



The more you know, the more you dare®

HEC / Master in Management

STATISTICS (AND BASIC ECONOMETRICS) CORRECTIONS OF THE EXERCISES

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

Contents

Chapter 1. Statistical thinking: sample the world!	5
Advanced exercises	6
Discounts to increase the number and amounts of orders	6
Car insurance company	7
Chapter 2. Confidence intervals: the basics	11
Elementary exercises	12
Advanced exercises	14
Discounts to increase the number and amounts of orders, continued	14
Car insurance company, continued	17
An advertisement featuring statistics	21
TwitterAudit	22
Chapter 3. Confidence intervals: advanced notions	25
Elementary exercises	26
Advanced exercises	29
Gender pay gap?	29
Sample size determination, in a different way though...	32
Budget planning for traveling costs	34
The effect of touch	36
Success rate of a new dating method	38
Car insurance company: planning	40
Value of a stock	41
Chapter 4. Hypothesis testing: methodology	45
Elementary exercises	46
Advanced exercises	50
Nicotine patches, continued	50
Managing customers' dissatisfaction	53
Chapter 5. One-sample tests	55
Elementary exercises	56
Advanced exercises	62
A controversial governmental reform, short version	62
Seizure of MegaUpload and side effects	64
Public health surveillance	66
"We look like our names"	69
A controversial governmental reform, long version	70
Walking many steps a day	72

Chapter 6. Two-sample tests	75
Elementary exercises	76
Advanced exercises	80
Separate or pooled marketing campaign?	80
Comparing prices in two local supermarkets	84
Gender pay gap, revisited	89
Alcohol consumption during the POWs at HEC, first version	91
Cash in the wallet, by country	94
Speedy self-assessment	95
The effect of touch, re-worked	96
Choosing between two gifts	98
Lead levels in children's blood	100
Chapter 7. χ^2 -tests	103
Elementary exercises	104
Advanced exercises	107
Reading habits per socio-professional category	107
The sinking of the Titanic	109
Hair color by gender	110
Call centers	112
"Junior enterprise" campaign at HEC Paris	114
Mendel's experiment	115
Alcohol consumption during the POWs at HEC, second version	117
M&M colors	118
Satisfaction survey	120
Births by month	122
Chapter 8. Simple linear regression	123
Elementary exercises	124
Advanced exercises	126
The historical example of regression	126
An example of a spurious correlation	128
Some French politics: demonstrations	130
Prices of ski passes	132
Chapter 9. Multiple linear regression	133
Excerpts of past exams	134
Wage discrimination?	134
Modeling life expectancy	138

Statistical thinking: sample the world!

Advanced exercises

Discounts to increase the number and amounts of orders

page 6

Car insurance company

page 7

Chapter 1 / Exercise: "Discounts to increase the number and amounts of orders"

Population: the hundreds of thousands of customers in the database

Sample: 1,000 such customers, drawn at random

Variable of interest:
 → first, whether each customer places an order or not
 → second, in case she/he does, the amount of the order

Parameters of interest:
 (= the aims of the statistical study)
 → first, the fraction p_0 of all customers that would place an order if they all got the discount
 → second, the average amount of such orders that would be placed, which we denote by μ_0

Data collected:
 (to estimate these parameters)
 → $x_1, \dots, x_{1000} \in \{0, 1\}$
 where $x_j = \begin{cases} 1 & \text{if the } j\text{-th customer placed an order} \\ 0 & \text{otherwise} \end{cases}$

→ $y_1, \dots, y_{170} \in (0, +\infty)$ is, positive numbers
 where y_k is the amount of the k -th order placed (does not necessarily correspond to the k -th customer in sample)

! There is a slight ambiguity in the statement; it is unclear whether the 73 corresponds to an average of 170 "real" orders, or if we count 0 for no order and 73 is then an average of 1,000 numbers.

I picked the first interpretation, as the second one would have had to be underlined and explained in the statement of the exercise...

Summary of the data:
 → a fraction $\bar{x}_{1000} = 17\%$ of the customers in the sample placed an order when getting the discount
 → the customers in the sample placed orders of an average amount of $\bar{y}_{170} = 73 \text{ €}$ (and with a standard deviation of $s_{y_{170}} = 8 \text{ €}$)

! The population proportion p_0 is close to but different from the sample proportion $\bar{x}_{1000} = 17\%$
 The same can be said for μ_0 and \bar{y}_{170}

Chapter 1 / Exercise: "Car insurance company"

1. How not to conduct the survey: e.g.,
- not via Face Book
 - not by selecting just 3, 4 or 10 French universities (there are 100 of them, not taking the schools like HEC into account) and by asking students at random there
 - not only in the Paris area
 - not on the phone: most students only have cell phones

It's actually a very difficult study to make!

My own solution would be: aggregate 10 samples of 500 students obtained in various ways, hoping that the biases will compensate each other.

Sample 1: at a major tollbooth; try various days and hours, pick 30 students each time; repeat 10 times

Sample 2: in front of movie theaters in major French cities; pick 10 cities (Paris, Lyon, ... as major ones, as well as smaller ones: Brest, Toulon, etc.).

Sample 3: in shopping malls (again 10 x 30 students)

Sample 4: in bars on a Friday night (same remark)

Sample 5: in students' parties (same remark)

Etc.

↳ Do you have any better idea? Please let us know! We will write them here.

2. We should first wonder why this study is conducted:
- first - to determine the size of the targeted market
 - second - to assess how much at risk the target population is

For the size of the market:

Suppose we know how many students there are in France (we do so: the Higher Education Ministry wrote in December 2016 that there were 2,55, 000 of them - this number even seems too precise to me ... let's say: 2,5 millions of students).

To determine the size of our market, we thus only need to know the proportion π_0 of them who have a car insurance in their names.

Population of interest: The 2,5 millions of students in France

Sample: (hopefully representative of the population) 3,000 students picked at random

Variable of interest: whether or not each student has a car insurance in her/his name

Parameter of interest: π_0 , the proportion of students that had a car insurance in their names in the past year
from which we get the total number of such students: $2,5 \times 10^6 \times \pi_0$

Data collected: $z_1, \dots, z_{3000} \in \{0,1\}$ where $z_j = \begin{cases} 1 & \text{if the } j\text{-th student sampled had a car insurance} \\ 0 & \text{otherwise} \end{cases}$

Summary of the data: a fraction $\bar{z}_{3000} = \frac{1472}{3000} \approx 49.1\%$ of the students in the sample had a car insurance in their own names in the past year.

For the NSR
assessment:

Population of interest:
(= our business target
in this case)

All French students with an insurance
in place in their names
(\neq all French students)

Sample:
(hopefully representative of the population)

1,472 such students
(we discard the uninsured students!)

Variables of interest:

→ first, whether a given student had
an accident for which she/he was
held responsible

↘ second, when this was the case, the
amount of expenses that had to be paid

Parameters of interest:
(= key numbers to assess
the viability of our
business idea)

→ first, the fraction p_0 of all insured
students that had an accident in the
past year (for which they were held resp.)

↘ second, the average amount μ_0 of
expenses these students with
an accident generated

Data collected:

→ first, $x_1, \dots, x_{1472} \in \{0, 1\}$ where

$$x_j = \begin{cases} 1 & \text{if the } j\text{-th insured student had an accident} \\ 0 & \text{otherwise} \end{cases}$$



second, y_1, \dots, y_{256} , where y_k denotes
the expenses generated
by the k -th student with an accident

Summary of the data:

→ a fraction $\pi_{1472} = \frac{256}{1472} \approx 17.4\%$ of
the students in the
sample had an accident in the past year



they generated an average amount of expenses
equal to $\bar{y}_{256} = 1,865 \text{ €}$

(with a sample standard deviation of
 $s_{y,256} = 524 \text{ €}$)

Confidence intervals: the basics

Elementary exercises	page 12
Advanced exercises	
Discounts to increase the number and amounts of orders, continued	page 14
Car insurance company, continued	page 17
An advertisement featuring statistics	page 21
TwitterAudit	page 22

Chapter 2 /

Correction of the elementary exercises

Exercise 2.1.

Calculation of an under-estimate:

$$\begin{aligned}\bar{x}_{100} - 1.65 \sqrt{\frac{\bar{x}_{100}(1-\bar{x}_{100})}{100}} \\ = 0.78 - 1.65 \sqrt{\frac{0.78(1-0.78)}{100}} \\ \approx 0.7116494 \geq 71\% \quad (\text{or even } 70\%) \end{aligned}$$

! you should round downwards to stay on the safe side

Her statement to her manager will be:

"The satisfaction rate among my portfolio of customers would (with high confidence) be larger than 70% if you interviewed them all to get their opinions."

(Of course the manager does not have the time nor the will to interview 1,536 customers.)

Exercise 2.2.

Calculation of a symmetric interval:

$$\begin{aligned}\bar{x}_{51} \pm 1.96 \frac{s_{x,51}}{\sqrt{51}} \\ = 16 \pm 1.96 \frac{5}{\sqrt{51}} \\ = 16 \pm 1.372274 \text{ minutes} \\ = 16 \pm 1 \text{ minute and } 22.33647 \text{ sec.} \\ \rightarrow 16 \pm 1 \text{ minute and } 30 \text{ sec.} \end{aligned}$$

! you should round the error margin upwards to stay on the safe side

$$0.372274 \times 60 \approx 22.33647$$

Final statement:

A statistical study reveals that the average waiting time on Saturday mornings at the Jany-en-Josas post office equals 16 minutes, a number we're highly confident in up to ± 1 minute and 30 sec.

Put differently, this average waiting time lies between 14 minutes and 30 sec. and 17 minutes and 30 sec. (with high confidence).

Exercise 2.3

Calculation of
an overestimate:

$$\begin{aligned} & \bar{x}_{1002} + 1.65 \sqrt{\frac{\bar{x}_{1002} (1 - \bar{x}_{1002})}{1002}} \\ &= 0.284 + 1.65 \sqrt{\frac{0.284 (1 - 0.284)}{1002}} \\ &= 0.3075053 \leq 31\% \end{aligned}$$

⚠ you should round
upwards to stay
on the safe side

↳ "This study proves that President Macron's popularity rating is, with high confidence, already lower than 31%."

