistics

Right-handed or left-handed		N	Mean	Std. Deviation	Std. Error Mean
Reaction time (sec.)	Right-handed	10	81,392	15,0696	4,7654
	Left-handed	10	94,958	8,3754	2,6485

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Reaction time (sec.)	Equal variances assumed	1,711	,207	-2,488	18	,023	-13,5659	5,4520	-25,0200	-2,1117
	Equal variances not assumed			-2,488	14,076	,026	-13,5659	5,4520	-25,2533	-1,8785

We denote by  $\mu_0^R$  and  $\mu_0^L$  the population average reaction times of right-handed and left-handed people (these averages are computed over millions of people).

1. Explain why we should test  $H_0 : \mu_0^R = \mu_0^L$  versus  $H_1 : \mu_0^R \neq \mu_0^L$ .
2. Which P-value do we get based on the data above? What is your conclusion?

**Elementary exercise 6.2.** [Independent data, proportions] An online seller is hesitating between two small gifts to boost its sales. He conducts a simultaneous test on two different samples of 300 customers picked independently at random; the customers of each sample are notified that they will get the corresponding small gift if they place an order within the next two weeks. Denote by  $p_0^1$  and  $p_0^2$  the order rates that would be achieved if the first and second small gifts respectively considered were offered to all the thousands of customers.

1. Explain why we should test  $H_0 : p_0^1 = p_0^2$  versus  $H_1 : p_0^1 \neq p_0^2$ .

Two weeks later, the experiment is complete and the online seller realizes that 125 and 143 customers of the first and second samples respectively placed an order.

2. Which P-value do we get based on the data above? What is your conclusion?

**Elementary exercise 6.3.** [Paired data, general means] Consider an online buying club: members have to place an order every trimester (otherwise, they get some product by default, like “the book of the trimester”). We want to assess the effect of a small gift and to that end follow 200 customers for two consecutive trimesters. On the first trimester they get no particular incentive for placing an order of a given amount (they just have to place any order as imposed by being a member of the club), while for the second trimester we offer a small gift whenever the order placed is above 30 euros. We are interested in determining whether the small gift impacts the amounts of the orders placed.

1. Explain why this is paired data. Explain why the parameter of interest studied here consists of the average increase  $\Delta_0$  in orders with the small gift, where the average is computed over all thousands of customers of the club.

The data collected can be summarized as follows:

Amounts ordered	Mean	Standard deviation
First trimester	27.56	10.2
Second trimester	29.14	11.5
Difference	1.58	13.8

Note: the standard deviation of the difference may look large to you; yet, this could be achieved on real data and is only due to the large range of values taken by the differences in amounts.

2. Which P-value do we get based on the data above? What is your conclusion?

**Note:** Pairing data is more efficient from a statistical viewpoint, but this requires here conducting the study during two trimesters, which is probably too long from a business viewpoint. Hence, most companies would go for a shorter study during a single trimester, based on independent data.

**Elementary exercise 6.4.** [Answering students' complaints] Students often complain that it is unfair that the same quiz statement is given to two consecutive groups with the same instructor, as some communication takes place during the break and some students of the second group may get valuable private information. Studies show that, on the contrary, such communication has a detrimental effect on performance—mostly due to the poor quality of feedback communicated and also because it prevents informed students to think about the exercises from scratch. During the Fall 2018 semester we wanted to convince students that this theoretical explanation is indeed taking place. To that end, we computed the following statistics for two consecutive groups with the same instructor.

Group Statistics					
	Group	N	Mean	Std. Deviation	Std. Error Mean
Grade	1pm	40	14,300	3,5641	,5635
	2.40pm	36	14,847	3,1933	,5322

Levene's Test for Equality of Variances				t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Grade	Equal variances assumed	,257	,614	-,702	74	,485	-,5472	,7797	-2,1007	1,0063
	Equal variances not assumed			-,706	73,999	,482	-,5472	,7751	-2,0917	,9973

State the hypotheses tested, indicate the outcome of the test, and conclude: should your instructors write two different quiz statements, or is a single one good enough?

## 6. More advanced exercises (quiz-like exercises)

**Advanced exercise 6.1** (Separate or pooled marketing campaign?). A fitness-machines producer doubts whether to create a single advertisement campaign for men and women, or two separate ones (and then, decide for each magazine, which is the most appropriate given the readers' average gender). Of course, having a single campaign would save money, but you would only consider this option if tastes of men and women are similar enough when it comes to the advertisement of fitness machines. The tentative pictures for the campaign are the following ones:



We will refer to them as the left and the right pictures, respectively. Men and women interested in using fitness machines are interviewed (e.g., when they get out of some randomly selected fitness centers!). We assume that we obtained a representative sample of the population going to gym facilities. The question asked is: “Which picture do you find the most inspiring for your body-shaping efforts?” (It is not only about bodybuilding, you rather target the customers interested in body shaping.) The collected data is summarized in the  $2 \times 2$  table below (called a contingency table).

Favorite picture	Left	Right	Total
Men	75	89	164
Women	51	54	105
Total	126	143	269

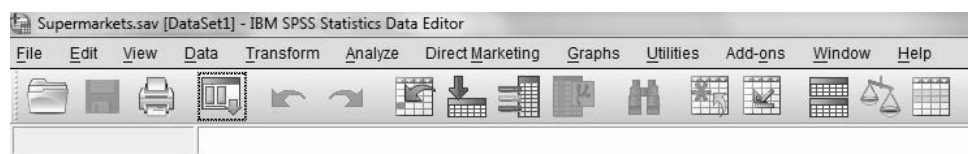
1. What do data say about the respective preferences of men and women? How many advertisement campaigns would you recommend?

Another question is the following; it actually does not use the methods seen in this chapter—try to think about it from scratch.

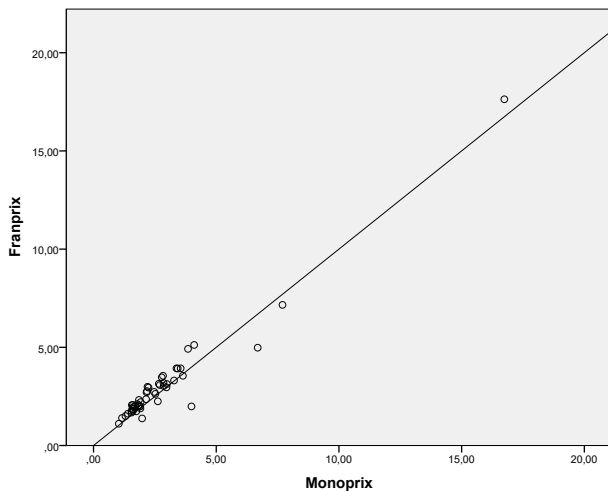
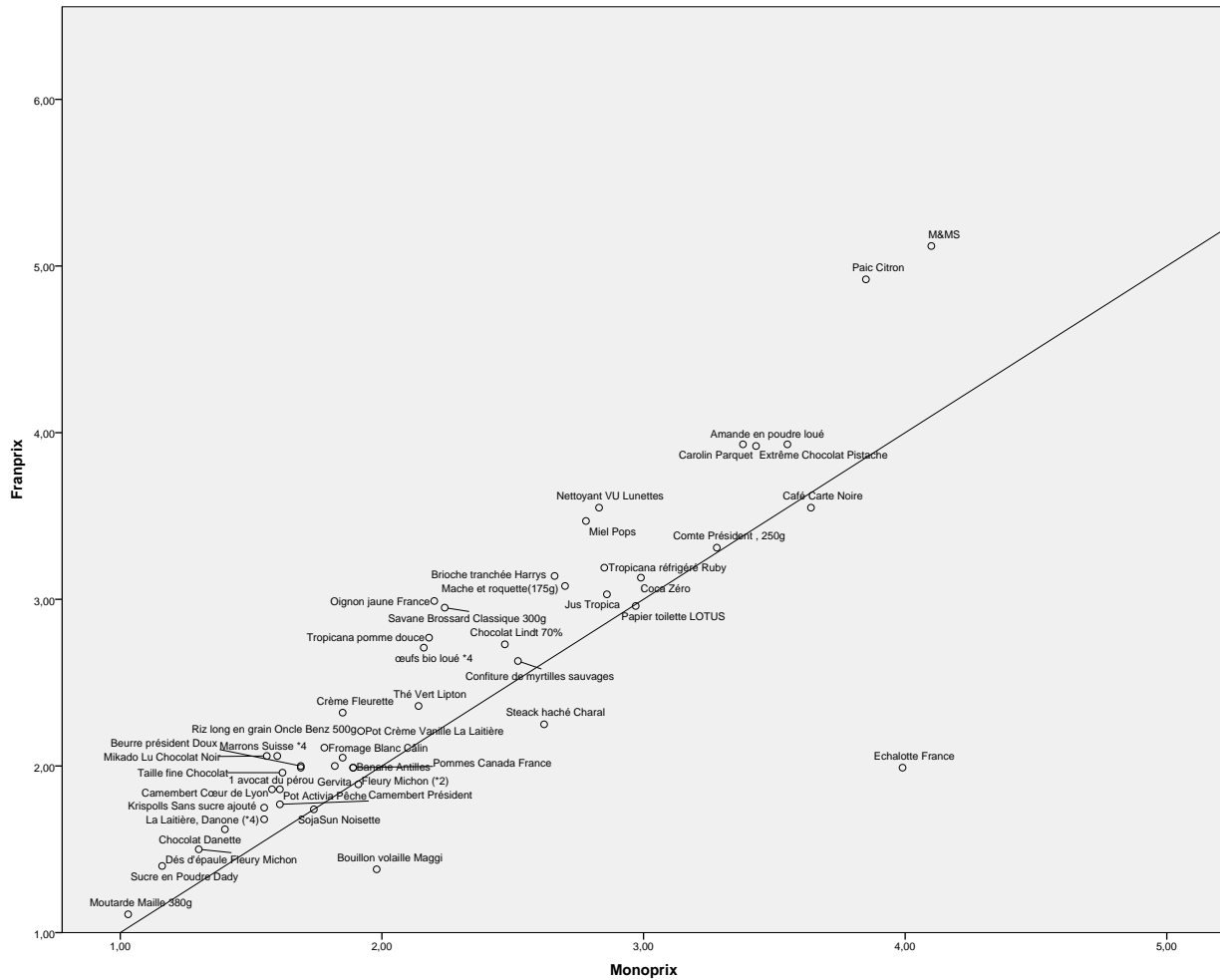
2. Is the subpopulation of people attending gym facilities comparable to the entire population of France, at least from a gender-ratio viewpoint? Put differently, do we observe the same proportions of men attending gym facilities among all men and of women attending gym facilities among all women? Useful piece of information: 51.4% of France's population are women.

**Advanced exercise 6.2** (Comparing prices in two local supermarkets). Benjamin Petiau, a dedicated instructor in our statistics group at HEC (he is teaching in L3 to French students), performed the following data collection in Fall 2015. He considered two local supermarkets of Versailles, where he usually does his grocery shopping: Franprix and Monoprix. He picked 51 everyday consumer products, which are partly listed in the data screenshot below. (Please accept our M1 apologies for not being dedicated enough to translate the names...) For each item, he calculated the arithmetic differences in prices (variable: Difference) as well as the log-ratio (natural logarithm of the ratio) of the prices (variable: LogRatio). Answer the following questions.

1. Which kind of data (one sample, two paired samples, two independent samples) is it?
2. Which is the question that Benjamin Petiau had in mind before collecting data? State the associated pair of hypotheses.
3. Based on the pictures (the second picture is a zoom of the first one), what would your impression be? We will quantify this impression in the questions below.
4. Extract the relevant statistical information for the variable Difference.
5. Perform a suitable test to answer the question Benjamin Petiau had in mind; do first the calculations by yourself and then, compare your results to the SPSS outputs.
6. Provide a conclusion that would be understandable by any Versailles resident, preferably accompanied with a well-chosen figure.
7. Repeat the previous three questions with the LogRatio variable.



	Product	Franprix	Monoprix	Difference	LogRatio
1	Vodka Polialov (1L)	17,63	16,74	,89	,05
2	Chocolat Lindt 70%	2,73	2,47	,26	,10
3	Coca Zéro	3,13	2,99	,14	,05
4	Papier toilette LOTUS	2,96	2,97	-,01	,00
5	Nettoyant VU Lunettes	3,55	2,83	,72	,23
6	Paic Citron	4,92	3,85	1,07	,25
7	Dés d'épaule Fleury Michon	1,50	1,30	,20	,14
8	Pot Crème Vanille La Laitière	2,21	1,92	,29	,14
9	Fromage Blanc Cälin	2,05	1,85	,20	,10
10	Pot Activia Pêche	1,86	1,61	,25	,14
11	Jus Tropica	3,03	2,86	,17	,06
12	Taille fine Chocolat	1,96	1,62	,34	,19
13	SojaSun Noisette	1,74	1,74	,00	,00
14	Gervita	2,00	1,82	,18	,09
15	Chocolat Danette	1,62	1,40	,22	,15
16	Steack haché Charal	2,25	2,62	-,37	-,15
17	Camembert Président	1,77	1,61	,16	,09
18	Crème Fleurette	2,32	1,85	,47	,23
19	Carte d'or Vanille	7,16	7,70	-,54	-,07
20	Fleury Michon (*2)	1,89	1,91	-,02	-,01
21	Krispolls Sans sucre ajouté	1,75	1,55	,20	,12
22	Sucre en Poudre Dady	1,40	1,16	,24	,19
23	Confiture de myrtilles sauvages	2,63	2,52	,11	,04
24	1 avocat du pérou	1,99	1,69	,30	,16
25	Mache et roquette(175g)	3,08	2,70	,38	,13
26	Tropicana réfrigéré Ruby	3,19	2,85	,34	,11
27	Echalotte France	1,99	3,99	-2,00	-,70
28	Oignon jaune France	2,99	2,20	,79	,31
29	Banane Antilles	1,99	1,89	,10	,05



Statistics

		Franprix	Monoprix	Difference	LogRatio
N	Valid	51	51	51	51
	Missing	0	0	0	0
Mean		2,9741	2,7667	,2075	,0901
Median		2,3200	2,1600	,2500	,1214
Std. Deviation		2,37978	2,34611	,53437	,17093
Minimum		1,11	1,03	-2,00	-,70
Maximum		17,63	16,74	1,07	,31

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Difference	2,772	50	,008	,20745	,0572	,3577
LogRatio	3,765	50	,000	,09012	,0420	,1382

**Advanced exercise 6.3** (Gender pay gap, revisited). We revisit the data of the gender-pay-gap exercise of page 43 and use SPSS to conduct a test. The output is reproduced below.

Group Statistics					
Gender		N	Mean	Std. Deviation	Std. Error Mean
Monthly net salary	Men	179	3431,46	3895,437	291,159
	Women	147	2434,90	1282,947	105,816

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
Monthly net salary	Equal variances assumed	8,195	,004	2,972	324	,003	996,560	335,371	Lower: 336,781 Upper: 1656,339
	Equal variances not assumed			3,217	223,375	,001	996,560	309,791	Lower: 386,074 Upper: 1607,047

What is your conclusion? Please report the P-value for a one-sided (not two-sided) test, in which the alternative hypothesis would be the existence of a detrimental effect for women.

#### Advanced exercise 6.4 (Alcohol consumption during the POWs at HEC).

This exercise is based on real data collected by our former colleague Veronika Czellar in Fall 2008. In that good old time, our statistics classes were taking place on Friday mornings, either at 8am or 10am (yes, indeed, at 8am!). We had groups of about 40 students but you will see in the data the low presence rates. We wanted to know more about the average alcohol consumption of the students present in each group (thus discarding the students that did not have the will or the energy to get up for their classes).

- Which hypotheses did we have in mind, according to you?  
(Determine in particular whether the alternative hypothesis is one-sided or two-sided.)

An excerpt of the data collected is reproduced on the right.

- Can you see some outliers / implausible values?
- Which kind of data (one sample, two paired samples, two independent samples) is it?

The next page displays the outputs when performing the adequate test. We first performed the test on the raw data, and then performed it again on a subsample of the data, by eliminating some outliers / implausible values. (Which ones?)

- What are the P-values associated with our hypotheses (beware! which may be one-sided, while SPSS only considers two-sided hypotheses).
- What should we conclude or not conclude?

	Group	NumberGlasses
16	1	6,0
17	1	3,0
18	1	2,0
19	1	2,5
20	1	,0
21	1	3,1
22	1	2,0
23	1	,0
24	2	18,7
25	2	5,0
26	2	8,0
27	2	3,0
28	2	5,5
29	2	3,0
30	2	2,0
31	2	4,0
32	2	7,0
33	2	2,0
34	2	5,0
35	2	5,0
36	2	1,0
37	2	3,0
38	2	10,0



**T-Test**

[DataSet1] D:\Personal\Enseignement\Cours HEC M1\Material\Tests3\SPSS\Alcohol\Alcohol-HEC.sav

**Group Statistics**

	Group	N	Mean	Std. Deviation	Std. Error Mean
Number of glasses	8am	23	4,448	3,0598	,6380
	10am	31	7,735	10,1658	1,8258

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Number of glasses	Equal variances assumed	5,005	,030	-1,498	52	,140	-3,2877	2,1944	-7,6911	1,1158
	Equal variances not assumed			-1,700	37,021	,098	-3,2877	1,9341	-7,2064	,6311

```
USE ALL.
COMPUTE filter_$=(NumberGlasses <= 20).
VARIABLE LABELS filter_$ 'NumberGlasses <= 20 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
T-TEST GROUPS=Group(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=NumberGlasses
  /CRITERIA=CI(.95).
```

**T-Test****Group Statistics**

	Group	N	Mean	Std. Deviation	Std. Error Mean
Number of glasses	8am	23	4,448	3,0598	,6380
	10am	29	5,441	4,3697	,8114

**Independent Samples Test**

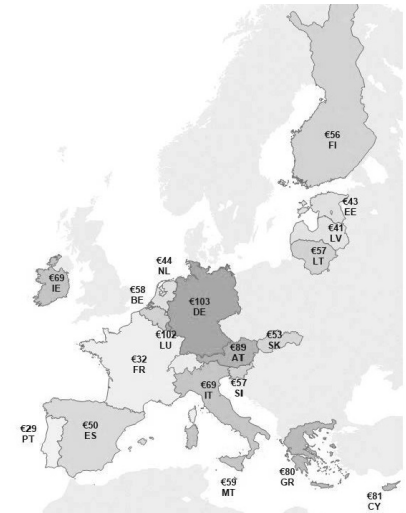
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Number of glasses	Equal variances assumed	2,268	,138	-,925	50	,360	-,9936	1,0746	-3,1520	1,1649
	Equal variances not assumed			-,963	49,328	,340	-,9936	1,0322	-3,0676	1,0804

**Other advanced exercises:** The next pages feature exercises extracted from past quiz statements

## Exercise 2 — Cash in the wallet, by country — 4 points

A 2017 study by researchers of the European Central Bank, Henk Esselink and Lola Hernández, titled *The use of cash by households in the euro area*, provided the following picture, where the value written on each country is the average amount of cash in the wallet reported by interviewees of the sample.

Since I did not get access to the original data, I invented some that is compatible with this picture. Let's focus on Germany and Austria.



Group Statistics

	Country	N	Mean	Std. Deviation	Std. Error Mean
Cash in wallet	Austria	857	89,13395283	19,95491575	,681646937
	Germany	1026	103,0778278	30,33950275	,947184927

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Cash in wallet	Equal variances assumed	136,018	,000	-11,531	1881	,000	-13,9438750	1,209231121	-16,3154505	-11,5722996
	Equal variances not assumed			-11,949	1787,518	,000	-13,9438750	1,166962653	-16,2326295	-11,6551205

- ☐ ☐ What hypotheses are tested here? (State them in words only, do not forget *important adjectives*.)  
 Circle on the SPSS output above where you read the final P-value for your hypotheses.  
 Do you reject or fail to reject your  $H_0$ ?

- ☐ ☐ Write a statistical conclusion quantifying the effect size. (Again, do not forget *important adjectives*.)

**Exercise 3 — Speedy self-assessment — 7 points**

A newly hired salesman was given the company's sales pitch, that is currently successfully used by dozens of salesmen, but thinks he would have such a better and more effective pitch. However, because he is prudent, he wants to test his idea first, before using it for an extended period of time. So, on day 1 of his first job, he does as he was asked to and uses the company's sales pitch. But on day 2, that ambitious and self-confident salesman uses his own pitch. Results are: on day 1, he talked over the phone to 534 persons, out of which 64 subscribed to the product; on day 2, he obtained 67 subscriptions out of 526 phone calls. What should he do?

- ☐ ☐ What hypotheses are tested here? State them *in words only* and carefully explain your choice.

Work out the test of the hypotheses

- ☐ by drawing a picture summarizing the expected behaviors of your test statistic under  $H_0$  and  $H_1$ ,
- ☐ by computing the numerical value of your test statistic (please spell out the calculation that you typed),
- ☐ by providing the associated P-value.

- ☐ Provide a conclusion, by circling one element in each of the two columns; it must be picked in accordance to your hypotheses and your P-value:

These data show that	both pitches work equally well
These data suggest that	the company's pitch is more effective
These data do not bring to light that	the salesman's personal pitch is more effective
These data cannot exclude that	the salesman's personal pitch is less effective

- ☐ Draw a strategic conclusion: which pitch should he use the next day?

## Exercise 1 — The effect of touch, re-worked (10 points)

It is well documented, e.g., in marketing studies (Jacob Hornik, “Tactile stimulation and consumer response”, *Journal of Consumer Research*, 1992) that light tactile contacts influence human beings in a subtle way towards the requests of the contact-maker. For instance, if a seller touches you lightly, you should be more inclined to buy a product.



We want to illustrate this fact by performing the following experiment. We consider two similar stores (e.g., two Hollister stores) and ask the sellers of the first store to avoid any physical contact with the customers, while the ones of the second store are asked to lightly touch the customers' arm. We are interested in the corresponding purchase rates, which we denote by  $p_0$  (without any contact) and  $q_0$  (with a light contact), respectively. Data collected are that 12 out of the 120 customers served without a contact purchased an item, while 23 out of the 120 served with such a contact did so.

We want to determine whether a light contact has a significative impact on the purchase rate.

### Two-sided test of $H_0 : p_0 = q_0$ against $H_1 : p_0 \neq q_0$

We first test  $H_0 : p_0 = q_0$  against  $H_1 : p_0 \neq q_0$  based on the data collected:

- ☐ draw a picture summarizing the expected behaviors of the test statistic of interest under  $H_0$  and  $H_1$ ,
- ☐ compute the numerical value of this test statistic (please spell out the calculation typed on your calculator),
- ☐ provide the associated P-value.

Write a conclusion consistent with the hypotheses and the P-value obtained, and which is the *most informative possible*. Do so by picking the beginning and the end of the sentence:

- [Beginning]
- A. The data collected cannot exclude that
  - B. The data collected suggest that
  - C. The data collected show that
  - D. The data collected fail to prove that

☐

- [End]
- 1. purchase rates are different with and without a light contact
  - 2. purchase rates are similar with and without a light contact
  - 3. the purchase rate increases with a light contact
  - 4. the purchase rate decreases with a light contact

**One-sided test of  $H_0 : p_0 = q_0$  against  $H_1 : p_0 < q_0$** 

We now test  $H_0 : p_0 = q_0$  against  $H_1 : p_0 < q_0$ ; to that end,

- ☐ draw a picture summarizing the expected behaviors of the test statistic of interest under  $H_0$  and  $H_1$ , provide the associated P-value.

(We do not ask for a conclusion in this case.)

**One-sided test of  $H_0 : p_0 = q_0$  against  $H_1 : p_0 > q_0$** 

We finally test  $H_0 : p_0 = q_0$  against  $H_1 : p_0 > q_0$ ; to that end,

- ☐ draw a picture summarizing the expected behaviors of the test statistic of interest under  $H_0$  and  $H_1$ ,
- ☐ provide the associated P-value.

Write conclusions consistent with the hypotheses and the P-value obtained, by using the same coding as above; two conclusions are possible here and we ask for both of them:

- |  |           |           |
|--|-----------|-----------|
| <input type="checkbox"/> First conclusion possible | Letter: _ | Number: _ |
| Second conclusion possible                         | Letter: _ | Number: _ |

**Who picks which hypotheses?**

Let us consider an academic researcher and a shopkeeper. Which of the three pairs of hypotheses above would they each consider?

- (Pair 1)  $H_0 : p_0 = q_0$  against  $H_1 : p_0 \neq q_0$   
(Pair 2)  $H_0 : p_0 = q_0$  against  $H_1 : p_0 < q_0$   
(Pair 3)  $H_0 : p_0 = q_0$  against  $H_1 : p_0 > q_0$

Just write the number, no explanation or justification is needed (for once):

- ☐ Academic researcher: Pair \_
- Shopkeeper: Pair \_

### Exercise 3 — Choosing between two gifts — 6 points / 15 minutes

Consider an online buying club: members have to place an order every trimester (otherwise, they get some product by default, like “the book of the trimester”). Typically, members were ordering for an average amount of  $\mu_{\text{ref}} = 165$  euros. The club wants to assess the effect of a small gift on its revenue but hesitates between two gifts. It conducts a simultaneous test on two different samples of 200 customers picked independently at random; the customers of each sample are notified that they will get the corresponding small gift if they place an order above 100 euros. Denote by  $\mu_0^1$  and  $\mu_0^2$  the average amounts of orders that would be achieved if the first and second small gifts considered were offered to the many customers of the club. We wonder which gift is the most effective in terms of total revenue (or equivalently, in terms of per customer average revenue).

Consider first the following SPSS output:

Group Statistics

Group		N	Mean	Std. Deviation	Std. Error Mean
Amount	1	200	166,076	22,8922	1,6187
	2	200	170,076	31,8717	2,2537

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Amount	Equal variances assumed	26,990	,000	-1,441	398	,150	-3,9992	2,7748	-9,4542	1,4558
	Equal variances not assumed			-1,441	361,167	,150	-3,9992	2,7748	-9,4559	1,4575

- ☐ What hypotheses are tested here? (State them in equations only.)  
 Circle on the SPSS output above where you read the final P-value for your hypotheses.  
 Do you reject or fail to reject your  $H_0$ ?

- ☐ Write a statistical conclusion (in plain words, that should be understandable by a layman).  
 Do these data, based on their treatment above, indicate *per se* which gift, if any, should be chosen?

We now attempt to follow an alternative approach on the same data. To that end, we consider the SPSS outputs of the next page:

**Group 1 / Comparison to the reference value 165****One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
Amount	200	166,076	22,8922	1,6187

**One-Sample Test**

	Test Value = 165					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Amount	,665	199	,507	1,0763	-2,116	4,268

**Group 2 / Comparison to the reference value 165****One-Sample Statistics**

	N	Mean	Std. Deviation	Std. Error Mean
Amount	200	170,076	31,8717	2,2537

**One-Sample Test**

	Test Value = 165					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Amount	2,252	199	,025	5,0755	,631	9,520

- ☐ What hypotheses are tested here?  
Which would we rather want to test? (State them in equations only.)
- ☐ Indicate the P-values associated with the hypotheses we would rather want to test.  
Do we reject or fail to reject  $H_0$  in each of the two cases?
- ☐ ☐ Do the answers to the question right above prove *per se* that one gift is superior to the other one? Explain.  
Do we get a contradiction with or a confirmation of the conclusion written on the previous page, or none of these?

If you are bored and solved all other questions, you may recompute the P-values of the outputs above. This may lead to 1 bonus point. (Do it on the bottom of this page.)

## Exercise 4 — Lead levels in children’s blood — 6 points / 15 minutes

The presentation of the data set considered here is extracted from an article written by Robert M. Pruzek and James E. Helmreich and published in the *Journal of Statistics Education*:

“[This exercise is] based on an observational study by Morton et al. Children of parents who had worked in a factory where lead was used in making batteries were matched by age, exposure to traffic, and neighborhood with children whose parents did not work in lead-related industries. Whole blood was assessed for lead content yielding measurements in mg/dl; results shown compare the exposed with control children.”

Reference: Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. and Saah, M.: Lead absorption in children of employees in a lead related industry. *American Journal of Epidemiology*, volume 115, pages 549–55, 1982.

Data is listed and plotted on the final page of this statement (for information only).

- ☐ Do we deal with one sample, two independent samples, or two paired samples?  
State accordingly the parameter of interest. (Only one single parameter of interest should be stated.)
- ☐ State relevant hypotheses to be tested. Explain with few words in brackets why you picked them.

Sample data is summarized here:

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Exposed	33	10	73	31,85	14,407
Control	33	7	25	15,88	4,540
Difference	33	-9	60	15,97	15,864
Valid N (listwise)	33				

- ☐ Which sample statistics in the table above will your calculations use? Circle them.  
Compute accordingly the numerical value of your test statistic (provide intermediary calculations).
- ☐ Then work out the test of your hypotheses, by drawing a picture summarizing the expected behaviors of your test statistic under  $H_0$  and  $H_1$  and by computing the associated P-value.



☐ Write a statistical conclusion (in plain words, that are understandable by a layman).

☐ To check your results with the following SPSS output, which *two* cells do you read?  
Are the two values thus read in line with the ones that you calculated?

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Exposed	31,85	33	14,407	2,508
	Control	15,88	33	4,540	,790

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Exposed - Control	15,970	15,864	2,762	10,345	21,595	5,783	32	,000

Pair	Exposed	Control	Difference
1	38	16	22
2	23	18	5
3	41	18	23
4	18	24	-6
5	37	19	18
6	36	11	25
7	23	10	13
8	62	15	47
9	31	16	15
10	34	18	16
11	24	18	6
12	14	13	1
13	21	19	2
14	17	10	7
15	16	16	0
16	20	16	4
17	15	24	-9
18	10	13	-3
19	45	9	36
20	39	14	25
21	22	21	1
22	35	19	16
23	49	7	42
24	48	18	30
25	44	19	25
26	35	12	23
27	43	11	32
28	39	22	17
29	34	25	9
30	13	16	-3
31	73	13	60
32	25	11	14
33	27	13	14

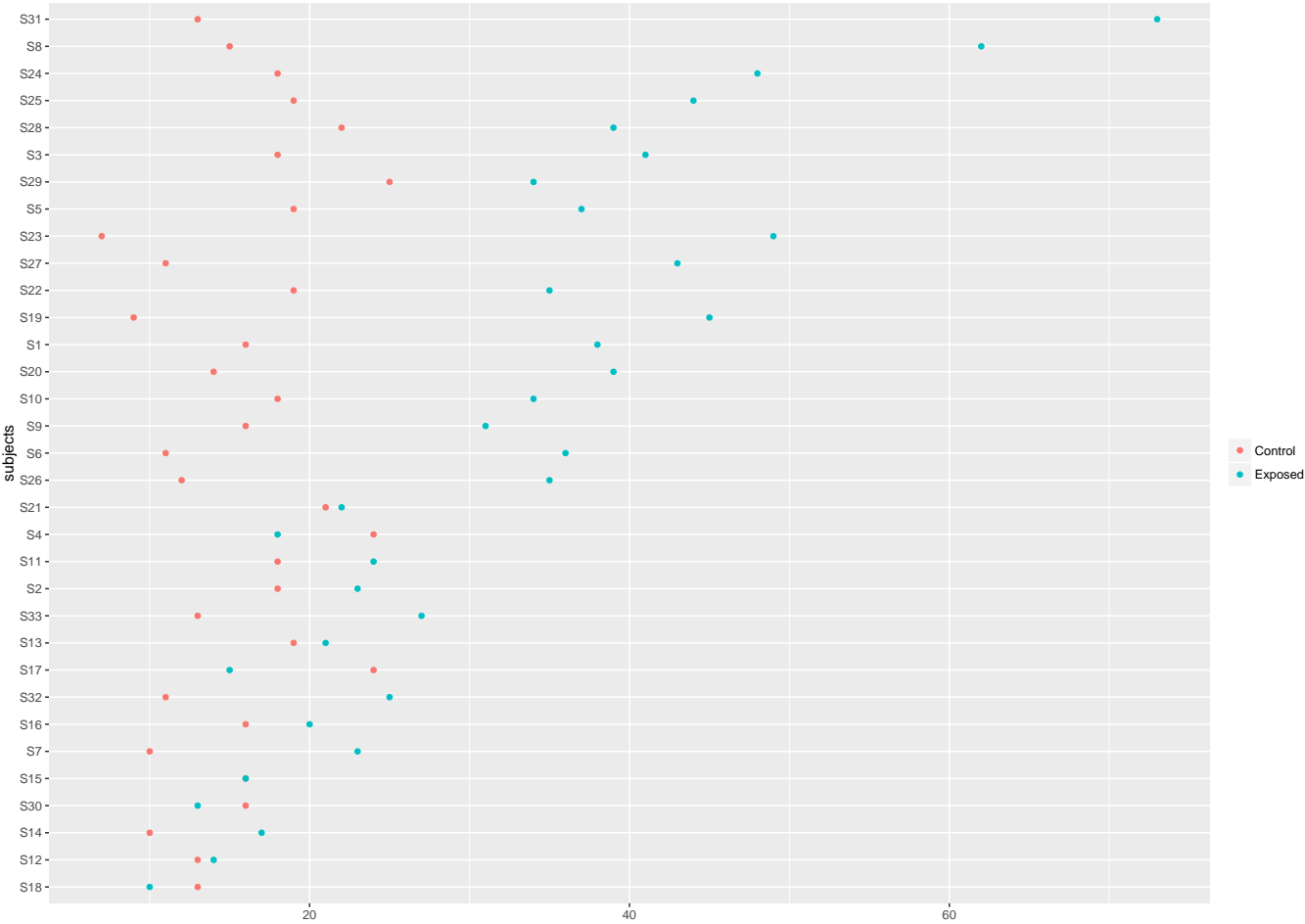


Figure 1: The considered data set (above) and a scatterplot of the data (below).

## $\chi^2$ -tests of independence and of goodness of fit

(Pronounce: “chi-square tests of independence and of goodness of fit”.)

In this chapter (and in this chapter only) we will deal with categorical variables: variables whose values are categories. For instance: gender; age range; socio-professional category; level of satisfaction; yes or no; etc. This case generalizes the binary yes-or-no case, which we studied extensively through inference on proportions.

Such variables are far from the “general quantitative” variables that we studied through inference on population means.

**Two tests, with different aims.** We will study two tests.

First, we study the  $\chi^2$ -test of goodness of fit, to determine whether data are distributed according to a reference distribution or not. This is useful in numerous ways:

- to detect manipulations of the data;
- to check whether some commercial objectives were reached or not;
- to assess whether a sample is representative of a population or not, given some criterion or some sets of criteria.

Second, we study the  $\chi^2$ -test of independence (also known as the  $\chi^2$ -test of homogeneity), to determine whether two variables are independent from each other or not. This is mostly useful in marketing, when you need

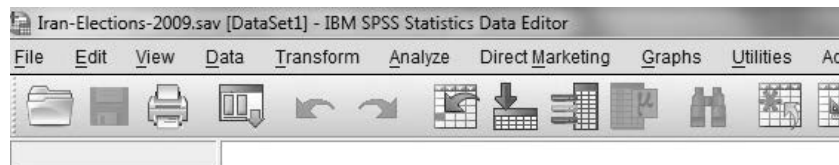
- to determine whether a population should be segmented or not;

that is, whether considered sub-populations share similar or different behaviors; e.g., are they sensitive to the same or to different advertisement campaigns?

## 1. $\chi^2$ -test of goodness of fit

We start with an example and then summarize what you need to know (basically: which hypotheses are considered and how to read the output of a statistical software performing a  $\chi^2$ -test of goodness of fit).

**Did Iran manipulate the results of the 2009 presidential elections?** We consider a political example, described on the right page<sup>1</sup>; an excerpt of the available data is reproduced below. Presidential elections took place in Iran in 2009, with four candidates (misters Ahmadinejad, Karroubi, Mousavi and Rezaee). Iran is made of 29 provinces. Scores of the 4 candidates in each of the 29 provinces are reported, which leads to  $4 \times 29 = 116$  data elements. These scores are large numbers but, as explained in the article on the right, if their orders of magnitudes and first digits heavily depend on the candidate's popularity and on the population of the province, this is not the case of their last digits. The last digits are basically noise and do not contain any information.



	Province	Candidate	Score	LastDigit
22	Mazandaran	Ahmadinejad	1289257	7
23	Qazvin	Ahmadinejad	498061	1
24	Qom	Ahmadinejad	422457	7
25	Semnan	Ahmadinejad	295177	7
26	Sistan and Baluchistan	Ahmadinejad	450269	9
27	Tehran	Ahmadinejad	3371523	3
28	Yazd	Ahmadinejad	337178	8
29	Zanjan	Ahmadinejad	444480	0
30	Ardabil	Karroubi	2319	9
31	Azarbaijan, East	Karroubi	7246	6
32	Azarbaijan, West	Karroubi	21609	9
33	Bushehr	Karroubi	3563	3
34	Chahar Mahaal and Bakhtiari	Karroubi	4127	7
35	Fars	Karroubi	16277	7
36	Gilan	Karroubi	7183	3
37	Golestan	Karroubi	10097	7

Now, in truthfully-reported election results, these last digits should be equally present in the data; in our case, we would have expected a fraction  $1/10$  of the total number of data elements to be a 0, or a 1, etc. That is, we would have expected the counts 11.6 for each of the 10 digits 0–9.

Human beings are bad at making up numbers and this is why in manipulated election results, the last digits may not be equally distributed. E.g., human beings tend to pick more often numbers as 3 and 7, and less often rounder numbers as 0 or 5. In any case, the distribution of the last digits would be different from the uniform distribution if a manipulation had taken place.

<sup>1</sup>The methodology followed in this article is incorrect (for reasons that we will explain in detail in class:  $\chi^2$  tests should have been used!). But the conclusion would have been interesting, if only it survived to a methodologically cleaner analysis: “Each of these two tests provides strong evidence that the numbers released by Iran’s Ministry of the Interior were manipulated.”