Chapter 2/

Exercise: "Discounts, continued"

0. How did we get the 3.64€ number?

$$\underset{\text{order rate}}{13\%} \times \underset{\text{average amount}}{70\text{€}} \times \underset{\text{average margin rate}}{40\%} = 3.64\text{€}$$

1. Summary of the relevant statistical information

Two parameters of interest:

p_0 , the fraction of all customers that would place an order if they all got the discount = the order rate with the new offer

μ_0 , the average amount of such orders that would be placed

We will deal with them based on sample data collected over 1,000 customers:

$\bar{z}_{1000} = 17\%$ of the customers in the sample placed an order when getting the discount

That is, 170 orders were placed.

$\bar{y}_{170} = 73\text{€}$ was the average amount of the 170 orders thus placed (and is associated with a standard deviation of $s_{y,170} = 8\text{€}$)

2. Confidence interval on p_0

Context:

we are testing a new offer and thus want to stay on the safe side, we don't want to make less money than usual with too favorable a discount.

you might get fired

↳ Putting in place a new offer that leads to less profit will be a visible mistake

you will just not get promoted

↳ Not putting in place a new discount that would have led to more profit than usual is an invisible mistake that no one will realize it

Therefore, we want to stay on the safe side and draw conclusions based on a pessimistic estimate of p_0 : an under-estimate.

That is, we are looking for a statement like:
"With high confidence, the order rate p_0 with the new offer would be at least [...]"

Calculation:

The underestimate equals

$$\begin{aligned}\bar{x}_{1000} - 1.65 \sqrt{\frac{\bar{x}_{1000}(1-\bar{x}_{1000})}{1000}} \\ = 0.17 - 1.65 \sqrt{\frac{0.17(1-0.17)}{1000}} \\ = 0.1504004 \geq 15.0\%\end{aligned}$$

which corresponds to a confidence interval of 15% - 100%

Conclusion:

With high confidence, the order rate p_0 with the new offer would be larger than 15.0%

3. Confidence interval on μ_0

Context: Same context, we want to stay on the safe side and we will output a pessimistic estimate (an under-estimate)

Calculation: this underestimate equals

$$\begin{aligned}\bar{y}_{170} - 1.65 \frac{s_{y,170}}{\sqrt{170}} \\ = 73 - 1.65 \frac{8}{\sqrt{170}} \\ = 71.98761 \geq 71.90 \text{ €}\end{aligned}$$

! we divide by $\sqrt{170}$ and not by $\sqrt{1000}$ as we have only 170 data points for amounts of orders...

Conclusion:

With high confidence, the average amount μ_0 of orders placed with the new offer would be larger than 71.90 €

4. General conclusion

Question is whether the new offer with a 5% overall discount is more profitable than the current situation.

To answer this question, we compute an underestimate of the gross margin per customer per trimester that we would get:

$$\begin{array}{ccccc} 15\% & \times & 71.90 \text{ €} & \times & ? \\ \text{order} & & \text{average} & & \text{new} \\ \text{rate} & & \text{amount} & & \text{margin rate} \end{array}$$

higher than in the current situation

smaller than in the current situation (because of the discount)

New margin rate:

- Currently, if 100 € of products are ordered, 100 € are paid, the products cost 60 € on average and 40 € of gross profit are made.
 - With the new offer, keeping in mind (see statement of the exercise) that all amounts were reported before the 5% discount is applied): if 100 € of products are ordered, the customer pays 95 € and the products cost 60 € on average, hence a gross profit of 35 €.
- ↳ That is, the new margin rate would be 35%

Calculation:

$$15\% \times 71.90 \text{ €} \times 35\% \approx 3.77 \text{ €}$$

General conclusion:

The test performed on 1,000 customers shows that a general 5% discount on all products would lead (with high confidence) to a 3.77 € gross margin per customer per trimester, which is larger (and more desirable) than the current 3.64 € such margin. We should generalize the new offer!

Note for later purpose:

As we combine here two confidence intervals of confidence level 95%, the final result only holds with confidence level 90%, as you will learn in the next chapter.

That is, the 3.77 € figure should be believed with a confidence level of 90% (it is still a high level of confidence).

Chapter 2 /

Exercise: "Car insurance company, continued"

1. Summary of the relevant statistical information.

Two parameters of interest:

- p_0 , the fraction of all insured students that had an accident in the past year (for which they were held responsible)
- μ_0 , the average amount of expenses these students generated

We will deal with them based on sample data collected over 1,472 insured students:

$\bar{x}_{1472} = 17.4\%$ of the insured students in the sample had an accident in the past year (for which they were held responsible)

These accidents (there were 256 of them) generated an average amount of expenses equal to $\bar{y}_{256} = 1865\text{€}$ (a number that is associated with a standard deviation of $s_{y,256} = 524\text{€}$).

Mindset #1: The entrepreneur

2. The statement of the exercise says that he just needs to get a better idea of the expenses at stake to cover his risks through the prudential rules, and in turn, of how much money to leverage.

(Of course, this all is a simulated situation, his business idea is just terrible!)

Thus, to stay on the safe side, he should overestimate the accident rate p_0 and the average expenses μ_0 suffered in case of an accident.

We get the following two statements:

- With high confidence, the accident rate within the population of insured students is smaller than

$$\begin{aligned} & \bar{x}_{1472} + 1.65 \sqrt{\frac{\bar{x}_{1472}(1-\bar{x}_{1472})}{1472}} \\ &= 0.174 + 1.65 \sqrt{\frac{0.174(1-0.174)}{1472}} \\ &= 0.190304 \leq 19.1\% \end{aligned}$$

- With high confidence, the average expenses for the insurance company per accident are smaller than

$$\begin{aligned} \bar{y}_{256} + 1.65 \frac{s_{256}}{\sqrt{256}} \\ = 1865 + 1.65 \frac{524}{\sqrt{256}} \\ = 1919.037 \leq 1920 \text{ €} \end{aligned}$$

3. Conclusion:

Based on these values, we deduce a bound on $\mu_{p_{10}}$, the average expenses per insured student:

$$\underbrace{19.1\%}_{\text{responsible accident rate}} \times \underbrace{1920 \text{ €}}_{\text{average expenses in case of a responsible accident}} = 366.72 \leq 367 \text{ €}$$

And get the following general conclusion:

- With high confidence, the average expenses per insured student observed in the past year were smaller than 367 €, which is the value we would keep in mind to evaluate these expenses for the year to come.

Again, confidence level of 90% given that we combine two confidence intervals

Mindset #2: The best friend.

- Having students share their driving risks is probably a terrible business idea. The best friend would probably want to convince the entrepreneur NOT to start this company and would have to prove that students would generate much more costs than the average standard customer in the industry.

His statements would be of the form

"Look my friend, with high confidence, the average expenses per insured student were larger than [...] €, which is much superior to the average expenses per standard policyholder."

That is, the best friend will exhibit underestimates and comment that even these underestimates take large values.

Calculations:

$$\begin{aligned} \bar{x}_{1472} - 1.65 \sqrt{\frac{\bar{x}_{1472}(1-\bar{x}_{1472})}{1472}} &= 0.174 - 1.65 \sqrt{\frac{0.174(1-0.174)}{1472}} \\ &= 0.157696 \geq 15.7\% \end{aligned}$$

$$\bar{y}_{256} - 1.65 \frac{s_{y_{256}}}{\sqrt{256}} = 1865 - 1.65 \frac{524}{\sqrt{256}} = 1810.963 \geq 1810 \text{ €}$$

That is, confidence intervals at the 95% level for p_0 and μ_0 would respectively be: $[15.7\%, 100\%]$ and $[1810\text{€}, +\infty[$

In plain words :

- With high confidence, the accident rate within the population of insured students is larger than 15.7%
- With high confidence, the average expenses for the insurance company per accident are larger than 1810 €

3. General conclusion: an underestimate of $p_0 \mu_0$ is

$$15.7\% \times 1810\text{€} = 284.17 \geq 284\text{€}$$

The friend's statement should be something around:

90% confidence level → With high confidence my dear, the average expenses per insured student were larger than 284 € in the past year, and I think we may take this value for the average expected expenses per insured student in the year to come. This is so much more than the expenses generated by the standard policyholder, not mentioning those of good drivers! You're going to pay a lot of money for your customers, I don't think this is a viable business idea...

Mindset #3: The banker

Consider the following mindset:

- " Bankers try to assess business plans in a rigorous and neutral way; they want to simultaneously
- control the risks (they do not want to lend money and not get it)
 - get an idea of the order of magnitude of the "profits" (bids)

In that case, we want to exhibit both under- and overestimate (to have, respectively, an idea of the risks and profits); that is, we should draw symmetric confidence intervals.

Calculations :

$$\bar{x}_{1472} \pm 1.96 \sqrt{\frac{\bar{x}_{1472}(1-\bar{x}_{1472})}{1472}} = 0.174 \pm 1.96 \sqrt{\frac{0.174(1-0.174)}{1472}} = 0.174 \pm 0.0193672$$

is included in the interval $17.4\% \pm 2.0\%$
 $= [15.4\%, 19.4\%]$

$$\bar{y}_{256} \pm 1.96 \frac{s_{y,256}}{\sqrt{256}} = 1865 \pm 1.96 \frac{524}{\sqrt{256}}$$

$$= 1865 \pm 64.19$$

is included in the interval $1865 \pm 65 = [1800, 1930]$

In plain words:

- With high confidence, the accident rate within the population of insured students $\left\{ \begin{array}{l} \text{lies between } 15.4\% \text{ and } 19.4\% \\ \text{or} \\ \text{equals } 17.4\% \text{ (up to an error margin of } \pm 2.0\%) \end{array} \right.$
- With high confidence, the average expenses suffered by the insurance company per accident $\left\{ \begin{array}{l} \text{lie between } 1800\text{€ and } 1930\text{€} \\ \text{or} \\ \text{equals } 1865\text{€ (up to an error margin of } \pm 65\text{€)} \end{array} \right.$

3. General conclusion

With high confidence, p_0 lies between 15.4% and 19.4%

μ_0 lies between 1800€ and 1930€

we compute the two most extreme values

thus

p_0, μ_0

$$15.4\% \times 1800\text{€}$$

$$= 277.20\text{€}$$

$$\geq 277\text{€}$$

$$19.4\% \times 1930\text{€}$$

$$= 374.42$$

$$< 375\text{€}$$

The banker should state:

- "With high (90%) confidence, the average expenses per insured student $\left\{ \begin{array}{l} \text{had in the past year (and) are thus likely to lie in} \\ \text{the upcoming year) between } 277\text{€ and } 375\text{€} \end{array} \right.$

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.

Of course, this number 89,019 was supposed to be humoristic! It is falsely accurate.



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?

An underestimate: we want to convey the idea that on average, stairs are used often, at least X many times on average. We hope to get an impressive number X in our calculations.

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):

With high confidence, people living in places with stairs walk them up and down at least

$$12.60 - 1.645 \frac{2.41}{\sqrt{534}} = 12.42844... \geq 12.42 \text{ times a day, on average.}$$

- ☐ 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):

$$19.34 - 1.645 \frac{4.35}{\sqrt{534}} = 19.0303... \geq 19.03 \text{ years, on average.}$$

- ☐ ☐ 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 90% and on average, you will walk up and down your stairs at least 86,000 times.

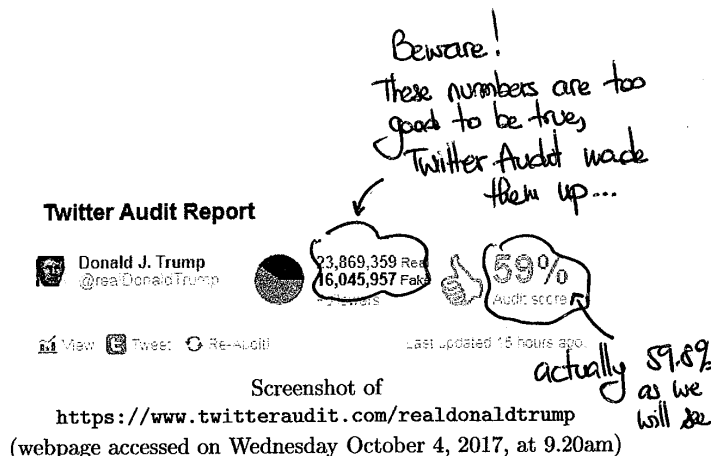
technical statement

Details of the calculation: $12.42 \times 19.03 \times 365.25 = 86,327.78715$, which we round downwards to 86,000 for readability. If you're concerned that we should still make a humoristic statement, pick rather 86,327 and explain that you are doing so for the sake of humor (and not because you do not realize that you should round off numbers for better readability).

TwitterAudit

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."



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.

All the followers of the @realDonaldTrump account: there are
23,869,359 + 16,045,957 = 39,915,316 of them

- Indicate the parameter of interest.

The proportion p_0 of real followers among this population
of 39,915,316 followers.
is unknown, despite the misleading information in the box!

□ 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%?).

Available data: $x_1 \dots x_{5000}$ where $x_j = \begin{cases} 1 & \text{if the } j\text{-th follower in the sample is real} \\ 0 & \text{if it is a fake account} \end{cases}$

Data summary: $\bar{x}_{5000} = 59.8\%$ of the followers in the sample were declared real followers.

We may reverse engineer the numbers:
 $23,869,359 / 39,915,316 \approx 59.8\%$

- Should we compute a symmetric confidence interval, an underestimate, or an overestimate? Explain.

Two answers are legitimate here, with a personal preference for the first one:

- Either: symmetric interval as Twitter Audit is a neutral third party that tries to provide an accurate and non-biased evaluation of a Twitter account

- Or: underestimate, as we should discard fake users (too many Twitter accounts try to boost their statistics!) and only focus on the real users: how many of them can we minimally guarantee?

Symmetric case: $1.96 \sqrt{\frac{0.598(1-0.598)}{5000}} \approx 1.36\%$

With high confidence, the proportion of Trump's real followers among all his followers equals 59.8% up to a $\pm 1.36\%$ margin of error.

□ Based on your answer to the previous point, perform the calculations (provide some intermediary details, not just the final answer).

Underestimate: $59.8\% - 1.65 \sqrt{\frac{0.598(1-0.598)}{5000}} \geq 58.6\% \quad \text{or} \quad 58.65\%$

With high confidence, at least 58.6% of Trump's followers are real followers.

□ We go back to the picture. The number 23,869,359 therein is misleading, isn't it? By which number or number range should it have been replaced?

↑ It looks accurate but it's not (it's a point estimate only)

← A number range obtained by multiplying the confidence interval on the proportion by the population count:

- Underestimate: at least $58.65\% \times 39,915,316 = 23,410,333$

which we round downwards to 23.4 M:
with high confidence, Trump has at least 23.4 M of real followers

□ All in all, provide a TwitterAudit box that would be both more accurate and more honest than the one shown above, while still containing the same information. To that end, just write in a box the 2/3 numbers that would be relevant to show, based on all calculations above.

DJ Trump	Audit score:
#Followers: 39,915,316	59.8%
Real: ≥ 23.4 M	on 5,000 followers

an important note!

- Symmetric number range and box (our favorite choice):
With high confidence, the total number of real followers lies between

$$(59.8\% - 1.36\%) \times 39,915,316 \geq 23.3 \text{ M}$$

$$\text{and } (59.8\% + 1.36\%) \times 39,915,316 \leq 24.5 \text{ M}$$

Similarly / Equivalently, the total number of fake followers lies between

$$(100\% - (59.8\% + 1.36\%)) \times 39,915,316 \geq 15.5 \text{ M}$$

$$\text{and } (100\% - (59.8\% - 1.36\%)) \times 39,915,316 \leq 16.6 \text{ M}$$

Hence the box:

(with a grey zone of 1.1 M followers)

DJ Trump

Followers: 39,915,316

Real: 23.3M - 24.5M

Fake: 15.5M - 16.6M

Audit score: 59.8% on 5,000 followers

Real	?	Fake
23.3M	1.1M	15.5M

Confidence intervals: advanced notions

Elementary exercises	page 26
Advanced exercises	
Gender pay gap?	page 29
Sample size determination, in a different way though...	page 32
Budget planning for traveling costs	page 34
The effect of touch	page 36
Success rate of a new dating method	page 38
Car insurance company: planning	page 40
Value of a stock	page 41

Chapter 3/

Elementary exercises -

Exercise 3.1

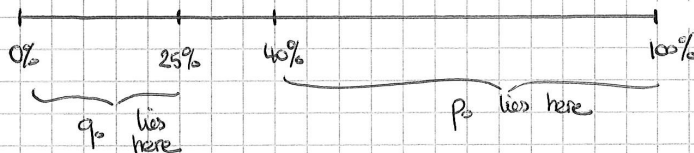
We are already given two confidence intervals,

- first, on p_0 , the purchase rate of a new product with a nice treatment
- second, on q_0 , the purchase rate of a new product in a neutral setting

These intervals are:

- With high confidence (confidence level of 95%),
 p_0 is larger than 40% (ie, lies between 40% and 100%)
- With high confidence (confidence level of 95%),
 q_0 is smaller than 25% (ie, lies between 0% and 25%)

Here is a picture:



With a confidence level of 90% (only), we have the simultaneous statements

$$\begin{cases} p_0 \text{ is larger than } 40\% \\ \text{AND} \\ q_0 \text{ is smaller than } 25\% \end{cases}$$

From which we deduce:

Either: "With a confidence level of 90%, a nice treatment leads to an at least 15 point increase in the purchase rate"

$$(15 = 40 - 25)$$

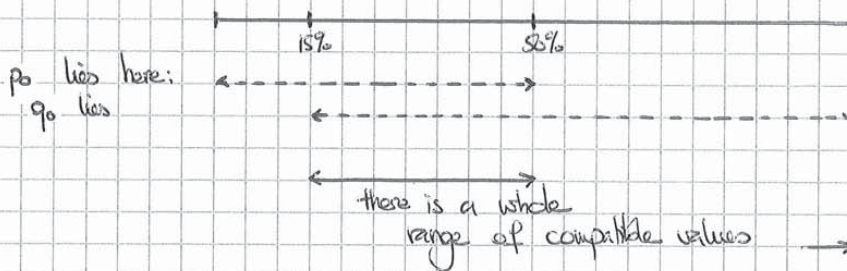
Or: "With a confidence level of 90%, a nice treatment leads to an increase in sales by 60%."
 $(40/25 - 1 = 0.6 = 60\%)$

Exercise 3.2.

First situation:

This is the kind of statements that some students would like to make to seemingly maximize the effect of the nice treatment, but they do not allow for any interesting conclusion!

Here is a picture:

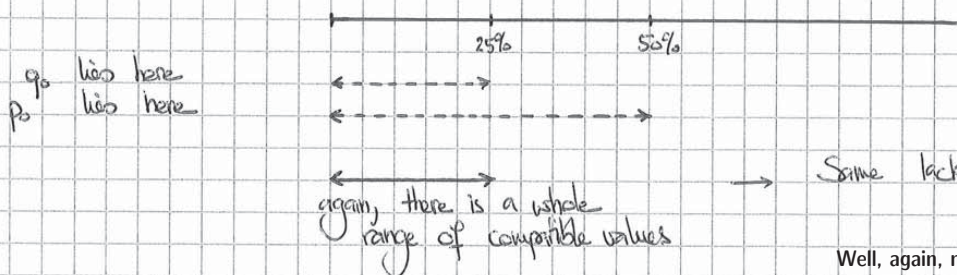


Actually we may still conclude that the increase in the purchase rates is with high confidence smaller than 35%.
 But from a business viewpoint this is not an interesting statement...

We cannot draw any conclusion, we could well have that $p_0 = q_0$, i.e., that the nice treatment has no effect!

Second situation:

Again, we will not be able to draw any conclusion.



Some lack of conclusion.