# The Washington Post

## The Devil Is in the Digits

D{"Dgtpf "Ddgt "cpf "Crgzcpf te"Uceeq  
Ucwtf c{"Lypg"42."422;"34-24"CO

Upeg"vj g"feretevqp"qh'O c j o qwf "Cj o cf lpglcf"u'itpf urf g"xlevqt {"lp"Kcp"u  
r tgu'f gp'vcln'grgevkqp."ceewuc'vqp'u'qh'itcw'f"j cxg"uy gngf OCi clpu'g'zr ge'vclpu'it qo  
r qm'ugtu'cpf "r wpf ku'c'ring."Cj o cf lpglcf "f k'f u'wtr tkupi n' "y gnilp'w'dcp'et'geu  
kpenw'fpi "Vgj tcp'/"y j gtg"j g'ku'vj qwi j v'q'dg"j k j n' "wpr qr wct'/"cpf "gxgp"Vcdtk  
y j g'ecr k'c'le'k'f {"qh'qr r qu'k'qp'ecpf k'f cvg'O k'J' wu'g'lp'O q'w'x'k'u'p'c'v'x'g'G'cu  
C'j ctd'clcp'r t'q'x'k'peg0

Qy gtu'j cxg'r q'lp'v'f "q"vj g'w'ur tkupi n' "r qqt'r gth'qto cpeg'qh'O g'f k'M'ctt'qwdk  
cpq'j gt'g'ht'o "ecpf k'f cvg."cpf "r ct'v'w'ct'n' {"lp"j ku'j qo g'r t'q'x'k'peg'qh'N'qt'gu'cp."y j gtg  
eq'p'ug'x'c'v'x'g'ecpf k'f cvg'u'ht'gf "r q'q't'n' {"lp"4227."dw'y j gtg"Cj o cf lpglcf "cngi g'f n'  
ecr w'at'g' "93'r gtegp'v'qh'vj g'x'q'v'0G'g'g'dt'qy u'j cxg'd'ggp't'c'k'ug'f "h'w'j gt'd' {"y j g't'g'v'x'g'eq'p'uk'v'gpe {"lp"Cj o cf lpglcf"u'x'q'v'uj ctg'cet'quu  
Kcp'u'r t'q'x'k'pegu "lp"u'r k'g'q'h'y k'f g'r t'q'x'k'peg'cl'x'c'v'k'qp'lp'r cu'v'g'ev'k'qp'u0

Vj g'ug'r l'ge'g'v'qh'vj g'w'ut {"r q'lp'v'lp"vj g't'f'k'g'ev'k'p'qh'itcw'f."q'v'd'g'w'ut'g'0Vj g {"j cxg'g'f "g'zr gt'w'v'q'ur ge'w'v'g'vj cv'vj g'g'ev'k'qp't'gu'w'u  
t'g'v'c'ug'f "d {"Kcp"u'O l'p'k'ut {"qh'vj g'f'p'v'g't'k'q't'j cf "d'ggp'c'ng't'g'f "d'g'j l'p'f "e'ng'ug'f "f q'q't'u'0Dw'y g'f q'p'v'j cxg'v'q't'g'f "q'p'w'i i g'u'x'g'g'x'k'f g'peg  
c'ng'p'0Y g'ecp'w'ug'unc'v'k'ne'u'o q't'g'u'f u'go c'v'ec'n' {"q"uj qy "y cv'vj ku'k'u'k'ng'f "y j cv'j crr g'p'g'f 0J g'g'u'j qy 0

Y g'h'le'q'peg'p't'c'v'g'q'p'x'q'v'eq'w'p'/"y j g'p'wo dgt'q'h'x'q'v'ut'g'eg'k'g'f "d {"f'k'ht'g'p'v'ecpf k'f cvg'u'lp'f'k'ht'g'p'v'r t'q'x'k'pegu'/"cpf "lp'r ct'v'w'ct  
y j g'w'ur'cpf "l'ge'q'p'f/v'q'w'ur'f k' ku'q'h'vj g'ug'p'wo dgt'u'0H'q't'g'zco r'ng."k'ic'ecpf k'f cvg't'g'eg'k'g'f "36.79;"x'q'v'g'u'lp'c'r t'q'x'k'peg"0 t'0M'ctt'q'wd'k'u  
ce'w'cl'x'q'v'eq'w'p'lp'k'k'z'j cp+ "y g'g'h'le'q'w'u'q'p'f k' ku'9'cpf"; 0

Vj ku'o c {"l'uggo "u'it'cpi g."d'ge'c'w'ug'vj g'ug'f k' ku'w'w'c'n' {"f q'p'v'ej cpi g'y j q'y l'p'u'0k'p'f'c'v'v'f'ku'f k' ku'lp'c'f'c'k'g'ev'k'qp'f q'p'v'g'm'w'u'cp {"y l'p'i  
cd'q'w'vj g'ecpf k'f cvg'u."y j g'o c'ng'w'r "qh'vj g'g'ev'q't'c'v'g'q't'vj g'le'q'p'v'z'v'qh'vj g'g'ev'k'qp'0Vj g {"ctg't'cp'f qo "p'q'k'ug'lp'vj g'ug'p'ug'vj cv'c'f'c'k'k'x'q'v'  
eq'w'p'ku'cu'k'ng'f {"q'g'p'f "lp'3'cu'k'ku'q'g'p'f "lp'4.'5.'6."q't'cp {"q'j gt'p'wo g't'c'd'0Dw'y cv'u'g'z'c'w'f "y j {"y j g'f "ecp'ug't'x'g'cu'c'f'k'o w'u'g'u'v'ht'q'  
g'ev'k'qp'f'itcw'f'0H'q't'g'zco r'ng."cp'g'ev'k'qp'lp'y j lej "c'o cl'q't'k'f "qh'r t'q'x'k'peg'cl'x'q'v'eq'w'p'v'g'p'f g'f "lp'7'y q'w'f "u'w'g'f "t'c'k'g't'g'f "h'v'i u'0

Y j {"y q'w'f "h'c'w'f w'g'p'v'p'wo dgt'u'iq'q'nic'p {"f'k'ht'g'p'v'AVj g't'g'cu'p'ku'vj cv'j wo c'pu'ct'g'd'ef "cv'o c'n'k'p'i "w'r "p'wo dgt'u'0E'qi p'k'k'g'r u'f e'j q'q'i k'u'u  
j cxg'h'q'w'p'f "y cv'w'w'f {"r ct'v'k'c'cp'u'lp'f'cd'g'zr g't'o g'p'u'c'ung'f "q'y j t'k'g'v'g'v'g'p'eg'u'q'h't'cp'f qo "f k' ku'y k'u'v'g'p'f "q'v'g'ev'v'q'o g'f k' ku'o q't'g  
h'g's w'g'p'v'f "y cp'q'j gtu'0

Uq'y j cv'ecp'y g'o c'ng'q'h'Kcp'u'g'ev'k'qp't'gu'w'u'AY g'w'ug'f "y j g't'gu'w'u't'g'v'c'ug'f "d {"y j g'O l'p'k'ut {"qh'vj g'f'p'v'g't'k'q't'cp'f "r w'd'k'ij g'f "q'p'vj g'y g'd  
u'k'g'q'h'r'g'u'VX."c'p'gy u'ej c'p'p'g'h'w'p'f g'f "d {"Kcp"u'j q'x'g't'p'o g'p'0Vj g'o l'p'k'ut {"r t'q'x'k'g'f "f'c'v'f'ht'4;"r t'q'x'k'pegu."cpf "y j g'g'zco l'p'g'f "y j g  
p'wo dgt'q'h'x'q'v'g'ce'j "qh'vj g'h'w'v'o cl'p'ecpf k'f cvg'/"Cj o cf lpglcf."O q'w'x'k'k'M'ctt'q'wd'k'cp'f "O q'j u'g'p'T'g'l c'k'/"k'u't'g'r q't'v'g'f "q'j cxg  
t'g'eg'k'g'f "lp'g'ce'j "qh'vj g'r t'q'x'k'pegu'/"c'q'v'cl'q'h'338'p'wo dgt'u'0

Vj g'p'wo dgt'u'iq'q'nic'w'ur l'ek'w'u'0Y g'h'p'f "q'q'o cp {"9'u'cp'f "p'q'v'g'p'q'w'j "7'u'lp'vj g'w'ur'f k' k'0Y g'g'zr ge'v'g'ce'j "f k' k'02.'3.'4."cp'f "u'q'p'v'q'  
crr g'ct'cv'vj g'g'p'f "qh'32'r gtegp'v'qh'vj g'x'q'v'eq'w'p'u'0Dw'lp'Kcp'u'r t'q'x'k'peg'cl'x'q'v'g'g'ev'k'qp'0Vj g'f k' k'9'crr g'ct'u'39'r gtegp'v'qh'vj g'w'o g."cp'f "q'p'v'  
6'r gtegp'v'qh'vj g't'gu'w'u'g'p'f "lp'vj g'p'wo dgt'70V'y q'w'ej "f gr c't'w'g'u'it'q'o "y j g'c'x'g't'c'i g'/"c'ur k'ng'qh'39'r gtegp'v'q't'o q't'g'lp'q'p'g'f k' k'cp'f "c'  
f t'q'r "q'6'r gtegp'v'q't'g'u'u'lp'cp'q'j gt'/"ctg'g'z'v'go g'f "w'p'k'ng'f 0H'q'g'gt'vj cp'h'w'f "lp'c'j"v'p'f t'g'f "p'q'p'f'itcw'f w'g'p'v'g'ev'k'qp'u'y q'w'f "r t'q'f w'eg  
u'w'ej "p'wo dgt'u'0

Cu'c'r q'lp'v'qh'eqo r c't'ku'p'f "y j g'ecp'c'p'c'n'f | g'vj g'w'v'g'd {"w'v'g'x'q'v'eq'w'p'u'ht'q'j p'O e'Ec'lp'cp'f "D'et'c'eni'Q'dco c' "lp'w'ur'f {g'ct'u'w'U'0  
r t'gu'f gp'v'cl'n'gr'ge'v'k'qp'0Vj g'f'g's w'g'p'el'g'u'q'h'w'ur'f k' ku'lp'vj g'ug'g'ev'k'qp't'g'w't'p'u'p'g'x'g't' "k'ug'cd'q'x'g'36'r gtegp'v'q't'f'c'm'd'g'm'y "8'r gtegp'v'c'  
r c'w'g't'p'y g'y q'w'f "g'zr ge'v'q'ug'g'lp'v'g'g'p'v'f "q'w'q'h'c'j"v'p'f t'g'f "f'c'k'g'ev'k'qp'u'0

Dw'yj cv'u'p'q'v'c'n'0Ru'f e'j q'q'i k'u'u'j cxg'c'nu'q'h'w'p'f "y j cv'j wo c'pu'j cxg't'q'w'd'g'f'g'p'g't'c'v'p'i "p'q'p'cf l'ce'g'p'v'f k' ku'w'ej "cu'86'q't'39."cu  
qr r q'ug'f "q'45+cu'ht'g's w'g'p'v'f "cu'q'p'g'y q'w'f "g'zr ge'v'lp'c'v'g's w'g'p'eg'q'h't'cp'f qo "p'wo dgt'u'0V'q'ej g'eni'ht'f "f g'x'c'v'k'qp'u'q'h'vj ku'v'f r'g."y j g  
g'zco l'p'g'f "y j g'r c't'u'q'h'w'ur'cp'f "l'ge'q'p'f/v'q'w'ur'f k' ku'lp'Kcp'u'x'q'v'eq'w'p'u'0Q'p'c'x'g't'c'i g."h'vj g't'gu'w'u'j cf "p'q'v'd'ggp'o c'p'k'w'v'g'f."92  
r gtegp'v'qh'vj g'ug'r c't'u'uj q'w'f "eq'p'uk'v'q'h'f k'u'p'ew'p'q'p'cf l'ce'g'p'v'f k' ku'0

P'q'v'q'p'lp'vj g'f'c'v'f'ht'q'o "Kcp'<Q'p'v' "84'r gtegp'v'qh'vj g'r c't'u'eq'p'v'cl'p'p'q'p'cf l'ce'g'p'v'f k' ku'0Vj ku'o c {"p'q'v'w'p'f "u'q'f'k'ht'g'p'v'it'q'o "92  
r gtegp'v'dw'vj g'r t'q'd'cd'k'k'f {"y j cv'c'f'c'k'k'g'ev'k'qp'y q'w'f "r t'q'f w'eg"c'f'k'ht'g'p'eg'vj ku'f'c't'i g'ku'g'u'v'j cp'60#r gtegp'v'0C'p'f "y j k'g'q'w'f'k'u'v'g'u'v  
/"x'c't'c'v'k'p'lp'w'ur'f k' k'ht'g's w'p'el'g'u'/"w'i i g'u'u'vj cv'T'g'l c'k'u'x'q'v'eq'w'p'u'ct'g'f g'o q'u'v'k't'g'i w'ct."y j g'f'c'eni'q'h'p'q'p'cf l'ce'g'p'v'f k' ku'0'o q'u'v  
u't'k'p'i "lp'vj g't'gu'w'u't'g'r q't'v'g'f "ht'q'j o cf lpglcf 0

G'ce'j "qh'vj g'ug'f y q'v'g'u'ur' t'q'x'k'f g'u'it'q'p'i "g'x'k'f g'peg'vj cv'vj g'p'wo dgt'u't'g'v'c'ug'f "d {"Kcp"u'O l'p'k'ut {"qh'vj g'f'p'v'g't'k'q't'j y gt'g'o c'p'k'w'v'g'f 0Dw  
w'c'ng'p'q'i g'y gt."y j g'f "g'c'x'g'x'g't {"k'w'g't'q'q'o "ht'f'g'c'u'p'cd'ng'f'q'w'd'0Vj g'r t'q'd'cd'k'k'f {"y j cv'c'f'c'k'k'g'ev'k'qp'y q'w'f "r t'q'f w'eg'd'q'y "q'q'f'g'y  
p'q'p'cf l'ce'g'p'v'f k' ku'cp'f "y j g'w'ur l'ek'w'u'f g'x'c'v'k'qp'u'lp'w'ur'f k' k'ht'g's w'p'el'g'u'f g'ue't'ld'g'f "g'c't'r'g't'k'u'g'u'v'j cp'0270k'p'q'j g't'y q't'f u."c'd'g'v'vj cv  
y j g'p'wo dgt'u'ct'g'eng'cp'ku'c'q'p'g'lp'v'y q'j w'p'f t'g'f "ht'q'i "uj q'u'0

*Bernd Beber and Alexandra Scacco, Ph.D. candidates in political science at Columbia University, will be assistant professors in New York University's Wilf Family Department of Politics this fall."*

We test the following hypotheses:

- $H_0$  [conformity]: The last digits are uniformly distributed (i.e., data seem authentic);
- $H_1$  [nonconformity]: The last digits follow another distribution (i.e., data must have been manipulated).

For  $\chi^2$  tests of goodness of fit, we do not get to determine whether the fit (also known as conformity) should be  $H_0$  or  $H_1$ : it has to be  $H_0$ , while  $H_1$  has to be the lack of fit (or nonconformity). The rules given in Chapter 4 for picking  $H_0$  and  $H_1$  (e.g., prudence vs. risky actions) do not apply here.

Let us consider the SPSS output for the  $\chi^2$  test of goodness of fit against a uniform distribution on our data.

LastDigit			
	Observed N	Expected N	Residual
0	9	11,6	-2,6
1	11	11,6	-,6
2	8	11,6	-3,6
3	9	11,6	-2,6
4	10	11,6	-1,6
5	5	11,6	-6,6
6	14	11,6	2,4
7	20	11,6	8,4
8	17	11,6	5,4
9	13	11,6	1,4
Total	116		

Test Statistics	
	LastDigit
Chi-Square	15,552 <sup>a</sup>
df	9
Asymp. Sig.	,077

a. 0 cells (0,0%) have expected frequencies less than 5.  
The minimum expected cell frequency is 11,6.

The first column indicates the observed counts: the counts that we read in our data. The second column indicates the expected counts: what, very ideally, we would have got on average if  $H_0$  was true. We have here 116 data elements, so a uniform distribution would have resulted on average on counts 11.6 for each 0–9 digit. (Of course, this is very idealized as counts should be integers!)

The high-level idea is the following: observed and expected counts are compared in some global way. We denote them by  $N_{\text{obs}}$  and  $N_{\text{exp}}$ . If they differ significantly, then data are said to severely contradict  $H_0$ , which in turn should be rejected. How is this global comparison conducted?

The following test statistic is computed:

$$D_n = \sum \frac{(N_{\text{obs}} - N_{\text{exp}})^2}{N_{\text{exp}}},$$

where the sum is over all categories (we have 10 categories in our case, the 0–9 digits).

For instance, our data lead to a global difference (called the  $\chi^2$ -divergence) equal to

$$D_{116} = \frac{(9 - 11.6)^2}{11.6} + \frac{(11 - 11.6)^2}{11.6} + \dots + \frac{(17 - 11.6)^2}{11.6} + \frac{(13 - 11.6)^2}{11.6} = 15.552,$$

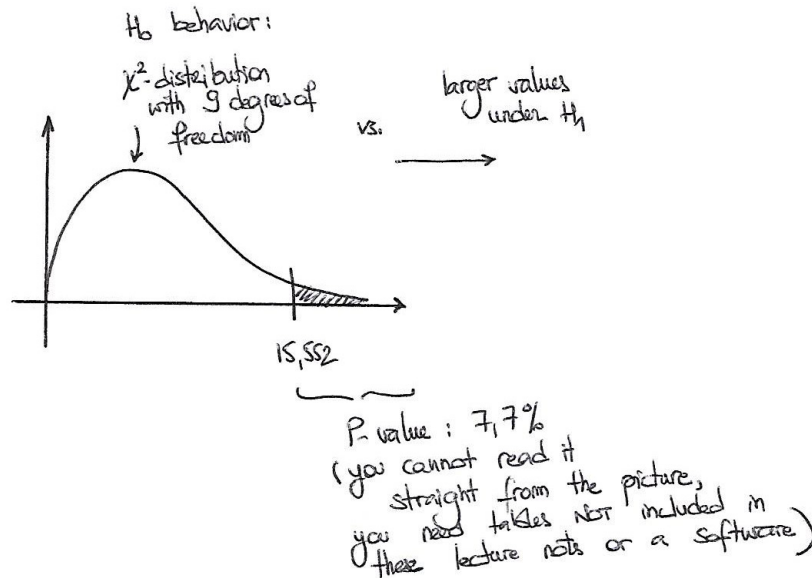
which can be read in the right table above, first line.

What was the expected behavior of this  $D_n$  statistic? Under  $H_0$ , it follows (approximately) a  $\chi^2$ -distribution with  $k - 1$  degrees of freedom<sup>2</sup>, where  $k$  is the number of categories. That is, in our case,

<sup>2</sup>You may wonder what the  $\chi^2$  distribution is... but we will not have time to dig into this matter.

$D_{116}$  was expected to follow a  $\chi^2$ -distribution with 9 degrees of freedom. The value 9 can be read in the right table above, second line, where df is a short-hand notation for “degrees of freedom”.

Finally,  $D_n$  is expected to take much larger values under  $H_1$ . We thus reject  $H_0$  when  $D_n$  is above some threshold. The following picture summarizes the discussion above and indicates how the P-value should be calculated. For you to calculate it by hand, we should provide tables for the  $\chi^2$ -distributions as we did for the normal distributions. We have a good news instead: you do not need to calculate them by hand, you simply need to be able to read them in the SPSS output—right table above, third line. We read a P-value of  $0.077 = 7.7\%$ .



The P-value is above 5% and thus, we fail to reject  $H_0$ . We cannot claim that these data are a smoking gun in proving that the Iranian government manipulated the elections results, we have no (blatant) evidence of a manipulation. (This is in contradiction with the findings stated in the Washington Post article: but remember, their methodology was incorrect!)

There is a final point to discuss: the table note under the right table stating that no cell has an expected count smaller than 5 (SPSS writes “frequency” instead of “count” but SPSS is wrong). Indeed, for the  $D_n$  statistic to approximatively follow the indicated  $\chi^2$  distribution, some conditions should be satisfied:

- The sample size  $n$  should be larger than 30 (which is the case here).
- All expected counts  $N_{exp}$  need to be larger than 5 (which is the case here).

If the latter point does not hold, then some categories need to be merged or suppressed. Examples of such treatments will be given in the next section, for  $\chi^2$ -tests of independence (homogeneity).

**$\chi^2$ -test of goodness of fit: what you should remember and be able to do.**

The  $\chi^2$  test of goodness of fit considers categorical variables. It tests whether observed data correspond, or not, to a reference distribution.

You only need to

- know which hypotheses are tested ( $H_0$  : conformity to the reference distribution versus  $H_1$  : nonconformity);
- be able to read software outputs of  $\chi^2$ -tests of goodness of fit, and in particular, read the P-value and verify the conditions for applying the test;
- in case  $H_0$  is rejected, be able to read in the data why this is the case (which categories are over- or under-represented; we have not illustrated this yet, but will do so in the exercises);
- ideally, be able to recompute the expected counts, if asked to do so.

The exercises will demonstrate the various uses of the  $\chi^2$ -test of goodness of fit:

- to detect manipulations of the data (as we already saw with the Iranian 2009 elections, and as we will further see with Mendel's experiments; another use, not documented in this book, is in auditing and accounting with Benford's law);
- to check whether some commercial objectives were reached or not (see an exercise about a call-center waiting time);
- to assess whether a sample is representative of a population or not, given a criterion or some sets of criteria (see an exercise about the pre-professional contest of the JE campaign at HEC Paris).

In any case, we will not teach you how to manipulate data but merely, how to detect manipulations. Were we at Hogwarts, this course could be called "Defence Against the Dark Arts"...

Also, it is our pleasure to remind you of this fake quote by UK Prime Minister Churchill (that the nazi propaganda wanted to attribute to him!):

Do not trust statistics that you did not fake yourself.



## 2. $\chi^2$ -test of independence (homogeneity)

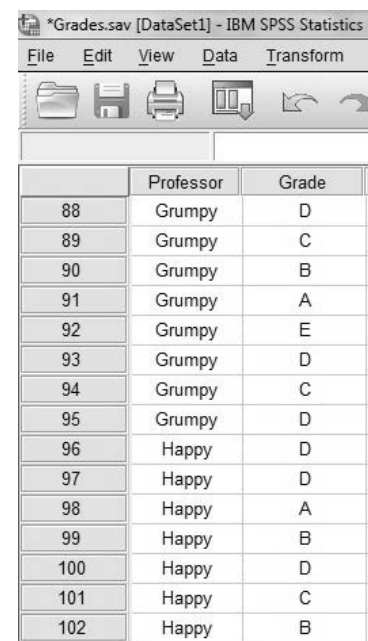
This test considers pairs of categorical variables (e.g., gender and satisfaction level) and studies whether the two variables are independent or not; or, to put it differently, whether there is some homogeneity in the data, or to the contrary, whether the distributions of the second variable given the values of the variable differ.

Again, we will first discuss an example and then summarize what you should be able to do in practice (and practice requires solving several exercises).

**How we ensured a fair grading among statistics instructors, in the good old past...** Before the current quota system was put in place for grading, we, statistics instructors, already checked that we were all using the same grading scales, i.e., that the distributions of grades were homogeneous among the groups or, to put it differently, that the grade distributions were independent of the professors.

Consider the following data (which we made up), about how two instructors, with respective nicknames Professor Grumpy and Professor Happy, graded an exam.

Grades	A	B	C	D	E	F	Total
Professor Grumpy	14	15	26	18	17	5	95
Professor Happy	21	18	24	19	15	2	99
Total	35	33	50	37	32	7	194



	Professor	Grade
88	Grumpy	D
89	Grumpy	C
90	Grumpy	B
91	Grumpy	A
92	Grumpy	E
93	Grumpy	D
94	Grumpy	C
95	Grumpy	D
96	Happy	D
97	Happy	D
98	Happy	A
99	Happy	B
100	Happy	D
101	Happy	C
102	Happy	B

What do you think: are the nicknames appropriate or are they only a consequence of a false impression?

An excerpt of the data is reproduced on the right and we will use SPSS to answer the question.

We first compute the grade distributions per professor.

Professor * Grade Crosstabulation									
			Grade						Total
			A	B	C	D	E	F	
Professor	Grumpy	Count	14	15	26	18	17	5	95
		Expected Count	17,1	16,2	24,5	18,1	15,7	3,4	95,0
		% within Professor	14,7%	15,8%	27,4%	18,9%	17,9%	5,3%	100,0%
	Happy	Count	21	18	24	19	15	2	99
		Expected Count	17,9	16,8	25,5	18,9	16,3	3,6	99,0
		% within Professor	21,2%	18,2%	24,2%	19,2%	15,2%	2,0%	100,0%
Total	Count	35	33	50	37	32	7	194	
	Expected Count	35,0	33,0	50,0	37,0	32,0	7,0	194,0	
	% within Professor	18,0%	17,0%	25,8%	19,1%	16,5%	3,6%	100,0%	

If the grading scales were independent of the professor (if they were homogeneous between the two professors), then we would roughly expect to read the global proportions (last line of the table) for each professor. This leads to the expected counts: for instance, the expected count for Professor Happy and the grade A is 18.0% (global proportion of A) times 99 (total number of students for Professor Happy). The value 18.0% itself corresponds to the ratio  $35/194 \approx 18.04\%$ , which is the

total count of A divided by the total number of students. All in all, the expected count for Professor Happy and the grade A equals

$$N_{\text{exp}} = \frac{35 \times 99}{194} = 18.04\% \times 99 \approx 17.9,$$

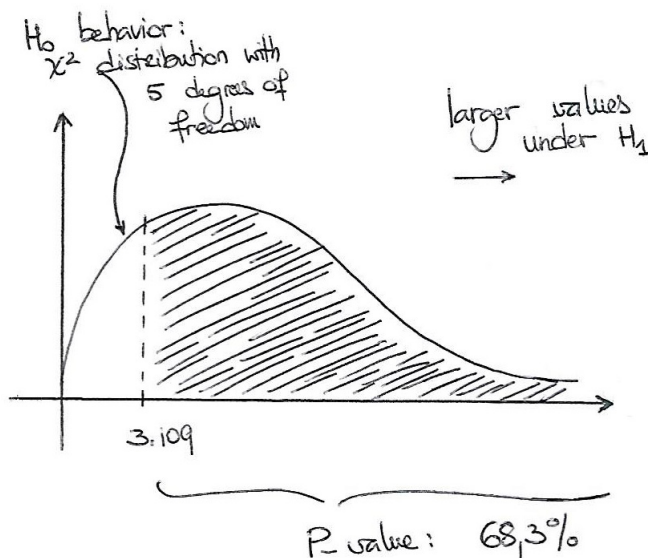
which is indeed the value we read in the corresponding cell. It should be compared to the observed value  $N_{\text{obs}} = 21$ . The comparison should be made for each grade-professor cell, in the same fashion as above: we compute the  $\chi^2$  divergence

$$D_{194} = \sum \frac{(N_{\text{obs}} - N_{\text{exp}})^2}{N_{\text{exp}}} = \frac{(14 - 17.1)^2}{17.1} + \frac{(15 - 16.2)^2}{16.2} + \dots + \frac{(2 - 3.6)^2}{3.6} \approx 3.109,$$

as we can read in the first set of tables on the right (in the smaller table, first line: “Pearson Chi-square”).

Now, we also need to state our hypotheses; in this case again, we have no choice, and  $H_0$  has to correspond to independence (homogeneity) and  $H_1$  to some dependency structure (to some lack of homogeneity). To put it differently, we consider here

- $H_0$  [independence/homogeneity]: The grading scales are identical for the two professors (i.e., in a more technical way: the grading distributions are independent of the professor);
- $H_1$  [some dependency/lack of homogeneity]: The grading scales are different (e.g., one of the two professors is stricter<sup>3</sup> than the other when it comes to grading, i.e., one of the two grading scales puts more mass on lower grades).



Under  $H_0$ , we expected the  $D_{194}$  statistic to approximately follow a  $\chi^2$  distribution, with  $(6 - 1) \times (2 - 1) = 5$  degrees of freedom (the product of the numbers of categories for each variable minus 1).

Under  $H_1$ , we expected larger values. This leads to the computation of a P-value, as illustrated in the picture below. Again, the picture only illustrates the principle of the calculation, the exact calculation would require tables that are not provided in these lectures notes or a statistical software; in our case, we read the P-value in the SPSS outputs (first set of tables on the right, “Pearson Chi-square” line in the smaller table, third column).

Wait — Can we conclude now? No! Do not forget that we need to check that some conditions are satisfied. First, the total sample size  $n$  should be larger than 30: this is the case here, as it equals 194. But not all expected counts are larger than 5: for the F cells, we read expected counts equal to 3.4 and 3.6. Something has to be done, we should not exploit this first series of outputs. Note that SPSS draws your attention to the issue, with a table note (“2 cells [...] have expected count less than 5”).

There are two ways of fixing this.

<sup>3</sup>But the scales could differ in other manners, e.g., one professor could mostly use grades B and C while the other ones gives lots of A and lots of E - F. The hypothesis  $H_1$  cannot be too specific on how the dependency takes place.

## On raw data

Professor \* Grade Crosstabulation

			Grade						Total
			A	B	C	D	E	F	
Professor	Grumpy	Count	14	15	26	18	17	5	95
		Expected Count	17,1	16,2	24,5	18,1	15,7	3,4	95,0
	Happy	Count	21	18	24	19	15	2	99
		Expected Count	17,9	16,8	25,5	18,9	16,3	3,6	99,0
Total	Count	35	33	50	37	32	7	194	
	Expected Count	35,0	33,0	50,0	37,0	32,0	7,0	194,0	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3,109 <sup>a</sup>	5	,683
Likelihood Ratio	3,161	5	,675
N of Valid Cases	194		

a. 2 cells (16,7%) have expected count less than 5. The minimum expected count is 3,43.

## Merging E and F grades into a single category

Professor \* Grade (merging E and F) Crosstabulation

			Grade (merging E and F)					Total
			A	B	C	D	E or F	
Professor	Grumpy	Count	14	15	26	18	22	95
		Expected Count	17,1	16,2	24,5	18,1	19,1	95,0
	Happy	Count	21	18	24	19	17	99
		Expected Count	17,9	16,8	25,5	18,9	19,9	99,0
Total	Count		35	33	50	37	39	194
	Expected Count		35,0	33,0	50,0	37,0	39,0	194,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,339 <sup>a</sup>	4	,674
Likelihood Ratio	2,350	4	,672
N of Valid Cases	194		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 16,16.

## Removing the F category

Professor \* Grade Crosstabulation

			Grade					Total
			A	B	C	D	E	
Professor	Grumpy	Count	14	15	26	18	17	90
		Expected Count	16,8	15,9	24,1	17,8	15,4	90,0
	Happy	Count	21	18	24	19	15	97
		Expected Count	18,2	17,1	25,9	19,2	16,6	97,0
Total	Count		35	33	50	37	32	187
	Expected Count		35,0	33,0	50,0	37,0	32,0	187,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,645 <sup>a</sup>	4	,801
Likelihood Ratio	1,653	4	,799
N of Valid Cases	187		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 15,40.

First, the best way of dealing with the issue is probably to merge the category F with another category, preferably with the closest category: the E. The second set of tables on the right illustrates this. The hypotheses  $H_0$  and  $H_1$  are still the same and we read a P-value of 67.4%. We cannot reject  $H_0$  and have no evidence to back up the nicknames! Perhaps we should stop using them, at least till we have better evidence?

A second way is to discard all F values, which leads to the third set of tables. We read a P-value of 80.1% and reach the same (lack of) conclusions as above.

$\chi^2$ -test of independence (homogeneity): what you should remember and be able to do.

The  $\chi^2$  test of independence considers pairs of categorical variables. It tests whether observed data pairs are compatible, or not, with the two variables of interest being independent (mathematically speaking), which corresponds to testing some homogeneity in behaviors.

You only need to

- know which hypotheses are tested ( $H_0$  : independence between the two variables / homogeneity of behaviors versus  $H_1$  : lack of independence / presence of some dependency / lack of homogeneity);
- be able to read software outputs of  $\chi^2$ -tests of independence, and in particular, read the P-value and check that the conditions for applying the test are indeed met;
- in case  $H_0$  is rejected, be able to read in the data why this is the case (which dependencies and lacks of homogeneity appear in the data; we have not illustrated this yet, but will do so in the exercises);
- ideally, be able to recompute the expected counts, if asked to do so.

The marketing use of the  $\chi^2$ -test of independence is to determine whether a population should be segmented into sub-populations (with significantly different behaviors) or can be considered as a whole. This affects, e.g., the number and type of advertisements campaigns to be considered.

### 3. Elementary exercises

**Elementary exercise 7.1.** [ $\chi^2$ -test of independence (also called: of homogeneity)] A manager supervises four sales representatives and conducts some satisfaction study among their customers to determine whether they all share the same satisfaction rate, or if one or more stand(s) out (in either way!). Thus, she tests

- $H_0$ : All four sales representatives share the same satisfaction rate
- $H_1$ : They do not, one or two of them stand(s) out in a good or in a bad way

She contacts about 30 customers of each sales representative and ask them about their opinion. She summarizes the data in the following table.

Sales representative		#1	#2	#3	#4
Customers	Satisfied	21	23	19	17
	Not satisfied	8	7	10	11

She gets the following SPSS output.

**Customers \* Sales representative Crosstabulation**

			Sales representative				Total
			1	2	3	4	
Customers	Satisfied	Count	21	23	19	17	80
		Expected Count	20,0	20,7	20,0	19,3	80,0
		% within Sales representative	72,4%	76,7%	65,5%	60,7%	69,0%
	Not satisfied	Count	8	7	10	11	36
		Expected Count	9,0	9,3	9,0	8,7	36,0
		% within Sales representative	27,6%	23,3%	34,5%	39,3%	31,0%
Total	Count		29	30	29	28	116
	Expected Count		29,0	30,0	29,0	28,0	116,0
	% within Sales representative		100,0%	100,0%	100,0%	100,0%	100,0%

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,044 <sup>a</sup>	3	,563
Likelihood Ratio	2,054	3	,561
Linear-by-Linear Association	1,420	1	,233
N of Valid Cases	116		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 8,69.

Answer the following questions based on this output.

- How did we get the expected count for satisfied customers / sales representative #1?
- Can the outcome of the test be validly exploited? Explain.
- Indicate the P-value obtained on the data. Should  $H_0$  be rejected or do we fail to reject it?
- State a statistical conclusion in plain words; in case  $H_0$  was rejected, explain, based on the data, which sales representative(s) stand(s) out.

**Elementary exercise 7.2.** [ $\chi^2$ -test of goodness of fit] A committee was set at HEC Paris around 2009 to determine a uniform grading policy. The decision taken and approved by the Dean was initially suggested by the Finance professors, who were already using it: consider minimum and maximum values for the cumulative distribution of grades. Specifically, assign

- A grades to at least 10% and at most 20% of the students;
- A or B grades to at least 20% and at most 40% of them;
- A, B or C grades to at least 40% and at most 70% of them.

In particular, at least 30% of the students must have a D, E or F. (See the introductory chapter about rules and evaluation.)

The suggestion of the Statistics professors was rejected: set target proportions for all the grades and check whether they are followed by implementing a  $\chi^2$  test of goodness of fit. My colleagues replied that this would be too complicated to understand for some HEC Paris instructors — even if this is a fundamental notion of our basic statistics course!

These target proportions could have been, for instance:

A	B	C	D-E-F
15%	15%	30%	40%

This will form our reference distribution. Suppose that two instructors gave the following grades:

	A	B	C	D	E	F	Total
Instructor 1	9	10	16	11	2	2	50
Instructor 2	9	9	21	8	1	1	49

Did they grade in accordance with the proportions prescribed? To answer this question, we perform a  $\chi^2$ -test of goodness of fit of the data of each instructor to the reference distribution. To do so we first group all D, E, F grades into a single D-E-F category:

	A	B	C	D-E-F	Total
Instructor 1	9	10	16	15	50
Instructor 2	9	9	21	10	49

Recall that we test

- $H_0$ : The grades given follow the distribution prescribed
- $H_1$ : They come from another distribution

We then use SPSS and get the two outputs reproduced on top of the next page. Answer the following questions for each output.

1. How did we get the expected count for grade A?
2. Can the outcome of the test be validly exploited? Explain.
3. Indicate the P-value obtained on the data. Should  $H_0$  be rejected or do we fail to reject it?
4. State a statistical conclusion in plain words; in case  $H_0$  was rejected, explain, based on the data, why this is so (which grade frequencies on the data particularly differ from the prescribed values).

**Instructor 1****Grades**

	Observed N	Expected N	Residual
A	9	7,5	1,5
B	10	7,5	2,5
C	16	15,0	1,0
D-E-F	15	20,0	-5,0
Total	50		

**Test Statistics**

	Grades
Chi-Square	2,450 <sup>a</sup>
df	3
Asymp. Sig.	,484

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 7,5.

**Instructor 2****Grades**

	Observed N	Expected N	Residual
A	9	7,4	1,7
B	9	7,4	1,7
C	21	14,7	6,3
D-E-F	10	19,6	-9,6
Total	49		

**Test Statistics**

	Grades
Chi-Square	8,143 <sup>a</sup>
df	3
Asymp. Sig.	,043

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 7,4.

**Elementary exercise 7.3.** [Answering students' complaints, revisited] We may revisit the data of Elementary exercise 6.4 (see page 104) and rather compare grade distributions between the two groups than just the means. We collect grades into well-chosen bins and may then apply a  $\chi^2$ -test of homogeneity.

**Group \* GradeCategory Crosstabulation**

			GradeCategory					
			0 - 11.5	12 - 13.5	14 - 15.5	16 - 17.5	18 - 20	Total
Group	1pm	Count	10	4	10	8	8	40
		Expected Count	8,4	6,3	9,5	7,4	8,4	40,0
		% within Group	25,0%	10,0%	25,0%	20,0%	20,0%	100,0%
	2.40pm	Count	6	8	8	6	8	36
		Expected Count	7,6	5,7	8,5	6,6	7,6	36,0
		% within Group	16,7%	22,2%	22,2%	16,7%	22,2%	100,0%
	Total	Count	16	12	18	14	16	76
		Expected Count	16,0	12,0	18,0	14,0	16,0	76,0
		% within Group	21,1%	15,8%	23,7%	18,4%	21,1%	100,0%

**Chi-Square Tests**

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2,638 <sup>a</sup>	4	,620
Likelihood Ratio	2,669	4	,615
Linear-by-Linear Association	,028	1	,866
N of Valid Cases	76		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,68.

Answer the following questions based on this output.

1. State the hypotheses tested (you may take much inspiration from the ones stated on page 126).
2. How did we get the expected count for 18–20 / 1pm group?
3. Can the outcome of the test be validly exploited? Explain.
4. Indicate the P-value obtained on the data. Should  $H_0$  be rejected or do we fail to reject it?
5. State a statistical conclusion in plain words; in case  $H_0$  was rejected, explain, based on the data, how the two grade distributions differ.

#### 4. More advanced exercises (quiz-like exercises)

The statements below will only describe the situation considered and provide the SPSS outputs. The list of questions will always be the same, and this is why we state them only once for good here:

1. Which test is worked out here, and what are the hypotheses considered? (Please write a complete, not too technical, sentence for each of the hypotheses.)
2. Can the outcome of the test be validly exploited? Explain. If not, describe how the issues were corrected and which outputs should be exploited instead.
3. Indicate the P-value obtained on the data for all valid applications of the tests.
4. State a statistical conclusion; if this conclusion is about the  $H_0$  phenomenon not taking place, then explain, based on the data, why this is so (which parts of the data severely contradict  $H_0$ ).
5. If applicable: state a business conclusion.
6. Subsidiary question: pick one expected count of your choice and explain how SPSS computed it.