Well, again, no interesting conclusion can be made. Of course, we can still say that the impact of the smile in the purchase rate lies, with high confidence, anywhere between +50% and -25%...

Exercise 3.3

The parameter of interest is μ_0 , the average homework time over the 300 students

- (1) The sample average equals $\bar{x}_{35} = 62$ min. (with a sample standard deviation of $s_{x,35} = 4$ min.)

The current error margin equals $\pm 1.96 \frac{4}{\sqrt{35}} \approx 1.33$ min

To have it of the order of 1 minute, we should base our calculation on a sample of n students, where

$$1.96 \frac{4}{\sqrt{n}} = 1, \quad \text{that is, } n = (1.96 \times 4)^2 \approx 62$$

↳ we should sample $62 - 35 = 27$ more students.

An alternative way to get this conclusion is the following. The error margin depends on $1/\sqrt{n}$ where n is the sample size.

Thus to decrease it by a factor γ , we need to multiply the sample size by γ^2 .

Here, we want to decrease the current error margin by a factor of $\gamma = 1.33$, thus the sample size 35 should be multiplied by $(1.33)^2$ to get

$$n = 35 \times (1.33)^2 \approx 62.$$

- (2) We repeat the same calculation with $s_{x,35} = 15$

The current error margin equals $\pm 1.96 \frac{15}{\sqrt{35}} = \pm 4.97$ min

To reduce it to ± 1 min, we should consider a total sample size of [first method] n such that $1.96 \frac{15}{\sqrt{n}} = 1$, that is, $n = (1.96 \times 15)^2 \approx 865$

[second method] $n = \underset{\substack{\uparrow \\ \text{current} \\ \text{sample} \\ \text{size}}}{35} \times \underset{\substack{\uparrow \\ \text{factor by which} \\ \text{the error} \\ \text{margin should be decreased}}}{(4.97)^2} = 865$

This is a large sample size! It is larger than the number of students we have in one year. We conclude that the desired precision level is too demanding and should be relaxed. (Note: The population is composed here of thousands of students as we are considering several years of students, but still, 865 may be a non-negligible fraction of the population, which would call for a correction of the confidence-interval formula, leading to a smaller interval, which is thus favorable.)

Chapter 3 / Exercise: "Gender pay gap?"

(0) Extraction of the relevant statistical information

Populations = $\begin{cases} \text{Men of the Ile-de-France region that are company executives} \\ \text{Women of the Ile-de-France region that are company executives} \end{cases}$

Samples = $\begin{cases} 179 \text{ such men} \\ 147 \text{ such women} \end{cases}$

Variable of interest = The monthly net salary

Parameters of interest = $\begin{cases} \mu_0^{\text{M}}, & \text{the average monthly net salary of all considered men} \\ \mu_0^{\text{F}}, & \text{the average monthly net salary of all considered women} \end{cases}$

Data collected = $\begin{cases} \text{for men, } x_1, \dots, x_{179} & \text{taking positive values, where } x_j \text{ is the monthly net salary of the } j\text{-th male respondent} \\ \text{for women, } y_1, \dots, y_{147} & \text{taking positive values, where } y_k \text{ is the monthly net salary of the } k\text{-th female respondent} \end{cases}$

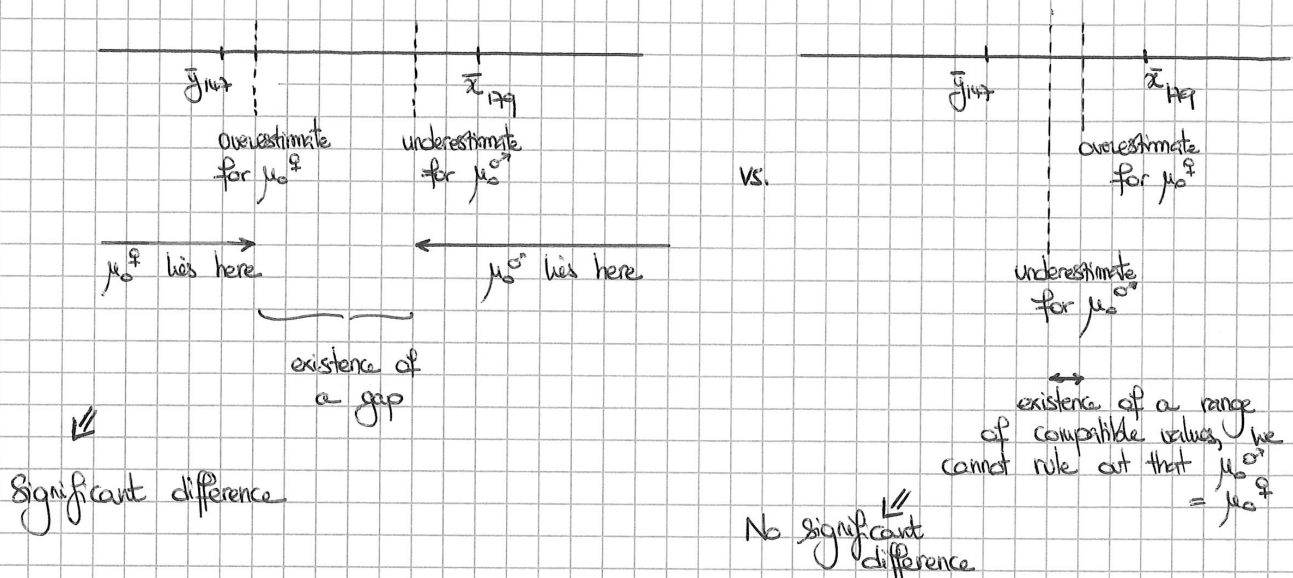
Summary of the data = $\begin{aligned} &\text{Sample average salaries of } \bar{x}_{179} = 3431.46 \text{ €} \\ &\text{and } \bar{y}_{147} = 2434.90 \text{ €} \\ &\text{respectively for men and women of the} \\ &\text{considered category (with respective sample standard} \\ &\text{deviations of } s_{x_{179}} = 3895.44 \text{ €} \\ &\text{and } s_{y_{147}} = 1282.95 \text{ €}) \end{aligned}$

(1) Existence of a "significant difference" between \bar{x}_{179} and \bar{y}_{147}
 \leftrightarrow Do \bar{x}_{179} and \bar{y}_{147} show that with high confidence, μ_0^{F} is smaller than μ_0^{M} ?

To that end, we will

- underestimate μ_0^{F}
- overestimate μ_0^{M}
- compare these two estimates

We will end up in one of the following two situations:



The data collected leads to the following estimates:

- With confidence 95%, μ_o^m , the average monthly net salary of all considered men, is larger than

$$\begin{aligned} \bar{x}_{179} - 1.65 \frac{s_{x,179}}{\sqrt{179}} &= 3,431.46 - 1.65 \frac{3895.44}{\sqrt{179}} \\ &= 2951.047... \\ &\geq 2,950 \text{ €} \end{aligned}$$

- With confidence 95%, μ_o^f , the average monthly net salary of all considered women, is smaller than

$$\begin{aligned} \bar{y}_{147} + 1.65 \frac{s_{y,147}}{\sqrt{147}} &= 2,434.90 + 1.65 \frac{1,282.95}{\sqrt{147}} \\ &= 2,609.4962... \\ &\leq 2,610 \end{aligned}$$

- Therefore, we are in the left situation above, where there is a significant difference:

With confidence 90%, the average monthly salary of women working as company executives in Ile-de-France is smaller (by at least $2,950 - 2,610 = 340\text{€}$) than the one of their male counterparts.

we actually even quantified here the gap at least guaranteed by these data

(2) Did we exhibit a gender pay gap?

→ Not yet, to do so, we should have isolated the effect of gender and have argued all things being equal (full-time vs. part-time working, exact education level, seniority, etc.)

One way of isolating the effect of a variable is to resort to linear regression (which we will study in the third part of this course).

(3) Getting error margins of ± 100 €

Current error margins for the estimation of μ_o^m and μ_o^f :

respectively, $\pm 1.96 \frac{s_{x,179}}{\sqrt{179}} = \pm 1.96 \frac{3,895.44}{\sqrt{179}} = \pm 570.6713 \dots$
 $\approx \pm 571$ €

and $\pm 1.96 \frac{s_{y,147}}{\sqrt{147}} = \pm 1.96 \frac{1,282.95}{\sqrt{147}} = \pm 207.3992 \dots$
 $\approx \pm 208$ €

To get ± 100 € error margins, we should increase the sample size by respective factors of

men → $(571/100)^2$ thus leading to sample size of $179 \times \left(\frac{571}{100}\right)^2 \approx 5837 \approx 5,850$
 women → and $(208/100)^2$ $147 \times \left(\frac{208}{100}\right)^2 \approx 636 \approx 650$

Alternative calculations (closer to the ones detailed in the textbook):

- For men: the desired sample size n satisfies

$$1.96 \frac{s_{x,179}}{\sqrt{n}} = 1.96 \frac{3,895.44}{\sqrt{n}} = 100$$

that is, $n = (1.96 \times 3,895.44 / 100)^2 \approx 5,830$ to be rounded up to 5,850

- For women:

the desired sample size equals $n = (1.96 \times 1,282.95 / 100)^2 \approx 633$ to be rounded up to 650

Chapter 3/

Exercise: "Sample size determination, in a different way though..."

Ante-scriptum: This is 3rd street, in Santa Monica, CA (greater Los Angeles)

(1) Extraction of the relevant statistical information, as far as the test is concerned:

Population = pedestrians/potential customers walking in 3rd street on Saturdays

Sample = 100 of them picked at random

Variable of interest = whether or not they accept to take the survey

Parameter of interest = p_0 , the participation rate that will be observed on the Saturday when the survey will be massively administered

Data collected = $x_1 \dots x_{100}$ where $x_j = \begin{cases} 1 & \text{if the } j\text{-th pedestrian asked accepted to take the survey} \\ 0 & \text{otherwise} \end{cases}$

Summary of the data = $\bar{x}_{100} = 15\%$ of the pedestrians in the sample accepted to take the survey

(2) Total number of pedestrians to be asked to get the 2000 respondents

By definition of p_0 , it is $N = 2000/p_0$

To be on the safe side, we should rather overestimate N
To do so, we need to underestimate p_0

I.e., we will issue statements like:
"with high confidence, we will get a participation rate of at least [...]%, so that it will be enough to interview [...] pedestrians."