**Advanced exercise 7.1** (Reading habits per socio-professional category). This exercise is based on true data collected by INSEE (France's National Institute for Statistics and Economic Studies). The socio-professional categories are thus divided according to the French typical classification. The study was on households' lifestyles and features many topics; we only consider the reading habits. SPSS outputs are reproduced on the next page.

Now, this exercise illustrates a possible need of segmenting a population. To that end, consider yourself a publisher: what could you possibly do, based on the data studied here? Which consumers should you target (and how): all or some specific categories? Remember: traditional publishers have a hard time to make ends meet and need to be creative to survive.



## Reading habits per socio-professional category: 1/2

Socio-professional category \* Reads Crosstabulation

			Reads			Total
			1 book per month, or more	Fewer than 1 book per month	No book in the entire year	
Socio-professional category	Farmers	Count	3	5	10	18
		Expected Count	2,8	7,6	7,6	18,0
	Craftsmen, shopkeepers, business owners	Count	6	18	25	49
		Expected Count	7,7	20,7	20,6	49,0
	Company executives, intellectual professions	Count	36	70	23	129
		Expected Count	20,3	54,5	54,3	129,0
	Associate / intermediate professions	Count	35	102	58	195
		Expected Count	30,6	82,3	82,1	195,0
	Employees	Count	37	117	94	248
		Expected Count	39,0	104,7	104,4	248,0
	Workers	Count	9	56	131	196
		Expected Count	30,8	82,7	82,5	196,0
	Retired	Count	76	162	221	459
		Expected Count	72,1	193,7	193,1	459,0
	Other non-economically active	Count	34	104	70	208
		Expected Count	32,7	87,8	87,5	208,0
Total	Count	236	634	632	1502	
	Expected Count	236,0	634,0	632,0	1502,0	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	121,562 <sup>a</sup>	14	,000
Likelihood Ratio	127,370	14	,000
Linear-by-Linear Association	10,555	1	,001
N of Valid Cases	1502		

a. 1 cells (4,2%) have expected count less than 5. The minimum expected count is 2,83.

## Reading habits per socio-professional category: 2/2

Socio-professional category \* Reads Crosstabulation

			Reads			Total
			1 book per month, or more	Fewer than 1 book per month	No book in the entire year	
Socio-professional category	Craftsmen, shopkeepers, business owners	Count	6	18	25	49
		Expected Count	7,7	20,8	20,5	49,0
	Company executives, intellectual professions	Count	36	70	23	129
		Expected Count	20,3	54,7	54,1	129,0
	Associate / intermediate professions	Count	35	102	58	195
		Expected Count	30,6	82,7	81,7	195,0
	Employees	Count	37	117	94	248
		Expected Count	38,9	105,1	103,9	248,0
	Workers	Count	9	56	131	196
		Expected Count	30,8	83,1	82,2	196,0
	Retired	Count	76	162	221	459
		Expected Count	72,1	194,5	192,4	459,0
	Other non-economically active	Count	34	104	70	208
		Expected Count	32,7	88,2	87,2	208,0
Total	Count	233	629	622	1484	
	Expected Count	233,0	629,0	622,0	1484,0	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	120,019 <sup>a</sup>	12	,000
Likelihood Ratio	125,622	12	,000
Linear-by-Linear Association	13,239	1	,000
N of Valid Cases	1484		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 7,69.

**Advanced exercise 7.2** (The sinking of the Titanic). Were all Titanic passengers equal in the eyes of death or did some fare better than others? If so, who fared better and why? (The answer to the “why?” part has nothing to do with statistics and should be based on common sense only.)

**Class \* Survived Crosstabulation**

			Survived		Total
			No	Yes	
Class	Crew members	Count	673	212	885
		Expected Count	599,1	285,9	885,0
	First class	Count	122	203	325
		Expected Count	220,0	105,0	325,0
	Second class	Count	167	118	285
		Expected Count	192,9	92,1	285,0
	Third class	Count	528	178	706
		Expected Count	477,9	228,1	706,0
Total	Count	1490	711	2201	
	Expected Count	1490,0	711,0	2201,0	

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	190,401 <sup>a</sup>	3	,000
Likelihood Ratio	180,901	3	,000
Linear-by-Linear Association	,000	1	,991
N of Valid Cases	2201		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 92,06.

**Advanced exercise 7.3** (Hair color by gender). Data below were collected by one of the founding fathers of statistics, Ronald Fisher, in a Scottish district. What do you think of them? Do they lead to any interesting conclusion? If so, try to identify where the phenomenon is the most significant.

**Hair color \* Gender Crosstabulation**

			Gender		Total
			Men	Women	
Hair color	Blond	Count	592	544	1136
		Expected Count	614,4	521,6	1136,0
	Red	Count	119	97	216
		Expected Count	116,8	99,2	216,0
	Chestnut	Count	849	677	1526
		Expected Count	825,3	700,7	1526,0
	Brown	Count	504	451	955
		Expected Count	516,5	438,5	955,0
	Jet-black	Count	36	14	50
		Expected Count	27,0	23,0	50,0
Total	Count	2100	1783	3883	
	Expected Count	2100,0	1783,0	3883,0	

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,467 <sup>a</sup>	4	,033
Likelihood Ratio	10,756	4	,029
Linear-by-Linear Association	1,722	1	,189
N of Valid Cases	3883		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 22,96.

**Advanced exercise 7.4 (Call centers).** Customer services are sometimes hard to reach; call centers are given the task to filter all and answer most of the requests, but call centers are not equally efficient. In France, Internet-access providers are notoriously hard to reach. Suppose that one of them wants to invest more on its call centers (more employees, so that waiting times decrease), because it wants to advertise its customer service. The promise it wants to make is that 50% of the customers will wait less than 2 minutes on the phone before their call is taken, and 90% of them less than 5 minutes. These objectives were given to two call centers some months ago and in the past week, the Internet-access provider performed random tests (at random hours) in the two call centers, to see whether the objectives are reached or not. The data obtained are summarized and studied below. What do you think? Hint: note that the limit distribution among the suitable distributions of waiting times is that 50% of the customers wait 2 minutes or less, 40% of them wait between 2 and 5 minutes, and 10% of them wait more than 5 minutes.

### Call center #1

Waiting time

	Observed N	Expected N	Residual
<= 2 min	58	55,5	2,5
> 2 min & <= 5 min	44	44,4	-,4
> 5 min	9	11,1	-2,1
Total	111		

Test Statistics

	Waiting time
Chi-Square	,514 <sup>a</sup>
df	2
Asymp. Sig.	,774

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 11,1.

### Call center #2

Waiting time

	Observed N	Expected N	Residual
<= 2 min	58	52,0	6,0
> 2 min & <= 5 min	44	41,6	2,4
> 5 min	2	10,4	-8,4
Total	104		

Test Statistics

	Waiting time
Chi-Square	7,615 <sup>a</sup>
df	2
Asymp. Sig.	,022

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,4.

This exercise illustrated how to check whether some commercial objectives are reached or not.

**Advanced exercise 7.5** (“Junior entreprise” campaign at HEC Paris). Every year, in October and November, the “Junior entreprise” campaign takes place at HEC Paris. Various events and activities need to be organized by the two lists, whose members also go through a pre-professional contest. In November 2012, the list led by Nicolas Hubert had to conduct a study on cultural activities at HEC Paris. To that end, they designed a survey and administrated it carefully (on paper, which remains the most effective way of administrating a survey! and then spent the necessary hours to enter data into a spreadsheet). To make sure that their findings would be valid, they wanted to check that their sample of 200 students was indeed representative of the population of full-time students on campus, which we can roughly divide by gender and program<sup>4</sup>. Collected data was distributed as follows:

	Men	Women	Total
MiM	78	73	151
MSc	17	12	29
MBA	10	7	17
PhD	1	2	3
Total	106	94	200

while HEC Paris administration indicated that the 2,983 full-time students that were on campus in 2012–13 were distributed as follows:

	Men	Women	Total
MiM	39.4%	31.3%	70.7%
MSc	9.4%	7.0%	16.4%
MBA	6.5%	3.6%	10.1%
PhD	1.5%	1.3%	2.8%
Total	56.8%	43.2%	100%

The data was used to perform tests, whose outputs are reproduced on the right page.

---

<sup>4</sup>MiM: master in management—your program!; MSc: one-year specialized master—the second year of MiM consists of a MSc; and of course, MBA and PhD programs, that do not need any further clarification.

## Raw data

Categories (all)

	Observed N	Expected N	Residual
MiM / Man	78	78,8	-,8
MiM / Woman	73	62,6	10,4
MSc / Man	17	18,8	-1,8
MSc / Woman	12	14,0	-2,0
MBA / Man	10	13,0	-3,0
MBA / Woman	7	7,2	-,2
PhD / Man	1	3,0	-2,0
PhD / Woman	2	2,6	-,6
Total	200		

Test Statistics

	Categories (all)
Chi-Square	4,364 <sup>a</sup>
df	7
Asymp. Sig.	,737

a. 2 cells (25,0%) have expected frequencies less than 5. The minimum expected cell frequency is 2,6.

## After some treatment

Categories (with some merging)

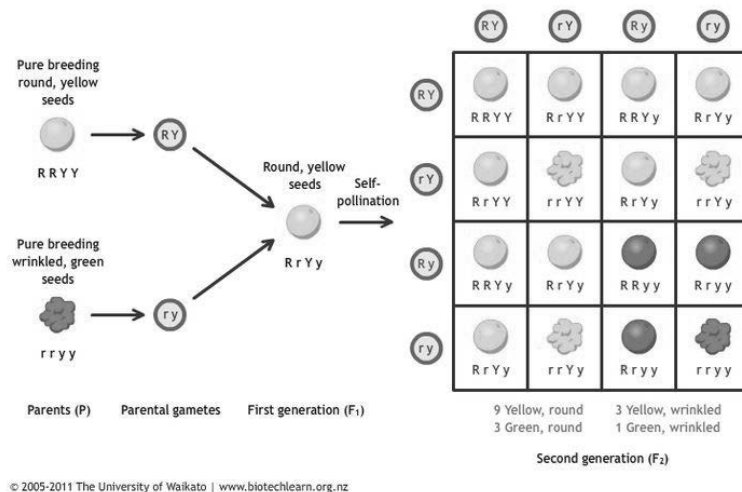
	Observed N	Expected N	Residual
MiM / Man	78	78,8	-,8
MiM / Woman	73	62,6	10,4
MSc / Man	17	18,8	-1,8
MSc / Woman	12	14,0	-2,0
MBA / Man	10	13,0	-3,0
MBA / Woman	7	7,2	-,2
PhD / Man or Woman	3	5,6	-2,6
Total	200		

Test Statistics

	Categories (with some merging)
Chi-Square	4,099 <sup>a</sup>
df	6
Asymp. Sig.	,663

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 5,6.

**Advanced exercise 7.6 (Mendel's experiment).** Mendel, a catholic monk and priest of the 19th century, was the first genetician. He had a theory that each human being had two versions of each gene (two “alleles”), one stemming from each parent. Some would be dominant, some would be recessive. And they would be transmitted at random to the children. The Church disliked this theory, because how would randomness fit in a world guided by God? Nowadays, theologians can concile God's plan and Mendel's theory; but not at that time. So, Mendel had to work in secret... in the monastery's garden. He conducted experiments on peas. A picture of a given experimental scheme is reproduced below.



Two traits, color and appearance, are considered; the dominant alleles are yellow and round, while green and wrinkled are both recessive. The expected distribution of pairs of traits is explained on the picture: in the second generation, we expect 9/16 of the pea plants to be yellow and round; 3/16 of them to be green and round; 3/16, yellow and wrinkled; and 1/16, green and wrinkled. Mendel cultivated 556 second-generation plants (what a work!) and obtained the following distribution of pairs of traits. At his time, there was no theory of statistics but we can conduct an appropriate test to see whether his data back up his theory. What do you think?

Phenotype			
	Observed N	Expected N	Residual
Yellow + Round	315	312,8	2,3
Green + Round	108	104,3	3,8
Yellow + Wrinkled	101	104,3	-3,3
Green + Wrinkled	32	34,8	-2,8
Total	556		

#### Test Statistics

	Phenotype
Chi-Square	,470 <sup>a</sup>
df	3
Asymp. Sig.	,925

0 cells (0,0%) have expected frequencies less than 5.  
The minimum expected cell frequency is 34,8

Historical note (and question): When studying any other data set that Mendel claimed to create, similar conclusions (with, in particular, the same orders of magnitude for P-values) are obtained. Fisher pointed out that this is implausible. Why? Fisher is right. So, how do you think Mendel cheated... and why did he do so?

**Exercise 2 — Alcohol consumption at HEC Paris — 4 points / 10 minutes**

We already studied this data in class, from a different angle. We collected data on HEC Paris students, on a Friday morning: how many glasses of alcohol they had the night before. We already showed in class that the average numbers of glasses per group of students (8am or 10am) were not significantly different. We now look at the same data but in the following way:

**Group \* Number of glasses Crosstabulation**

			Number of glasses			Total
			3 glasses or less	3 - 6 glasses	6 glasses or more	
Group	8am	Count	8	10	5	23
		Expected Count	8,8	7,5	6,6	23,0
	10am	Count	12	7	10	29
		Expected Count	11,2	9,5	8,4	29,0
Total		Count	20	17	15	52
		Expected Count	20,0	17,0	15,0	52,0

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,335 <sup>a</sup>	2	,311
Likelihood Ratio	2,343	2	,310
Linear-by-Linear Association	,072	1	,789
N of Valid Cases	52		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 6,63.

- ☐ What do we compare here, given that we are not comparing average numbers of glasses? Carefully state the corresponding hypotheses.
- ☐ Which is the complete name of the test worked out here? Can the outcome of the test be validly exploited? Explain.
- ☐ What P-value do you read, and do you reject or fail to reject  $H_0$ ? State a statistical conclusion (in plain words, that should be understandable by a layman).
- ☐ What calculations led to the expected count of 6.6 in the top-right cell? To which observed value should it be compared?

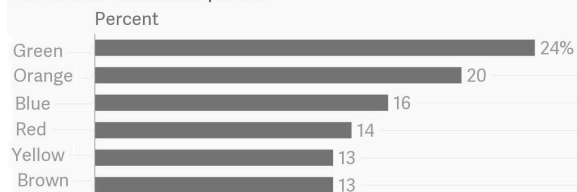
## Exercise 1 — M&M colors — 6 points

This exercise is based on real data linked to M&Ms, which are famous chocolate candies that come in various colors. Rick Wicklin is a computer programmer and statistician at SAS, the company that owns and develops the statistical software of the same name. Rick Wicklin also spends a lot more time than most people do in proximity to M&Ms: his employer is the biggest corporate consumer of M&Ms. Indeed, its CEO Jim Goodnight instituted “M&Ms Wednesday” upon the company’s founding in 1976, after falling in love with the snack during one late-night work session. Ever since, bowls in every SAS office are refilled once a week with the candies. Given the chocolatey bounty of his workplace, Wicklin had plenty of opportunities to ponder the statistical distribution of M&M colors. Then inspiration struck. The first step was to collect his data: two scoops of M&Ms a week from a jar in the closest break room over several weeks in late 2016 and early 2017. He eventually collected 712 candies, or about 1.5 pounds. Then he got counting. The breakdown of the colors in his sample was: 139 green, 133 orange, 133 blue, 108 red, 103 yellow, and 96 brown candies.

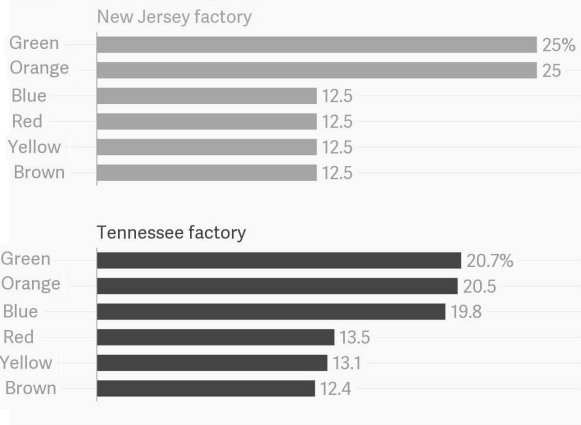


Some breakdowns to compare his sample to include: the latest color distribution available on Mars’ website (was published in 2008, did not get updated since then, was actually erased from the website meanwhile, but Rick Wicklin could get it from a Google search); the color distributions of the US M&M factories as sent by Mars to Rick Wicklin upon his request early 2017.

M&Ms color distribution, c. 2008



M&Ms color distribution, c. 2017



He then conducted a series of tests, whose results are reported on the next page.

- ☐ What are the hypotheses tested in each of these tests?
  - A.  $H_0$ : independence between two variables vs.  $H_1$ : some dependency
  - B.  $H_0$ : lack of conformity to some distribution vs.  $H_1$ : conformity
  - C.  $H_0$ : some dependency between two variables vs.  $H_1$ : independence
  - D.  $H_0$ : conformity to some distribution vs.  $H_1$ : lack of conformity
- ☐ Circle all P-values and indicate below each table which tests reject  $H_0$  and which fail to reject  $H_0$ .
- ☐ Title each test with the name of color distribution considered.  
Carefully explain (on the next page, below the set of tables) how you obtained the assignment.
- ☐ Write a one-sentence-long conclusion, that is understandable by a layman (i.e., avoid statistical jargon!).



**Color**

	Observed N	Expected N	Residual
Green	139	118,7	20,3
Orange	133	118,7	14,3
Blue	133	118,7	14,3
Red	108	118,7	-10,7
Yellow	103	118,7	-15,7
Brown	96	118,7	-22,7
Total	712		

**Test Statistics**

	Color
Chi-Square	14,303
df	5
Asymp. Sig.	,014

**Color**

	Observed N	Expected N	Residual
Green	139	178,0	-39,0
Orange	133	178,0	-45,0
Blue	133	89,0	44,0
Red	108	89,0	19,0
Yellow	103	89,0	14,0
Brown	96	89,0	7,0
Total	712		

**Test Statistics**

	Color
Chi-Square	48,483
df	5
Asymp. Sig.	,000

**Color**

	Observed N	Expected N	Residual
Green	139	170,9	-31,9
Orange	133	142,4	-9,4
Blue	133	113,9	19,1
Red	108	99,7	8,3
Yellow	103	92,6	10,4
Brown	96	92,6	3,4
Total	712		

**Test Statistics**

	Color
Chi-Square	11,764
df	5
Asymp. Sig.	,038

**Color**

	Observed N	Expected N	Residual
Green	139	147,4	-8,4
Orange	133	146,0	-13,0
Blue	133	141,0	-8,0
Red	108	96,1	11,9
Yellow	103	93,3	9,7
Brown	96	88,3	7,7
Total	712		

**Test Statistics**

	Color
Chi-Square	5,235
df	5
Asymp. Sig.	,388

## Exercise 4 — Satisfaction survey — 4 points

Assume that you want to conduct a survey on academic satisfaction at HEC Paris, and that you want to do it in a clean way (i.e., unlike <http://qpvhec.fr/2018/satisfaction-generale/> which simply collected as many responses as possible). You identify at random 100 French-only students that are taking or took the pre-MiM program, 100 international students that joined HEC for the 1st year of the MiM program, and 100 students that joined HEC for a 1-year specialized master (French or international ones). Each of these sets of 100 students is decomposed between 50 students that are currently taking the program, and 50 recent alumni. You pick them at random based on lists of students and then chase them till they answer. (Of course, a few of them remain unreachable.) Your single question was: on a 1 (lowest satisfaction) to 5 (highest satisfaction) scale, how do you rate HEC's academic curriculum?

At the end of the day, the data collected look like that:

			Category			
			Pre-MiM students	1st year of MiM students	Specialized master students	Total
Satisfaction	5 (highest)	Count	21	34	32	87
		% within Category	22,1%	35,1%	33,3%	30,2%
	4	Count	16	21	26	63
		% within Category	16,8%	21,6%	27,1%	21,9%
	3	Count	13	10	11	34
		% within Category	13,7%	10,3%	11,5%	11,8%
	2	Count	20	14	13	47
		% within Category	21,1%	14,4%	13,5%	16,3%
	1 (lowest)	Count	25	18	14	57
		% within Category	26,3%	18,6%	14,6%	19,8%
Total		Count	95	97	96	288
		% within Category	100,0%	100,0%	100,0%	100,0%

Assume that you have to comment on these results.

- ☐ Give a quick example of numbers (percentages) that you would highlight.  
Does the table above prove that satisfaction varies among the three subpopulations considered? Explain.

Next you perform, on second thoughts, the treatment reproduced on the next page.

			Category			
			Pre-MiM students	1st year of MiM students	Specialized master students	Total
Satisfaction	5 (highest)	Count	21	34	32	87
		Expected Count	28,7	29,3	29,0	87,0
	4	Count	16	21	26	63
		Expected Count	20,8	21,2	21,0	63,0
	3	Count	13	10	11	34
		Expected Count	11,2	11,5	11,3	34,0
	2	Count	20	14	13	47
		Expected Count	15,5	15,8	15,7	47,0
	1 (lowest)	Count	25	18	14	57
		Expected Count	18,8	19,2	19,0	57,0
Total		Count	95	97	96	288
		Expected Count	95,0	97,0	96,0	288,0

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11,302 <sup>a</sup>	8	,185
Likelihood Ratio	11,342	8	,183

☐ What are the hypotheses tested?

- A.  $H_0$ : independence between two variables vs.  $H_1$ : some dependency
- B.  $H_0$ : lack of conformity to some distribution vs.  $H_1$ : conformity
- C.  $H_0$ : some dependency between two variables vs.  $H_1$ : independence
- D.  $H_0$ : conformity to some distribution vs.  $H_1$ : lack of conformity

☐ Is the test validly applied? Recall all elements to check to that end.

☐ Does the table above prove that satisfaction varies among the three subpopulations considered? Explain.

## Exercise 1 — Births by month — 4 points / 10 minutes

*This exercise is inspired by real data.* Births used to occur with some seasonality: for instance, there was a significant peak in April (due to having much free time 9 months before in the summer). But the modern lifestyle allows for spare time on a more regular basis throughout the year: did it affect the seasonality of births? We study data collected in 2010 from some large maternity center, consisting in the number of births per month:

Month			
	Observed N	Expected N	Residual
January	667	667,9	-,9
February	611	667,9	-56,9
March	660	667,9	-7,9
April	640	667,9	-27,9
May	667	667,9	-,9
June	655	667,9	-12,9
July	697	667,9	29,1
August	687	667,9	19,1
September	679	667,9	11,1
October	702	667,9	34,1
November	668	667,9	,1
December	682	667,9	14,1
Total	8015		

Test Statistics	
	Month
Chi-Square	10,395 <sup>a</sup>
df	11
Asymp. Sig.	,495

a. 0 cells (0,0%) have expected counts less than 5.

- ☐ Which is the complete name of the test worked out here?  
Can the outcome of the test be validly exploited? Explain.
- ☐ What are the hypotheses considered? (State them in plain words only.)  
What P-value do you read, and do you reject or fail to reject  $H_0$ ?
- ☐ State a statistical conclusion (in plain words, that are understandable by a layman).
- ☐ What calculations led to the expected count for births occurring in May?  
To which observed value should it be compared?

## Simple linear regression

We consider pairs of quantitative data  $(x_j, y_j)$ , where  $j = 1, \dots, n$ , and think that a relationship of the form

$$y_j = a + bx_j + e_j \quad (8.1)$$

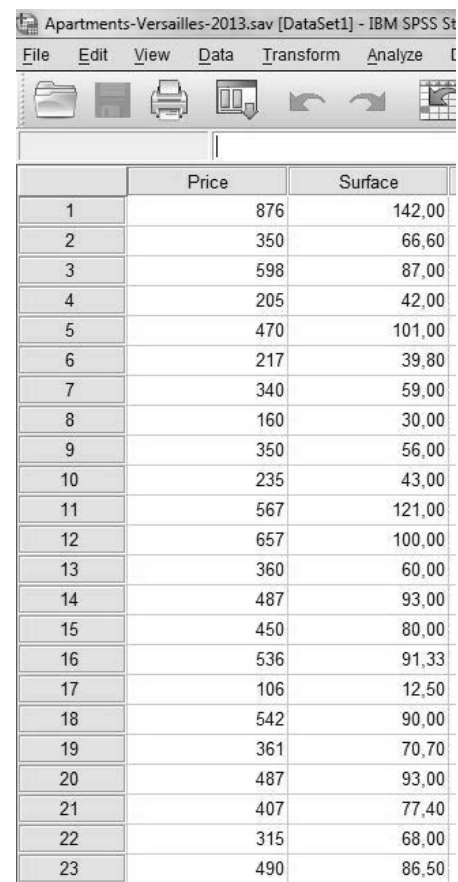
may exist. The coefficients  $a$  and  $b$  are the same for all pairs and define a line: the regression line.

We want to “explain” (in a statistical way) the  $y_j$  in terms of the  $x_j$ . In the example on the right, the  $y_j$  are the prices (in kEuros) of some apartments in Versailles, while the  $x_j$  are their surfaces (in squared meters:  $m^2$ ). As people typically think in terms of price per squared meter, you would indeed expect a linear relationship between the prices  $y_j$  and the surfaces  $x_j$ ; i.e., you would expect  $a = 0$  and  $b$  equal to the said price per squared meter.

Now, in (8.1), we also read some  $e_j$ . These  $e_j$  quantities are called residuals and account for the fact that the surface does not explain all of the price of the apartment; a fraction of the price is explained by other criteria: the neighborhood, the charm, the exposure, the presence of a balcony, of an underground garage, etc.

For the sake of concreteness, let us consider data collected by one of the instructors of this course, namely, Benjamin Petiau<sup>1</sup>. In 2013 (when considering moving and buying an apartment), he collected the following data, on housing prices in Versailles (the prices are the prices initially posted, not the ones that resulted from the negotiations between the buyer and the seller).

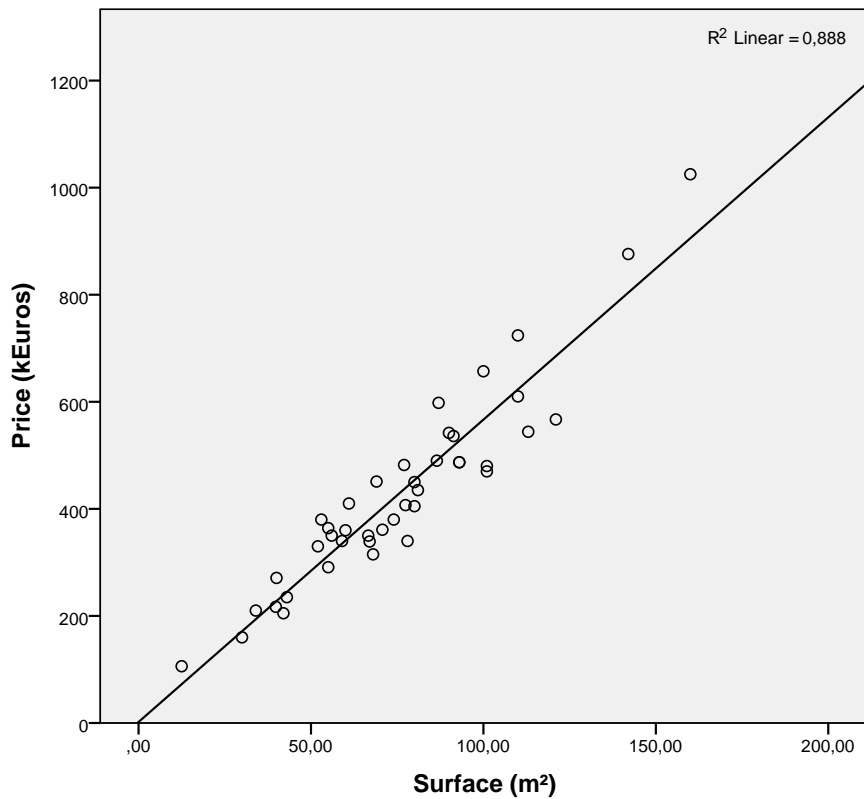
The data pairs are represented on the next page, with a regression line (of equation  $a + bx$ ).



	Price	Surface
1	876	142,00
2	350	66,60
3	598	87,00
4	205	42,00
5	470	101,00
6	217	39,80
7	340	59,00
8	160	30,00
9	350	56,00
10	235	43,00
11	567	121,00
12	657	100,00
13	360	60,00
14	487	93,00
15	450	80,00
16	536	91,33
17	106	12,50
18	542	90,00
19	361	70,70
20	487	93,00
21	407	77,40
22	315	68,00
23	490	86,50

<sup>1</sup>Again! See page 106.

The (surface, price) pairs  $(x_j, y_j)$  are depicted each by a small  $\circ$  while the regression line (with equation  $a + bx$ ) is the solid line. Almost no pair  $(x_j, y_j)$  lies on the regression line: there exist residual terms  $e_j$ . These terms are sometimes positive, sometimes negative.



Here are the corresponding regression summaries as output by SPSS.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,942 <sup>a</sup>	,888	,885	60,665

a. Predictors: (Constant), Surface (m²)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1163593,829	1	1163593,829	316,171	,000 <sup>b</sup>
	Residual	147210,576	40	3680,264		
	Total	1310804,405	41			

a. Dependent Variable: Price (kEuros)

b. Predictors: (Constant), Surface (m²)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,861	25,805		,072	,943
	Surface (m²)	5,648	,318	,942	17,781	,000

a. Dependent Variable: Price (kEuros)

Your mission will be to make sense of all this information.

We will henceforth call the  $y_j$  the dependent variables and the  $x_j$ , the explanatory (or independent) variables. We want to statistically explain  $y_j$  in terms of a linear function of the  $x_j$  plus a constant term (plus if needed, a residual term):  $y_j = a + bx_j + e_j$ , where the  $a$  (intercept) and  $b$  (slope) coefficients are independent of  $j$ .

The questions to be answered are the following ones and in this order; we will go over each of them in detail in the next pages.

1. Is the linear relationship significant, i.e., do the  $x_j$  significantly contribute to the statistical explanation of the  $y_j$ , or to put it differently: is the slope coefficient  $b$  significantly different from 0? If not, we stop studying this linear regression. The above linear regression is significant, as we will see.
2. Otherwise, we wonder how well this linear regression explains / models existing data; an index between 0% and 100% called the  $r^2$  will measure the quality of the model. For the above regression, we have  $r^2 = 88.8\%$ , meaning that 88.8% of the (variations of the) prices are captured by the (variations of the) surfaces, which is truly excellent!
3. We then write the linear regression read, e.g., in our example:

$$\begin{aligned}\text{Price (in kEuros)} &= 1.861 \\ &+ 5.648 \times \text{Surface (in m}^2\text{)} \\ &+ \text{Residual term (with standard deviation: 60.665)}\end{aligned}$$

We should not forget neither the units (kEuros,  $\text{m}^2$ ) nor the residual term (as the linear regression does not explain 100% of the phenomenon, but only 88.8% of it).

4. We then have to interpret the coefficients and see whether the linear relationship makes sense (from an economic, political, or common-sense viewpoint). Here, the slope coefficient of 5.648 can be interpreted as the average price per squared meter in Versailles. The relationship makes sense because the price increases with the surface. The intercept term of 1.861 corresponds to a fixed, base amount of 1.861 kEuros = 1,861 euros to be added to the price. This term might be difficult to interpret but actually, it is not significantly different from 0 (we will see how to make sense of that). We should recompute the linear relationship by enforcing a null intercept term (SPSS can do it and we will see the result).
5. Finally, we may then want to predict future values (what should be the price of a 76- $\text{m}^2$  apartment that was not present in the data set above?) and/or detect outliers (among all apartments in the data set, which are way too expensive given the average prices, or way too cheap?).

These five questions will be the questions asked in each of the exercises of this chapter.

Now, before we dig into the details, we provide regressions summaries output by two other softwares: Microsoft Excel and R. You will realize that all these softwares compute the same quantities (and that they also compute lots of numbers, just as SPSS did!).

Your ultimate mission is to be able to extract from any such regression summary (SPSS, Microsoft Excel, R, or any other statistical software) the relevant information:

- the P-value for the test of a significant relationship;
- the  $r^2$  measure of quality of the regression;
- the intercept and slope coefficients, together with their significance tests, as well as the standard deviation of the residual term.

In the exam, we will only provide outputs from SPSS. What you will have to use in later courses at HEC or during your career, we do not know...

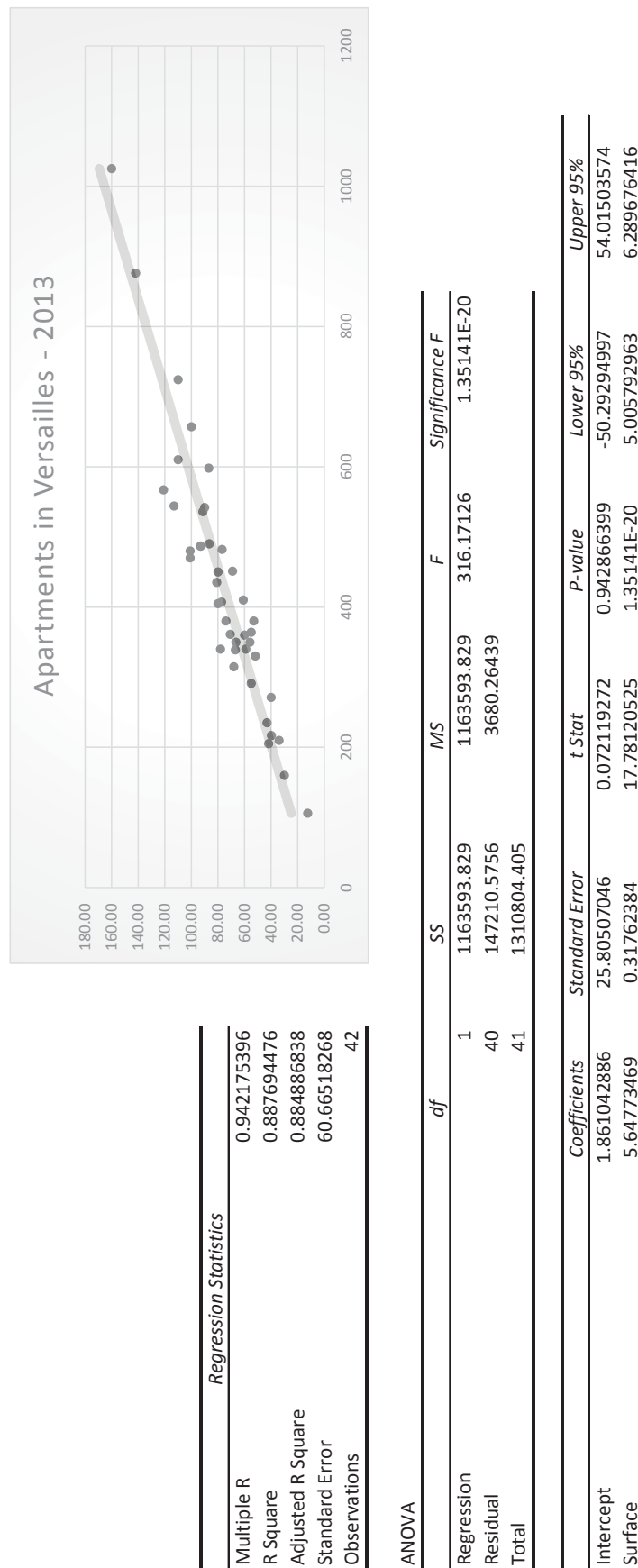


Figure 8.1: Simple linear regression: output obtained with Microsoft Excel.



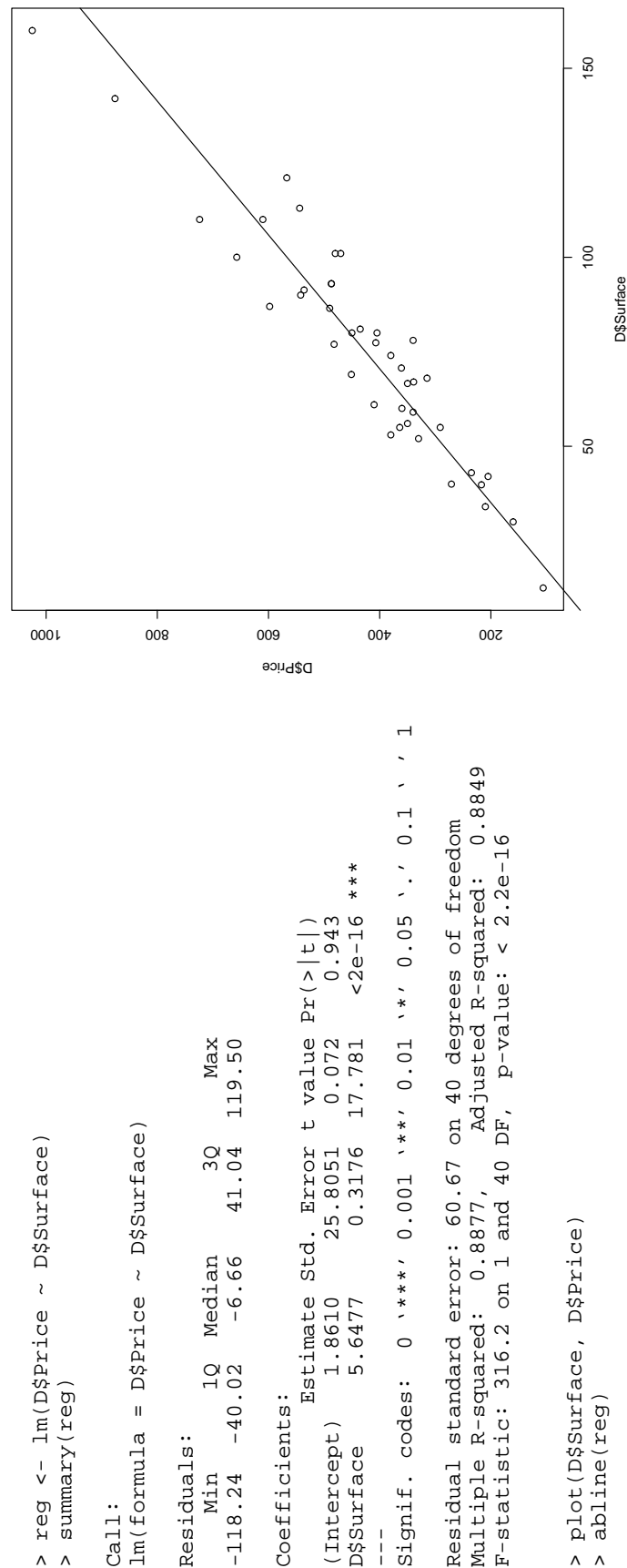


Figure 8.2: Simple linear regression: output obtained with R, an open-source statistical software used in particular by engineers and scientists.

## 1. Basic notions, descriptive statistics, and a first significance test

**Coefficients.** The most important explanation is how to compute the intercept and slope coefficients  $a$  and  $b$ . We do so by minimizing least squares, i.e., we take  $a$  and  $b$  such that the following quantity

$$\sum_{j=1}^n (y_j - (a' + b'x_j))^2$$

is minimum when  $a'$  and  $b'$  describe all possible pairs of real numbers. It sounds very technical but think of it as the best line fit when errors (residuals: distances to the putative line) are measured in terms of their squares (and then summed).

For us, anyway, the important takeaway is that there are closed-form expressions for  $a$  and  $b$ , so that a statistical software can give us their values. (See the appendix for these expressions, but beware, you will not learn anything from studying them.)

**Residuals.** For a given value  $x_j$  of the explanatory variable, the regression line proposes the value  $\hat{y}_j = a + bx_j$  for the dependent variable.

The residuals  $e_j = y_j - \hat{y}_j = y_j - (a + bx_j)$  measure the quality of the fit: the distance of the actual observation  $y_j$  to the value  $\hat{y}_j$  proposed by the model.

An interesting observation (that can be proved via a generalization of the Pythagorean theorem) is that

$$\sum_{j=1}^n e_j = 0, \quad \text{that is,} \quad \frac{1}{n} \sum_{j=1}^n \hat{y}_j = \frac{1}{n} \sum_{j=1}^n y_j = \bar{y}_n.$$

A consequence of this fact is that the regression line goes through the point  $(\bar{x}_n, \bar{y}_n)$ .

**Coefficient of determination  $r^2$ .** It indicates the proportion of the variability in the dependent variable that the models recovers (“explains” in a statistical way) from the explanatory variable.

Another consequence of the “interesting observation” above is that

$$\underbrace{\sum_{j=1}^n (y_j - \bar{y}_n)^2}_{\stackrel{\text{not.}}{=} \Sigma_T} = \underbrace{\sum_{j=1}^n (\hat{y}_j - \bar{y}_n)^2}_{\stackrel{\text{not.}}{=} \Sigma_E} + \underbrace{\sum_{j=1}^n (y_j - \hat{y}_j)^2}_{\stackrel{\text{not.}}{=} \Sigma_R},$$

where

- $\Sigma_T$  (T for total) measures the total variability of the instances  $y_j$  of the dependent variable, around their average  $\bar{y}_n$ ;
- $\Sigma_E$  (E for explained) measures the total variability of outputs  $\hat{y}_j$  of the model, around their average which is also equal to  $\bar{y}_n$ ;
- $\Sigma_R$  (R for residual) measures the total variability of the residuals  $e_j = y_j - \hat{y}_j$ , around their average, which is null.

The coefficient of determination  $r^2$  is then defined as

$$r^2 = \frac{\Sigma_E}{\Sigma_T},$$

and represents the fraction of the total variability explained (recovered) by the model.

In our apartment-data example, we have  $r^2 = 88.8\%$ , meaning that 88.8% of the (variations of the) prices are captured by the (variations of the) surfaces. This is truly excellent!

Indeed, at least in economics, the  $r^2$  is already considered good when it is of order 20%, and very good around 40%.

**Significant linear model?** However, to determine whether a linear model is significant, it does not suffice to discuss the value of  $r^2$ . Actually it has to be compared to something of the order of  $1/\sqrt{n}$ , where  $n$  is the sample size.

The P-value of the test  $H_0$  : no significant linear regression against  $H_1$  : significant linear regression is performed in the last column of the ANOVA table. It relies on Fisher's statistic:

$$F_n = (n - 2) \frac{r^2}{1 - r^2} = \frac{\Sigma_E}{\Sigma_R / (n - 2)}.$$

An equivalent test for the significance of the linear relationship will be seen in the next section. But before we do so, let us summarize the quantities we have already discussed and how they appear in the SPSS outputs.

**Summary of what we have seen so far.** The SPSS outputs provide the following elements:

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,942 <sup>a</sup>	,888	,885	60,665

a. Predictors: (Constant), Surface (m<sup>2</sup>)

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1163593,829	1	1163593,829	316,171	,000 <sup>b</sup>
	Residual	147210,576	40	3680,264		
	Total	1310804,405	41			

a. Dependent Variable: Price (kEuros)

b. Predictors: (Constant), Surface (m<sup>2</sup>)

**Model Summary**

R	R Square	Adjusted R Square	Std. Error of the Estimate
$\sqrt{r^2}$	$r^2$	[see next chapter]	[see next section]

**ANOVA**

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	$\Sigma_E$	1	$\Sigma_E$	$\Sigma_E / (\Sigma_R / (n - 2))$	P-value for $H_0$ : not significant linear model
Residual	$\Sigma_R$	$n - 2$	$\Sigma_R / (n - 2)$		
Total	$\Sigma_T$	$n - 1$			

**An important comment: linear dependence versus causality.** In this chapter we view the linear relations from a statistical viewpoint. We discuss whether there is, or not, a linear dependence between two variables. But we do not discuss causality! What we take as the explanatory variable is not necessarily the true cause of the variations of the dependent variable. There might exist a third variable (called a latent variable), which influences linearly both the explanatory and the dependent variables. Exercise 8.2 will illustrate this.

**Other technical comments.** We gather here important but advanced comments (they are meant for your general culture only).

First, the dependency between the dependent and the explanatory variables could be of a different nature than a linear relationship; e.g.,

$$y_j = a + b \ln(x_j) + e_j \quad \text{or} \quad y_j = a + bx_j^2 + e_j.$$

How do we know? You first have to plot the data on a scatter plot. Then you will see which type of dependency is expected. By transforming the  $x_j$  into  $x'_j$  of a suitable form, e.g.,  $x'_j = \ln(x_j)$  or  $x'_j = x_j^2$ , you will be back to a linear regression of the  $y_j$  in terms of the  $x'_j$ .

Where does the name “regression” come from? Indeed, in English, “regression” means returning to a former state. Exercise 8.1 discusses the first known example of a linear regression, as studied by sir Galton, and therein, the former state is given by an average state, namely, the average height of fathers. Read and solve the exercise to know more about this!

Finally, why do we consider a least-square criterion? Why don't we minimize, e.g., the sum of the absolute errors instead of the squared errors? There are two reasons. First, there are closed-forms solutions for the minimizers, see Section 5. That was important in the past; nowadays, we have good numerical solvers that can give you in no time the solutions to other ways of measuring errors (as long as these measures are convex). Second, we can study what the distributions of these closed-form expressions for the coefficients, etc., are, and can derive tests and confidence intervals for them. Actually, these tests and confidence intervals are the topic of the next section; we will omit the closed-form formulas in the next section, but we will provide them in Section 5 for those who would like to check why we do not provide them in the main text...

## 2. Inferential statistics: significance tests on the coefficients

Beware! This is a very technical section. Read and understand as much as you can. We promise that the next section will be easier to read.

We assume in this section that there exists an underlying (true) model of the form

$$y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j,$$

where the (random) residuals  $\varepsilon_j$  all follow a normal distribution with zero mean and standard deviation  $\sigma_0$ . This means that  $\varepsilon_j/\sigma_0$  follows a standard normal curve, and in particular, is 95% of the time between  $-1.96$  and  $+1.96$ .

The true coefficients  $\alpha_0$  and  $\beta_0$  are estimated by the least-square coefficients  $a$  and  $b$  discussed in the previous section, while an estimate of the standard deviation  $\sigma_0$  is given by  $\hat{\sigma}_n = \sqrt{\Sigma_R/(n-2)}$ .

Now, we can go a step further and provide confidence intervals on  $\alpha_0$  and  $\beta_0$ , as well as tests of

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} \quad \text{and} \quad \begin{cases} H_0 : \alpha_0 = 0 \\ H_1 : \alpha_0 \neq 0 \end{cases}$$

To that end, we introduce two quantities  $S_{a,n}$  and  $S_{b,n}$  (explicit formulas in Section 5), so that the 95% confidence intervals are (approximately) given by

$$[a \pm 1.96 S_{a,n}] \quad \text{and} \quad [b \pm 1.96 S_{b,n}],$$

while the test statistics equal  $a/S_{a,n}$  and  $b/S_{b,n}$ , with (approximately) a normal-curve behavior under  $H_0$ , and larger or smaller values under  $H_1$ .

**Significance test on the slope coefficient.** The test of  $H_0 : \beta_0 = 0$  against  $H_1 : \beta_0 \neq 0$  is exactly a test of significance for the linear regression; indeed,  $\beta_0 = 0$  if and only if the explanatory variable does not contribute to the linear modeling.

This is why in the SPSS outputs, we read in the ANOVA table the same P-value as in the line for the slope coefficient. To put it differently, the test based on Fisher's  $F_n$  statistic leads to the same P-value as the test above on whether  $\beta_0 = 0$  or not.

### 3. General summary: how to read SPSS outputs

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,942 <sup>a</sup>	,888	,885	60,665

a. Predictors: (Constant), Surface (m<sup>2</sup>)

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1163593,829	1	1163593,829	316,171	,000 <sup>b</sup>
	Residual	147210,576	40	3680,264		
	Total	1310804,405	41			

a. Dependent Variable: Price (kEuros)

b. Predictors: (Constant), Surface (m<sup>2</sup>)

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1,861	25,805		,072	,943	-50,293	54,015
	Surface (m <sup>2</sup> )	5,648	,318	,942	17,781	,000	5,006	6,290

a. Dependent Variable: Price (kEuros)

**Model Summary**

R	R Square	Adjusted	Std. Error of the
		R Square	Estimate
$\sqrt{r^2}$	$r^2 = \Sigma_E / \Sigma_T$	[see next chapter]	$\hat{\sigma}_n$

**ANOVA**

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	$\Sigma_E$	1	$\Sigma_E$	$\Sigma_E / (\Sigma_R / (n - 2))$	P-value for $H_0 : \beta_0 = 0$
Residual	$\Sigma_R$	$n - 2$	$\Sigma_R / (n - 2)$		(i.e.: $H_0$ : not significant model)
Total	$\Sigma_T$	$n - 1$			

**Coefficients**

Model	Unstandardized coefficients		Standardized Coefficients	t	Sig.	95% Confidence intervals	
	B	Standard error	Beta			Lower	Upper
Intercept	a	$S_{a,n}$		$a / S_{a,n}$	P-value for $H_0 : \alpha_0 = 0$	$[a \pm 1.96 S_{a,n}]$	
Explanatory	b	$S_{b,n}$	[discard]	$b / S_{b,n}$	P-value for $H_0 : \beta_0 = 0$	$[b \pm 1.96 S_{b,n}]$	

We can now go back to the methodology stated on page 147, items 1 to 4, which we reproduce below. We want to exploit this regression summary and discuss the following points:

1. existence of significant linear relationship;
2. quality of the relationship;
3. writing of the relationship;
4. interpretation of the coefficients.

To that end, we exploit only the cells with a value on the right page (where in each cell, we wrote the item of the discussion where the cell is used); we left blank the other cells.

For instance, on the price-of-apartments example, for items 1 to 4, we proceed as follows.

1. Existence of a significant linear relationship: indeed, the P-value for a null slope coefficient,  $H_0 : \beta_0 = 0$  against  $H_1 : \beta_0 \neq 0$  is very small<sup>2</sup>, we thus reject  $H_0$  and conclude a significant linear relationship.
2. Quality of the relationship: we read  $r^2 = 88.8\%$ , which means that 88.8% of the (variations of) prices can be explained by the (variations of) surfaces. This is a high and truly excellent value for  $r^2$ .
3. Writing of the relationship:

$$\begin{aligned} \text{Price (in kEuros)} &= 1.861 \\ &+ 5.648 \times \text{Surface (in m}^2\text{)} \\ &+ \text{Residual term (with standard deviation: 60.665)} \end{aligned}$$

We should not forget neither the units (kEuros,  $\text{m}^2$ ) nor the residual term (as the linear regression does not explain 100% of the phenomenon, but only 88.8% of it).

4. Interpretation of the coefficients: the slope coefficient of 5.648 can be interpreted as the average price per squared meter in Versailles: 5.648 kEuros = 5,648 euros per squared meter. The relationship makes sense because the price increases with the surface. The intercept term of 1.861 corresponds to a fixed, base amount of 1.861 kEuros = 1,861 euros to be added on the price. This term might be difficult to interpret but actually, it is not significantly different from 0, as the P-value for its test of nullity equals 94.3% (as we can read in the third table, first row). We can recompute the linear relationship by enforcing a null intercept term; SPSS can do it and here is what we get for the new slope coefficient:

**Coefficients<sup>a,b</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 Surface (m <sup>2</sup> )	5,669	,114	,992	49,811	,000

a. Dependent Variable: Price (kEuros)

b. Linear Regression through the Origin

We will see the interpretation and the meaning of the residual term in the next section, where we discuss item 5 of reading a linear relationship, namely: predict future values (on average or individual values) and/or detect outliers.

<sup>2</sup>A P-value is never equal to 0: SPSS writes .000 only because it rounds off the number; it would be safer to write instead  $< .001$ , which some softwares do indeed.



**Model Summary**

R	R Square	Adjusted R Square	Std. Error of the Estimate
	$r^2$ [for 2]		$\hat{\sigma}_n$ [for 3]

**ANOVA**

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression					P-value [for 1]
Residual					
Total					

**Coefficients**

Model	Unstandardized coefficients B	Standard error	Standardized Coefficients Beta	t	Sig.	95% Confidence intervals Lower	Upper
Intercept	a [for 3]				P-value [for 4]		
Explanatory	b [for 3]				P-value [for 1 and 4]		

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1		,888		60,665

a. Predictors: (Constant), Surface (m<sup>2</sup>)**ANOVA<sup>a</sup>**

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression					,000 <sup>b</sup>
Residual					
Total					

a. Dependent Variable: Price (kEuros)

b. Predictors: (Constant), Surface (m<sup>2</sup>)**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients B	Standard error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	Upper Bound
1 (Constant)	1,861				,943		
Surface (m <sup>2</sup> )	5,648				,000		

a. Dependent Variable: Price (kEuros)

## 4. Prediction of new values, detection of outliers

So, we read the linear relationship

$$\begin{aligned}\text{Price (in kEuros)} &= 1.861 \\ &+ 5.648 \times \text{Surface (in m}^2\text{)} \\ &+ \text{Residual term (with standard deviation: 60.665)}\end{aligned}$$

and we now proceed to item 5: prediction of new values (individual values or on average) and detection of outliers. For all these purposes we need to understand what the

Residual term (with standard deviation: 60.665)

means. Remember from the (very technical!) beginning of Section 2 that the residual term followed a normal distribution, with a given standard deviation.

This has two consequences.

- Residuals are centered: their average value is 0.
- Residuals are typically (i.e., 95% of the time) between  $\pm 1.96$  times their standard deviation.

Let us consider a general relationship, of the form

$$y = a + bx + \text{residual term (with standard deviation } s\text{)};$$

you should never write such a “dry” relationship but it is easier for us to explain how to proceed by getting back to this “dry” relationship (and to the heavy formulas of Section 2...).

**Predictions “on average”.** Consider a new value  $x_{n+1}$  for the explanatory variable. (Remember: the linear model was constructed based on  $y_1, \dots, y_n$  and  $x_1, \dots, x_n$ .) What value(s) do we expect for  $y_{n+1}$  on average<sup>3</sup>? The true underlying model says  $\alpha_0 + \beta_0 x_{n+1}$ .

The point estimate we would propose is of course  $\hat{y}_{n+1} = a + b x_{n+1}$ , but as usual, we would prefer to provide a confidence interval. An approximate formula for this confidence interval is

$$a + b x_{n+1} \pm 2 \frac{s}{\sqrt{n}} \quad \text{and even, to be on the safe side} \quad a + b x_{n+1} \pm 4 \frac{s}{\sqrt{n}}.$$

The actual formula is provided in Section 5.

**Individual predictions.** We now ask ourselves “What value(s) do we expect for  $y_{n+1}$ ?” Note that we do not add anymore “on average”. Thus we are shooting for an interval containing all plausible values for  $y_{n+1}$ : an interval such that about 95% of the possible values of  $y_{n+1}$  would lie inside it. We call such an interval a prediction interval.

Such an interval is essentially driven by the residual term with standard deviation  $s$ , which most of the time (about 95% of the time) lies between  $\pm 2s$ .

An approximate formula for this prediction interval is

$$a + b x_{n+1} \pm 2s.$$

The actual formula is provided in Section 5.

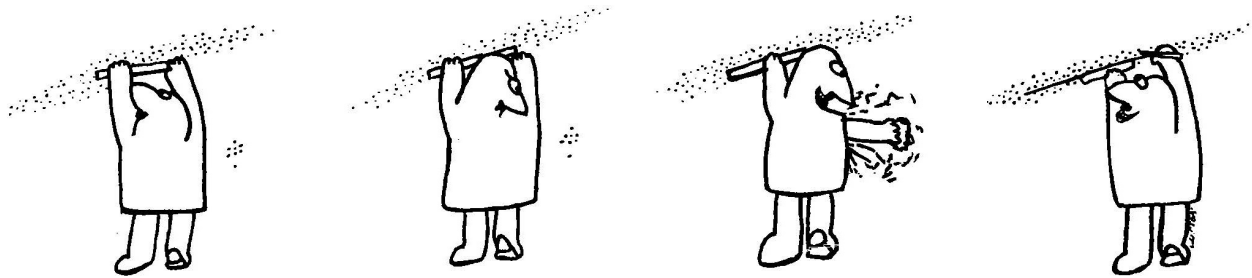
---

<sup>3</sup>This means: if we were averaging out over many possible values of  $y_{n+1}$  based on the same  $x_{n+1}$  considered, what would we get?

**Outliers.** A pair  $(x_j, y_j)$  is an outlier to the model when it does not lie in its own prediction interval, that is, when

$$|e_j| = |\hat{y}_j - y_j| > 2s.$$

If there are a priori and extra-statistical reasons for excluding these outliers from the data set (e.g., in our apartment example: luxury apartments with a view on the Versailles castle), then we can recalculate the linear regression without them. But only in this case, otherwise, we would be cheating! See picture below.

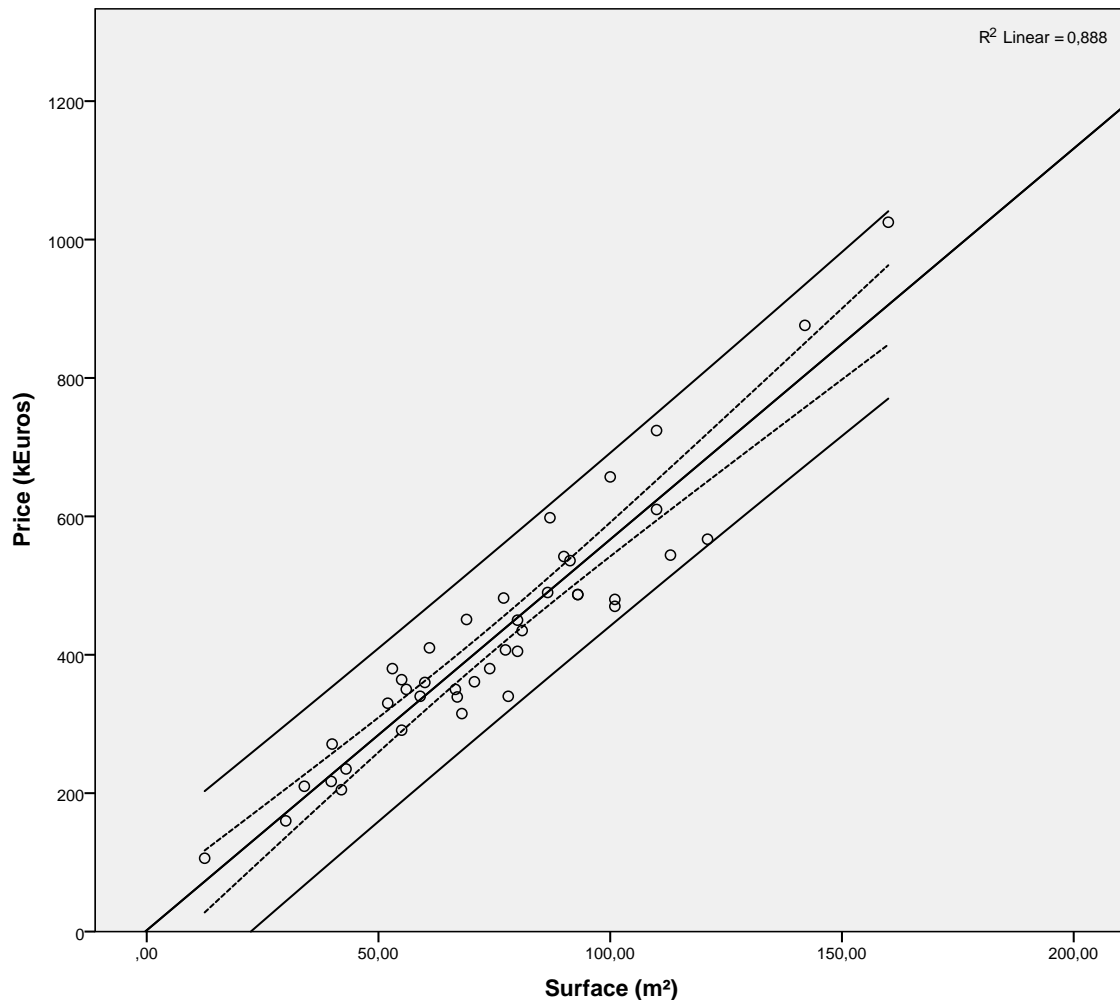


### Application on our data set

[See next page.]

We asked SPSS to draw

- the upper and lower bounds of the 95% confidence intervals on the average prices, as functions of the surface (dashed lines),
- as well as the upper and lower bounds of the 95% prediction interval on the individual prices, as functions of the surface (solid line).



Here, there is no outlier: all points are between the two bounds for prediction intervals.

Let us now compute confidence and prediction intervals for a given value of  $x$ , say, with  $x = 100\text{m}^2$ . We recall that  $a = 1.861$ ,  $b = 5.648$ , and  $s = 60.665$  on this example.

- The point estimate equals  $a + b x = 1.861 + 5.648 \times 100 = 566.661$  kEuros (to be rounded off).
- The confidence interval for  $x = 100$  equals approximately  $566.661 \pm 2 \times 60.665 / \sqrt{42} \approx 566.661 \pm 18.72 \approx 566 \pm 20$ .
- The prediction interval for  $x = 100$  equals approximately  $566.661 \pm 2 \times 60.665 \approx 566.661 \pm 121.33 \approx 566 \pm 122$ .

That is,

- with high confidence,  $100\text{-m}^2$  apartments in Versailles cost  $566 \pm 20$  kEuros on average;
- about 95% of all  $100\text{-m}^2$  apartments in Versailles cost  $566 \pm 122$  kEuros each.

This is indeed what we read on the picture above!

## 5. Mathematical appendix: formulas

We provide here various closed-form expressions omitted from the main text. You should not learn them. Actually, you should even not be reading this section. All that follows is for completeness and to highlight the importance to use a statistical software: using the closed-form expressions and performing the numerical applications with calculators would be a nightmare!

### 5.1. Useful short-hand notation

Useful notation include the  $x$ -sample variance, as well as the  $(x, y)$ -sample covariance,

$$\text{Var}(x_1^n) = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}_n)^2 \quad \text{and} \quad \text{Cov}(x_1^n, y_1^n) = \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y}_n)(x_j - \bar{x}_n).$$

### 5.2. Coefficients

The intercept coefficient  $a$  and the slope coefficient  $b$  are then given by

$$b = \frac{\text{Cov}(x_1^n, y_1^n)}{\text{Var}(x_1^n)} \quad \text{and} \quad a = \bar{y}_n - b\bar{x}_n.$$

### 5.3. Standard errors for the coefficients

The standard errors  $S_{a,n}$  and  $S_{b,n}$  for the estimated coefficients  $a$  and  $b$  equal

$$S_{a,n} = \sqrt{\frac{\hat{\sigma}_n^2}{n} \left( 1 + \frac{(\bar{x}_n)^2}{\text{Var}(x_1^n)} \right)} \quad \text{and} \quad S_{b,n} = \sqrt{\frac{\hat{\sigma}_n^2}{n \text{Var}(x_1^n)}}.$$

### 5.4. Confidence interval on $\alpha_0 + \beta_0 x$

We provided the approximate formula  $a + b x \pm 2s/\sqrt{n}$ , but the actual, more complex, formula is

$$\left[ a + b x \pm t_{n-2, 97.5\%} \sqrt{\frac{\hat{\sigma}_n^2}{n} h_{x,n}} \right] \quad \text{where} \quad h_{x,n} = 1 + \frac{1}{\text{Var}(x_1^n)} (x - \bar{x}_n)^2$$

and where  $t_{n-2, 97.5\%}$  denotes the Student's quantile of order 97.5% (approximately equal to 1.96).

In practice (on real data), the quantity  $h_{x,n}$  typically lies between 1 and 4 when no extrapolation is performed, i.e., when  $x$  belongs to the interval generated by the  $x_j$ . The quantity  $\pm t_{n-2, 97.5\%} \sqrt{h_{x,n}/n}$  thus lies between  $\pm 2/\sqrt{n}$  and  $\pm 4/\sqrt{n}$ , hence the formulas for approximate confidence intervals on page 158.

Note that we use somewhat indifferently the notation  $s$  or  $\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2}$  for the estimated standard deviation of the residuals.

### 5.5. Prediction interval on $\alpha_0 + \beta_0 x + \varepsilon$

With the notation above, a 95%-probability prediction interval is given by

$$\left[ a + b x \pm t_{n-2, 1-97.5\%} \sqrt{\hat{\sigma}_n^2 \left( 1 + \frac{h_{x,n}}{n} \right)} \right].$$

The factor  $1 + h_{x,n}/n$  is not much larger than 1 and  $t_{n-2, 97.5\%}$  approximately equals 1.96, hence the margin  $\pm 2s$  used in the approximate prediction intervals on page 158.

## 6. Elementary exercises

**Elementary exercise 8.1.** You learned in elementary school that each year, a tree forms new cells, arranged in concentric circles called annual rings or annual growth rings. Therefore, we should observe a linear increase of the circumferences of trees over time. Is it the case indeed? The data set of the next page recorded the growth of five orange trees at seven given points in time. Units are days for the age and millimeters for the circumferences.

We want to explain the circumference (y variable, also called dependent variable, in millimeters) as an affine function of the age (x variable, also called independent or explanatory variable, in days).

1. Start with the exploitation of the relevant regression-summary output. Replicate the exploitation performed on page 147 for the apartment-price data set, as follows.

- (a) Show that the linear regression computed is statistically significant: indicate the P-value for the test of  $H_0$  : nullity versus  $H_1$  : non-nullity of the slope coefficient.
- (b) What  $r^2$  do you read? Provide a number between 0% and 100%. Is it a good or a bad value? Write a complete sentence containing this number and explaining what it means in terms of explained variations.
- (c) Write the linear relationship read in the SPSS outputs, by filling the following formula:

$$\begin{aligned}\text{Circumference (in millimeters)} &= \dots \\ &+ \dots \times \text{Age (in days)} \\ &+ \text{Residual term (with standard deviation: } \dots)\end{aligned}$$

- (d) Would a positive intercept be easy to interpret? (Recall that the intercept is the additive constant in the right-hand side affine relationship.) Is the intercept written in the previous question significantly different from 0?

On the other hand, what does the slope coefficient quantify? Is it significantly different from 0?

2. From the last two sub-questions, it should be easy to answer the following question: “What is the approximate average growth rate in centimeters per year?” Note that the adjective “average” is used here, and that we changed the units (why don’t we stick to the original units, that is, to millimeters per day?).

Provide also a more precise answer featuring a confidence interval on this average growth rate.

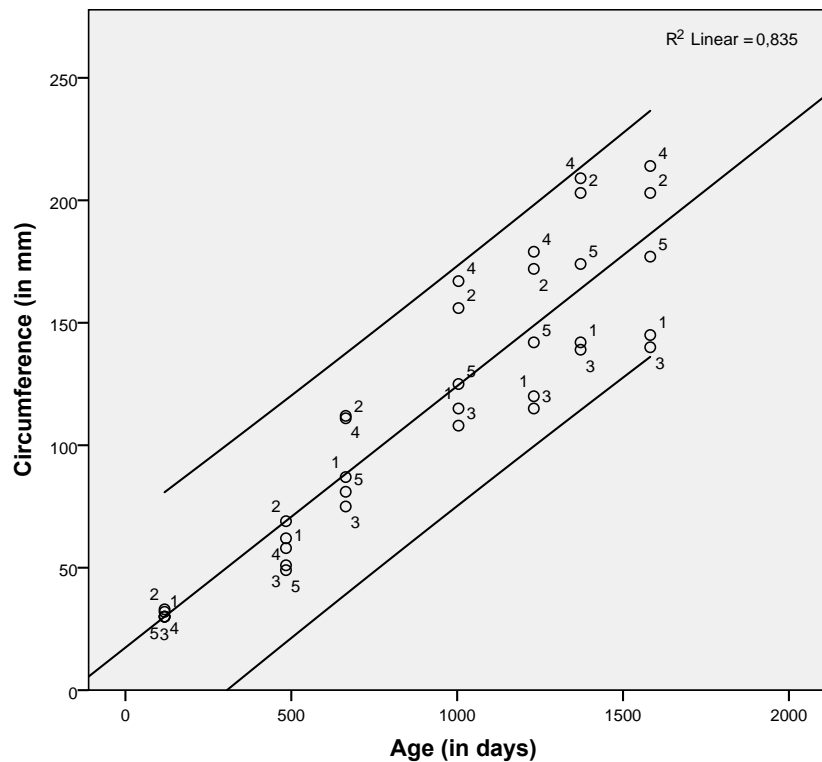
3. According to this model, what should the circumference of a 2-year-and-4-month-old orange tree be? Provide an answer “on average” first, and then about “95% of such trees”, by answering the following sub-questions and by considering that such a tree is aged  $852 \approx (2 + 1/3) \times 365$  days.

- (a) Provide a point estimate on the expected circumference of such a tree; do so by discarding the residual term in the linear relationship above.
- (b) Provide now a confidence interval on the average expected circumference of a tree of that age; to do so, consider the residual term and use the  $\pm 2s/\sqrt{n}$  margin of error. Provide an answer in plain words containing the word “average.”
- (c) Provide finally a prediction interval on the individual expected circumference of a tree of that age; this time, resort to the  $\pm 2s$  correction. Provide an answer in plain words with “95% of the trees” as a subject.

4. Take a look at the scatter plot: what do the upper and lower solid lines correspond to? What can you say in terms of outliers in the data?

Orange.sav [DataSet1] - IBM SPSS Statistics Data

	Tree	Age	Circumference
1	1	118	30
2	1	484	58
3	1	664	87
4	1	1004	115
5	1	1231	120
6	1	1372	142
7	1	1582	145
8	2	118	33
9	2	484	69
10	2	664	111
11	2	1004	156
12	2	1231	172
13	2	1372	203
14	2	1582	203
15	3	118	30
16	3	484	51
17	3	664	75
18	3	1004	108
19	3	1231	115
20	3	1372	139
21	3	1582	140
22	4	118	32
23	4	484	62
24	4	664	112
25	4	1004	167
26	4	1231	179
27	4	1372	209
28	4	1582	214
29	5	118	30
30	5	484	49
31	5	664	81
32	5	1004	125
33	5	1231	142
34	5	1372	174
35	5	1582	177



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,914 <sup>a</sup>	,835	,830	23,738

a. Predictors: (Constant), Age (in days)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	93771,541	1	93771,541	166,416	,000 <sup>b</sup>
	Residual	18594,744	33	563,477		
	Total	112366,286	34			

a. Dependent Variable: Circumference (in mm)

b. Predictors: (Constant), Age (in days)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	17,400	8,623		2,018	,052
	Age (in days)	,107	,008	,914	12,900	,000

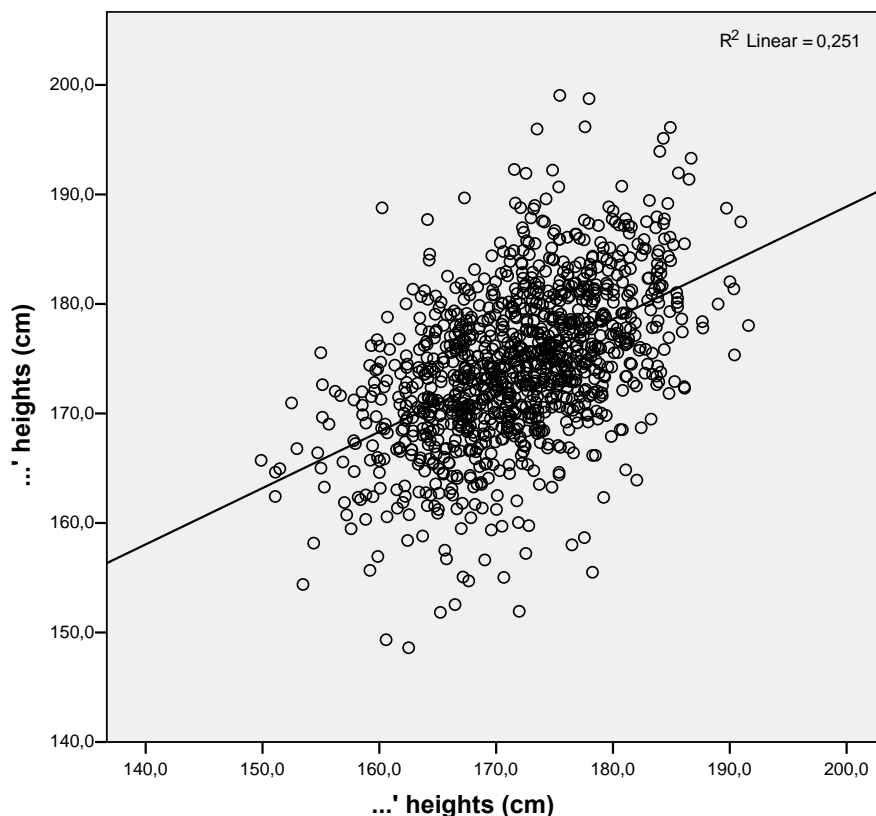
a. Dependent Variable: Circumference (in mm)

## 7. More advanced exercises (quiz-like exercises)

**Advanced exercise 8.1** (The historical example of regression). Where does the name “regression” come from? From a regression toward an average value, as illustrated in this exercise (based on real, historical data): Sir Galton, a British scientist (1822–1911) studied the heights of sons in relation to their fathers' heights. He noticed in some intuitive, empirical way a regression toward the average height: the taller (respectively, shorter) fathers had sons shorter (respectively, taller) than them. We provide below a scatter plot of the data, as well as some descriptive statistics. On the next page, we also reproduce two regression summaries (but only one of them should be exploited, see below).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Fathers' heights (cm)	1078	149,9	191,6	171,925	6,9720
Sons' heights (cm)	1078	148,6	199,0	174,458	7,1493
Valid N (listwise)	1078				



1. What is the dependent variable, what is the independent (explanatory) one? Set the titles of the axes accordingly and determine which regression summary to read.
2. Exploit this regression summary (existence of significant linear relationship, quality of the relationship, writing of the relationship; can you find any obvious interpretation of the coefficients?).
3. Transform the above linear relationship for it to match Galton's intuition of a regression toward an average height. Was Galton right?

Hint: use the fact that the regression line goes through the point  $\bar{x}$ ,  $\bar{y}$  given by the averages of each variable.



## Regression output #1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,501 <sup>a</sup>	,251	,251	6,0353

a. Predictors: (Constant), Sons' heights (cm)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13157,942	1	13157,942	361,235	,000 <sup>b</sup>
	Residual	39193,204	1076	36,425		
	Total	52351,146	1077			

a. Dependent Variable: Fathers' heights (cm)

b. Predictors: (Constant), Sons' heights (cm)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	86,633	4,491		19,289	,000	77,820	95,446
	Sons' heights (cm)	,489	,026	,501	19,006	,000	,438	,539

a. Dependent Variable: Fathers' heights (cm)

## Regression output #2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,501 <sup>a</sup>	,251	,251	6,1889

a. Predictors: (Constant), Fathers' heights (cm)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13835,971	1	13835,971	361,235	,000 <sup>b</sup>
	Residual	41212,829	1076	38,302		
	Total	55048,800	1077			

a. Dependent Variable: Sons' heights (cm)

b. Predictors: (Constant), Fathers' heights (cm)

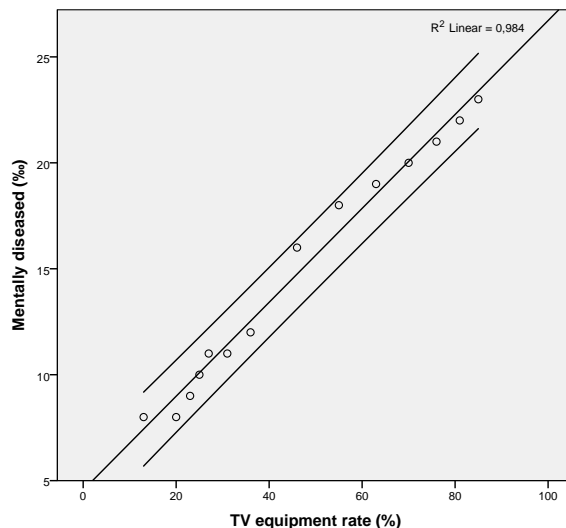
Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	86,072	4,654		18,493	,000	76,940	95,204
	Fathers' heights (cm)	,514	,027	,501	19,006	,000	,461	,567

a. Dependent Variable: Sons' heights (cm)

**Advanced exercise 8.2** (An example of a spurious correlation). Can too much television make people mentally ill? We consider data triplets for Great Britain: year, corresponding equipment rate in TV sets (in %) and rate of mentally diseased persons (in ‰). The whole data set is reproduced below; it features 14 data points (triplets).