Underestimate of p_0 :

$$\begin{aligned} \bar{x}_{100} - 1.65 \sqrt{\frac{\bar{x}_{100}(1-\bar{x}_{100})}{100}} \\ = 0.15 - 1.65 \sqrt{\frac{0.15(1-0.15)}{100}} &= 0.091083... \\ &\geq 9.1\% \end{aligned}$$

Overestimate of N :

$$\begin{aligned} 2000 / 9.1\% &= 21,978.02... \\ \text{which we will round up to } &22,000 \end{aligned}$$

With high confidence (confidence level of 95%), it will be enough to ask 27,000 pedestrians for participation.

(3) Total (wo)man hours needed to administrate the survey

So we want to make sure that 27,000 pedestrians be asked (takes 1 min. each to do so) and that 2,000 surveys be fully conducted (takes 5 minutes each to do so).

Total time needed: $27,000 + 2,000 \times 5 = 37,000$ minutes
 $= 533,33... \text{ hours}$

Let's hire 534 (wo)man hours: if each survey administrator works 8h that day (say, 9.30 am - 6pm with 30 min of lunch break), you will need to recruit

$$534 / 8 = 66.75 \approx 67 \text{ pollsters}$$

It's quite an army! You need to plan well this recruitment.

Also, this number will be key in determining how much to charge for your study.

And please note: to determine this number of 67 pollsters to recruit, you only needed to use 100 data points, which took 100 minutes of your time!

Of course, by collecting more data points you would have got more accurate estimates of ρ and might have ended up with fewer pollsters to recruit. It might be worth investing 100 additional minutes of your time to increase the size of your sample!

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 be assured 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):

$$39.25 + 1.65 \frac{18.92}{\sqrt{30}}$$

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

44.949...

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

45 €

(we need to round upwards since we're speaking of an overestimate)

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):

$$41.54 + 1.65 \frac{14.24}{\sqrt{30}}$$

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

45.8297...

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

46 €

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.):

With high confidence, Gilles will pay on average 91 € per week for his train trip and accommodation in the upcoming months.

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

90%

(simultaneous use of two 95% - confidence intervals)

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) ± 5 euros?

Current error margin: $\pm 1.96 \times \frac{18.92}{\sqrt{30}} \approx 6.78 \text{ €}$

Target: $\pm 5 \text{ €}$

Need a total sample size of $30 \times \left(\frac{6.78}{5}\right)^2 \approx 55.16208$
rounded upwards to 56

That is: Gilles will need to wait 26 more weeks.

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 slightly, 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):

$$\bar{x}_{p_0} = 10\% \quad \text{thus} \quad \left[10\% \pm 1.96 \sqrt{0.1 \frac{(1-0.1)}{120}} \right]$$

- ☐ Provide your final numerical value(s),

– as read on your calculator (no need for rounding yet):

$$10\% \pm 5.367\%$$

– after rounding the error margin to a X.X% format:

$$10\% \pm 5.4\% \quad \text{that is,}$$

- ☐ 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 between 276 and 924 purchases on average.

- ☐ 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.

Current margin of 5.4% : needs to be improved by a factor of 2.7 to get the desired $\pm 2\%$ margin

Thus the sample size should have been: $120 \times (2.7)^2 \approx 875$ customers

Numbers obtained as:

$$276 = 6000 \times (10\% - 5.4\%)$$

$$924 = 6000 \times (10\% + 5.4\%)$$

Could have been rounded off to 270 and 930, eg.

ie, we cannot rule out a detrimental effect of a light touch!

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):

[12.1% to 26.3%]

- ☐ Same question for $q_0 - p_0$ (with some calculation details or with a picture):

∇ We have overlapping intervals:



Confidence interval for $q_0 - p_0$:

$$[12.1\% - 15.4\% \text{ to } 26.3\% - 4.6\%]$$

$$= [\text{from } -3.3\% \text{ to } 21.7\%]$$

- ☐ What is the confidence level of the interval calculated in the previous question?

90% confidence only

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:

Symmetric interval (cf. neutral viewpoint: ready to believe that a touch could have a positive or a negative effect)

Underestimate: wants to see how effective a light touch is so as to determine whether all the sellers should act so

↳ hopes to reach a positive underestimate on $q_0 - p_0$

↳ Needs an underestimate on q_0 overestimate on p_0

↳ Updating calculations, we get:

$$q_0 \geq 13.2\% \text{ and } p_0 \leq 14.6\%$$

with high confidence and still cannot conclude to a significant effect of a light touch!

Details for
[19.2% ± 7.1%]

$$\Rightarrow \left[\frac{23}{120} \pm 1.96 \sqrt{\frac{23/120(1-23/120)}{120}} \right] = [19.166\% \pm 7.042\%]$$

$$= [12.124\% \text{ to } 26.208\%]$$

included in the range

$$[19.2\% \pm 7.1\%] = [12.1\% \text{ to } 26.3\%]$$

Chapter 3/

Exercise: "Success rate of a new dating method"

(1) Extraction of the relevant statistical information (even if not specifically asked for)

Population = All the users of online dating services (several thousands, even hundreds of thousands)

Samples = 2×200 users picked at random,
 the first sample of 200 users resorted to a traditional online service
 the second sample of 200 users resorted to our new online dating method

Variable of interest = the time elapsed between registration to the service and the first serious relationship (the statement of the exercise implicitly assumes that every user gets eventually a serious relationship, I'm not so sure that this is a realistic assumption!)

Parameters of interest = $\left\{ \begin{array}{l} \mu_0^T, \text{ the average waiting time before a serious relationship with traditional online services (average computed over all the many users)} \\ \mu_0^N, \text{ this average waiting time with the new concept} \end{array} \right.$

Data collected = $\left\{ \begin{array}{l} x_1 \dots x_{200} \text{ the waiting times observed on the first sample} \\ y_1 \dots y_{200} \text{ the waiting times observed on the second sample} \end{array} \right.$

Summary of the data = $\left\{ \begin{array}{l} \text{Sample average waiting time } \bar{x}_{200} = 78 \text{ days with the traditional online services (a number which is associated with a standard deviation of } s_{x,200} = 18 \text{ days)} \\ \text{Sample average waiting time } \bar{y}_{200} = 46 \text{ days with the new online method (standard deviation: } s_{y,200} = 23 \text{ days)} \end{array} \right.$

(2) Answer to "By which guaranteed percentage is your method more effective than traditional methods?"

This final sentence of a long statement is about the quantity $1 - \frac{\mu_0^N}{\mu_0^T}$. That is, we want to say that the new average waiting time μ_0^N is shorter than μ_0^T by $100 \times (1 - \mu_0^N / \mu_0^T) \%$.

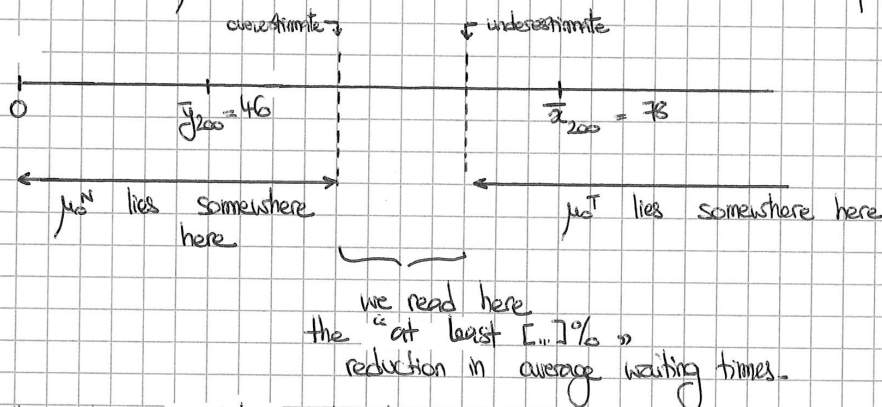
As some estimation will be needed, our final statement will rather be of the form: "The average waiting time μ_0^N is shorter than the traditional one μ_0^T by at least [...] %".

That is, because we are honest in our communication, we want to stay on the safe side and underestimate the reduction in time $1 - \mu_0^N / \mu_0^T$.

To that end, we will

- underestimate μ_0^T
- overestimate μ_0^N

as is illustrated on the picture below:



Computation of the two confidence intervals:

- With confidence 95%, μ_0^T is longer than $\bar{x}_{200} - 1.65 \frac{s_{x,200}}{\sqrt{200}}$

$$= 78 - 1.65 \frac{18}{\sqrt{200}} = 75.899... \geq 75.8$$

- With confidence 95%, μ_0^N is shorter than $\bar{x}_{200} + 1.65 \frac{s_{x,200}}{\sqrt{200}}$

$$= 46 + 1.65 \frac{23}{\sqrt{200}} = 48.6834... \leq 48.7$$

(We will round numbers at the very end!)

Therefore, with a high confidence level of 90%, the minimally guaranteed reduction in average waiting times is:

We can use these numbers (35% or $\frac{1}{3}$) in our ads!

$$1 - \frac{48.7}{75.8} = 0.357519... \geq 35\% \text{ or even } \frac{1}{3}$$

Car insurance company: planning

This exercise is a continuation of previous exercises. 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 ± 15 euros?

We want to improve accuracy by a factor $f = 65/15$. We based our calculations so far on $m = 256$ students held responsible for an accident. We would thus need in total

$$m f^2 = 256 \left(\frac{65}{15} \right)^2 \approx 4,808 \quad (\text{rounding this number upwards})$$

students held responsible for an accident, that is, $4,808 - 256 = 4,552$ more such students.

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?

The question can be reformulated as follows: how many additional students with a car insurance should we interview to guarantee that among them, there will be 4,552 students held responsible for an accident?