1. Consider first that the dependent variable is the mental-disease rate and that the independent (explanatory) variable is the TV-equipment rate. Exploit the relevant regression outputs (scatter plot, regression summary): existence of significant linear relationship, quality of the relationship, writing of the relationship, interpretation of the coefficients.
2. Can we / should we conclude that television makes people mentally ill? Show that some third variable explains both variables considered in the previous question. (This third variable is called a latent variable.)
3. Read and enjoy the supplementary material about spurious correlations: the blog post “does eating chocolate make you a serial killer?” explaining that people have been making fun at a research article in medicine showing a linear relationship between Nobel prizes and chocolate consumption (see the scatter plot extracted from the said article). Other fun spurious correlations are listed on the next pages.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,992 <sup>a</sup>	,984	,983	,728

a. Predictors: (Constant), TV equipment rate (%)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	393,361	1	393,361	742,976	,000 <sup>b</sup>
	Residual	6,353	12	,529		
	Total	399,714	13			

a. Dependent Variable: Mentally diseased (‰)

b. Predictors: (Constant), TV equipment rate (%)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,552	,425		10,707	,000
	TV equipment rate (%)	,222	,008	,992	27,258	,000

a. Dependent Variable: Mentally diseased (‰)

TV.sav [DataSet3] - IBM SPSS Statistics Data Editor				
	Year	TV	Diseased	
1	1970	13	8	
2	1971	20	8	
3	1972	23	9	
4	1973	25	10	
5	1974	27	11	
6	1975	31	11	
7	1976	36	12	
8	1977	46	16	
9	1978	55	18	
10	1979	63	19	
11	1980	70	20	
12	1981	76	21	
13	1982	81	22	
14	1983	85	23	
15				

**Explanation: part 1/2****Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,982 <sup>a</sup>	,964	,960	1,103

a. Predictors: (Constant), Year

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	385,125	1	385,125	316,780	,000 <sup>b</sup>
	Residual	14,589	12	1,216		
	Total	399,714	13			

a. Dependent Variable: Mentally diseased (‰)

b. Predictors: (Constant), Year

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2556,765	144,487		-17,695	,000
	Year	1,301	,073	,982	17,798	,000

a. Dependent Variable: Mentally diseased (‰)

**Explanation: part 2/2****Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,986 <sup>a</sup>	,972	,969	4,361

a. Predictors: (Constant), Year

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7781,232	1	7781,232	409,057	,000 <sup>b</sup>
	Residual	228,268	12	19,022		
	Total	8009,500	13			

a. Dependent Variable: TV equipment rate (%)

b. Predictors: (Constant), Year

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-11512,767	571,530		-20,144	,000
	Year	5,848	,289	,986	20,225	,000

a. Dependent Variable: TV equipment rate (%)

21st November 2012

## Does eating chocolate make you a serial killer?

You may have seen the news item on the BBC website

[[http://www.bbc.co.uk/news/magazine-](http://www.bbc.co.uk/news/magazine-20356613)

[20356613](http://www.bbc.co.uk/news/magazine-20356613)] and several

other news outlets that suggested that eating chocolate may make you more intelligent!

This originates from a paper by Franz Messerli that was published in the New England Journal of Medicine [1

[<http://www.nejm.org/doi/full/10.1056/nejmon1211064>]

[1056/nejmon1211064]

]. He obtained data from several countries on the number of Nobel laureates per-capita and plotted these data in relation to the annual per-capita chocolate consumption. Why, you ask? Well, there is limited evidence that dietary flavonoids improve cognitive function and a subgroup of flavonoids known as flavanols are widely present in cocoa, green tea, red wine, and some fruits.



[<http://www.bbc.co.uk/news/magazine-20356613>]

There was a significant linear correlation ( $r=0.791$ ,  $P<0.0001$ ) between chocolate consumption and the number of Nobel laureates. The good doctor says, "Switzerland was the top performer in terms of both the number of Nobel laureates and chocolate consumption. The slope of the regression line allows us to estimate that it would take about 0.4 kg of chocolate per-capita per year to increase the number of Nobel laureates in a given country by 1". His conclusion was that, "Chocolate consumption enhances cognitive function, which is a *sine qua non* for winning the Nobel Prize, and it closely correlates with the number of Nobel laureates in each country. It remains to be determined whether the consumption of chocolate is the underlying mechanism for the observed association with improved cognitive function".

Now call me a skeptic if you like but I just don't believe this is true. So I was really pleased that after just a little searching I came across a blog article by James Winters and Seán Roberts [<http://replicatedtypo.com/chocolate-consumption-traffic-accidents-and-serial-killers/5718.html>], entitled "Chocolate Consumption, Traffic Accidents and Serial Killers". They reproduced Messerli's findings but also showed that chocolate consumption per-capita is significantly correlated with the (log-transformed) number of serial and rampage killers per-capita ( $r = 0.52$ ,  $p=0.02$ ). Are we to infer that all this chocolate consumption in Switzerland, Germany and the UK is causing some people to loose the plot? Maybe not.

More importantly, Winters and Roberts showed that when they controlled for per-capita GDP and mean air temperature, chocolate consumption was not a significant predictor of the number of Nobel laureates. Countries with higher GDP and lower mean outside air temperatures tended to have higher numbers of Nobel laureates per capita. So a better explanation for these observations might be that Nobel laureates tend to work in affluent northern hemisphere countries that can afford to support research and where the population buys lots of chocolate.

## Reference

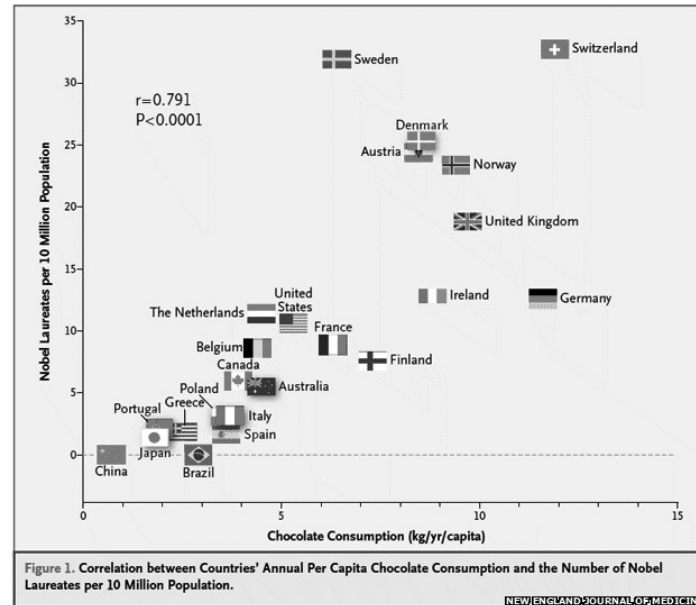
1. Franz H. Messerli FH. (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates [<http://www.nejm.org/doi/full/10.1056/nejmon1211064>] N Engl J Med; 367:1562-1564 October 18, 2012 DOI: 10.1056/NEJMon1211064

Posted 21st November 2012 by John Cherrie

Location: Edinburgh, City of Edinburgh, UK

Retrieved from

<http://johncherrie.blogspot.fr/2012/11/does-eating-chocolate-make-you-serial.html>



Other examples at <https://www.tylervigen.com/spurious-correlations>

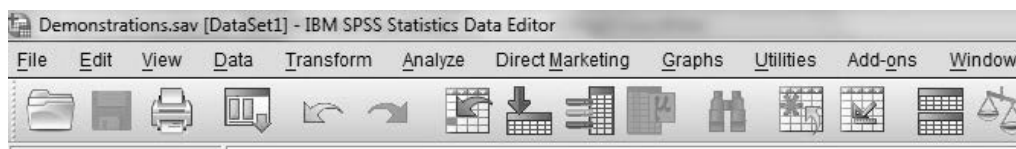
■ TABLE 11.2 Examples of Spurious Relationships

Observed Spurious Relationship*	Reason for the Relationship (the Third Variable)
Amount of ice cream sold and deaths by drownings (Moore, 1993)	Season: Ice cream sales and drownings tend to be high during the warm months of the year.
Size of left hand and size of right hand	Genetics: The size of both hands is due to genetic makeup.
Height of sons and height of daughters (Davis, 1985)	Genetics: Heights of sons and daughters are both due to their parents' genetic makeup.
Ministers' salaries and price of vodka	Area (i.e., urban or rural): In urban areas, prices and salaries tend to be higher.
Shoe size and reading performance for elementary school children	Age: Older children have larger shoe sizes and read better.
Number of doctors in region and number of people dying from disease	Population density: In highly dense areas, there are more doctors and more people die.
Number of police officers and number of crimes (Glass & Hopkins, 1996)	Population density: In highly dense areas, there are more police officers and more crimes.
Number of homicides and number of churches	Population density: In highly dense areas, there are more homicides and more churches.
Number of storks sighted and the population of Oldenburg, Germany, over a six-year period (Box, Hunter, & Hunter, 1978)	Time: Both variables were increasing over time.
Number of public libraries and the amount of drug use	Time: Both were increasing during the 1970s.
Teachers' salaries and the price of liquor (Moore and McCabe, 1993)	Time: Both tend to increase over time.
Tea drinking and lung cancer	Smoking: Tea drinkers have a lower risk only because they smoke less.

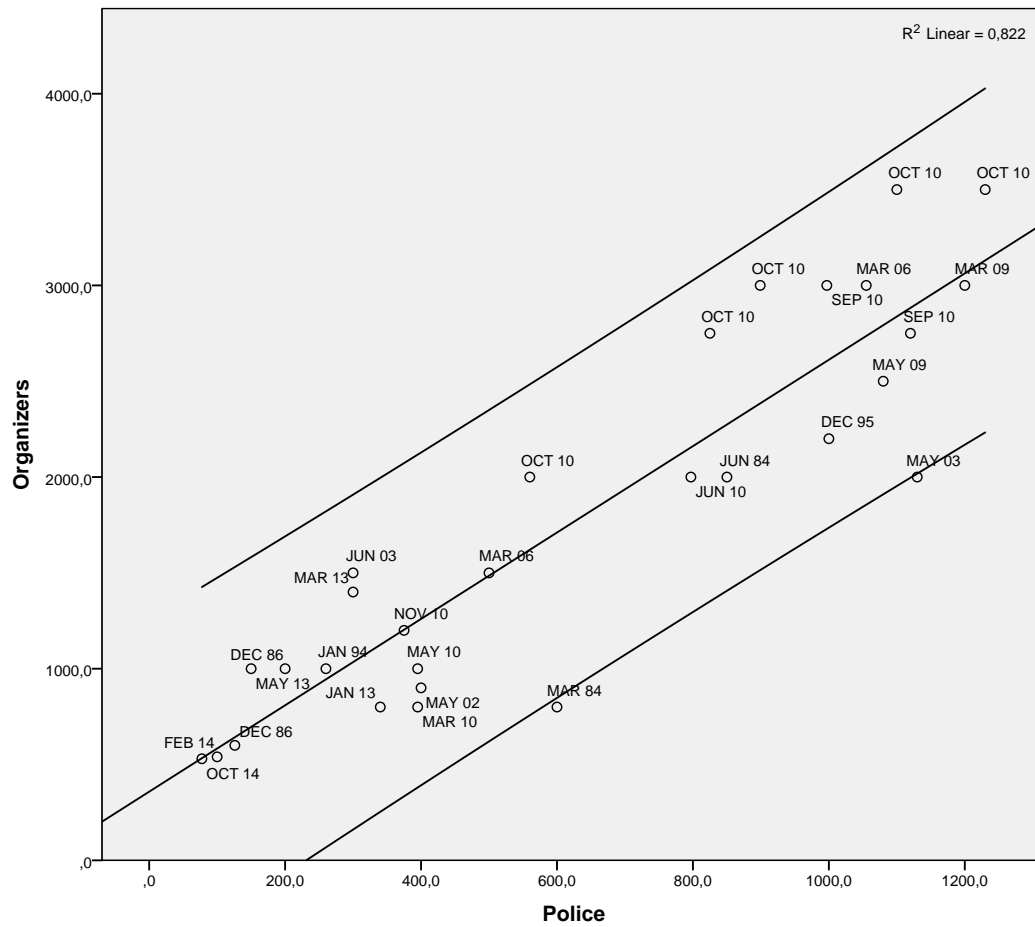
\*All but one of the spurious relationships in the first column shows a positive relationship. That is, as one of the variables increases, the other variable also increases. The one negative relationship is the relationship between tea drinking and lung cancer.

**Advanced exercise 8.3** (Some French politics: demonstrations). The data below concern some mass demonstrations organized in France in the 1984 – 2014 period (thus not including the 2016 demonstrations against the labor-law reform, which anyway were not as massive as the ones detailed below). For each demonstration, we report the number of participants as counted by the organizers and by the police (in France, it is the police's mission to count the participants on behalf of the government). The units are (obviously) thousands of people.

1. Exploit the relevant regression outputs (scatter plot, regression summary) in the usual way: existence of significant linear relationship, quality of the relationship, writing of the relationship, interpretation of the coefficients, etc.
2. Pay particular attention to the slope coefficient: can you prove that there is a significant difference in the two sets of counts?
3. How many participants do we expect the organizers to report when the police indicates that there were 500,000 demonstrators? Provide an answer “on average” first, and then about “95% of such demonstrations”.



	Date	Cause	Police	Organizers
1	MAR 84	Private school system	600	800
2	JUN 84	Private school system	850	2000
3	DEC 86	University reform	200	1000
4	DEC 86	Death of Malik Ousekine	126	600
5	JAN 94	Private school system (again)	260	1000
6	DEC 95	Juppé's reform on pension system	1000	2200
7	MAY 02	Le Pen at 2nd round of presidential elections	400	900
8	MAY 03	Fillon's reform on pension system	1130	2000
9	JUN 03	Fillon's reform on pension system	300	1500
10	MAR 06	Lower hourly wage for young adults [CPE]	500	1500
11	MAR 06	Lower hourly wage for young adults [CPE]	1055	3000
12	MAY 09	Cost of life	1080	2500
13	MAR 09	Cost of life	1200	3000
14	MAR 10	Woerth's reform on pension system	395	800
15	MAY 10	Woerth's reform on pension system	395	1000
16	JUN 10	Woerth's reform on pension system	797	2000
17	SEP 10	Woerth's reform on pension system	1120	2750
18	SEP 10	Woerth's reform on pension system	997	3000
19	OCT 10	Woerth's reform on pension system	899	3000
20	OCT 10	Woerth's reform on pension system	1230	3500
21	OCT 10	Woerth's reform on pension system	825	2750
22	OCT 10	Woerth's reform on pension system	1100	3500
23	OCT 10	Woerth's reform on pension system	560	2000
24	NOV 10	Woerth's reform on pension system	375	1200
25	JAN 13	Manif Pour Tous (= against equal marriage)	340	800
26	MAR 13	Manif Pour Tous (= against equal marriage)	300	1400
27	MAY 13	Manif Pour Tous (= against equal marriage)	150	1000
28	FEB 14	Manif Pour Tous (= against equal marriage)	100	540
29	OCT 14	Manif Pour Tous (= against equal marriage)	78	530



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,906 <sup>a</sup>	,822	,815	413,4445

a. Predictors: (Constant), Police

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21243842,21	1	21243842,21	124,279	,000 <sup>b</sup>
	Residual	4615281,926	27	170936,368		
	Total	25859124,14	28			

a. Dependent Variable: Organizers

b. Predictors: (Constant), Police

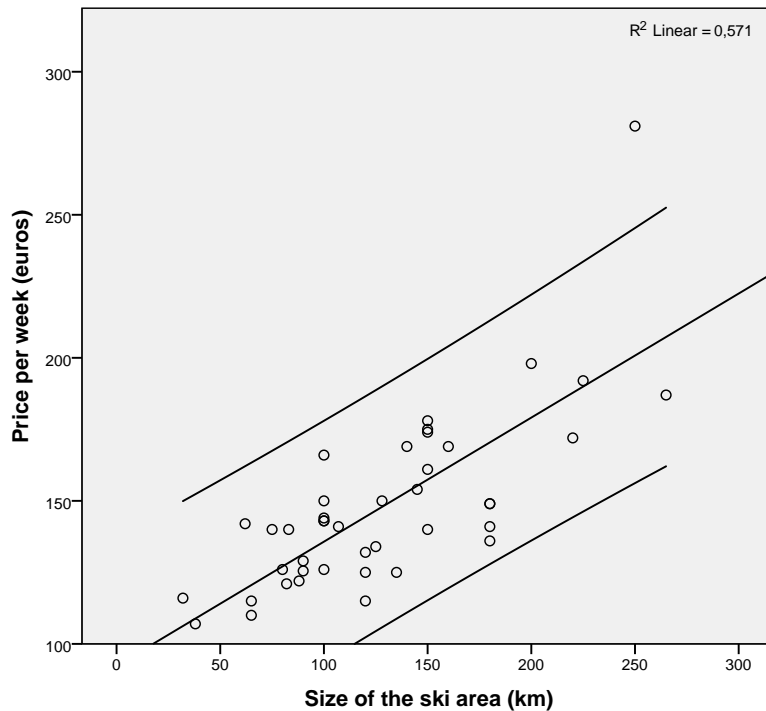
Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	358,197	149,261		2,400	,024	51,938	664,455
	Police	2,254	,202	,906	11,148	,000	1,839	2,669

a. Dependent Variable: Organizers

**Advanced exercise 8.4 (Prices of ski passes).** Data collected in this exercise by a dedicated statistics instructor, namely, Xavier Boute, date back to 2008. For 42 Alpine ski resorts, he noted the size of the ski area (in kilometers) and the price of the weekly ski pass (in euros). The regression summary is provided below.

1. Exploit this regression summary: existence of significant linear relationship, quality of the relationship, writing of the relationship, interpretation of the coefficients.
2. Show by means of a calculus that the Serre Chevalier resort (size: 250 km, price: 281 euros) is an outlier. Actually, in view of the prediction intervals, plotted in solid line on the scatter plot, it is the only outlier.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,756 <sup>a</sup>	,571	,560	20,594

a. Predictors: (Constant), Size of the ski area (km)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22596,541	1	22596,541	53,280	,000 <sup>b</sup>
	Residual	16964,418	40	424,110		
	Total	39560,958	41			

a. Dependent Variable: Price per week (euros)

b. Predictors: (Constant), Size of the ski area (km)

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	92,366	8,247		11,200	,000
	Size of the ski area (km)	,434	,059	,756	7,299	,000

a. Dependent Variable: Price per week (euros)



## Multiple linear regression

In this chapter, we extend the results of the previous chapter to the case where several independent (explanatory) variables can be considered, which we call multiple linear regression.

That is, we consider observations  $y_1, \dots, y_n$  of a dependent variable, and corresponding values for the dependent variables. A first series of independent variables consists of  $x_1^{(1)}, \dots, x_n^{(1)}$ ; a second such series consists of  $x_1^{(2)}, \dots, x_n^{(2)}$ ; the  $k$ -th and last series of them is  $x_1^{(k)}, \dots, x_n^{(k)}$ . We now still have one intercept coefficient  $\alpha_0$ , one residual term, but  $k$  slope coefficients  $\beta_0^{(k)}$  to form our linear relationship. More precisely, we are thinking of a statistical model of the form, for  $j = 1, \dots, n$ ,

$$y_j = \alpha_0 + \beta_0^{(1)} x_j^{(1)} + \beta_0^{(2)} x_j^{(2)} + \dots + \beta_0^{(k)} x_j^{(k)} + \varepsilon_j,$$

where the residuals  $\varepsilon_j$  all follow a normal distribution with standard deviation  $\sigma_0$ .

**Disclaimer.** We are so sorry for the heavy notation! You should know that in most textbooks, it becomes even worse then, because matrices are introduced to write in a compact way the  $n$  equalities stated above. This is not a direction that we are ready to consider in this statistics course built for students with no background in mathematics.

**Example.** Our running example in these notes considers the price of the weekly ski pass for 98 ski resorts located in France. We want to model that price as a linear function of three characteristics of the resorts, indicated in the second part of the following table.

Variable	Definition	Units
ResortName	Resort name	[N/A]
SkiPassPrice	Ski pass price (for 7 days)	euros
MaxAltitude	Maximum altitude of the resort	meters
NumberSkiLifts	Number of ski lifts	[N/A]
NumberSlopes	Number of slopes	[N/A]

An excerpt of the data is reproduced at the top of the next page.

As in the previous chapter, our main aim is to explain how to read the multiple-regression outputs, like the one reproduced on the next page. We will mostly do so by explaining the additional concepts with respect to simple-regression outputs.

Ski-short.sav [DataSet1] - IBM SPSS Statistics Data Editor

	ResortName	SkiPassPrice	MaxAltitude	NumberSkiLifts	NumberSlopes
1	Alpe d'Huez	197	3330	85	108
2	Alpe du Grand Serre	107	2200	20	34
3	Arches Beaufort	110	2300	15	30
4	Auris	104	2175	15	19
5	Auron	120	2450	27	38
6	Aussois	90	2750	11	21
7	Autrans	78	1710	15	18
8	Avoriaz	168	2460	39	44
9	Bessans	70	2200	4	4
10	Beuil	130	2100	27	59
11	Bonneval sur arc	110	3050	10	18
12	Chamonix	241	3840	49	69
13	Chamrousse	136	2255	26	36
14	Chatel	153	2200	39	40
15	Combloux	110	1850	80	35
16	Courchevel	192	3200	67	102
17	Crest Voland Corennoz	102	1950	17	26
18	Eaux Bonnes Courette	100	2400	23	30

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,847 <sup>a</sup>	,717	,708	20,500

a. Predictors: (Constant), Number of slopes, Maximum altitude of the resort, Number of ski lifts

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	100300,605	3	33433,535	79,553	,000 <sup>b</sup>
	Residual	39505,283	94	420,269		
	Total	139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort, Number of ski lifts

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,626	9,685		2,956	,004
	Maximum altitude of the resort	,031	,005	,437	6,657	,000
	Number of ski lifts	-,029	,071	-,032	-,410	,683
	Number of slopes	,630	,101	,548	6,265	,000

a. Dependent Variable: Ski pass price (for 7 days)

## 1. What to read in a given, single output

The exploitation of a regression output still follows the same 5-step methodology, but some steps, like steps 1 and 4, require more work:

1. assessment of the statistical validity of the model,
  - (a) global validity of the model (i.e., at least one variable is useful);
  - (b) marginal validity of each variable of the model (i.e., the model is parsimonious enough, each variable is useful);
2. quality of the model;
3. writing the model;
4. economic/political validity of the model: interpretation of the (signs of the) coefficients;
5. outlier detection, prediction of new values.

What we read where and for which step is summarized below (and discussed on the next page). The (true, underlying) coefficients  $\alpha_0$  and  $\beta_0^{(1)}, \dots, \beta_0^{(k)}$ , as well as the standard deviation  $\sigma_0$ , are respectively estimated by  $a$  and  $b^{(1)}, \dots, b^{(k)}$ , as well as  $s$ .

**Model Summary**

R	R Square	Adjusted R Square	Std. Error of the Estimate
	$r^2$ [for 2]		$s$ [for 3 and 5]

**ANOVA**

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression		$k$			P-value [for 1(a)]
Residual		$n - 1 - k$			
Total		$n - 1$ [for 5]			

**Coefficients**

Model	Unstandardized coefficients	Standardized Coefficients	t	Sig.	95% Confidence intervals
	B	Standard error	Beta		Lower Upper
Intercept	$a$ [for 3]			P-value [for 4]	
Explanatory 1	$b^{(1)}$ [for 3]			P-value [for 1(b)]	
Explanatory 2	$b^{(2)}$ [for 3]			P-value [for 1(b)]	
...	...			...	
Explanatory k	$b^{(k)}$ [for 3]			P-value [for 1(b)]	

We first mention that the coefficients  $\alpha_0$  and  $\beta_0^{(1)}, \beta_0^{(2)}, \dots, \beta_0^{(k)}$  are estimated by minimizing a squared error, just like in the case of simple linear regression. This leads to coefficients estimated on the data and denoted, as we indicated, by  $a$  and  $b^{(1)}, b^{(2)}, \dots, b^{(k)}$ . This also allows to get some estimation  $s$  of the standard deviation  $\sigma_0$ .

**Step 1: Statistical validity of the model — global and marginal.** This step is subdivided into two steps. If any substep the hypothesis  $H_0$  is not rejected, then we do not consider the model statistically valid and we do not further study it.

**Step 1(a): Global validity of the model.** It is about testing  $H_0$  : no explanatory variable contributes significantly to the linear explanation of the dependent variable, versus  $H_1$  : at least one does. In more technical terms, this corresponds to testing that  $H_0$  : all slope coefficients are null, versus  $H_1$  : at least one is different from zero,

$$\begin{cases} H_0 : \beta_0^{(1)} = \beta_0^{(2)} = \dots = \beta_0^{(k)} = 0 \\ H_1 : \beta_0^{(1)} \neq 0, \text{ or } \beta_0^{(2)} \neq 0, \dots, \text{ or } \beta_0^{(k)} \neq 0 \end{cases}$$

If we fail to reject  $H_0$ , then no explanatory variable contributes significantly to the linear modeling and we do not further study the model.

**Step 1(b): Marginal validity of each variable of the model.** For each fixed explanatory variable, we test  $H_0$  : this variable provides no useful contribution to the multiple-linear model with respect to the model in which it would be omitted, versus  $H_1$  : it does. We thus do not test the intrinsic contribution of a variable, but its incremental contribution to the model, given the other explanatory variables. We test here that the model is parsimonious enough: that it does not contain unnecessary variables. This is important as we always shoot for simplicity when exploiting and interpreting models. More formally, we could say that we test, for each explanatory variable  $i = 1, \dots, k$ :

$$\begin{cases} H_0^{(i)} : \beta_0^{(i)} = 0 & \text{in the model } y_j = \alpha_0 + \beta_0^{(1)} x_j^{(1)} + \beta_0^{(2)} x_j^{(2)} + \dots + \beta_0^{(k)} x_j^{(k)} + \varepsilon_j \\ H_1^{(i)} : \beta_0^{(i)} \neq 0 & \text{in this model} \end{cases}$$

If any of the  $k$  hypotheses  $H_0^{(i)}$  fails to be rejected, then the model is not parsimonious enough, and at least one of the explanatory variables should be omitted. We therefore do not further study this model; we will discuss in the second part of these notes how to then simplify the model and study instead a model with fewer explanatory variables.

There are two main reasons why a variable could bring no significant contribution to the linear modeling given the other variables:

- an intrinsic lack of significance (the variable has nothing to do with the problem, or at least, has no linear influence on the dependent variable);
- a redundancy issue (the variable is highly correlated with one or several other explanatory variables). In this case, given that one of them is in the model, it is unnecessary to consider the other variable(s). But at least one of them should probably be considered, just not all of them at the same time.

Because of the possible redundancy issue, the simplification of a non-marginally valid model should be done with care, as we will explain below (basically: variable after variable, never by dropping two variables or more at a time).

**Step 2: Quality of the model.** It is still measured by the coefficient of determination  $r^2$ , which is still defined as the ratio  $\Sigma_E / \Sigma_T$  of the squared variations  $\Sigma_E$  of the observations as reconstructed (explained) by the model, to the squared variations  $\Sigma_T$  of the original observations.

In a much less technical manner, the coefficient of determination  $r^2$  represents the fraction of the total variability explained (recovered) by the model.

**Step 3: Writing the relationship proposed by the model.** We recommend you do it as follows, because this way it is understandable even by colleagues that are not familiar with statistics. And we also recommend on the other hand that you do not use technical notation as  $y_j$  or  $x_j^{(i)}$ : it is tempting to write the results in a mathy way, but it is most likely that you will not do it in an accurate way and/or that the colleagues you are writing it for will not understand what you write!

$$\begin{aligned} \text{Dependent variable (write its units)} &= a \\ &+ b^{(1)} \times \text{Explanatory variable \#1 (write its units)} \\ &+ b^{(2)} \times \text{Explanatory variable \#2 (write its units)} \\ &+ \dots \\ &+ b^{(k)} \times \text{Explanatory variable \#k (write its units)} \\ &+ \text{Residual term (with standard deviation: } s) \end{aligned}$$

Do not forget: the units and the residual term. (If you omit the residual term, then you are stating a deterministic model, while our regression models are only statistical models, that come with coefficients of determination  $r^2$  not equal to 100%.)

**Step 4: Economic (or political) validity of the model.** In this step, we interpret the coefficients. Interpreting the exact value of the coefficients may be challenging in multiple regression, but we will always perform at least a sanity check on the signs of the coefficients. Does common (economic or political) sense indicate that the dependent variable should increase or decrease as the  $i$ -th explanatory variable increases and thus, does it agree with the sign of the coefficient  $b^{(i)}$  in the regression equation?

The coefficients are not always in line with common sense. When it is not the case, we should not exploit the model. This is why we will often ask you in exam statements to determine economic (or political) validity before even writing the model, so as to save time. This will require from you to ask yourself which sign (positive, negative) you expected for each coefficients and whether this expected sign is the one obtained on data.

**Step 4, continued: Most influential variables.** Suppose we want to rank the independent variables by order of importance. We should not do so based on the nominal values of the coefficients (because these nominal values highly depend on the units: they are not intrinsic measures). We should do so based on the P-values for the nullity tests: the smaller the P-value, the most influential the variable.

To compare P-values that look like .000 and are all almost null, we should go back to the  $t$  column, in which we read the values of the test statistics. The higher values therein (in absolute values), the smaller the P-values.

**Step 5: Detection of outliers, prediction of future values.** This step is similar to the corresponding one for simple regression. Point estimates are calculated based on the first part of the linear relationship (without the residual term). Then, we use the following (very) approximate formulas:

- for confidence intervals on the average values associated with a set of new values for the independent variables, we add  $\pm 2s/\sqrt{n}$ , or even  $\pm 4s/\sqrt{n}$ , to these point estimates;
- for prediction intervals (intervals in which 95% of the observations will lie), we add  $\pm 2s$ .

Outliers are observations of the dependent variable that do not lie within their prediction intervals.

### Example

We go back to our running example. The multiple regression model of page 174 is not statistically valid, because of the lack of marginal validity of the variable “number of ski lifts” (we read a P-value of 68.3% for the nullity test of its coefficient). We should not exploit this output.

We recompute the regression by omitting this variable and get the following output.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,847 <sup>a</sup>	,717	,711	20,411

a. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	100229,959	2	50114,979	120,298	,000 <sup>b</sup>
	Residual	39575,929	95	416,589		
	Total	139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,252	9,599		2,943	,004
	Maximum altitude of the resort	,031	,005	,441	6,825	,000
	Number of slopes	,602	,074	,524	8,114	,000

a. Dependent Variable: Ski pass price (for 7 days)

**Step 1:** The model is globally valid and marginally valid for each independent variable. Indeed, the P-value read in the second table (almost null) reveals that at least one variable is useful in the linear model, while the P-values read in the third table (there are two to be read and both are almost null as well) indicate that both independent variables have a significant incremental contribution with respect to each other.

**Step 2:** We see that  $r^2 = 71.7\%$  of the (variations in the) ski pass prices are explained, in a statistical way, by (the variations of) the maximum altitudes of the ski resorts and (of) the numbers of slopes. This model is a good model, that catches an overwhelming part of the phenomenon.

**Step 3:** The proposed model is

$$\begin{aligned}
 \text{Ski pass price (in euros)} &= 28.252 \\
 &+ 0.031 \times \text{Maximum altitude of the resort (in meters)} \\
 &+ 0.602 \times \text{Number of slopes (no unit)} \\
 &+ \text{Residual term (with standard deviation: 20.411)}
 \end{aligned}$$

**Step 4:** The model suggests that the ski pass price increases as the maximum altitude increases (the first slope coefficient is positive), and it also increases as the number of slopes does (the second slope coefficient is positive).

This is in line with common (economic) sense: customers are indeed ready to pay more to access a larger ski resort, which is higher and wider (with more slopes).

The most influential variable is the number of slopes. Not because of the value of its coefficient, but because the t statistic based on which the P-value of its test of nullity is calculated is the largest in absolute value (8.114, versus 6.825 for the maximum altitude).

To better see why the coefficients should never be used for the ranking, consider the following equivalent model, where we simply changed the units of the altitude:

$$\begin{aligned}\text{Ski pass price (in euros)} &= 28.252 \\ &+ 31 \times \text{Maximum altitude of the resort (in km)} \\ &+ 0.602 \times \text{Number of slopes (no unit)} \\ &+ \text{Residual term (with standard deviation: 20.411)}\end{aligned}$$

Since the two models are equivalent, the number of slopes is still the most influential variable; yet, it is not associated with the largest coefficient this time. Coefficients highly depend on the units of measurement, they are not normalized, unlike t statistics and P-values, which are normalized quantities.

**Step 5:** What prices are expected for the resorts with maximum altitude of 2,200 meters and 34 slopes?

The point estimate of price equals

$$28.252 + 0.031 \times 2,200 + 0.602 \times 34 = 116.92 \text{ euros.}$$

We also read in the table that  $n - 1 = 97$ , that is,  $n = 98$  ski resorts were considered. Thus, with high confidence, the average price of ski passes of such resorts should lie within the interval

$$116.92 \pm 2 \times 20.411 / \sqrt{98} = 116.92 \pm 4.13, \quad \text{which we round off to} \quad [112.50, 121.50] \text{ euros.}$$

Now, the typical prices of ski passes of these resorts (i.e., about 95% of these prices) lie in the interval

$$116.92 \pm 2 \times 20.411 \approx [76, 158] \text{ euros.}$$

The Alpe du Grand Serre resort has these characteristics; its ski pass price equals 107 euros, which is compatible with the model (its price is not an outlier value).

## 2. Model comparison, variable selection

Based on a set of explanatory variables, there are many multiple-regression models that you can compute. Which of them should you use? This is a question that is answered best through examples (see the excerpts of past exams); but there are some general rules that we can explain here.

**General aim.** We aim here to perform a good trade-off between getting

- a good model (that fits well the data: that has large value of  $r^2$  — actually  $r_{\text{adj}}^2$ , see below);
- an interpretable model (that does not contain too many explanatory variables).

**Comparing two models.** We of course mean here the comparison of two statistically and economically valid models (if one of them is not valid, it should not be considered at all).

The  $r^2$  always increases as the set of explanatory variable increases. Therefore, to compare two models with different numbers of explanatory variables, one should not compare their  $r^2$  but rather (and equivalently)

- the standard deviations  $s$  of their residual terms (the smaller, the better);
- the adjusted coefficient of determination  $r_{\text{adj}}^2$  (the larger, the better); the adjusted<sup>1</sup> coefficient of determination  $r_{\text{adj}}^2$  is indicated in the first table of the outputs, between the (non-adjusted)  $r^2$  and the standard deviation  $s$ .

**Variable selection.** We will discuss two automatic selection methods in this course. Below, “P-value(s)” refers in short to the P-value(s) of the nullity tests of the coefficients.

The backward selection method incrementally simplifies the model, as long as such a simplification is needed. More formally, we start with all explanatory variables in the model; then we repeat:

1. Remove the explanatory variable with the highest P-value (if this P-value is greater than 5%);
2. Recompute the model;
3. If at least one P-value is above 5%, go to step 1, and otherwise, stop here.

The forward selection method incrementally enriches the model, as long as such an enrichment is possible. More formally, we start with the best simple linear regression (the one with the best  $r^2$  among all that are statistically valid, if any exists); then we repeat:

1. For each explanatory variable currently not in the model, check its P-value if it was added to the model and only consider it for the next step if this P-value is  $< 5\%$ .
2. Pick the best explanatory variable for extending the model: the one still considered at the end of the previous step that leads to the extended model with largest  $r^2$ .
3. Repeat until no new explanatory variable can be added.

Both methods are interesting but they lead to different models in general. Usually backward selection is more convenient. However, with many explanatory variables, and if you expect that a small number of them only should be relevant, then forward selection could be simpler.

There are other variable selection methods, such as stepwise selection (which alternates backward and forward steps). In any case, this issue of variable selection requires some practice! That is why we will mostly discuss it in specific examples. It is time to get some practice!

### 3. Regression problems extracted from past exams

We now consider two excerpts of past exams (two second halves of past exam statements), both based on real data:

- Wage discrimination? (from page 181 onwards)
- Modeling life expectancy (from page 195 onwards)

---

<sup>1</sup>The  $r_{\text{adj}}^2$  is a slight correction of the (non-adjusted)  $r^2$  which penalizes for considering more variables; the exact form of the correction was set for the  $r_{\text{adj}}^2$  to be a monotonic function of  $s$ , hence guaranteeing the equivalence between comparing models based on the standard deviations  $s$  or on the  $r_{\text{adj}}^2$ . For those who are curious, the exact formula is

$$r_{\text{adj}}^2 = 1 - \frac{n-1}{n-(k+1)}(1-r^2).$$



## Problem II: Wage discrimination? (30 points)

A company has operated for a number of years, and in recent years there has been an increasing number of complaints about the salaries paid to various workers. One particular complaint of great concern to the management is that female workers are paid less than male workers with the same experience and skill level. The company gathers “base” workers and workers at two higher (junior, senior) levels. Assume that you are an external auditor (from a consulting firm) or an internal auditor (from the HR department of the company). The cases of the present 150 workers were studied and a data set containing the following variables for each worker was created:

Variable	Definition	Units
<b>Salary</b>	Present annual salary	dollars
<b>Age</b>	Age of the worker	years
<b>Total experience</b>	Total experience at the firm	years
<b>Year junior</b>	Time spent in a junior level position	years
<b>Years senior</b>	Time spent in a senior level position	years
<b>Gender</b>	0 for a man, 1 for a woman	[no units]
<b>Skill</b>	1 if specialized skill with high market value, 0 otherwise	[no units]

The last variable describes whether the worker has a specialized skill that is particularly well valued on the labor market. Your task is to evaluate whether there is a wage discrimination or whether there is nothing to hold against the company. This wage discrimination could take place in two manners concerning:

- the starting salary;
- the salary increases.

### Preliminary question (1 point)

1. Look at the histograms describing the distributions of each variable by gender. Explain in one sentence why there might be a feeling of wage discrimination but also why this feeling is not so clear. (*1 point*)

### Simple linear regressions (7 points)

We study first the simple linear regressions of the **Salary** variable over each individual explanatory variable.

2. Consider first the matrix of scatterplots of the non-binary variables: which of the four explanatory variables seems the best variable to explain the dependent variable in a linear fashion? Explain which cells of the matrix you compare and how you do so. (*1 point*)
3. Then consider the SPSS table outputs pertaining to simple linear regressions, and indicate which of the models studied therein are statistically valid, and which are also economically valid. For the latter answer, explain first what economic validity consists in. (*3 points*)
4. Which one is the best single explanatory variable for the annual salary? Write and interpret the corresponding linear model. (*3 points*)

**A first series of multiple linear regressions (17 points)**

We now move to a first series of multiple linear regressions.

5. What are the actual names of the automatic methods 1 and 2 (whose outputs are titled “Automatic method 1” and “Automatic method 2”)? Briefly explain, in one sentence, how each method proceeds. Do they recommend the same model? (*2 points*)
6. Explain carefully (based on factual elements, not just on guesses) the deep reason why the **Age** variable was suppressed by the “Automatic method 1”. (*2 points*)
7. Study the regression output titled “Regression of Question 7”: existence of a significant and parsimonious enough linear relationship, quality of the model, writing of the relationship, economic validity of the relationship. You do not need, however, to comment on the specific value of each coefficient. (*4 points*)
8. Did we bring to light and prove some gender wage discrimination? (Contrast your answer with the one for Question 1.) If a gender wage discrimination was proved: quantify its salary impact, i.e., which average salary difference can be minimally proved, with high confidence, based on the available data? (I.e.: on which amount would a court base its ruling?) How does this average difference compare to the orders of magnitude of the annual salaries? (*3 points*)
9. Compare and contrast your answers to the previous question with a comparison between the models titled “Regression of Question 7” and “Regression of Question 9”; a short and concise sentence only is expected here. (*2 points*)

We thus have built two models for the salary. We now apply them to a specific case.