We proceed carefully and must first determine the minimally guaranteed share of students with a car insurance that were held responsible for an accident: with high confidence, it is larger than

$$\frac{256}{1,472} - 1.645 \sqrt{\frac{(256/1,472)(1 - 256/1,472)}{1,472}} \geq 15.76\%$$

(no need for proper rounding for now). Therefore, with high confidence, it should suffice to interview

$$\frac{4,552}{15.76\%} \approx 28,884$$

students with a car insurance. This is of course unfeasible (too costly).

Note: it was unsafe to directly use the sample proportion $256/1,472 \approx 17.39\%$ and recommend to interview $4,552/17.39\% \approx 26,176$ students with a car insurance. The sample proportion needed to be corrected into a (safe) underestimate.

Chapter 3 / Exercise:

Value of a stock

- (1) She audits all references of Category 2 because:
- it's feasible (it takes less than 4 hours)
 - it matters (these references account for almost half of the stock value according to Table 1)

For Category 1, an exhaustive study would take about 30h to 5 persons, i.e. 150 (or) (no) man hours, which is probably too expensive given the power of statistics — see the needed time in question (8) below.

(2) Population = The 1,532 references with small unit cost (less than 10€)

Sample = 50 such references picked at random

Variable of interest = The actual value of a reference in the stock
(= actual price \times actual number of units)

Parameter of interest = Eventually, the total actual value of all these 1,532 references

Which we will get access to through the average actual value μ_0 of all these 1,532 references

The total actual value then equals $1,532 \mu_0$

Data collected = x_1, \dots, x_{50} taking positive values, where x_j denotes the actual value of the stock for the j -th reference checked

Summary of the data = The sample average of the actual values of the 50 references checked equals $\bar{x}_{50} = 2,304.10 \text{ €}$ (with an associated standard deviation of $s_{x,50} = 753.74 \text{ €}$)

- (3) We first compute a confidence interval on μ_0 and then provide a confidence interval on the total value $1,532 \mu_0$

We resort to symmetric confidence intervals because we conduct a neutral study (a financial audit); see also the statements of Questions (4) and (8), which confirm that we are interested in symmetric intervals.

$$\begin{aligned} \text{Calculation \#1: } \bar{x}_{50} \pm 1.96 \frac{s_{x,50}}{\sqrt{50}} &= 2,304.10 \pm 1.96 \frac{753.74}{\sqrt{50}} \\ &= 2,304.10 \pm 208.926 \dots \end{aligned}$$

which we round off to $2,304.10 \pm 208.93$
(we will make a final rounding and get more readable numbers below).

Calculation #2: $1,532 \times 2,304.10 \pm 1,532 \times 208.93$
(for 1,532%)
 $= 3,529,881.20 \pm 320,080.76$

(8) Always make the confidence interval larger)
is included in: $3,530,000 \pm 325,000$ for instance
or even: $3.5 \text{ ME} \pm 350 \text{ KE}$
95% confidence level

Conclusion: With high confidence, the total actual value of the stock equals 3.5 ME up to a 350 KE error margin.

Please note: This interval estimate seems to be of the correct order of magnitude, given that the current total value of the stock equals $3,366,495 \text{ €}$ in the ledger account.

Comment on the accuracy: The accuracy obtained with 50 observations is $\pm 350 \text{ KE}$, which represents slightly more than $\pm 10\%$ of the current total value of the stock.

(4) As indicated above, the obtained accuracy is not sufficient; we basically want to divide the error margin by 10 and will thus have to multiply the sample size by $10^2 = 100$, thus leading in principle to an increased sample size of the order of 5,000... which is larger than the population size.

Thus (at least unless we proceed in a more clever way, see Questions (5) to (9) to do so) we would have to perform an exhaustive study to get the desired $\pm 1\%$ accuracy.

Note: The exact calculation to get the increased sample size n is:

$$1,532 \times 1.96 \frac{2,350}{\sqrt{n}} = 1,532 \times 1.96 \frac{753.74}{\sqrt{n}} = 1\% \times 3,366,495$$

that is, $n = \left(1,532 \times 1.96 \times 753.74 / 33,664.95 \right)^2 = 4,519.775...$

↳ we should take $n = 4,520$
but as discussed above, this would be ridiculous!

The conclusion is that we need a SMARTER way to proceed, and we will do so in Questions (5)-(9).

(5) We consider here paired data: for each reference j in the sample, we have:

- its current value y_j in the ledger account
- its actual value x_j , which we recomputed with some efforts

from which we are able to compute:

- its accounting gap $g_j = x_j - y_j$

This leads to the following extraction of relevant statistical information:

Population = Still the 1,532 references with small unit cost

Sample = Still 50 such references picked at random

Variable of interest = This time, the difference between the actual value of a reference in the stock and its current value in the accounts

Parameter of interest = Eventually, the sum of all differences, that is, by how much the current total value 3,386,495 € is incorrect

Which we will get access to through the average difference over all the 1,532 references, which we denote by Δ_0 .

The sum of all differences then equals $1,532 \Delta_0$

Data collected = e_1, \dots, e_{50} taking positive, negative or null values, where e_j denotes the difference computed for the j -th reference checked

Summary of the data = The sample average of the differences in values of the 50 references checked equals $\bar{e}_{50} = -11.73$ € (with an associated standard deviation of $s_{e,50} = 110.32$ €)

(6) We proceed similarly to Question (3):

$$\text{95\% confidence interval on } \Delta_0: \quad \bar{e}_{50} \pm 1.96 \frac{s_{e,50}}{\sqrt{50}} = -11.73 \pm 1.96 \frac{110.32}{\sqrt{50}} = -11.73 \pm 30.58$$

$$\text{95\% confidence interval on } 1,532 \Delta_0: \quad 1,532 \times (-11.73) \pm 1,532 \times 30.58 = -17,970.36 \pm 46,848.56$$

$$\begin{aligned} \text{which is included in } & -18,000 \pm 47,000 \\ & = -18 \text{ k€} \pm 47 \text{ k€} \\ & = -65 \text{ k€ to } +29 \text{ k€} \end{aligned}$$

Conclusion: With high confidence (95% confidence level), the total differences between the actual values and the current accounting values of the 1,532 references at hand equals -18k€ up to a ± 47 k€ margin of error.

- (7) The statement of course means the actual total stock value for the 1,532 references corresponding to unit costs less than 10€, we were not given the results of the exhaustive study on Category 2 items as discussed in Question (1).

With high confidence (95% confidence level), the actual total stock value for the considered 1,532 references equals

$$\begin{aligned} & 3,366,495 - 18,000 \pm 47,000 \\ & = 3,348,495 \pm 47,000 \text{ €} \end{aligned}$$

The current accuracy of this estimate is at $\pm 100 \times (47,000 / 3,366,495) \%$
 $\approx \pm 1.4 \%$ of the current stock value

- (8) We proceed as in Question (4): we want to reduce the error margin by a factor of 1.4 and thus have to consider a sample size larger by a multiplicative factor of $1.4^2 = 1.96 \approx 2$.

That is, we need a sample size of $50 \times 2 = 100$ and just need to check 50 other references on top of the 50 references already checked. This will only take 1 hour!

- (9) High-level reason: The method considered in Questions (2)-(4) started from scratch. In Questions (5)-(8) we do exploit available data given by the current accounting values; we use more information, we use existing information and that's always a plus.

How this high-level reason is reflected in our data:

- the current or actual values y_i and x_i are given by large numbers, with large associated standard deviations
- *while* we only have few and probably not so big accounting mistakes: the differences e_i are orders of magnitude smaller than the x_i or y_i (and many may be null), so that their standard deviation $s_{e,50}$ is much smaller than $s_{x,50}$ and $s_{y,50}$.

Hypothesis testing: methodology

Elementary exercises	page 46
Advanced exercises	
Nicotine patches, continued	page 50
Managing customers' dissatisfaction	page 53

Chapter 4 / Elementary exercises

Exercise 4.1

There are two possible errors:

- not offering a discount that would have been profitable for the company (because much more sales would have been performed)
 ↳ this results in some invisible shortfall (in some invisible loss of additional benefits)
 the associated risk is moderate:
 you will just not get promoted.
- offering a discount that turns out to generate less benefits than the current situation (because the increase in sales does not compensate for the reduced margin of profit)
 ↳ this results in tangible losses of benefits, everybody will realize how terrible your idea was
 the associated risk is severe:
 you might get fired!

Hence:

- the prudent viewpoint (H_0) is: H_0 : The discount is not profitable
- the risky viewpoint (H_1) is: H_1 : The discount is profitable

Based on sample data (to be collected), H_0 will be rejected if and only if these sample data show with high confidence that the discount should be profitable.

Minor / thin evidence of profitability will not be enough to reject H_0 ; we need some strong evidence.

Exercise 4.2

The statement explains that she would be ready to implement the name change if and only if there is enough support from the customers.

This is sound and rational:

- A name change is costly (in particular, in terms of advertisement!) and may confuse the customers.
 - ↳ There are high, short-term risks associated with a name change.
- But on the long run, if supported by customers, a name change can be very effective.
 - ↳ There are moderate, long-term benefits associated with a name change.

Hence,

- The prudent viewpoint (H_0) is to think that there is not enough support for the name change and that it should not be implemented.
- The risky viewpoint (H_1) is to think that there is enough support for the name change and that it should be implemented.

Mathematically speaking, deciding by

- p , the proportion of customers that would prefer the new name,
- $p_{ref} = 50\%$, the reference proportion,

we should test: $H_0: p \leq 50\%$ vs. $H_1: p > 50\%$

! Always put a non-strict inequality or an equality for H_0 .
 a strict inequality or a "different from" symbol \neq for H_1 .

The limit case of this set of hypotheses is: $H_0: p = 50\%$ vs. $H_1: p > 50\%$.

Exercise 4.3

(We are considering that the bank will announce the signal publicly, if she finds out that such a signal; Risk study is not done for internal purposes only.)

1. Sending a signal about a disaster to take place (here, an economic crisis) must be done with care: if the disaster didn't take place, your credibility will be affected for ever.

↳ that is a high risk!

While not sending out a signal when it should have been issued is a less severe mistake (many other institutions will probably have missed the early warnings as well).

↳ That is a moderate risk.