10. Is the salary of the first woman in the excerpt of the data (Figure 1) compatible with the model titled “Regression of Question 9”? And with the one titled “Regression of Question 7”? (*4 points*)

**In-depth study: on the salary increases (5 points)**

We create two new explanatory variables from two existing ones:

$$\begin{aligned}\text{Total experience for men} &= \text{Total experience} \times (1 - \text{Gender}) \\ \text{Total experience for women} &= \text{Total experience} \times \text{Gender}\end{aligned}$$


Thus, the variable **Total experience for men** takes the values of the **Total experience** variable when the underlying worker is a man, and whenever the worker is a woman, it equals 0. In particular,

$$\text{Total experience} = \text{Total experience for men} + \text{Total experience for women}$$

11. Consider the two models titled “Regression on total experience, separately for men and women – version 1” and “Regression on total experience, separately for men and women – version 2”. Explain why these seemingly different models actually correspond to the same underlying linear modeling of the data: do so by interpreting carefully each of the coefficients of the two relationships. (*2 points*)
12. Based on the output titled “Regression of Question 12”, explain which wage discriminations among starting salaries, on the one hand, and salary increases, on the other hand, can be brought to light and proven, based on the studied data (and which do not seem to be well grounded). Quantify their minimum guaranteed impact. (*3 points*)

Salary.sav [Jeu\_de\_données1] - IBM SPSS Statistics Editeur de données

Fichier Edition Affichage Données Transformer Analyse Marketing direct Graphiques Utilitaires Fenêtre Aide



	Salary	Age	TotalExperience	YearsJunior	YearsSenior	Gender	Skill
1	51825	57	30	19	0	0	0
2	45225	57	26	13	0	0	0
3	51000	47	20	9	0	0	0
4	39000	36	8	2	0	0	0
5	32250	30	5	0	0	1	0
6	44025	38	15	5	0	0	0
7	46500	47	15	7	0	1	0
8	32700	36	5	0	0	1	0
9	43800	45	18	0	0	1	0
10	32700	42	5	0	0	0	0
11	41250	46	13	2	0	0	0
12	63450	56	26	4	16	0	0
13	60000	63	35	9	18	0	0
14	53400	51	23	7	0	0	0
15	48900	50	22	11	0	0	0
16	55725	52	24	5	0	0	0
17	49500	59	27	22	0	0	0
18	52875	46	19	8	3	0	0
19	38475	44	11	0	0	1	0
20	40200	40	13	0	0	1	0
21	37500	46	7	0	0	1	0
22	70800	59	34	10	20	0	1
23	54525	49	16	7	2	0	1
24	54525	48	21	11	3	0	1
25	40350	32	4	0	0	0	1
26	43800	40	10	2	0	0	1

Figure 1: An excerpt of the data set (SPSS screenshot).

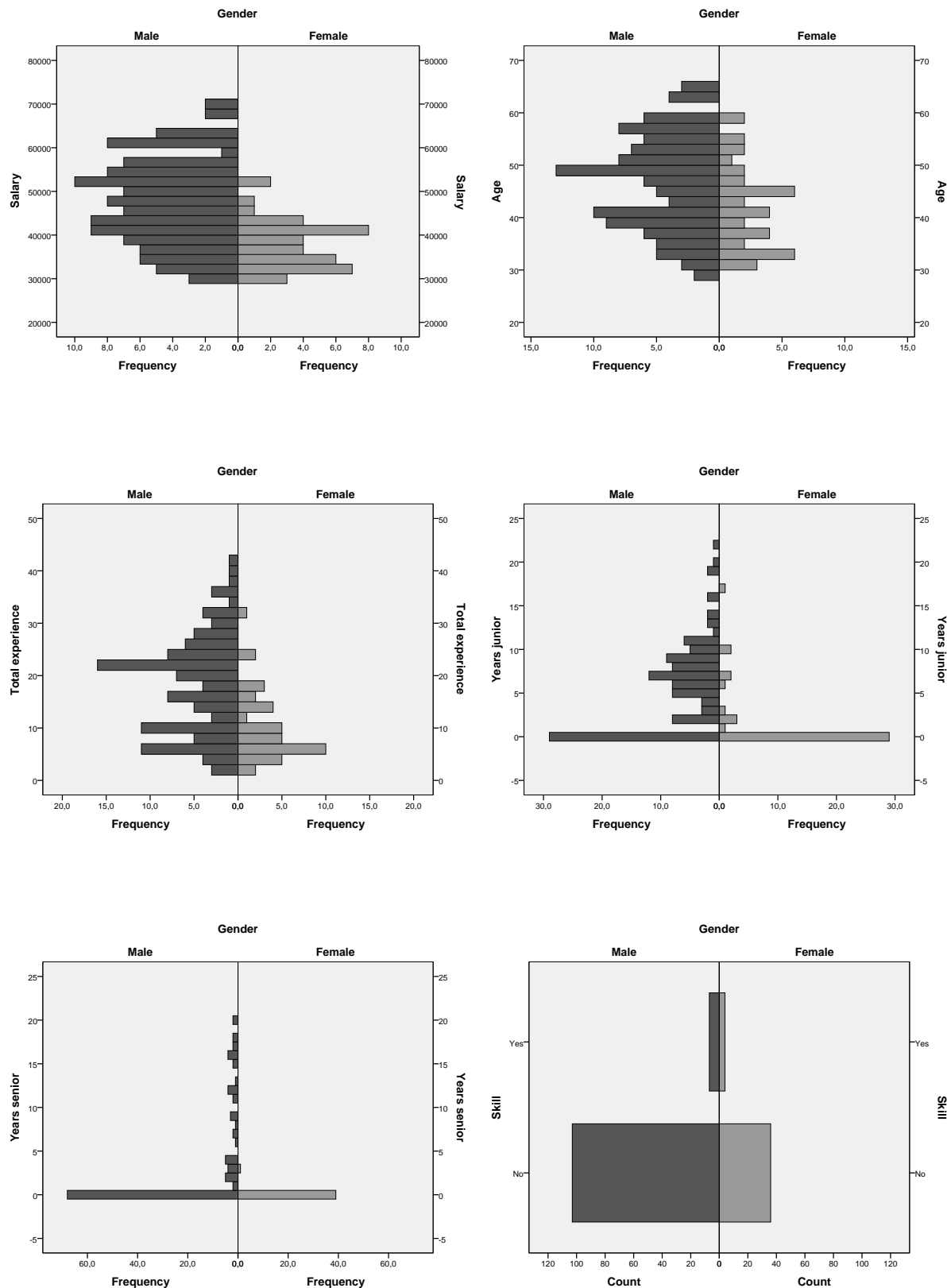


Figure 2: Histograms of each variable by gender. Note: SPSS incorrectly writes “Frequency”, it actually displays counts (not frequencies). It turns out that SPSS is unable to normalize histograms to show frequencies, its developers forgot to implement this option, as incredible as it sounds!

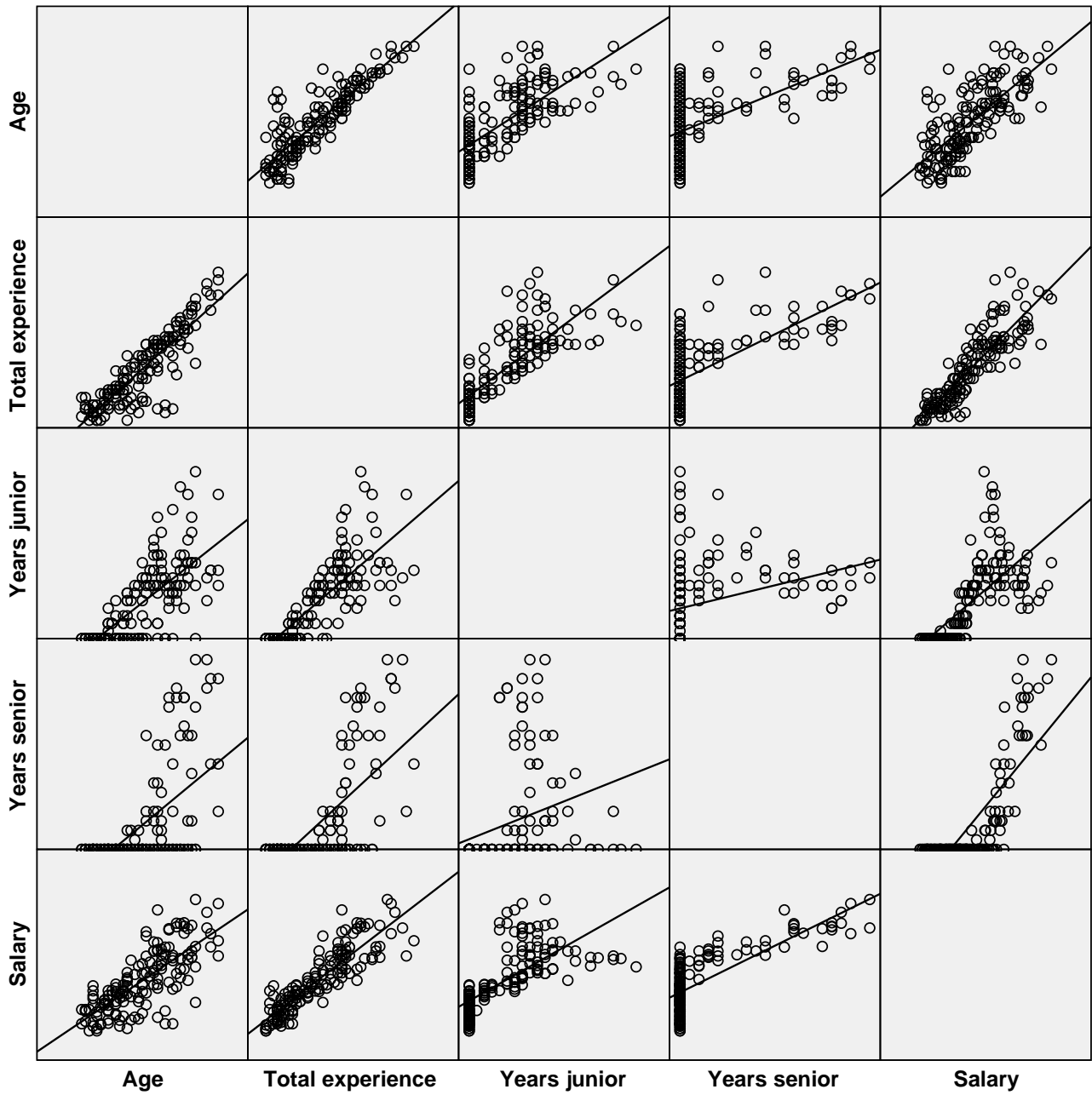


Figure 3: Scatterplots of the non-binary variables.

# Simple linear regressions

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,749 <sup>a</sup>	,560	,557	6615,801

a. Predictors: (Constant), Age

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8255723849	1	8255723849	188,621	,000 <sup>b</sup>
	Residual	6477786189	148	43768825,60		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Age

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8709,621	2712,838		3,211	,002
	Age	809,935	58,973	,749	13,734	,000

a. Dependent Variable: Salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,883 <sup>a</sup>	,780	,779	4676,830

a. Predictors: (Constant), Total experience

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,150E+10	1	1,150E+10	525,602	,000 <sup>b</sup>
	Residual	3237165373	148	21872739,00		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Total experience

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	30816,788	735,252		41,913	,000
	Total experience	924,169	40,311	,883	22,926	,000

a. Dependent Variable: Salary

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,698 <sup>a</sup>	,487	,483	7147,487

a. Predictors: (Constant), Years junior

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7172696755	1	7172696755	140,403	,000 <sup>b</sup>
	Residual	7560813282	148	51086576,23		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Years junior

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	38739,553	799,892		48,431	,000
	Years junior	1359,849	114,763	,698	11,849	,000

a. Dependent Variable: Salary

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,777 <sup>a</sup>	,604	,601	6279,105

a. Predictors: (Constant), Years senior

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8898290097	1	8898290097	225,689	,000 <sup>b</sup>
	Residual	5835219940	148	39427161,76		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Years senior

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41385,155	572,765		72,255	,000
	Years senior	1479,310	98,470	,777	15,023	,000

a. Dependent Variable: Salary

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,429 <sup>a</sup>	,184	,178	9013,937

a. Predictors: (Constant), Gender

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2708354305	1	2708354305	33,333	,000 <sup>b</sup>
	Residual	1,203E+10	148	81251052,25		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Gender

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	47783,864	859,445		55,599	,000
	Gender	-9608,864	1664,308	-,429	-5,773	,000

a. Dependent Variable: Salary

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,026 <sup>a</sup>	,001	-,006	9974,080

a. Predictors: (Constant), Skill

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10133753,98	1	10133753,98	,102	,750 <sup>b</sup>
	Residual	1,472E+10	148	99482272,19		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Skill

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	45148,381	845,990		53,367	,000
	Skill	997,073	3124,027	,026	,319	,750

a. Dependent Variable: Salary



# Automatic method 1

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,937 <sup>a</sup>	,879	,874	3533,827
2	,937 <sup>b</sup>	,878	,873	3539,786

a. Predictors: (Constant), Skill, Years senior, Gender, Years junior, Age, Total experience

b. Predictors: (Constant), Skill, Years senior, Gender, Years junior, Total experience

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,295E+10	6	2157955916	172,803	,000 <sup>b</sup>
	Residual	1785774544	143	12487933,88		
	Total	1,473E+10	149			
2	Regression	1,293E+10	5	2585835652	206,370	,000 <sup>c</sup>
	Residual	1804331777	144	12530081,79		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Skill, Years senior, Gender, Years junior, Age, Total experience

c. Predictors: (Constant), Skill, Years senior, Gender, Years junior, Total experience

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	36813,013	2223,725		16,555	,000
	Age	-82,012	67,277	-,076	-1,219	,225
	Total experience	530,308	96,165	,507	5,515	,000
	Years junior	409,157	113,556	,210	3,603	,000
	Years senior	754,091	89,510	,396	8,425	,000
	Gender	-1574,142	734,892	-,070	-2,142	,034
	Skill	4822,898	1131,364	,127	4,263	,000
2	(Constant)	34282,779	799,297		42,891	,000
	Total experience	461,729	78,124	,441	5,910	,000
	Years junior	401,043	113,552	,206	3,532	,001
	Years senior	751,522	89,636	,395	8,384	,000
	Gender	-1775,308	717,332	-,079	-2,475	,014
	Skill	4837,344	1133,209	,127	4,269	,000

a. Dependent Variable: Salary

## Automatic method 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,883 <sup>a</sup>	,780	,779	4676,830
2	,917 <sup>b</sup>	,841	,839	3992,238
3	,927 <sup>c</sup>	,859	,856	3776,573
4	,934 <sup>d</sup>	,872	,869	3601,799
5	,937 <sup>e</sup>	,878	,873	3539,786

- a. Predictors: (Constant), Total experience  
b. Predictors: (Constant), Total experience, Years senior  
c. Predictors: (Constant), Total experience, Years senior, Skill  
d. Predictors: (Constant), Total experience, Years senior, Skill, Years junior  
e. Predictors: (Constant), Total experience, Years senior, Skill, Years junior, Gender

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,150E+10	1	1,150E+10	525,602	,000 <sup>b</sup>
	Residual	3237165373	148	21872739,00		
	Total	1,473E+10	149			
2	Regression	1,239E+10	2	6195314381	388,714	,000 <sup>c</sup>
	Residual	2342881276	147	15937967,87		
	Total	1,473E+10	149			
3	Regression	1,265E+10	3	4217061342	295,675	,000 <sup>d</sup>
	Residual	2082326011	146	14262506,93		
	Total	1,473E+10	149			
4	Regression	1,285E+10	4	3213107837	247,677	,000 <sup>e</sup>
	Residual	1881078690	145	12972956,48		
	Total	1,473E+10	149			
5	Regression	1,293E+10	5	2585835652	206,370	,000 <sup>f</sup>
	Residual	1804331777	144	12530081,79		
	Total	1,473E+10	149			

- a. Dependent Variable: Salary  
b. Predictors: (Constant), Total experience  
c. Predictors: (Constant), Total experience, Years senior  
d. Predictors: (Constant), Total experience, Years senior, Skill  
e. Predictors: (Constant), Total experience, Years senior, Skill, Years junior  
f. Predictors: (Constant), Total experience, Years senior, Skill, Years junior, Gender

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	30816,788	735,252		41,913	,000
	Total experience	924,169	40,311	,883	22,926	,000
2	(Constant)	32833,242	682,921		48,078	,000
	Total experience	689,220	46,560	,659	14,803	,000
	Years senior	634,559	84,713	,333	7,491	,000
3	(Constant)	32000,026	674,800		47,421	,000
	Total experience	725,017	44,834	,693	16,171	,000
	Years senior	595,147	80,666	,313	7,378	,000
	Skill	5147,187	1204,254	,135	4,274	,000
4	(Constant)	33520,114	750,424		44,668	,000
	Total experience	461,080	79,492	,441	5,800	,000
	Years senior	781,778	90,354	,411	8,652	,000
	Skill	4763,014	1152,657	,125	4,132	,000
	Years junior	448,530	113,880	,230	3,939	,000
5	(Constant)	34282,779	799,297		42,891	,000
	Total experience	461,729	78,124	,441	5,910	,000
	Years senior	751,522	89,636	,395	8,384	,000
	Skill	4837,344	1133,209	,127	4,269	,000
	Years junior	401,043	113,552	,206	3,532	,001
	Gender	-1775,308	717,332	-,079	-2,475	,014

- a. Dependent Variable: Salary

## Regression of Question 7

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,937 <sup>a</sup>	,878	,873	3539,786

a. Predictors: (Constant), Skill, Years senior, Gender, Years junior, Total experience

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,293E+10	5	2585835652	206,370	,000 <sup>b</sup>
	Residual	1804331777	144	12530081,79		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Skill, Years senior, Gender, Years junior, Total experience

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	34282,779	799,297		42,891	,000	32702,908	35862,650
	Total experience	461,729	78,124	,441	5,910	,000	307,311	616,147
	Years junior	401,043	113,552	,206	3,532	,001	176,599	625,486
	Years senior	751,522	89,636	,395	8,384	,000	574,349	928,695
	Gender	-1775,308	717,332	-,079	-2,475	,014	-3193,168	-357,447
	Skill	4837,344	1133,209	,127	4,269	,000	2597,471	7077,217

a. Dependent Variable: Salary

## Regression of Question 9

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,934 <sup>a</sup>	,872	,869	3601,799

a. Predictors: (Constant), Skill, Years senior, Years junior, Total experience

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,285E+10	4	3213107837	247,677	,000 <sup>b</sup>
	Residual	1881078690	145	12972956,48		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Skill, Years senior, Years junior, Total experience

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	33520,114	750,424		44,668	,000	32036,931	35003,298
	Total experience	461,080	79,492	,441	5,800	,000	303,966	618,193
	Years junior	448,530	113,880	,230	3,939	,000	223,451	673,608
	Years senior	781,778	90,354	,411	8,652	,000	603,196	960,360
	Skill	4763,014	1152,657	,125	4,132	,000	2484,834	7041,194

a. Dependent Variable: Salary

## Regression on total experience, separately for men and women --- version 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,891 <sup>a</sup>	,793	,790	4552,784

a. Predictors: (Constant), Total experience women, Total experience men

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,169E+10	2	5843258879	281,904	,000 <sup>b</sup>
	Residual	3046992280	147	20727838,64		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Total experience women, Total experience men

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	31499,900	750,440		41,975	,000	30016,855	32982,945
	Total experience men	914,724	39,365	1,041	23,237	,000	836,929	992,520
	Total experience women	706,472	81,886	,387	8,627	,000	544,646	868,299

a. Dependent Variable: Salary

## Regression on total experience, separately for men and women --- version 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,891 <sup>a</sup>	,793	,790	4552,784

a. Predictors: (Constant), Total experience men, Total experience

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,169E+10	2	5843258879	281,904	,000 <sup>b</sup>
	Residual	3046992280	147	20727838,64		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Total experience men, Total experience

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	31499,900	750,440		41,975	,000	30016,855	32982,945
	Total experience	706,472	81,886	,675	8,627	,000	544,646	868,299
	Total experience men	208,252	68,753	,237	3,029	,003	72,380	344,124

a. Dependent Variable: Salary

## Regression of Question 12

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,938 <sup>a</sup>	,880	,875	3517,647

a. Predictors: (Constant), Total experience women, Skill, Years junior, Years senior, Gender, Total experience

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,296E+10	6	2160675143	174,616	,000 <sup>b</sup>
	Residual	1769459178	143	12373840,41		
	Total	1,473E+10	149			

a. Dependent Variable: Salary

b. Predictors: (Constant), Total experience women, Skill, Years junior, Years senior, Gender, Total experience

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	34905,350	876,607		39,819	,000	33172,568	36638,133
	Years junior	433,790	114,515	,223	3,788	,000	207,429	660,151
	Years senior	805,048	94,610	,423	8,509	,000	618,033	992,064
	Gender	-3520,606	1260,549	-,157	-2,793	,006	-6012,323	-1028,888
	Skill	4920,511	1127,211	,129	4,365	,000	2692,361	7148,660
	Total experience	404,892	84,697	,387	4,780	,000	237,472	572,311
	Total experience women	166,220	99,013	,091	1,679	,095	-29,499	361,939

a. Dependent Variable: Salary

## Problem II: Modeling life expectancy (10 points)

This exercise is based on real data (from 2012), which we collected on various websites and then aggregated into a single data set. The aim is to model life expectancy, by country, as a function of other variables. These other variables are considered twice: first, in their nominal forms; and second, through their (natural) logarithms. The reason why we compute and use the logarithmic values of the variables should hopefully become clearer later in this problem.

Variable	Definition	Units
<b>Country</b>	Country name (114 countries considered in total)	
<b>LifeExp</b>	Life expectancy	years
<b>GDP</b>	Per capita gross domestic product, in thousands of dollars	K\$ per capita
<b>Alcohol</b>	Average annual consumption of (pure) alcohol, per adult	liters per adult
<b>Tobacco</b>	Average annual number of cigarettes smoked, per adult	cigarettes per adult
<b>IQ</b>	Average intelligence quotient of the inhabitants (base: 100)	no units
<b>Democr</b>	Democratization index (grade between 0 and 10)	no units
<b>LnGDP</b>	$= \ln(\text{GDP})$	
<b>LnAlcohol</b>	$= \ln(\text{Alcohol})$	
<b>LnTobacco</b>	$= \ln(\text{Tobacco})$	
<b>LnIQ</b>	$= \ln(\text{IQ})$	
<b>LnDemocr</b>	$= \ln(\text{Democr})$	

An excerpt of the data set is reproduced in appendix.

### A first look at the data (0.5 point)

- You should not devote more than a few minutes to answer these questions that only intend to familiarize you with the data and the meaning of all variables (and their limitations). Base your answers on the excerpt of the data set.
  - Countries can be divided into three groups, as far as the **Alcohol** variable is concerned: in which type of countries are the average consumptions respectively high, medium and almost null? (Do not list individual countries, but try to find their common denominator.)
  - For which type of countries does the **IQ** variable take values around 100, for which type of countries is it significantly lower? Do you think the **IQ** is a “universal” indicator, or is it rather a biased indicator whose relevance is extremely questionable?

### Simple linear regressions based on GDP variables (2 points)

- Which is the best individual explanatory variable between **GDP** and **LnGDP**? Base your answer first on well-chosen scatter plots and then on well-chosen numerical results. (0.5 point)

3. About the simple linear regression model associated with this best explanatory variable: is it statistically and economically valid? Explain. Then write and interpret the model. (1.5 points)

**About the other simple linear regressions (2 points)**

4. Among the eight other simple linear regressions: which ones are statistically valid? (0.5 point)
5. Is it clear that the models `LifeExp / Tobacco` and `LifeExp / Alcohol` are economically valid? Elaborate on your doubts, if any, based on a third, latent, variable. The latent variable can be a variable already present in the data set or you can guess what it could be. (1 point)
6. According to you, which of the 10 simple linear regression models achieves the best trade-off between statistical interest and economic meaning? (0.5 point)

**About the multiple linear regressions (3 points)**

7. Explain why we only considered the following five variables when producing our multiple regression outputs: `LnGDP`, `Alcohol`, `LnTobacco`, `IQ` and `Democr`. (0.5 point)
8. Do you think highly of the linear regression model based on the 5 variables selected in the previous question? Why and how did we come up with the model with 4 variables reproduced right below it? (0.5 point)
9. Show that this model with 4 variables is statistically and economically valid; write and interpret its relation. (1.5 points)
10. Which selection method is worked out under the title “Automatic method”? Explain briefly how it works. Which model does it recommend? (0.5 point)

**Exploitation of the model (1.5 points)**

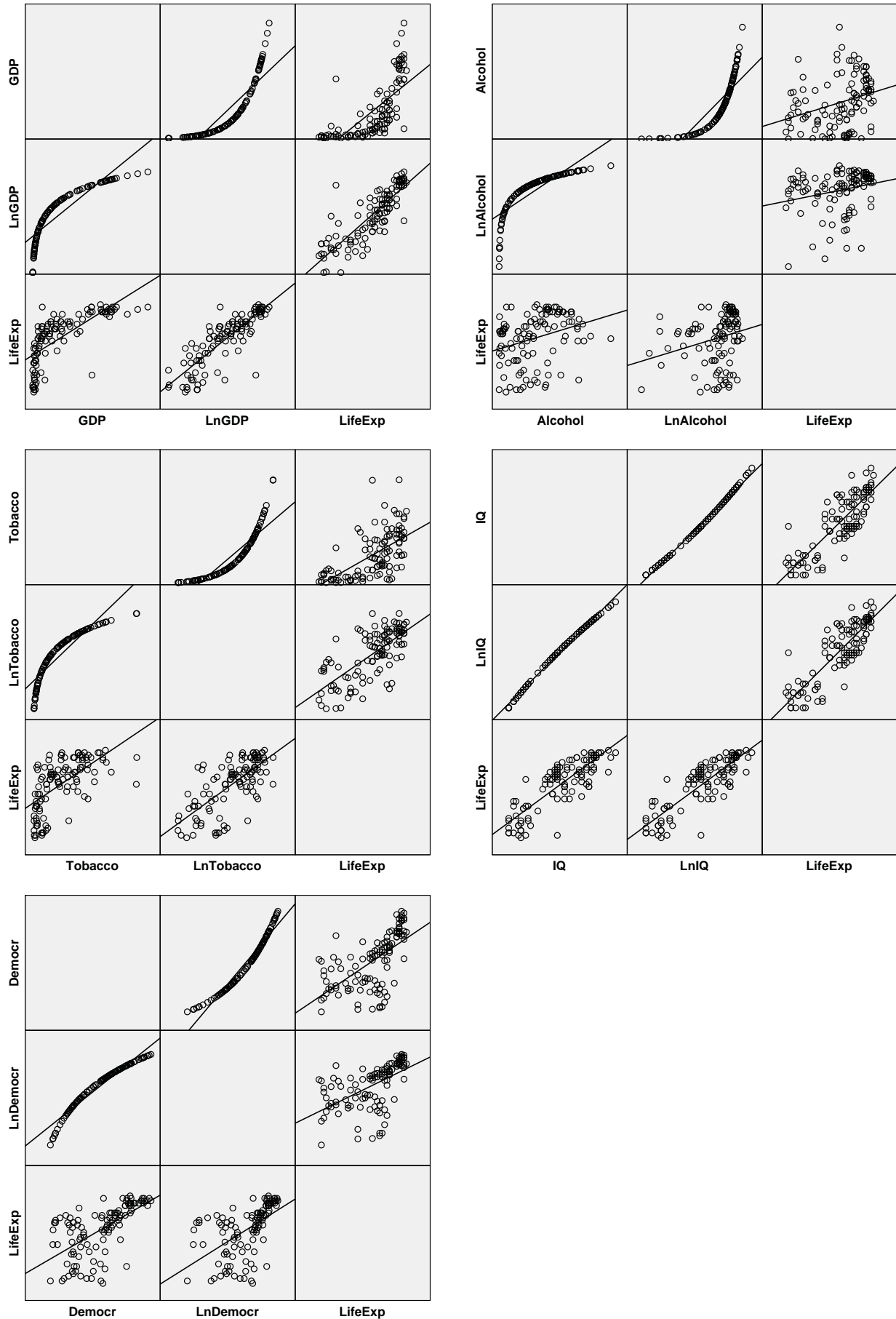
11. Is the life expectancy observed in France compatible with the model constructed in question 9? (1.5 points)

**Search for an alternative model (1 point)**

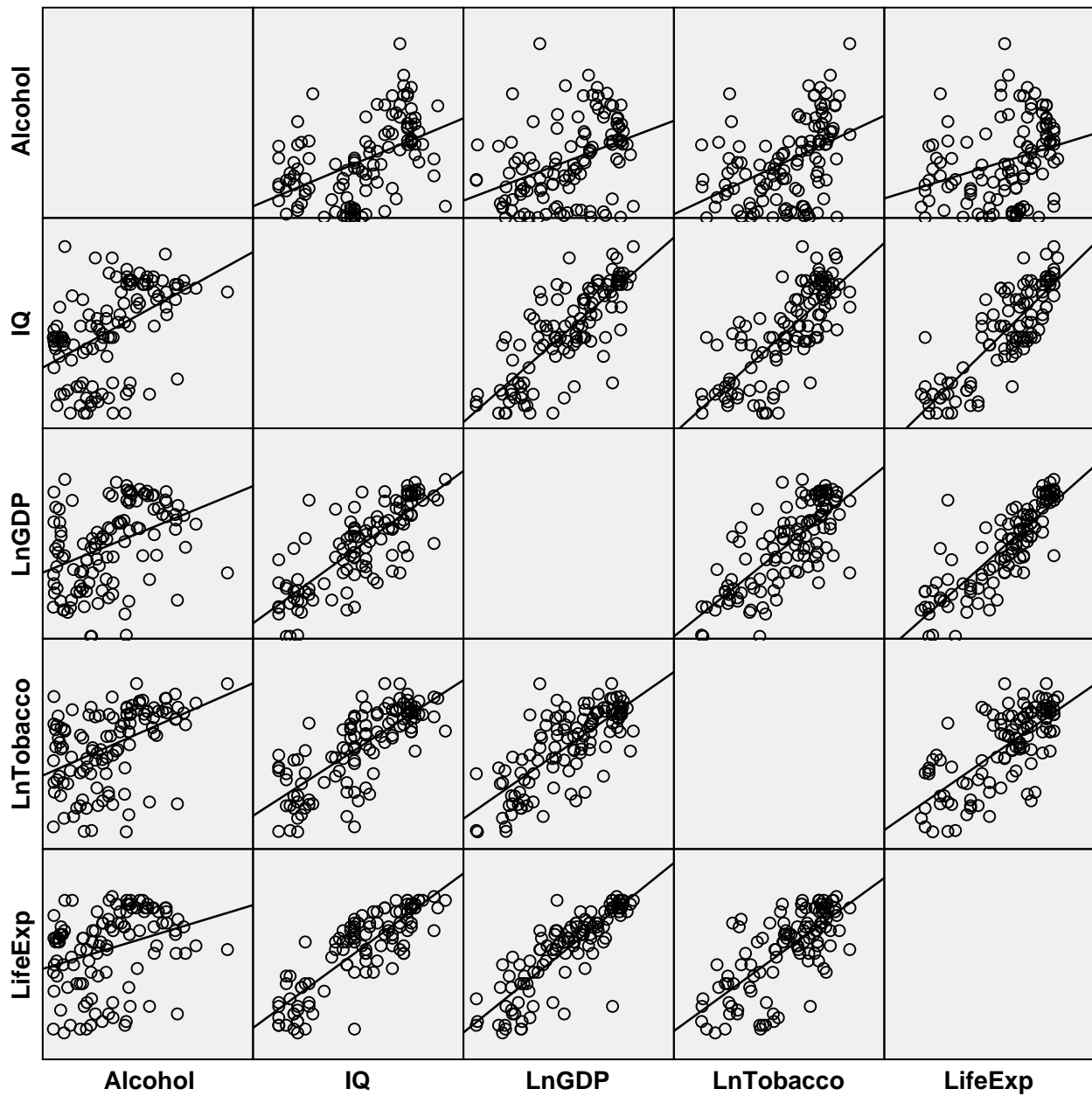
12. You should exploit the SPSS outputs titled “alternative model” on the last page.
  - (a) What was the logic behind trying the alternative model 1/2 (what puzzled us in the model of question 9)? What do you think of this first alternative model?
  - (b) Similarly, why did we try the alternative model 2/2 and what do you think of it?
  - (c) Conclude: among all regression models considered in this problem (simple as well as multiple regression models), which one do you *personally* prefer? Explain carefully the reasons of your preferences. You should not apply general rules in a mechanical way: we really ask about your personal but well-grounded opinion.



EspVie.sav [DataSet1] - IBM SPSS Statistics Data Editor													
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help													
Visible: 12 of 12													
	Country	LifeExp	GDP	Alcohol	Tobacco	IQ	Democr	LnGDP	LnAlcohol	LnTobacco	LnIQ	LnDemocr	
34	Ethiopia	54	1,014	4,10	87	64	3,68	,014	1,41	4,47	4,16	1,30	
35	Finland	80	34,402	13,10	1351	99	9,19	3,538	2,57	7,21	4,60	2,22	
36	France	81	34,092	12,48	2058	98	7,77	3,529	2,52	7,63	4,58	2,05	
37	Gambia	60	1,973	3,58	171	66	3,38	,679	1,28	5,14	4,19	1,22	
38	Germany	80	35,930	12,14	1702	99	8,38	3,582	2,50	7,44	4,60	2,13	
39	Ghana	60	1,610	3,11	161	71	6,02	,476	1,13	5,08	4,26	1,80	
40	Greece	80	28,834	11,01	4313	92	7,92	3,362	2,40	8,37	4,52	2,07	
41	Guatemala	69	4,871	7,10	609	79	6,05	1,583	1,96	6,41	4,37	1,80	
42	Haiti	62	1,122	5,92	172	67	4,00	,115	1,78	5,15	4,20	1,39	
43	Honduras	69	4,405	4,43	595	81	5,76	1,483	1,49	6,39	4,39	1,75	
44	Hungary	74	18,816	16,12	3265	98	7,21	2,935	2,78	8,09	4,58	1,98	
45	Iceland	82	36,681	7,38	1915	101	9,65	3,602	2,00	7,56	4,62	2,27	
46	India	65	3,291	2,69	129	82	7,28	1,191	,99	4,86	4,41	1,99	
47	Iran	73	11,025	1,03	765	84	1,94	2,400	,03	6,64	4,43	,66	
48	Iraq	66	3,599	,47	1430	87	4,00	1,281	-,76	7,27	4,47	1,39	
49	Ireland	80	38,816	14,92	2236	92	8,79	3,659	2,70	7,71	4,52	2,17	
50	Israel	82	29,405	2,52	2162	95	7,48	3,381	,92	7,68	4,55	2,01	
51	Italy	82	29,418	9,72	1901	102	7,83	3,382	2,27	7,55	4,62	2,06	
52	Jamaica	71	8,812	5,17	735	71	7,21	2,176	1,64	6,60	4,26	1,98	
53	Japan	83	33,828	7,79	3023	105	8,08	3,521	2,05	8,01	4,65	2,09	
54	Jordan	71	5,659	,65	1832	84	3,74	1,733	-,43	7,51	4,43	1,32	
55	Kazakhstan	64	12,402	11,10	2160	94	3,30	2,518	2,41	7,68	4,54	1,19	
56	Kenya	60	1,784	3,88	200	72	4,71	,579	1,36	5,30	4,28	1,55	
57	Kuwait	78	38,293	,10	3062	86	3,88	3,645	-2,30	8,03	4,45	1,36	
58	Kyrgyzstan	66	2,162	4,72	1886	90	4,31	,771	1,55	7,54	4,50	1,46	
59	Lao People Democ...	63	2,435	6,99	400	89	2,10	,890	1,94	5,99	4,49	,74	
60	Liberia	56	,396	5,07	89	67	5,07	-,925	1,62	4,49	4,20	1,62	
61	Libyan Arab Jama...	72	14,878	,10	1482	83	1,94	2,700	-2,30	7,30	4,42	,66	
62	Madagascar	65	,911	1,32	315	82	3,94	-,093	,28	5,75	4,41	1,37	
63	Malawi	47	,909	1,44	123	69	5,84	-,096	,36	4,81	4,23	1,76	
64	Malaysia	73	14,603	,87	910	92	6,19	2,681	-,14	6,81	4,52	1,82	
65	Mali	53	1,207	,99	223	69	6,01	,188	-,01	5,41	4,23	1,79	
66	Mauritania	58	2,100	,11	317	76	3,86	,742	-2,21	5,76	4,33	1,35	
67	Mexico	76	14,266	8,55	754	88	6,93	2,658	2,15	6,63	4,48	1,94	



## Matrix of scatter plots



## Simple regression #1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,712 <sup>a</sup>	,507	,503	7,296

a. Predictors: (Constant), GDP

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6137,292	1	6137,292	115,282	,000 <sup>b</sup>
	Residual	5962,568	112	53,237		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), GDP

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	61,991	,964		64,277	,000
	GDP	,522	,049	,712	10,737	,000

a. Dependent Variable: LifeExp

## Simple regression #2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,846 <sup>a</sup>	,716	,714	5,539

a. Predictors: (Constant), LnGDP

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8664,039	1	8664,039	282,428	,000 <sup>b</sup>
	Residual	3435,821	112	30,677		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnGDP

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	56,054	,943		59,411	,000
	LnGDP	6,718	,400	,846	16,806	,000

a. Dependent Variable: LifeExp

## Simple regression #3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,306 <sup>a</sup>	,093	,085	9,896

a. Predictors: (Constant), Alcohol

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1131,310	1	1131,310	11,552	,001 <sup>b</sup>
	Residual	10968,550	112	97,933		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), Alcohol

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	64,736	1,631		39,687	,000
	Alcohol	,607	,179	,306	3,399	,001

a. Dependent Variable: LifeExp

## Simple regression #4

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,249 <sup>a</sup>	,062	,053	10,068

a. Predictors: (Constant), LnAlcohol

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	747,280	1	747,280	7,372	,008 <sup>b</sup>
	Residual	11352,580	112	101,362		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnAlcohol

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66,404	1,423		46,658	,000
	LnAlcohol	1,903	,701	,249	2,715	,008

a. Dependent Variable: LifeExp

## Simple regression #5

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,619 <sup>a</sup>	,383	,378	8,163

a. Predictors: (Constant), Tobacco

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4637,609	1	4637,609	69,605	,000 <sup>b</sup>
	Residual	7462,251	112	66,627		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), Tobacco

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	60,955	1,259		48,424	,000
	Tobacco	,007	,001	,619	8,343	,000

a. Dependent Variable: LifeExp

## Simple regression #6

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,710 <sup>a</sup>	,504	,499	7,321

a. Predictors: (Constant), LnTobacco

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6096,721	1	6096,721	113,746	,000 <sup>b</sup>
	Residual	6003,138	112	53,599		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnTobacco

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	20,395	4,636		4,399	,000
	LnTobacco	7,276	,682	,710	10,665	,000

a. Dependent Variable: LifeExp

## Simple regression #7

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,849 <sup>a</sup>	,721	,718	5,495

a. Predictors: (Constant), IQ

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8718,467	1	8718,467	288,777	,000 <sup>b</sup>
	Residual	3381,392	112	30,191		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), IQ

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,279	3,744		1,677	,096
	IQ	,735	,043	,849	16,993	,000

a. Dependent Variable: LifeExp

## Simple regression #8

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,856 <sup>a</sup>	,733	,731	5,369

a. Predictors: (Constant), LnIQ

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8870,864	1	8870,864	307,692	,000 <sup>b</sup>
	Residual	3228,996	112	28,830		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnIQ

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-202,970	15,530		-13,070	,000
	LnIQ	61,295	3,494	,856	17,541	,000

a. Dependent Variable: LifeExp

## Simple regression #9

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,625 <sup>a</sup>	,391	,385	8,113

a. Predictors: (Constant), Democr

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4727,277	1	4727,277	71,814	,000 <sup>b</sup>
	Residual	7372,583	112	65,827		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), Democr

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	52,671	2,104		25,033	,000
	Democr	2,904	,343	,625	8,474	,000

a. Dependent Variable: LifeExp

## Simple regression #10

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,556 <sup>a</sup>	,309	,302	8,643

a. Predictors: (Constant), LnDemocr

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3733,883	1	3733,883	49,988	,000 <sup>b</sup>
	Residual	8365,977	112	74,696		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnDemocr

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	48,479	3,054		15,875	,000
	LnDemocr	12,591	1,781	,556	7,070	,000

a. Dependent Variable: LifeExp



## Regression with the 5 variables selected

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,911 <sup>a</sup>	,831	,823	4,355

a. Predictors: (Constant), LnTobacco, Democr, Alcohol, IQ, LnGDP

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10051,761	5	2010,352	106,010	,000 <sup>b</sup>
	Residual	2048,099	108	18,964		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnTobacco, Democr, Alcohol, IQ, LnGDP

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	19,277	4,846		3,978	,000
	Alcohol	-,388	,097	-,196	-4,019	,000
	IQ	,430	,065	,497	6,656	,000
	Democr	,841	,262	,181	3,211	,002
	LnGDP	2,780	,624	,350	4,454	,000
	LnTobacco	,853	,713	,083	1,196	,234

a. Dependent Variable: LifeExp

## Regression on 4 of these variables only

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,910 <sup>a</sup>	,828	,822	4,363

a. Predictors: (Constant), LnGDP, Alcohol, Democr, IQ

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10024,644	4	2506,161	131,635	,000 <sup>b</sup>
	Residual	2075,216	109	19,039		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), LnGDP, Alcohol, Democr, IQ

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,067	4,256		5,184	,000
	Alcohol	-,361	,094	-,182	-3,838	,000
	IQ	,461	,060	,532	7,730	,000
	Democr	,752	,252	,162	2,988	,003
	LnGDP	3,107	,562	,391	5,527	,000

a. Dependent Variable: LifeExp

## Automatic method

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,849 <sup>a</sup>	,721	,718	5,495
2	,894 <sup>b</sup>	,799	,796	4,676
3	,902 <sup>c</sup>	,814	,809	4,518
4	,910 <sup>d</sup>	,828	,822	4,363

- a. Predictors: (Constant), IQ  
b. Predictors: (Constant), IQ, LnGDP  
c. Predictors: (Constant), IQ, LnGDP, Alcohol  
d. Predictors: (Constant), IQ, LnGDP, Alcohol, Democr

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8718,467	1	8718,467	288,777	,000 <sup>b</sup>
	Residual	3381,392	112	30,191		
	Total	12099,860	113			
2	Regression	9672,900	2	4836,450	221,201	,000 <sup>c</sup>
	Residual	2426,960	111	21,865		
	Total	12099,860	113			
3	Regression	9854,623	3	3284,874	160,935	,000 <sup>d</sup>
	Residual	2245,237	110	20,411		
	Total	12099,860	113			
4	Regression	10024,644	4	2506,161	131,635	,000 <sup>e</sup>
	Residual	2075,216	109	19,039		
	Total	12099,860	113			

- a. Dependent Variable: LifeExp  
b. Predictors: (Constant), IQ  
c. Predictors: (Constant), IQ, LnGDP  
d. Predictors: (Constant), IQ, LnGDP, Alcohol  
e. Predictors: (Constant), IQ, LnGDP, Alcohol, Democr

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,279	3,744		1,677	,096
	IQ	,735	,043	,849	16,993	,000
2	(Constant)	26,521	4,420		6,000	,000
	IQ	,414	,061	,478	6,793	,000
	LnGDP	3,692	,559	,465	6,607	,000
3	(Constant)	23,845	4,364		5,464	,000
	IQ	,468	,062	,541	7,599	,000
	LnGDP	3,732	,540	,470	6,909	,000
	Alcohol	-,277	,093	-,140	-2,984	,004
4	(Constant)	22,067	4,256		5,184	,000
	IQ	,461	,060	,532	7,730	,000
	LnGDP	3,107	,562	,391	5,527	,000
	Alcohol	-,361	,094	-,182	-3,838	,000
	Democr	,752	,252	,162	2,988	,003

- a. Dependent Variable: LifeExp

**Alternative model 1/2****Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,857 <sup>a</sup>	,734	,727	5,404

a. Predictors: (Constant), Democr, Alcohol, LnGDP

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8887,070	3	2962,357	101,426	,000 <sup>b</sup>
	Residual	3212,790	110	29,207		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), Democr, Alcohol, LnGDP

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	53,769	1,410		38,135	,000
	LnGDP	6,072	,509	,765	11,933	,000
	Alcohol	-,164	,112	-,083	-1,467	,145
	Democr	,837	,311	,180	2,689	,008

a. Dependent Variable: LifeExp

**Alternative model 2/2****Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,854 <sup>a</sup>	,729	,724	5,432

a. Predictors: (Constant), Democr, LnGDP

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8824,198	2	4412,099	149,510	,000 <sup>b</sup>
	Residual	3275,662	111	29,510		
	Total	12099,860	113			

a. Dependent Variable: LifeExp

b. Predictors: (Constant), Democr, LnGDP

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	53,573	1,411		37,971	,000
	LnGDP	5,971	,507	,752	11,783	,000
	Democr	,691	,296	,149	2,330	,022

a. Dependent Variable: LifeExp

Pourquoi d'ailleurs supprime-t-on (ou ajoute-t-on) seulement une variable à la fois ? Car un manque de significativité individuelle peut venir de deux causes :

- une inutilité intrinsèque (si on prenait pour variable explicative du prix du forfait l'âge du maire de la commune, elle ne jouerait pas) ;
- une redondance (une variable peut mesurer la même information qu'une autre variable ou une combinaison de variables).

A cause de la possibilité de redondance, il arrive que deux variables ne soient pas simultanément utiles dans le même, mais qu'il soit utile d'en garder une des deux. Dans le modèle complet, les deux variables seraient non individuellement significatives, mais dès qu'une d'elle est supprimée, l'autre devient individuellement significative.

Voici ce que donne la sélection « forward » sur notre exemple :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,760 <sup>a</sup>	,578	,574	24,787
2	,847 <sup>b</sup>	,717	,711	20,411

a. Predictors: (Constant), Number of slopes

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

ANOVA<sup>a</sup>

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 80825,821	1	80825,821	131,558	,000 <sup>b</sup>
	Residual 58980,067	96	614,376		
	Total 139805,888	97			
2	Regression 100229,959	2	50114,979	120,298	,000 <sup>c</sup>
	Residual 39575,929	95	416,589		
	Total 139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes

c. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

Coefficients<sup>a</sup>

Model		B	Std. Error	Unstandardized Coefficients		Beta	t	Sig.
				Standardized	Coefficients			
1	(Constant)	88,459	4,596				19,248	,000
	Number of slopes	,873	,076				11,470	,000
2	(Constant)	28,252	9,599				2,943	,004
	Number of slopes	,602	,074				8,114	,000
	Maximum altitude of the resort	,031	,005				6,825	,000

a. Dependent Variable: Ski pass price (for 7 days)

La valeur minimale (approximative et moyenne) pour le prix du forfait résultant de ce modèle est donc de

$$28.252 + 0.031 \times 1100 + 0.602 \times 4 = 64.76 \text{ euros.}$$

La valeur minimale de 70 euros lue dans les données est compatible avec cette valeur minimale moyenne (approximative) de 64.76 euros, voir point suivant.  
En ajoutant et en soustrayant (opération blanche!)  $0.031 \times 1100 + 0.602 \times 4$  à la relation précédente-ment écrite, on pourrait la réécrire comme :

$$\begin{aligned} \text{Prix du forfait (euros)} &= 64.76 \\ &+ 0.031 \times (\text{Altitude maximale du domaine [mètres]} - 1100) \\ &+ 0.602 \times (\text{Nombre de pistes [sans unités]} - 4) \\ &+ \text{Résidu (d'écart-type : 20.411 euros)} \end{aligned}$$

Les interprétations des coefficients de pente 0.031 et 0.602 restent identiques, et 64.76 euros apparaît directement comme la valeur minimale moyenne (approximative) du prix du forfait.

**Etape 5 : Prévision / Compatibilité.** On calcule un intervalle de prévision (où les valeurs individuelles doivent se trouver, avec grande confiance) comme en régression simple : en remplaçant le terme de résidu d'écart-type  $s$  par  $\pm 2s$ .

**Exemple.** Dans les données, la station de Bessans propose un forfait à 70 euros, a une altitude maximale de 2200 euros et dispose de 4 pistes. Le prix de son forfait est-il compatible avec le modèle ? Le modèle propose une fourchette de prix de forfait de

$$28.252 + 0.031 \times 2200 + 0.602 \times 4 \pm 2 \times 20.411 = 98.86 \pm 40.822,$$

qui contient bien la valeur 70. Le prix du forfait à Bessans est compatible avec le modèle.

## 2. Comparaison de modèles / sélection de modèles

**Comparer deux modèles entre eux.** Le  $r^2$  augmente mécaniquement avec l'ajout de variables explicatives. Pour comparer de manière plus juste deux modèles avec des nombres de variables explicatives différents, il faut pondérer le  $r^2$  par le nombre de variables. C'est ce que font le  $r^2_{adj}$  (plus il est grand, meilleur est le modèle) et l'estimation de l'écart-type des résidus (plus elle est petite, meilleur est le modèle). Vous modulerez toutefois cela en pratique à l'aune de considérations « stratégiques » (= de commentaires subjectifs sur la manière dont vous recevez et pensez pouvoir exploiter les modèles).

**Méthodes de sélection automatique.** Les sujets d'examen reportent souvent le résultat de deux méthodes de sélection.

La méthode « forward » part de la meilleure variable explicative prise isolément (celle pour laquelle la régression linéaire simple a le plus grand  $r^2$ ) et enrichit itérativement le modèle d'une variable à la fois, la meilleure addition possible, tant que cela est possible : tant que la variable ajoutée reste individuellement significative.

La méthode « backward » part du modèle complet (= avec toutes les variables explicatives) et le simplifie itérativement en supprimant une variable à la fois, la moins individuellement significative (= celle avec la plus grande P-valeur), tant que cela est nécessaire.



**\*Not statistically valid**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.847 <sup>a</sup>	.717	.708	20.500

a. Predictors: (Constant), Number of slopes, Maximum altitude of the resort, Number of ski lifts

ANOVA<sup>a</sup>

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 100300,605	3	33433,535	79,553	,000 <sup>b</sup>
	Residual 39505,283	94	420,269		
	Total 139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort, Number of ski lifts

Coefficients<sup>a</sup>

Model	1	(Constant)		Maximum altitude of the resort		Number of ski lifts		Number of slopes	
		B	Std. Error	Unstandardized Coefficients	Standardized Coefficients	Beta	t	Sig.	
		28,626	9,685	,005	,437	-,032	,548	6,265	,000
		-,029	,071	,005	,437	-,032	,548	6,265	,000
		-,029	,071	,005	,437	-,032	,548	6,265	,000

a. Dependent Variable: Ski pass price (for 7 days)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.847 <sup>a</sup>	.717	.711	20,411

a. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

ANOVA<sup>a</sup>

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 100229,959	2	50114,979	120,298	,000 <sup>b</sup>
	Residual 39575,929	95	416,589		
	Total 139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

Coefficients<sup>a</sup>

Model	1	(Constant)		Maximum altitude of the resort		Number of slopes	
		B	Std. Error	Unstandardized Coefficients	Standardized Coefficients	Beta	t
		28,252	9,599	,005	,074	,441	8,114
		,031	,005	,005	,074	,441	8,114
		,031	,005	,005	,074	,441	8,114

a. Dependent Variable: Ski pass price (for 7 days)

## Synthèse — Régression linéaire multiple

Nous prenons comme exemple filé tout au long de ce chapitre la modélisation de prix de forfaits de ski (variable à expliquer) en fonction de diverses caractéristiques du domaine (variables explicatives) : altitude maximal, nombre de remontées, nombre de pistes.

### 1. Etude d'un modèle donné

Nous commençons par l'étude d'un modèle donné—voir les deux sorties de la page suivante—et suivons la même analyse en 5 étapes, avec quelques changements. Ce qui suit va surtout souligner ces changements et passera plus rapidement en revue ce qui ne change pas.

**Étape 1 : Validité statistique.** Elle se décompose ici en deux sous-points : (1.a) validité globale et (1.b) validités marginales.

La validité globale consiste à s'assurer que ( $H_1$ ) : au moins une variable explicative est utile pour la modélisation linéaire en rejetant ( $H_0$ ) : aucune variable n'est utile. La P-valeur pour ce test se lit dans le deuxième tableau.

Les tests de validité marginale sont menés dans le troisième tableau : pour chaque variable explicative on teste si ( $H_0$ ) : elle ne contribue pas significativement à la modélisation linéaire étant donné les autres variables déjà présentes dans le modèle contre ( $H_1$ ) : elle y contribue significativement. On teste ici l'apport incrémental d'une variable. Si au moins une variable ne contribue pas significativement au modèle linéaire, ce dernier est plus compliqué que nécessaire et pourrait être simplifié.

Si le modèle n'est pas globalement valide (rare) ou si des problèmes de validités individuelles se posent pour une ou plusieurs variables (plus fréquent), on déclare le modèle statistiquement non valide et on s'arrête là, on ne continue pas l'étude.

**Exemple.** La première régression linéaire multiple (avec les trois variables) n'est pas statistiquement valide : si elle est bien globalement valide, un problème de validité marginale se pose pour la variable Nombre de remontées (P-valeur 68.3%). Ce modèle est utile mais plus compliqué que nécessaire, on ne le retient pas.

En revanche, on peut le simplifier et supprimer la variable dont la contribution marginale n'est pas significative. On calcule ainsi le modèle Prix du forfait / Altitude maximal, Nombre de pistes et qu'au niveau des validités marginales (toutes les P-valeurs sont quasi-nulles). C'est ce second modèle, et lui seul, que nous exploiterons dans la suite.



taille d'échantillon  $n$  est grande. Pour savoir si un modèle est ou non statistiquement valide, on lit la  $P$ -valeur indiquée plus haut.

### Etape 3 : Ecriture de la relation.

Ici, c'est :

$$\text{Prix forfait (euros)} = 92.366 + 0.434 \times \text{Taille domaine (km)} + \text{résidu (d'écart-type 20.594)}$$

Il ne faut pas oublier les unités ni surtout, le terme de résidu (car la régression n'explique pas tout le phénomène, juste 57.1% du phénomène ici ; ce terme de résidu correspond aux 42.9% restants).

### Etape 4 : Validation économique / interprétation.

La validation économique consiste à se demander si la relation semble logique ; plus précisément, on se demande si le signe du coefficient de pente est logique. Ici, il est positif, ce qui veut dire que plus le domaine est de grande taille, plus le forfait hebdomadaire est cher, ce qui est parfaitement logique.

L'interprétation des coefficients consiste à commenter leur valeur numérique. Le coefficient 0.434 indique que lorsque la taille du domaine skiable augmente de 1 km, le prix du forfait augmente de 0.434 euros (approximativement et en moyenne). Pour le coefficient 92.366, on ne peut pas l'interpréter en soi, car la taille du domaine ne peut être nulle ; quand on regarde les données, on lit une taille minimale de domaine autour de 25 km. On refuse alors d'interpréter 92.366 et on interprète plutôt  $92.366 + 0.434 \times 25 = 103.216$  comme la valeur minimale du forfait de ski (approximative et en moyenne).

### Etape 5 : Prévission / Compatibilité.

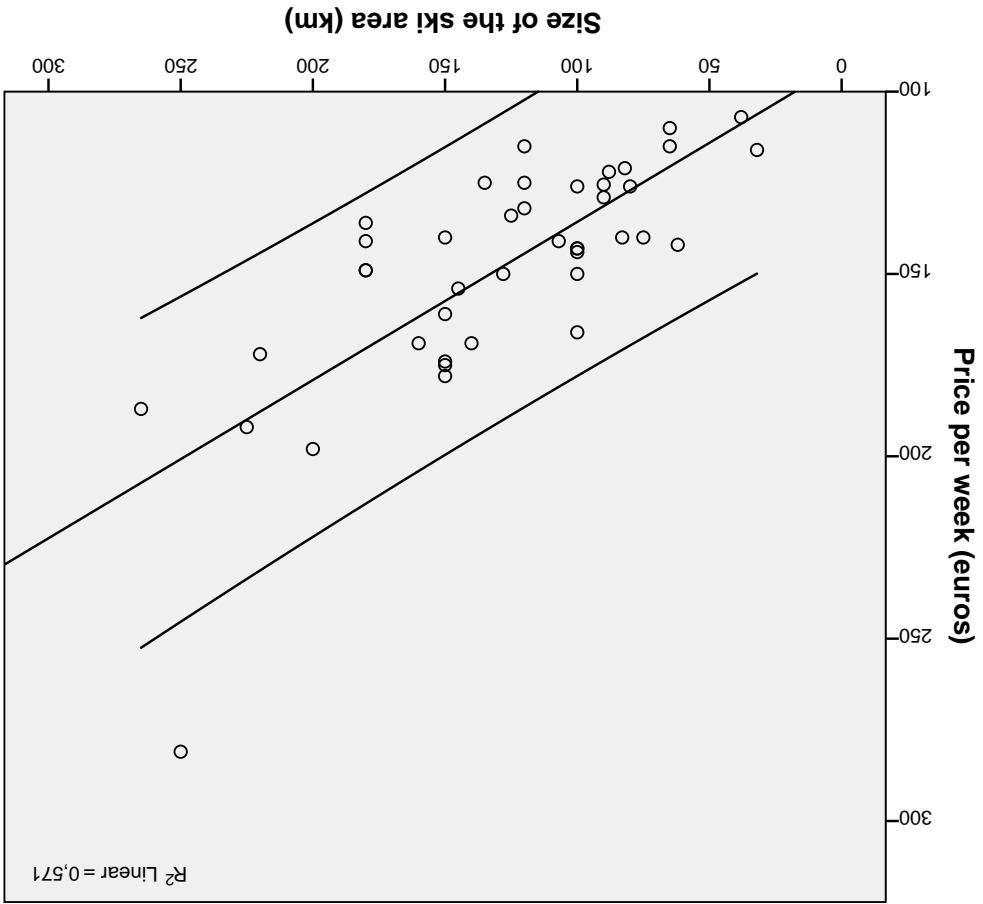
Cette dernière étape s'attache au terme de résidu : c'est une erreur, un écart au modèle déterministe formé par la droite de régression, mais dont on connaît l'ordre de grandeur de la taille. En effet, un résidu d'écart-type  $s$  prend typiquement (disons 95% du temps) des valeurs individuelles entre  $-2$  et  $+2s$ .

Cette dernière étape procure donc des fourchettes de valeurs plausibles pour des valeurs individuelles de forfaits de ski (on parle d'intervalle de prévission : car il s'agit de valeurs individuelles, par opposition aux intervalles de confiance qui portent sur des moyennes).

Par exemple, quelle est la fourchette de tarifs attendue pour un domaine skiable de taille 100 km ? C'est

$$92.366 + 0.434 \times 100 \pm 2 \times 20.594 = 135.766 \pm 41.188 \approx 136 \pm 42 \text{ euros.}$$

Le test d'une compatibilité consiste à se demander si une observation  $y$  se situe bien dans l'intervalle de prévission  $92.366 + 0.434x \pm 41.188$  associé à la valeur  $x$  de la variable explicative. Si oui, c'est parfait, l'observation est déclarée compatible avec le modèle et sinon, on parle de valeur atypique. Sur le nuage de points où nous avons représenté la droite de régression  $y = 92.366 + 0.434x$ , nous avons également représenté les droites  $y = 92.366 + 0.434x - 41.188$  et  $y = 92.366 + 0.434x + 41.188$ . Nous voyons une observation située hors de son propre intervalle de prévission (tandis que toutes les autres observations sont compatibles avec le modèle).



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,756 <sup>a</sup>	,571		20,594

a. Predictors: (Constant), Size of the ski area (km)

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22596,541	1	22596,541	
	Residual	16964,418	40	424,110	
	Total	39560,958	41	53,280	,000 <sup>b</sup>

a. Dependent Variable: Price per week (euros)

b. Predictors: (Constant), Size of the ski area (km)

Model	Unstandardized Coefficients		Std. Error	Beta	t	Sig.
	Standardized Coefficients					
1	(Constant)		92,366		11,200	
	Size of the ski area (km)		,434		7,299	
			,059		,756	
			8,247		11,200	
						,000

a. Dependent Variable: Price per week (euros)

## 8

## Synthèse — Régression linéaire simple

Dans ce chapitre, nous étudions l'explication statistique d'une variable dite variable à expliquer en fonction d'une variable explicative, selon une relation du type  $y = a + bx + \text{erreur}$ .

**Exemple :** Prix du forfait de ski en fonction de la taille du domaine. L'exemple suivi tout au long de ce chapitre est reproduit à la page suivante : la variable explicative est la taille du domaine skiable en km, et la variable à expliquer le prix du forfait en euros. On verra que la relation proposée ici est statistiquement significative, de très bonne qualité, et s'écrit :

$$\text{Prix forfait (euros)} = 92.366 + 0.434 \times \text{Taille domaine (km)} + \text{résidu (d'écart-type 20.594)}$$

**Vue d'ensemble.** Il vous suffira, à l'examen, de savoir réaliser l'analyse de régression en 5 points :

1. Validité statistique du modèle
2. Qualité du modèle
3. Ecriture de la relation proposée
4. Validation économique / Interprétation de la relation proposée
5. Préviation d'une nouvelle valeur / Vérification de compatibilité d'une observation avec le modèle

**Etape 1 : Validité statistique.** Il s'agit de voir si le coefficient de pente 0.434 est  $H_0$  : Pas significativement différent de 0 contre  $H_1$  : Significativement différent de 0. La P-valeur se lit dans la colonne "Sig." de la ligne du coefficient et est bien  $< 5\%$  (elle est même quasi-nulle). On rejette l'hypothèse de nullité du coefficient et on a bien prouvé que ce coefficient 0.434 était significativement différent de 0. Si cela n'avait pas été le cas, on aurait déclaré le modèle non statistiquement valide et on aurait arrêté là l'étude (le reste des étapes est sans objet).

**Etape 2 : Qualité du modèle.** On lit l'indicateur  $r^2 = 0.571 = 57.1\%$  dans le premier tableau. La phrase exprimant le sens de cette valeur est :

57.1% (des variations) du prix de forfait de ski sont expliquées par (les variations de) la taille du domaine skiable.

Un  $r^2$  supérieur à 20% est bon en régression linéaire simple, surtout s'agissant de phénomènes économiques.

Attention ! La validité statistique ne s'apprécie pas au vu de la valeur nominale du  $r^2$  ; des  $r^2$  bas de l'ordre de 2% ou 3% peuvent correspondre à des modèles statistiquement valides, lorsque la

Exemple :  $\chi^2$  d'indépendance (d'homogénéité).  
Considérons la sortie statistique suivante.

Group \* Number of glasses Crosstabulation

Group	8am	Count	Expected Count	Number of glasses		Total
				3 glasses or less	6 glasses or more	
10am	Count	12	11,2	7	10	29
	Expected Count	11,2	9,5	8,4	29,0	52
Total	Count	20	17	15	52,0	52
	Expected Count	20,0	17,0	15,0	52,0	52,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,335 <sup>a</sup>	2	,311
Likelihood Ratio	2,343	2	,310
Linear-by-Linear Association	,072	1	,789
N of Valid Cases	52		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 6,63.

On dispose de couples de données : 52 couples avec le nombre de verres bus la veille et le groupe de cours (8h ou 10h). Les données portent sur 23 étudiants du groupe de 8h et 29 étudiants de celui de 10h. On compare les profils de consommation d'alcool (= les lois de répartition du nombre de verres bus) par groupe. On se demande si  $H_0$  [indépendance/homogénéité] : les profils de consommation sont identiques entre les deux groupes (i.e., la consommation est indépendante du groupe) contre  $H_1$  [dépendance/absence d'homogénéité] : les profils de consommation sont différents dans les deux groupes (i.e., la consommation dépend du groupe).

Des effectifs attendus sont calculés sous l'hypothèse d'homogénéité, et confrontés aux effectifs observés. Le test est validement appliqué car la taille totale d'échantillon vaut  $n = 52 \geq 30$  et tous les effectifs attendus sont plus grands que 5. On lit la P-valeur (et la divergence du  $\chi^2$ , qui est une mesure de l'écart) dans la ligne "Pearson Chi-Square" :  $0,311 = 31,1\%$ . Elle est plus grande que 5%, on conserve  $H_0$  et on conclut :

Ces données ne mettent en évidence aucune différence significative entre les profils de consommation d'alcool des deux groupes. / Ces données ne permettent pas d'exclure que la consommation d'alcool soit la même dans les deux groupes.

Exemple :  $\chi^2$  d'ajustement.

Considérons la sortie statistique suivante.

Waiting time			
	Observed N	Expected N	Residual
<= 2 min	58	52,0	6,0
> 2 min & <= 5 min	44	41,6	2,4
> 5 min	2	10,4	-8,4
Total	104		

Test Statistics

Waiting time	Chi-Square	df	Asymp. Sig.
	7,615 <sup>a</sup>	2	,022

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,4.

On dispose d'une seule série de données : 104 temps d'attente téléphonique. La loi de référence (qui était donnée dans l'énoncé correspondant) est que 50% de ceux-ci sont compris entre 0 et 2 minutes, 40% entre 2 et 5 minutes, et 10% au-delà de 5 minutes. D'où les effectifs attendus :  $52 = 104 \times 50\%$ , puis  $41,6 = 104 \times 40\%$  et enfin,  $10,4 = 104 \times 10\%$ .

On teste donc  $H_0$  : les temps d'attente observés se répartissent selon cette loi de référence contre  $H_1$  : ils suivent une autre loi. On confronte les effectifs attendus aux effectifs observés, valant respectivement 58, 44 et 2. On le fait validement car la taille totale d'échantillon  $n = 104$  est plus grande que 30 et que tous les effectifs attendus sont plus grands que 5. (Le logiciel le signale d'ailleurs en note de bas de tableau.) On fait cela en calculant un écart, selon une formule compliquée appelée la divergence du  $\chi^2$  et valant ici 7,615. La loi limite est une loi du  $\chi^2$  à 2 degrés de liberté ("df" veut dire "degrees of freedom"). On en déduit la P-valeur 0.022 = 2.2%. Elle est plus petite que 5%, on rejette  $H_0$ . Les données ne suivent pas la loi de référence. Attention ! on compare alors les effectifs observés et attendus pour comprendre où se situent les plus grandes déviations. On se rend compte qu'il y a plus de temps d'attente courts qu'attendus, et moins de temps longs qu'attendus. On en conclut : Les temps d'attente observés ne suivent pas la loi de référence prescrite, en l'occurrence, les consommateurs attendent moins longtemps que loi de référence.

A l'examen—Nous vous demanderons simplement de

- formuler  $H_0$  et  $H_1$ ,
- constater que les conditions d'application sont vérifiées,
- lire la P-valeur
- et d'écrire une conclusion non technique et compréhensible par le grand public (et indiquant par ailleurs où les écarts significatifs à la loi de référence ont lieu lorsqu'il y en a).

Synthèse — Tests du  $\chi^2$ 

[Prononcer « ki-deux » ou « ki-carré » et écrire avec la lettre grecque  $\chi$ ]

**Hypothèses testées.** Nous voyons deux tests du  $\chi^2$  : celui d'ajustement à une loi de référence (cette loi étant précisée dans l'énoncé) et celui d'indépendance, aussi appelé test d'homogénéité. Les hypothèses qu'ils testent respectivement sont les suivantes.

**Test d'ajustement**—On dispose d'une seule série de données. On teste  $H_0$  [conformité : les données recueillies sont réparties selon la loi de référence contre  $H_1$  [non-conformité : les données sont réparties selon une autre loi.

**Test d'indépendance/homogénéité**—On dispose de couples de données, typiquement, l'indication de la catégorie à laquelle appartient l'individu statistique, d'une part, et d'autre part, une variable d'intérêt. On teste  $H_0$  [indépendance/homogénéité : la valeur de la variable d'intérêt ne dépend pas de la catégorie, i.e., les lois de répartition de la variable d'intérêt sont homogènes entre les catégories contre  $H_1$  [dépendance/absence d'homogénéité : la catégorie influe sur la valeur de la variable d'intérêt, i.e., les lois de répartition de la variable d'intérêt ne sont pas toutes égales selon les catégories.

Nous vous donnons ci-dessous des traductions concrètes de ces phrases générales sur des exemples.

**Conditions d'application.** Les tests du  $\chi^2$  requièrent que la taille totale d'échantillon  $n$  soit plus grande que 30 et que tous les effectifs attendus (calculés par le logiciel statistique) soient plus grands que 5. Une note de bas de tableau vous indique si la seconde condition est vérifiée, mais n'oubliez pas d'indiquer également la première.

[Exemples : voir pages suivantes]

ID	Variable 1	Variable 2	Différence
1	16	12	4
2	18	16	2
3	30	29	1
4	29	32	-3
5	18	17	1
6	22	25	-3
7	46	42	4
8	32	38	-6
9	33	34	-1
...	...	...	...

**Exemple :** Reprenons notre exemple sur les temps moyens d'écrans. Les individus statistiques

sont données par les 150 familles considérées. On effectue des mesures initiales  $x_1, \dots, x_{150}$  de temps d'écrans ; on obtient une moyenne de  $\bar{x}_{150} = 103$  minutes (et un écart-type de  $s_{x,150} = 37$  minutes). Les parents effectuent un stage de sensibilisation aux côtés néfastes des écrans, et on effectue une seconde campagne de mesures de temps d'écrans, donnant lieu aux observations  $y_1, \dots, y_{150}$ , de moyenne  $\bar{y}_{150} = 99$  minutes (et d'écart-type  $s_{y,150} = 34$  minutes). On calcule les écarts  $z_j = x_j - y_j$  : leur moyenne est  $\bar{z}_{150} = 4$  minutes et leur écart-type vaut  $s_{z,150} = 15$  minutes. Le stage a-t-il été efficace ? Ici, on considère un test unilatère  $H_0$  : moyennes pas significativement différentes contre  $H_1$  : la moyenne a significativement baissé après le stage (personne n'ira penser que le stage de sensibilisation puisse encourager les temps d'écrans !). C'est-à-dire que l'on va tester  $H_0 : \mu_0 = 0$  contre  $H_1 : \mu_0 > 0$ , où  $\mu_0$  est le temps moyen (sur l'ensemble des familles de France) de la différence du temps d'écrans qu'on observerait grâce au stage (dans le sens « avant moins après »). La statistique de test suit une loi normale sous  $H_0$ , elle prend des valeurs plus grandes sous  $H_1$ , et sa valeur numérique est

$$t_{150} = \frac{\sqrt{150}}{103 - 99} \frac{15}{15} \approx 3.26.$$

La P-valeur est très faible, on rejette fermement  $H_0$ , et on conclut : Ces données montrent que le stage de sensibilisation réduit bien le temp moyen passé devant les écrans.

4. Si besoin, un intervalle de confiance symétrique sur la différence des moyennes est fourni dans la partie droite du tableau.

**Exemple :** Les tableaux reproduits traitent des salaires moyens pour les hommes et les femmes dans la CSP 37 (cadres supérieurs) et en région parisienne. On se fonde sur deux échantillons indépendants, composés respectivement de 179 hommes et de 147 femmes. Les salaires moyens lus dans l'échantillon sont de  $\bar{x}_{179} = 3431.46$  euros pour les hommes et  $\bar{y}_{147} = 2434.90$  euros pour les femmes. Pour voir s'ils sont significativement différents, on effectue donc un pré-test d'égalité des variances (F-valeur de 0.4%, on rejette l'hypothèse d'égalité des variances et on lit la seconde ligne du tableau), puis le test d'égalité des moyennes proprement dit. Sa P-valeur est de 0.1%, on rejette donc l'hypothèse  $H_0$  d'égalité des moyennes, c'est-à-dire que l'on a montré que les salaires moyens des hommes et des femmes sont significativement différents (on a mis en évidence un effet). En l'occurrence, lorsque l'on regarde les salaires moyens d'échantillon, on a prouvé que les hommes gagnaient davantage en moyenne que les femmes dans cette catégorie (c'est ce que l'on appelle une direction d'effet). On pourrait même ajouter une phrase mettant en jeu l'intervalle de confiance [386.074 – 1607.047], en arrondissant ce dernier en l'intervalle [350 – 1650] pour plus de lisibilité; cette quantification est ce qu'on appelle une mesure de la taille d'effet. Résumons notre conclusion :

Ces données montrent que les salaires moyens des hommes et des femmes de cette catégorie sont différents [effet], en l'occurrence, les hommes gagnent davantage en moyenne [direction d'effet] : avec grande confiance, l'écart moyen de salaires mensuels se situe quelque part entre 350 et 1650 euros [taille d'effet].

### 3. Moyennes appariées

Deux échantillons sont appariés lorsqu'ils sont obtenus à partir des mêmes individus statistiques, sur lesquels deux mesures ont été effectuées. Ces deux mesures sont très liées. On pourrait par exemple : — Comparer les prix de produits (= individus statistiques) dans deux magasins différents ; les deux prix relevés sont très liés, car il y a des ordres de grandeur raisonnables pour le prix d'un pack de yaourts ou d'une bouteille de vodka.

— Comparer les salaires de l'enfant aîné et du suivant : les individus statistiques sont les frères, la première mesure est le salaire de l'aîné, la seconde mesure, celui du cadet. Les deux mesures sont liées car l'aîné et le cadet sont extraits du même milieu social et ont donc des trajectoires sociales comparables.

— De même pour la comparaison des salaires femmes-hommes au sein des couples hétérosexuels (= individus statistiques).

Puisque les mesures sont très liées, le mieux pour les comparer est de s'intéresser à leurs différences. Ainsi, on peut mettre en évidence l'effet des conditions (magasin 1 vs. magasin 2, aîné vs. cadet) par rapport à l'ordre de grandeur commun. On rajoute donc une colonne dans le tableau de principe des données appariées, qui était en première page de ce chapitre (voir figure en haut de la page suivante). Et on va tester si la moyenne des différences est significativement différente de 0 ou non. Comme on n'a plus qu'une seule série de données, à savoir les  $z_j = x_j - y_j$  pour  $j = 1, \dots, n$ , il suffit d'appliquer un test de comparaison à une valeur de référence, en l'occurrence,  $\mu_{ref} = 0$ . On se ramène au chapitre précédent ! En particulier, la statistique de test considérée est

$$t_n = \sqrt{n} \frac{\bar{z}_n - 0}{\frac{s_{z,n}}{\sqrt{n}}} = \frac{\bar{x}_n - \bar{y}_n}{\frac{s_{x-y,n}}{\sqrt{n}}};$$

la normalisation est ici donnée par  $s_{x-y,n}/\sqrt{n}$ , où  $s_{x-y,n}$  est l'écart-type  $s_{x-y,n}$  de la suite des différences  $z_j = x_j - y_j$ . (Sa valeur sera donnée par l'énoncé.)



Version complète en anglais

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Monthly net salary	Men	179	3431.46	3895.437	291.159
	Women	147	2434.90	1282.947	105.816

		Levene's Test for Equality of Variances		t-test for Equality of Means						
				F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
		Lower	Upper							
Monthly net salary	Equal variances assumed	8,195	,004	2,972	324	,003	996,560	335,371	336,781	1656,339
	Equal variances not assumed			3,217	223,375	,001	996,560	309,791	386,074	1607,047

Version simplifiée en français

Catégorie	Tailles d'échant.	Moyennes d'échantillon		
1ère	179	3431,46		
2nde	147	2434,90		

	Pré-test d'égalité des variances (H0 : égalité / H1 : var. différentes)		Test d'égalité des moyennes (H0 : égalité / H1 : moyennes différentes)					
	P-valeur	t	P-valeur (bilatère)	Intervalle de confiance à 95% sur la différence des moyennes				
Variances pas significativement différentes (H0 conservée)	,004 (soit 0.4%)	Statistique de test	,003 soit 0.3%					
Variances significativement différentes (H0 rejetée)			,001 soit 0.1%					

la population a-t-elle changé ? » et n'indiquant pas d'a priori, nous allons recourir à une hypothèse  $H_1$  bilatère et tester  $H_0 : p_1 = p_2$  contre  $H_1 : p_1 \neq p_2$ , où  $p_1$  et  $p_2$  sont respectivement les parts d'habitants satisfaits de la gestion de la première et de la deuxième crise. La proportion commune sous  $H_0$  admet pour estimation numérique

$$\hat{p} = \frac{530 + 492}{1000 + 1002} = 51.05\%,$$

de sorte que la statistique de test prend la valeur

$$t_{1000,1002} = \frac{530/1000 - 492/1002}{0.5105(1 - 0.5105)(1/1000 + 1/1002)} \approx 1.74.$$

Un schéma (que nous vous laissons dessiner) montre que la P-valeur vaut alors

$$P = 2 \times (100\% - 95.91\%) \approx 8.2\% > 5\%.$$

On conserve  $H_0 : p_1 = p_2$ . On traduit cela en une phrase « grand public » ; voici deux exemples de telles phrases :

Ces données ne permettent pas d'exclure que les taux de satisfaction soient identiques lors de la gestion des deux crises.  
Les taux de de satisfaction des actions de gestion de crise ne sont pas significativement différents entre les deux crises.