Hence,

- The prudent viewpoint (H_0) is to think that the current delinquency rate is the standard one: $H_0: p_0 = 8.5\%$
- The risky viewpoint (H_1) is to think that we have enough evidence to claim that we have an early signal of a crisis, $H_1: p_0 > 8.5\%$

Another justification could be that $H_1: p_0 > 8.5\%$ is what we want to possibly prove from the data: sample data might indicate some deviation from the 8.5% reference value, but we will want to know whether such deviation can be explained by chance only or if it was so unlikely that it actually carries the signal of a crisis taking place.

2. On our sample, we have a delinquency rate of $\bar{x}_{50} = 14\%$ and question is (see above) whether those 14% of delinquency rate are significantly larger than the reference rate of 8.5%.

(Of course they are larger, question is whether they are significantly larger.)

Test statistic:
$$t_{50} = \sqrt{50} \frac{\bar{x}_{50} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}} = \sqrt{50} \frac{\bar{x}_{50} - 0.085}{\sqrt{0.085(1-0.085)}}$$

Behavior expected under H_0 :

normal curve
(ie, standard Gaussian distribution)

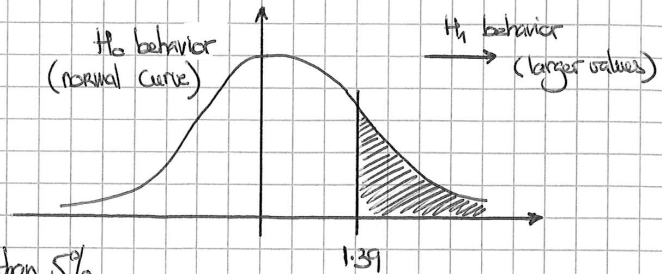
Behavior expected under H_1 :

larger values than under H_0
(why? because \bar{x}_{50} is expected to be close to p_0 , which is 7 pef)

Value on data:

sample average of $\frac{7}{50} = 14\%$
 thus t_{50} takes the value $\sqrt{50} \frac{0.14 - 0.085}{\sqrt{0.085(1-0.085)}} \approx 1.39$

Summary of all these elements
 on a picture:



P-value $\approx 8\%$, in particular, (slightly) larger than 5%

P-value
 $= 100\% - 91.77\% = 8.23\%$

Statistical conclusion:

Based on these data, we fail to reject H_0 , that is, the evidence gathered with this sample of 50 ffs is not strong enough to declare that a crisis is about to take place.

Business conclusion:
 (was not specifically asked for)

Wait one or two more months - then stronger signals may be observable.

All in all, it would also be a good idea to use a larger (eg, 100 or 200 ffs) sample; 50 is not so large a sample size.

Chapter 4 / Exercise "Nicotines patches, continued"

Question 1

Can we prove that the population efficiency rate is larger than 55%?
(Given that we could not prove it to be larger than 60%.)

We are considering the hypotheses

H_0 [prudent viewpoint]: The desired efficiency is not achieved,
ie, $p_0 \leq 55\%$

H_1 [what we want to prove]: The desired efficiency is achieved,
ie, $p_0 > 55\%$

We take the limit case: $H_0: p_0 = 55\%$ versus $H_1: p_0 > 55\%$

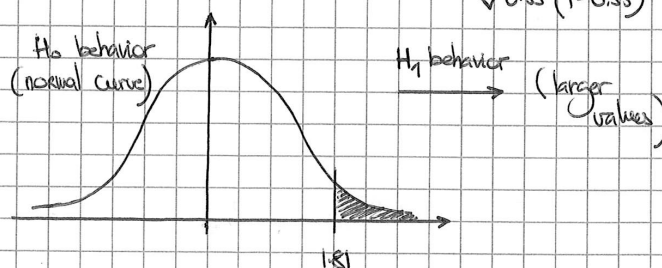
Test statistic: $t_{100} = \frac{\bar{x}_{100} - 0.55}{\sqrt{0.55(1-0.55)}}$ (now, the reference proportion equals 55%)

Behavior expected under H_0 : normal curve
(ie, standard Gaussian distribution)

Behavior expected under H_1 : larger values than under H_0

Value on data: sample average of $\bar{x}_{100} = 64\%$, thus
 t_{100} takes the value $\frac{\sqrt{100} (0.64 - 0.55)}{\sqrt{0.55(1-0.55)}} \approx 1.81$

Summary on a picture:



P-value $< 5\%$ (of the order of 3.5%)

$$P\text{-value} = 100\% - 96.49\% = 3.51\%$$

Statistical conclusion:

We reject H_0 . That is, the evidence gathered on this sample of 100 volunteers is sufficient to guarantee (with a small risk of error/ with high confidence) that the population efficiency rate is indeed larger than 55%.

Put differently, the sample proportion of 64% is

significantly larger than 55%.

Business conclusion:

Go and advertise
this 55% efficiency
rate!

Now, let's move to another
(and actually, dishonest) viewpoint
on the same data...

Question 2.

Let us consider the hypotheses $H_0: p_0 \geq 65\%$ (high efficiency rate)
 $H_1: p_0 < 65\%$ (low rate)

They are not well chosen (see comments after the statistical conclusion)
but let us proceed with the calculations despite all...

We take the limit case $H_0: p_0 = 65\%$ versus $H_1: p_0 < 65\%$

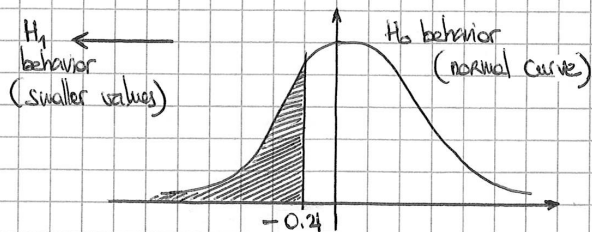
Test statistic: $t_{100} = \frac{\bar{x}_{100} - 0.65}{\sqrt{0.65(1-0.65)}}$ (now, the reference
proportion equals 65%)

Behavior expected under H_0 : normal curve
(ie, standard Gaussian distribution)

Behavior expected under H_1 : smaller values than under H_0
(y. \bar{x}_{100} should be close to p_0 , where
under H_1 , $p_0 < 65\%$)

Value on data: $\frac{0.64 - 0.65}{\sqrt{0.65(1-0.65)}} \approx -0.21$

Summary on a picture:



P-value $\approx 40\%$,
much larger than 5%

$$P\text{-value} = 100\% - 58.32\% = 41.68\%$$

Statistical conclusion:

Based on these data, we fail to reject H_0 , ...
which simply means that these data, while
contradicting H_0 (we read a sample average
proportion that is smaller than 65%), does not
contradict it severely enough for H_0 to be
rejected.

Note that we did not accept or prove H_0 : we could simply not reject it based on these data.

Business conclusion: There's none... Because:
The statistical conclusion we could draw is useless!

Indeed (as indicated in the textbook), based on the same data, we simultaneously failed to reject the following two statements:

$$p_0 \leq 60\% \quad \text{and} \quad p_0 \geq 65\%$$

One of them must be incorrect, though.

Failing to reject $H_0: p_0 \geq 65\%$ does not mean at all that we could prove that the efficiency rate is indeed larger than 65%. We don't know whether this statement is true or false, there is no conclusion we can draw from this failure of rejection.

Managing customers' dissatisfaction.

Population : Customers having recently bought or about to buy a product from our company

Population parameter : p_0 , the current or future dissatisfaction rate

Reference parameter : p_{ref} , the typical dissatisfaction rate observed before the change in customers' relation (before hiring sales advisers)

Hypotheses :

don't write such a lengthy justification !

The manager wants to see whether her/his policy change was effective. It is prudent (H_0) to think that it had no impact — at least, this is the viewpoint that the manager's managers would start with. Now, the manager would like to prove (H_1) that she/he has had an impact on the (dis)satisfaction rate.

this would be sufficient on your answer sheet !

That is :

H_0 [prudent] : similar dissatisfaction rate
 H_1 [what should be proved] : lower such rate

In equations :

H_0 : $p_0 = p_{ref}$
 H_1 : $p_0 < p_{ref}$

Data collected :

* We discard the customers with no opinion ! *

We thus only collected x_1, \dots, x_{459} taking 0/1 values and where

$$x_j = \begin{cases} 1 & \text{if the } j\text{-th customer in the sample intends to complain} \\ 0 & \text{otherwise} \end{cases}$$

Data summary : A fraction $\bar{x}_{459} = \frac{32}{459} = 7.0\%$ of the customers of the sample intend to complain.

Test statistic :

$$t_{459} = \frac{\bar{x}_{459} - p_{ref}}{\sqrt{p_{ref}(1 - p_{ref})}}$$

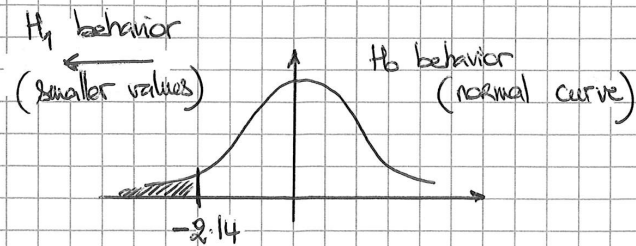
Expected H_0 behavior : normal curve

Expected H_1 behavior : smaller values

Value on data :

$$t_{459} = \frac{0.07 - 0.10}{\sqrt{0.10(1 - 0.10)}} \approx -2.14$$

Picture summary:



$$P\text{-value} = 100\% - 98.38\% = 1.62\% < 5\%$$

Statistical conclusion:

We reject H_0 , that is, these data show that the dissatisfaction rate dropped due to the new policy of better guiding customers.

Business conclusion:

Ask for a raise or a promotion — in any case, advertise the improvement and get some congratulations.