## 2. Moyennes indépendantes

On discute ici de deux échantillons de données obtenus indépendamment mais visant à comparer non plus des proportions mais des moyennes. Ce qui suit explique comment tester  $H_0 : \mu_1 = \mu_2$  contre  $H_1 : \mu_1 \neq \mu_2$ .

### Principe (plus compliqué).

Le test repose toujours sur une statistique de test de la forme

$$t_{n,m} = \frac{\bar{x}_n - \bar{y}_m}{\text{normalisation}},$$

mais la normalisation est beaucoup plus compliquée à énoncer ; sa formule dépend du fait que les écarts-types (variances) des deux échantillons sont égaux ou différents. Dans le premier cas, on effectuera une estimation conjointe de la valeur commune ; dans le second cas, on estimera chaque écart-type (chaque variance) séparément. Dans les deux cas, les formules sont longues, compliquées et impossibles à mettre en œuvre à la main : nous avons besoin de l'assistance d'un logiciel statistique ! Qui fournit des sorties similaires à celles reproduites à la page suivante. Les tableaux du haut sont les tableaux originaux et les tableaux du bas expliquent quelles cases sont utiles.

Notez en particulier que nous avons changé le texte de nombreuses cases pour mieux expliquer (en français) ce qui y est reporté. En l'occurrence, on lit, dans cet ordre :

1. La P-valeur du pré-test d'égalité des variances (pour savoir quelle formule utiliser ensuite) ;
2. Si cette P-valeur est supérieure à 5%, les variances ne sont pas significativement différentes ; si elle est inférieure à 5%, les variances sont significativement différentes ; pour la suite, on utilise la bonne ligne (parmi les deux lignes) du grand tableau ;
3. On passe au test d'égalité des moyennes proprement dit : on lit la P-valeur dans la colonne dédiée ;

Nous allons traiter trois cas seulement (car le quatrième cas, avec les proportions apparées, est trop compliqué à notre niveau) :

- les proportions indépendantes,
- les moyennes indépendantes,
- les moyennes apparées.

Dans tous ces cas, la statistique de test est donnée par la différence  $\bar{x}_n - \bar{y}_m$  entre les deux moyennes d'échantillon, correctement renormalisée. Plus précisément, elle sera de la forme

$$t_{n,m} = \frac{\bar{x}_n - \bar{y}_m}{\text{normalisation}},$$

où [normalisation] est une quantité telle que la statistique de test  $t_{n,m}$  suive (approximativement) une loi normale ou une loi proche sous  $H_0$ . Nous allons bien entendu vous dire dans chaque cas quelle est sa formule.

## 1. Proportions indépendantes

**Rappel.** Dans le cas d'une série de données et d'une comparaison à une proportion de référence, nous utilisons la statistique de test

$$t_n = \sqrt{n} \frac{\bar{x}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})}} = \frac{\bar{x}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})/n}}.$$

**Extension au test de deux proportions indépendantes.** Ici, nous allons remplacer le numérateur par  $\bar{x}_n - \bar{y}_m$  et adapter le dénominateur pour tenir compte du fait que nous n'avons pas de

proportion de référence  $p_{\text{ref}}$  et que nous avons deux tailles d'échantillon  $n$  et  $m$  : la statistique de test sera donnée par

$$t_{n,m} = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\hat{p}(1 - \hat{p})(1/n + 1/m)}},$$

où  $\hat{p}$  est une proportion que nous définissons maintenant. Nous allons considérer l'hypothèse  $H_0$  :  $p_1 = p_2$ . Sous cette hypothèse, la meilleure manière d'estimer la proportion de population commune  $p_1 = p_2$ , c'est de mettre toutes les données dans le même sac :

$$\hat{p} = \frac{x_1 + x_2 + \dots + x_n + y_1 + y_2 + \dots + y_m}{n + m}.$$

Le principe du test est conclu par l'indication que sous  $H_0$  :  $p_1 = p_2$ , la statistique de test suit approximativement une loi normale, tandis que sous  $H_1$  :  $p_1 > p_2$ , elle prend de grandes valeurs positives, sous  $H_1$  :  $p_1 < p_2$ , de grandes valeurs négatives, et sous  $H_1$  :  $p_1 \neq p_2$ , de grandes valeurs négatives ou positives.

**Exemple :** Nous reprenons notre exemple des sondages politiques sur la gestion de crise. La perception de la population a-t-elle changé entre les deux crises ? Un premier sondage, effectué lors de la crise de mars-avril, indique que sur  $n = 1000$  habitants sondés, une part de  $\bar{x}_{1000} = 530/1000 = 53\%$  pense que cette première crise est bien gérée. (Ce sont les données que nous considérons jusqu'à présent.) Un second sondage, effectué 7 mois plus tard lors de la crise d'octobre-novembre, indique que sur  $m = 1002$  habitants sondés, une part de  $\bar{y}_{1002} = 492/1002 \approx 49.1\%$  pense que cette deuxième crise est bien gérée. Notez bien que les deux échantillons utilisés pour les sondages n'ont rien à voir entre eux et ont été choisis indépendamment au sein de la population ; c'est cela qui justifie le caractère de données indépendantes. La question initialement posée étant neutre (« La perception de

6

Synthèse — Tests de comparaison de deux échantillons

On part de deux séries de données  $x_1, \dots, x_n$  et  $y_1, \dots, y_m$ , parfois de même longueur, parfois de longueurs différentes, et on veut déterminer si les moyennes d'échantillon  $\bar{x}_n$  et  $\bar{y}_m$  sont significativement différentes, c'est-à-dire si au vu de leurs valeurs, on peut conclure que les moyennes de population  $\mu_1$  et  $\mu_2$ , ou les proportions de population  $p_1$  et  $p_2$  correspondantes sont différentes.

Types de données.

Comme précédemment, les données peuvent porter sur des proportions ou des moyennes générales. Une seconde distinction est également à opérer, selon que les données sont — apparées, c'est-à-dire obtenues en effectuant deux mesures  $x_j$  et  $y_j$  sur chaque individu statistique  $j$  ; dans ce cas, on a nécessairement  $n = m$  ; — indépendantes, c'est-à-dire que les individus choisis pour obtenir les  $x_j$  n'ont rien à voir avec les individus choisis pour obtenir les  $y_j$  ; dans ce cas, les deux séries de données peuvent en particulier être de longueurs  $n$  et  $m$  différentes.

Dans un tableau, les données se présenteraient ainsi :

Données apparées			vs.			Données indépendantes		
ID	Variable 1	Variable 2	ID	Group	Variable	ID	Group	Variable
1	16	12	1	0	12	1	0	12
2	18	16	2	1	16	2	1	16
3	30	29	3	0	29	3	0	29
4	29	32	4	0	32	4	0	32
5	18	17	5	1	17	5	1	17
6	22	25	6	1	25	6	1	25
7	46	42	7	0	42	7	0	42
8	32	38	8	0	38	8	0	38
9	33	34	9	1	34	9	1	34
...	...	...	...	...	...	...	...	...
Ex : ID = Produits			Ex : ID = Clients			Group = Avec ou sans promotion		
Variable 1 = prix dans magasin 1			Variable = Montant d'achat					
Variable 2 = prix dans magasin 2								
→ On compare des colonnes						→ On compare des lignes		

peut calculer comme valant 2.87% à l'aide de la table de la loi normale standard en dernière page du polycopié. Cette P-valeur est inférieure à 5% : on rejette  $H_0$  et on a établi  $H_1$ , ce que l'on traduit en clair de la manière suivante :

Ces données montrent que plus de 50% de la population trouve que la crise est bien gérée.

#### Cas d'un parti d'opposition.

Il veut montrer que la crise est mal gérée et choisit, avant d'effectuer le sondage, les hypothèses  $H_0 : p_0 \geq 50\%$  contre  $H_1 : p_0 < 50\%$ . Son espoir est que les données permettent de rejeter  $H_0$  et de passer à  $H_1$ . Pour traiter ces hypothèses, on passe au cas limite et on se ramène à tester  $H_0 : p_0 = 50\%$  contre  $H_1 : p_0 < 50\%$ . Alors seulement on effectue le sondage et on recueille les données. Zut, ces dernières montrent que sur l'échantillon, une proportion  $\bar{x}_{1000} = 53\%$  des sondés trouvent la crise bien gérée. Les données d'échantillon ne sont pas en désaccord avec l'hypothèse  $H_0 : p_0 \geq 50\%$  et on ne pourra donc pas la rejeter, il faudra la conserver. Le mieux est de dire cela et de le traduire en langage grand public par :

Ces données ne permettent pas d'exclure que plus de 50% de la population trouve que la crise est bien gérée.

Si l'on veut toutefois faire les calculs et obtenir une P-valeur, c'est possible : la figure du milieu indique que l'on obtient alors une P-valeur de 97.13%. Cela confirme la conservation de  $H_0$  (mais on ne valide pas pour autant  $H_0$ , même avec cette grande P-valeur : nous sommes liés par les hypothèses choisies avant le recueil des données).

**Cas d'un observateur neutre, ouvert à tous les possibles.** Un observateur neutre est ouvert à toutes les alternatives possibles : que le taux de satisfaction  $p_0$  dans la population soit plus grand ou plus petit que 50%. Il n'en privilégie aucune (ce qui, d'ailleurs, lui évitera d'être frustré s'il s'est lié par avance avec des hypothèses unilatérales et le regrette). Il considérera donc toujours l'hypothèse  $H_0 : p_0 = 50\%$  mais cette fois-ci contre l'hypothèse bilatère  $H_1 : p_0 \neq p_{\text{ref}}$ . Dit autrement, on se demande si la proportion d'échantillon  $\bar{x}_{1000}$  est significativement différente de 50% (alors que dans les cas précédents, on se demandait respectivement si elle était significativement supérieure ou inférieure à 50%). Les déviations sous  $H_1$  ont lieu des deux côtés et il faut en tenir compte ; la figure du bas l'illustre. On obtient une P-valeur doublée par rapport au premier cas, de 5.74%. Elle est très légèrement supérieure à 5%, nous sommes vraiment dans une situation limite où nous pourrions dire que les données prouvent presque que le taux  $p_0$  est différent de 50%, en l'occurrence (vu la valeur de  $\bar{x}_{1000}$ ), qu'il est supérieur à 50%. Bien évidemment, si à l'examen vous dites simplement que la P-valeur est supérieure à 5% et que donc on conserve  $H_0$ , et que rien ne montre que le taux de satisfaction soit différent de 50%, vous aurez tous les points. Mais ici, grâce à la P-valeur, on voit que cette conservation se fait sur le fil. La P-valeur est une information plus riche sur l'accord entre  $H_0$  et les données qu'une simple réponse binaire de conservation ou rejet de  $H_0$ .

## 4. Test de comparaison à une moyenne de référence

Les choses se déroulent de manière tout à fait similaire que pour le cas d'une comparaison à une proportion. On teste  $H_0 : \mu_0 = \mu_{\text{ref}}$  contre une hypothèse  $H_1$  de la forme indiquée page 7. La seule différence par rapport au cas des proportions est la statistique de test, dont la formule est

$$t_n = \sqrt{n} \frac{\bar{x}_n - \mu_{\text{ref}}}{s_{x,n}},$$

où  $s_{x,n}$  désigne l'écart-type des données (sa valeur toujours indiquée dans l'énoncé).

correspondant à  $H_1$  et déterminée par la valeur numérique de  $t_n$ . Voir les dessins (que nous commentons

plus en détails ci-dessous).

Il faut voir la P-valeur comme un index de crédibilité de  $H_0$  au vu des données et de  $H_1$  : si la P-valeur est faible, c'est que la valeur de la statistique de test  $t_n$ , qui est rappelez-le, une mesure normalisée de l'écart de  $\bar{x}_n$  à  $p_{ref}$ , est trop grande pour être une valeur probable sous  $H_0$ . C'est donc que l'écart entre  $\bar{x}_n$  et  $p_{ref}$  est significatif. On rejettera donc  $H_0$ . Au contraire, une P-valeur plus grande indiquera un accord ou un désaccord léger mais pas rhédictoire entre  $H_0$  et les données : l'écart normalisé  $t_n$  n'est pas une valeur suffisamment improbable. On conservera donc  $H_0$ . En pratique, et au moins dans le monde du business, on détermine le côté probable / improbable au seuil de 5%. Une P-valeur plus grande que 5% invite à la conservation de  $H_0$ , une P-valeur plus petite que 5% invite au rejet de  $H_0$ .

### Écriture des conclusions.

Lorsque la P-valeur est plus grande que 5%, on conserve  $H_0$  et la conclusion doit traduire qu'on conserve l'hypothèse de départ sans toutefois la valider. On écrira une phrase sur les modèles suivants (à adapter à chaque cas d'espèce en remplaçant  $H_0$  ou  $H_1$  par des mots) :

Ces données ne permettent pas d'exclure  $H_0$ .

Ces données ne permettent pas de mettre en évidence  $H_1$ .

Il s'agit donc de conclusions prudentes et vagues, souvent formulées avec des doubles négations, et peu exploitables côté business.

Lorsque la P-valeur est plus petite que 5%, on rejette  $H_0$  et la conclusion doit traduire que nous sommes départis de l'hypothèse initiale, que nous l'avons invalidée avec une certitude raisonnable. On peut cette fois-ci être affirmatif :

Ces données excluent  $H_0$ .

Ces données mettent en évidence  $H_1$ .

Il s'agit donc de conclusions nettes et exploitables côté business.

## 3. Test de comparaison à une proportion de référence : exemple

Nous reprenons notre exemple fil rouge de satisfaction face à la gestion d'une crise et allons varier les commanditaires de l'étude. Ci-dessus, nous avions déjà discuté le cas du parti au pouvoir.

### Cas du parti au pouvoir.

Il veut montrer que la crise est bien gérée et choisit, avant d'effectuer le sondage, les hypothèses  $H_0 : p_0 \leq 50\%$  contre  $H_1 : p_0 > 50\%$ . Son espoir est que les données permettent de rejeter  $H_0$ . Pour traiter ces hypothèses, on passe au cas limite et on se ramène à tester  $H_0 : p_0 = 50\%$  contre  $H_1 : p_0 > 50\%$ . On rappelle qu'on interroge  $n = 1000$  sondés. Le comportement de la statistique de test

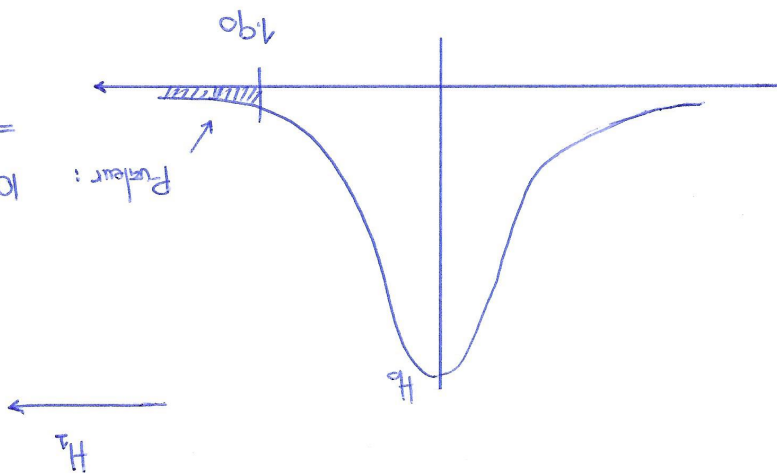
$$t_{1000} = \frac{\sqrt{1000}(\bar{x}_{1000} - 0.5)}{\sqrt{0.5(1 - 0.5)}}$$

est représentée sur la figure du haut : loi normale standard sous  $H_0$ , grandes valeurs positives sous  $H_1$ . On calcule sa valeur numérique :

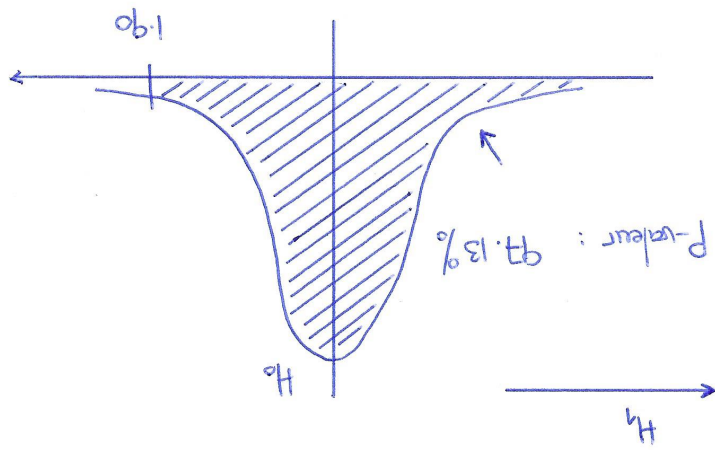
$$t_{1000} = \frac{\sqrt{1000}(0.53 - 0.5)}{\sqrt{0.5(1 - 0.5)}} \approx 1.90.$$

Dit autrement, l'écart « brut » de  $\bar{x}_{1000} - p_{ref} = 53\% - 50\% = 3\%$  est normalisé en un écart valant 1.90. La P-valeur est ici la probabilité qu'une loi normale standard soit plus grande que 1.90 et qu'on

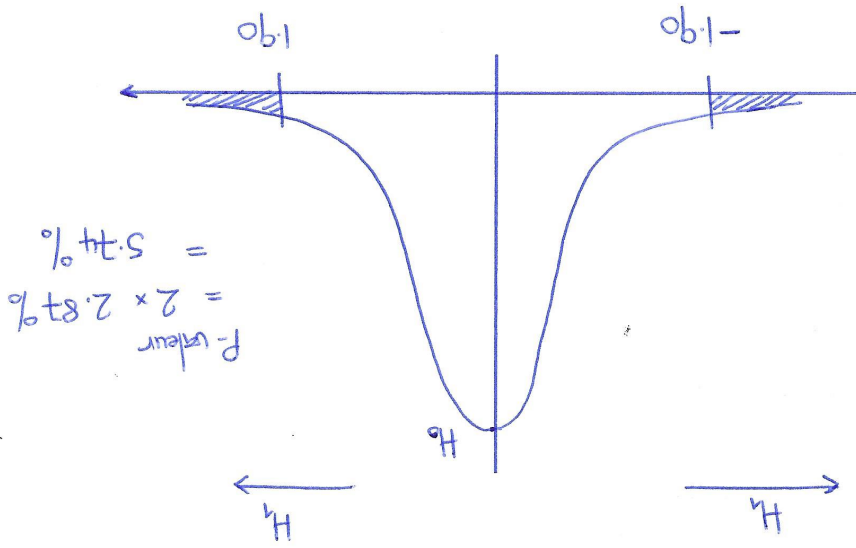
Cas du parti  
au pouvoir



Cas d'un  
parti d'opposition



Cas d'une  
observation  
neutre



**Cas limite.** Tester  $H_0 : p_0 \leq p_{ref}$  contre  $H_1 : p_0 > p_{ref}$  revient à tester  $H_0 : p_0 = p_{ref}$  contre  $H_1 : p_0 > p_{ref}$ . Et de même pour les autres cas. C'est ce que l'on appellera passer au cas limite : se ramener à une égalité pour  $H_0$ .

### Exemple de mise en œuvre du choix d'hypothèses

Partons d'une situation abstraite dans laquelle on veut effectuer un test sur une proportion (inconnue)  $p_0$  de la population. Des raisons externes nous amènent à considérer une proportion de référence  $p_{ref}$ . Supposons qu'on teste  $H_0 : p_0 = p_{ref}$  contre  $H_1 : p_0 > p_{ref}$ . On fixe ces hypothèses et on collecte ensuite (et ensuite seulement!) des données  $x_1, \dots, x_n$  sur  $n$  sondés. Notons  $\bar{x}_n$  la proportion d'échantillon observée. On se demande si la valeur  $\bar{x}_n$ , qui est proche de  $p_0$ , est compatible avec  $p_{ref}$  (aucun cas on conservera  $H_0$ ) ou si elle est le signe que  $p_0$  est plus grande que  $p_{ref}$  (aucun cas on rejette  $H_0$ ).

**Exemple :** On s'interroge sur la proportion  $p_0$  de la population trouvant que la crise est bien gérée et on se demande si cette proportion est plus grande que la valeur de référence  $p_{ref} = 50\%$ . Si l'on est le parti au pouvoir, on choisit les hypothèses  $H_0 : p_0 \leq 50\%$  contre  $H_1 : p_0 > 50\%$  parce que l'on veut montrer que la majorité de la France est derrière nous (ce qui correspond à  $H_1$ ). Les données de sondage sur  $n = 1\,000$  sondés sont que  $\bar{x}_{1000} = 53\%$  des sondés trouvent la crise bien gérée. On se demande si cette valeur d'échantillon de  $53\%$  nous permet d'affirmer qu'en effet, la majorité de la France est derrière nous, c'est-à-dire de rejeter l'hypothèse  $H_0 : p_0 \leq 50\%$ . On peut reformuler cela comme : on se demande si la valeur d'échantillon  $\bar{x}_{1000} = 53\%$  est significativement supérieure à  $50\%$ , auquel cas on rejette  $H_0$ , ou si au contraire, l'écart de 3 points peut n'être dû qu'à l'aléa d'échantillonnage (l'aléa ayant donné lieu au recrutement de cet échantillon-là et pas un autre). Dans ce dernier cas, on conservera  $H_0$  et on dira que l'écart de 3 points n'est pas significatif. Il y aurait alors contradiction entre les données et l'hypothèse  $H_0$ , mais trop légère pour qu'on rejette  $H_0$ .

## 2. Test de comparaison à une proportion de référence : théorie

**Statistique de test = écart renormalisé.** On s'intéresse à ce que l'on appelle une statistique de test, qui normalise l'écart entre  $\bar{x}_n$  et  $p_{ref}$  afin de lui donner un comportement plus universel : selon une loi normale standard sous l'hypothèse  $H_0 : p_0 = p_{ref}$ . Cette statistique de test a pour expression

$$t_n = \sqrt{n} \frac{\bar{x}_n - p_{ref}}{\sqrt{p_{ref}(1 - p_{ref})}}.$$

Son comportement sous  $H_0$  est représenté sur les dessins par une courbe en cloche, qui représente les probabilités des valeurs prises par la loi normale standard. Cette statistique de test  $t_n$  prend donc sous  $H_0$  des valeurs autour de 0, typiquement entre  $-2$  et  $+2$ .

Sous  $H_1$ , elle suit un comportement différent (qui provient du fait que  $\bar{x}_n$  est proche de  $p_0$ ) :

- Sous  $H_1 : p_0 > p_{ref}$ , la statistique de test prend des valeurs positives grandes. On place une flèche  $\rightarrow$  sur le dessin permettant de calculer la P-valeur, de même sens que le signe  $>$  dans l'hypothèse  $p_0 > p_{ref}$ .
- Sous  $H_1 : p_0 < p_{ref}$ , ce sont des valeurs négatives, représentées par une flèche  $\leftarrow$  de même sens que le signe  $<$ .
- Sous  $H_1 : p_0 \neq p_{ref}$ , les deux sont possibles, on place deux flèches  $\leftarrow$  et  $\rightarrow$ .

**P-valeur.** Pour calculer la P-valeur, on calcule d'abord la valeur numérique (sur les données) de la statistique de test. La P-valeur est alors la probabilité sous la loi normale standard de la zone





## Synthèse — Tests d'hypothèses : généralités et comparaison à une valeur de référence

**Dissymétrie des hypothèses en jeu.** Les tests d'hypothèses permettent de répondre à une question. Mais la réponse n'est pas strictement binaire : soit on pourra exclure l'assertion de départ (on la tiendra pour fausse et on validera l'hypothèse alternative) ; soit on ne pourra pas l'exclure (mais sans la valider pour autant : on ne saura pas si cette assertion de départ est vraie ou fausse, simplement, avec les données disponibles, on ne peut pas prouver qu'elle soit fausse). On note donc une grande dissymétrie entre l'hypothèse de départ  $H_0$ , que l'on ne pourra jamais valider, et l'hypothèse alternative  $H_1$ , qu'on pourra valider. Le vocabulaire est éclairant : soit on rejette  $H_0$ , soit on conserve  $H_0$  (mais peut-être faute d'avoir suffisamment de données pour la rejeter). Les tests ont fondamentalement tendance à conserver  $H_0$  ; ils ne la rejettent que si elle est gravement contredite par les données.

### 1. Choix des hypothèses

Plusieurs critères peuvent être mis en jeu, tous vont dans le sens de l'introduction ci-dessus mais certains s'appliquent mieux que d'autres pour chaque cas concret :

- $H_0$  est l'hypothèse de laquelle il est raisonnable de partir ou à laquelle on est attaché (l'opinion commune, l'avis du manager, « comme on a toujours pensé », etc.), et  $H_1$  est l'hypothèse contraire, qui susciterait l'étonnement ou la réflexion.
- Lorsque l'on veut établir un fait,  $H_1$  est ce que l'on veut prouver, et on prend simplement pour  $H_0$  le contraire de  $H_1$ .
- Quand une action risquée est envisagée et qu'évidemment, on ne veut l'entreprendre qu'à bon escient, on prend pour  $H_1$  la situation correspondant à cette action risquée et pour  $H_0$ , son contraire, qui correspond à une vision prudente de la situation.

**Attention à bien placer l'égalité dans  $H_0$  !** Le cas d'égalité est toujours dans  $H_0$ , pour des raisons mathématiques profondes. En pratique, nos  $H_0$  seront toujours d'une des formes suivantes :

- $H_0 : p_0 = p_{ref}$  ou  $H_0 : \mu_0 = \mu_{ref}$
  - $H_0 : p_0 \leq p_{ref}$  ou  $H_0 : \mu_0 \leq \mu_{ref}$
  - $H_0 : p_0 \geq p_{ref}$  ou  $H_0 : \mu_0 \geq \mu_{ref}$
- Les hypothèses  $H_1$  seront elles toujours d'une des formes suivantes :
- $H_1 : p_0 \neq p_{ref}$  ou  $H_1 : \mu_0 \neq \mu_{ref}$
  - $H_1 : p_0 > p_{ref}$  ou  $H_1 : \mu_0 > \mu_{ref}$
  - $H_1 : p_0 < p_{ref}$  ou  $H_1 : \mu_0 < \mu_{ref}$

**Exemple :** Pour l'estimation de la proportion de population  $p_0$  de l'ensemble des habitants trouvant que la crise est bien gérée, on a, avec les  $m = 1\,000$  sondés actuels, une marge d'erreur à  $\pm 3.1\%$ . Supposons que l'on veuille plutôt une précision de l'estimation à  $\pm 1.5\%$ , c'est-à-dire, une division de la marge d'erreur par un facteur  $a = 3.1/1.5$ . Cela requerrait au total

$$n = 1\,000 a^2 = 1\,000 (3.1/1.5)^2 = 4\,271.11 \dots \approx 4\,272 \quad \text{sondés,}$$

soit  $4\,272 - 1\,000 = 3\,272$  sondés de plus.

## Synthèse — Intervalles de confiance, résultats plus avancés

### 1. Intervalles de confiance simultanés

Toute exploitation simultanée de deux intervalles de confiance au niveau 95% (ce qui est le cas de tous les intervalles donnés ci-dessus) vaut avec confiance seulement 90%.

**Exemple :** En reprenant les exemples du chapitre précédent, avec grande confiance (mais de niveau seulement 90%), on peut simultanément dire que la part  $p_0$  de l'ensemble des habitants trouvant que la crise est bien gérée vaut  $53\% \pm 3.1\%$  et que le temps moyen  $\mu_0$  passé devant les écrans par l'ensemble des enfants de 6 ans est de  $103 \pm 6$  minutes.

### 2. Planification

Le problème de la planification consiste à se demander quelle taille d'échantillon constituer pour parvenir à une précision donnée. On procède en deux temps : on effectue un premier coup de sonde (sur  $m \geq 30$  individus), et on calcule ensuite le nombre de données  $n - m$  supplémentaires à récupérer. On applique les formules d'intervalles de confiance avec une marge d'erreur (les formules symétriques avec  $\pm \dots$ ), et uniquement celles-ci.

Plus précisément, avec les  $m \geq 30$  données préliminaires, on constate qu'on a pour l'instant une précision (une marge d'erreur) de

$$\gamma = 1.96 \frac{\sqrt{s_{x,m}}}{m} \quad \text{ou} \quad \gamma = 1.96 \sqrt{\frac{x_m(1 - x_m)}{m}}.$$

On se fixe un vœu de précision à  $\pm \epsilon$ , c'est-à-dire qu'il faut réduire la précision actuelle (la marge d'erreur actuelle) d'un facteur  $\alpha = \gamma/\epsilon$ . Pour cela, il faudra un échantillon de taille au total de  $n = m\alpha^2$ .

En clair, toute division de la marge d'erreur d'un facteur  $\alpha$  requiert une multiplication de la taille d'échantillon par un facteur  $\alpha^2$ .

**Exemple :** Pour l'estimation du temps moyen  $\mu_0$  passé chaque jour par l'ensemble des enfants de 6 ans devant des écrans, on a, avec les  $m = 150$  enfants observés actuels, une marge d'erreur à  $\pm 6$  minutes. Supposons que l'on veuille plutôt une précision de l'estimation à  $\pm 3$  minutes, c'est-à-dire, une division de la marge d'erreur par un facteur  $\alpha = 2$ . Cela requerrait au total  $n = 150 \times 2^2 = 600$  observations d'enfants, soit  $600 - 150 = 450$  de plus qu'actuellement.

**Estimation basse.** Avec grande confiance, la moyenne de population  $\mu_0$  est plus grande que

$$\bar{x}_n - 1.645 \frac{s_{x,n}}{\sqrt{n}}.$$

**Exemple :** Avec  $\bar{x}_{150} = 103$  et  $s_{x,150} = 37$ , obtenus pour un échantillon de taille  $n = 150$ , on déduit qu'avec grande confiance, le temps moyen  $\mu_0$  passé devant les écrans par l'ensemble des enfants de 6 ans est au moins de

$$103 - 1.645 \frac{37}{\sqrt{150}} = 98.0303 \approx 98 \text{ minutes}.$$

**Estimation haute.** Avec grande confiance, la moyenne de population  $\mu_0$  est inférieure à

$$\bar{x}_n + 1.645 \frac{s_{x,n}}{\sqrt{n}}.$$

**Exemple :** Avec  $\bar{x}_{150} = 103$  et  $s_{x,150} = 37$ , obtenus pour un échantillon de taille  $n = 150$ , on déduit qu'avec grande confiance, le temps moyen  $\mu_0$  passé devant les écrans par l'ensemble des enfants de 6 ans est au plus de

$$103 + 1.645 \frac{37}{\sqrt{150}} = 107.9696... \approx 108 \text{ minutes}.$$

**Estimation basse.** Avec grande confiance, la proportion de population  $p_0$  est plus grande que

$$\bar{x}_n - 1.645 \frac{\sqrt{\bar{x}_n(1 - \bar{x}_n)}}{\sqrt{n}}.$$

Exemple : Avec  $n = 1000$  et  $\bar{x}_{1000} = 53\%$ , on déduit qu'avec grande confiance, la part  $p_0$  de l'ensemble des habitants trouvant que la crise est bien gérée est au moins de

$$53\% - 1.645 \frac{\sqrt{0.53(1 - 0.53)}}{\sqrt{1000}} = 0.50403... = 50.403...\% \approx 50.4\%.$$

Note : on arrondit toujours les estimations basses à la baisse (ainsi, on ne fait qu'agrandir légèrement l'intervalle calculé).

**Estimation haute.** Avec grande confiance, la proportion de population  $p_0$  est inférieure à

$$\bar{x}_n + 1.645 \frac{\sqrt{\bar{x}_n(1 - \bar{x}_n)}}{\sqrt{n}}.$$

Exemple : Avec  $n = 1000$  et  $\bar{x}_{1000} = 53\%$ , on déduit qu'avec grande confiance, la part  $p_0$  de l'ensemble des habitants trouvant que la crise est bien gérée est au plus de

$$53\% + 1.645 \frac{\sqrt{0.53(1 - 0.53)}}{\sqrt{1000}} = 0.55596... = 55.596...\% \approx 55.6\%.$$

Note : on arrondit toujours les estimations hautes à la hausse (ainsi, on ne fait qu'agrandir légèrement l'intervalle calculé).

### 3. Formules dans le cas d'une moyenne générale

**Conclusion neutre.** Avec grande confiance, la moyenne de population  $\mu_0$  vaut

$$\bar{x}_n \quad \text{à une marge d'erreur de} \quad \pm 1.96 \frac{s_{x,n}}{\sqrt{n}} \quad \text{près.}$$

Exemple : Avec  $\bar{x}_{150} = 103$  et  $s_{x,150} = 37$ , obtenus pour un échantillon de taille  $n = 150$ , on déduit qu'avec grande confiance, le temps moyen  $\mu_0$  passé devant les écrans par l'ensemble des enfants de 6 ans est de

$$103 \text{ min} \quad \text{avec une marge d'erreur de} \quad \pm 1.96 \frac{37}{\sqrt{150}} = \pm 5.9212... \approx \pm 6 \text{ min.}$$

En clair, on propose l'intervalle  $[103 - 6, 103 + 6] = [97, 109]$  comme intervalle de valeurs plausibles pour  $\mu_0$ .  
Note : on ne dit pas du tout que 95% des enfants de 6 ans regardent quotidiennement des écrans pendant un temps compris entre 97 et 109 minutes ; l'assertion d'intervalle de confiance ne porte que sur le temps moyen  $\mu_0$  (apprécié sur l'ensemble des enfants), qui n'est pas connu avec certitude mais dont on pense avec grande confiance qu'il se situe entre 97 et 109 minutes.

Synthèse — Intervalles de confiance, résultats élémentaires

1. Forme des intervalles de confiance

Les intervalles de confiance expliquent comment généraliser la proportion ou la moyenne d'échantillon  $\bar{x}_n$  à l'ensemble de la population.  
Pour bien choisir la forme, il faut penser à la phrase de conclusion qu'on veut écrire : « Avec grande confiance, la proportion  $p_0$  de (ou la moyenne  $\mu_0$  sur) l'ensemble de la population est (ou serait, ou sera) ... »

**Conclusion neutre.** Dans ce cas, on dit juste que cette proportion  $p_0$  ou cette moyenne  $\mu_0$  sont égales à  $\bar{x}_n$  modulo une certaine marge d'erreur, ce qu'on note  $\bar{x}_n \pm \dots$  (et qu'on prononce : « plus ou moins »).

**Estimation basse.** On veut ici dire que  $p_0$  ou  $\mu_0$  seront supérieures à une certaine valeur (elles vaudront au moins telle valeur). Ainsi, on corrige  $\bar{x}_n$  à la baisse pour être sûr avec grande confiance que  $p_0$  ou  $\mu_0$  soient supérieures à cette valeur basse.

**Estimation haute.** Même chose avec pour objectif cette fois que  $p_0$  ou  $\mu_0$  soient inférieures à une certaine correction à la hausse de  $\bar{x}_n$  (elles vaudront au plus telle valeur).

2. Formules dans le cas d'une proportion

**Conclusion neutre.** Avec grande confiance, la proportion de population  $p_0$  vaut

$$\bar{x}_n \quad \text{à une marge d'erreur de} \quad \pm 1.96 \frac{\sqrt{\bar{x}_n(1 - \bar{x}_n)}}{\sqrt{n}} \quad \text{près.}$$

**Exemple :** Avec  $n = 1000$  et  $\bar{x}_{1000} = 53\%$ , on déduit qu'avec grande confiance, la part  $p_0$  de l'ensemble des habitants trouvant que la crise est bien gérée vaut

$$53\% \quad \text{avec une marge d'erreur de} \quad \pm 1.96 \frac{\sqrt{0.53(1 - 0.53)}}{\sqrt{1000}} = \pm 0.03093\dots = \pm 3.093\dots\% \approx \pm 3.1\%.$$

En clair, on propose l'intervalle  $[53\% - 3.1\%, 53\% + 3.1\%] = [49.9\%, 56.1\%]$  comme intervalle de valeurs plausibles pour  $p_0$ .  
Note : on arrondit toujours les marges d'erreur à la hausse (ainsi, on ne fait qu'agrandir légèrement l'intervalle calculé).



Synthèse — Modélisation

**Partie commune.** On considère des individus statistiques : leur ensemble est la population, et on en tire un échantillon aléatoire de taille  $n$ . Sur chaque individu de l'échantillon, on effectue une mesure de la variable d'intérêt, cela conduit aux données  $x_1, \dots, x_n$ , qu'on peut résumer par la moyenne d'échantillon  $\bar{x}_n$ .

**Cas d'une proportion.** La variable d'intérêt est binaire,  $x_j \in \{0, 1\}$ . Le paramètre d'intérêt est  $p_0$ , la proportion de la population pour qui la variable vaut 1.

**Exemple :** Un sondage politique sur  $n = 1000$  habitants montre que  $\bar{x}_{1000} = 53\%$  des sondés pensent qu'une crise est bien gérée. La variable d'intérêt est l'opinion positive, codée par 1, négative, de chaque individu. Le résultat  $\bar{x}_{1000} = 53\%$  du sondage nous permettra d'estimer (avec des intervalles de confiance) la proportion  $p_0$  de l'ensemble des habitants (qui peuvent être plusieurs millions) à penser que la crise est bien gérée. Cette proportion de population  $p_0$  forme notre paramètre d'intérêt.

**Cas d'une variable quantitative générale.** On résume alors les données  $x_1, \dots, x_n$ , qui sont des nombres réels, par leur moyenne  $\bar{x}_n$  et leur écart-type  $s_{x,n}$ . Le paramètre d'intérêt est  $\mu_0$ , la moyenne de ce qu'on mesure sur l'ensemble de la population. Cette moyenne de population  $\mu_0$  forme notre paramètre d'intérêt.

**Exemple :** On veut connaître le temps moyen passé par les enfants de 6 ans devant des écrans les jours de semaines. On sélectionne au hasard 150 familles avec un enfant de 6 ans, qu'on observe chacun pendant une journée. Le temps moyen observé  $\bar{x}_{150} = 103$  minutes sur l'échantillon estime le temps moyen  $\mu_0$  passé par jour devant des écrans par l'ensemble des enfants de 6 ans. On détermine aussi à partir des données un écart-type de  $s_{x,150} = 37$  minutes.





# STATISTIQUE : FICHES DE SYNTHÈSE / COURS CONDENSE

Coordinateur du cours :  
Gilles Fortin-Stoltz (stoltz@hec.fr)

Equipe enseignante :  
Benjamin Petiau (benjamin.petiau@gmail.com)  
Dinah Rosenberg (rosenberg@hec.fr)  
Dan Xie (dan.xie@hec.edu)