(If you have any more creative business conclusion, send it to me: stoltz@hec.fr)

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

Elementary exercises	page 56
Advanced exercises	
A controversial governmental reform, short version	page 62
Seizure of MegaUpload and side effects	page 64
Public health surveillance	page 66
“We look like our names”	page 69
A controversial governmental reform, long version	page 70
Walking many steps a day	page 72

Chapter 5 / Elementary exercises

Exercise 5.1:

1. The statement points out that we have no preliminary opinion on the question; our reasonable starting point will thus be that average heights are somewhat similar around the world and thus that the average height of male ABS students equals the average height of the country where the study is taking place, that is, the Netherlands.

thus: H_0 [reasonable viewpoint]: same average heights, $\mu_0^o = 181$
 H_1 : different average heights, $\mu_0^o \neq 181$,
 that is, either $\mu_0^o > 181$ or $\mu_0^o < 181$.

2. Note that we collect data only after the hypotheses were set.

The sample data features 253 data values x_1, \dots, x_{253} , where x_j is the height measured on the j -th male ABS student in the sample.

It can be summarized as follows: the sample average height equals $\bar{x}_{253} = 181.24$ (with an associated standard deviation of $s_{x,253} = 8.11$).

The P-value is obtained as follows:

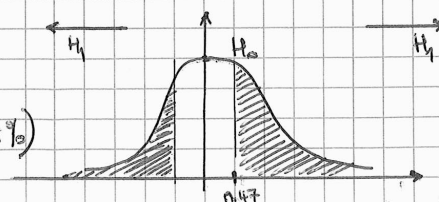
Test statistic
$$t_{253} = \frac{\sqrt{253} \cdot \bar{x}_{253} - 181}{s_{x,253}}$$

Under H_0 : normal-curve behavior

Under H_1 : large positive or large negative values (any of the two deviations $\mu_0^o > 181$ or $\mu_0^o < 181$ could take place)

Numerical value:
$$\frac{\sqrt{253} \cdot 181.24 - 181}{8.11} = 0.47$$

Summary picture:



The P-value is obtained as the sum of two identical probabilities.

$$P\text{-value} = 2 \times (100\% - 68.08\%) = 63.84\%$$

The P-value is (much) larger than 5%, we fail to reject H_0 .

In plain words: these data do not suggest that there is a difference in average heights between male ASS students and Dutch male adults.

↑
(Write a cautious and inconclusive conclusion, as always when we stick to H_0)

Note: we can already answer partly to question 1 of the third exercise:

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

↑ numerical value of t_{253}

P-value ↓ 63.7%

We find again our results, up to some minor rounding differences.

Exercise 5.2

1. We consider the hypotheses

H_0 [reasonable viewpoint]: same average heights for women,
 $\mu_0^f = 169$

H_1 : different average heights, $\mu_0^f \neq 169$

We will compute the P-value based on data y_1, \dots, y_{223} , where y_k denotes the height of the k -th ABS female student in the sample.

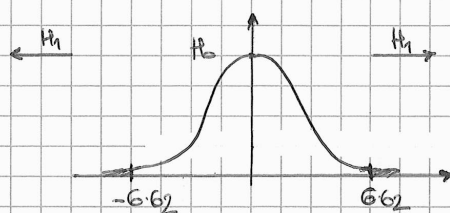
Data can be summarized as: average sample height $\bar{y}_{223} = 166.03$
 (with an associated standard deviation of $s_{y,223} = 6.70$).

Test statistic: $t_{223} = \frac{\sqrt{223} (\bar{y}_{223} - 169)}{s_{y,223}}$

Since H_0, H_1 behaviors as in the previous exercise

Numerical value: $\sqrt{223} \frac{166.03 - 169}{6.70} = -6.62$

Picture:



P-value is almost null
 (values above 5 are out of the table).

We strongly reject H_0 : we have indeed $\mu_0^f \neq 169$.
 Looking at the data, we get: $\mu_0^f < 169$

In plain words: these data show that the average height of ABS female students is (significantly) smaller than the average height of the general female Dutch population.

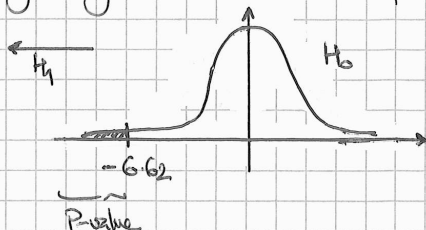
2. Given the context, we would this time resort to

H_0 [reasonable]: same average heights, $\mu_0^f = 169$

H_1 [what is to be proved]: smaller average height at ABS than in the Netherlands, $\mu_0^f < 169$

where "reasonable" stands for the viewpoint of people with no preconception

The only difference wrt question 1 is the behavior expected under H_1 : now, only large negative values are expected, so that the picture is



The P-value gets twice smaller, and remains almost null of course. We then reach the same conclusions.



Because the P-value depends on H_1 , the hypotheses H_0, H_1 have to be set before data is collected, based on the preconceptions or lack of preconceptions of the statistician.

3. We may check our results (numerical value + P-value) on the SPSS outputs: see next exercise.

Exercise 5.3.

We solve it only for women data; the solution is similar for men data.

1. Finding again all numerical values.

Test (for women)

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

Sample size \rightarrow N
 \bar{y}_{223} \rightarrow Mean
 $s_{y, 223}$ \rightarrow Std. Deviation

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

reference value $\mu_{ref} = 169$ \rightarrow Test Value = 169

numerical value of t_{223} \rightarrow t

P-value in the $H_1: \mu \neq 169$ situation (the "2-tailed" stands for that: the P-value is computed by considering the two tails of the normal curve) \rightarrow Sig. (2-tailed)

2. What other values correspond to!

(Note that 0 does not belong to the computed confidence interval for $\mu_0^? - 169$, which is yet another, equivalent, proof that $\mu_0^? \neq 169 \dots$)

Standard error on the mean:

$$s_{y,223} / \sqrt{223}$$

that is, the half-width of a symmetric confidence interval up to the 1.96 (quantile) factor

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

sample size - 1

SPSS actually uses the Student distribution with 222 degrees of freedom (df) instead of the standard Gaussian distribution; both are almost indistinguishable, though

$$\bar{x}_{223} - \mu_{ref} = 166,03 - 169$$

Symmetric confidence interval at the 95% confidence level for $\mu_0^? - 169$, that is:

$$\begin{aligned} & \bar{y}_{223} - 169 \pm 1.96 \frac{s_{y,223}}{\sqrt{223}} \\ & = -2.969 \pm 1.96 \times 0.449 \end{aligned}$$

EXERCISE 5.1 — A controversial governmental reform

1. Population: All inhabitants of France (mature enough to actively participate to political activities, say: older than 18 years)

Parameter: The fraction p_0 of the population that would actively fight the reform if it was indeed proposed

Hypotheses: The prudent behavior is to think that there will be mass political fights; the risky action is to propose the reform.

H_0 [prudent]: active political fights, i.e., $p_0 \geq 15\%$
 H_1 [risky]: little active opposition, i.e., $p_0 < 15\%$

with limit case:
 $H_0: p_0 = 15\%$
 $H_1: p_0 < 15\%$

2. Data collected: * We discard the no-opinion interviewees! *
 We collected x_1, \dots, x_{980} taking 0/1 values and where

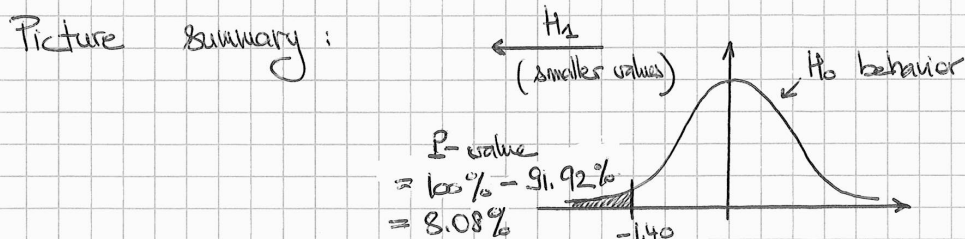
$$x_j = \begin{cases} 1 & \text{if the } j\text{-th interviewee is ready to actively fight the reform} \\ 0 & \text{otherwise} \end{cases}$$

Data summary: A fraction $\bar{x}_{980} = \frac{131}{980} \approx 13,4\%$ of the citizens of the sample are ready to actively fight the reform.

Test statistic:
$$t_{980} = \frac{\sqrt{980} (\bar{x}_{980} - p_{ref})}{\sqrt{p_{ref}(1-p_{ref})}}$$
 where $p_{ref} = 15\%$

Expected behaviors → under H_0 : normal curve
 ↘ under H_1 : smaller values

Value on the data:
$$t_{980} = \frac{\sqrt{980} (0.134 - 0.15)}{\sqrt{0.15(1-0.15)}} = -1.40$$



Statistical conclusions:
Political

The conventional risk level is 5%.
Here, we obtain a superior P-value
(about 8%).

The risk-averse ministers
will fail to reject H_0 .

They will say that data cannot
exclude some mass opposition.

These timid ministers will not open
Pandora's box and will not
give their names to a bill. They will
not remain in the annals but
will keep their good salaries for
many more months.

The adventurous ministers
will note that the P-value
is quite close to 5%.

(If only 128 instead of 131
interviewees had been declaring
that they would actively fight
the reform, then
the P-value would have
been smaller than 5%.)

They will say that for them,
data can lead to rejecting
 H_0 .

They will propose a bill in
Parliament with their name on it.

They take some risk but oh well,
life is full of (controlled) risks.

Alternative solution suggested by Fall 2018 students

You could argue that politicians are not afraid at all to take risks and that they love to take actions. Furthermore, they are stubborn. Hence, their starting point could be $H_0 : p_0 \leq 15\%$ and their alternative hypothesis could be $H_1 : p_0 > 15\%$. They would only be ready to consider H_1 and depart from what they had carefully planned if data strongly show that this is needed.

The test statistic would still equal -1.40 but the P-value would be 91.92%. Hypothesis H_0 would not be rejected. The conclusion in plain words would be: "Data collected do not rule out that little opposition would take place". (But data do not prove so, they only do not rule out this possibility.)

EXERCISE 52

Seizure of MegaUpload and side effects

1. Parameter of interest: μ_0 , the average share of Internet users visiting Gplay.fr during the period of time considered (January 20 to February 29, 2012)

↳ This parameter may be difficult to grasp but suppose that alexa.com is measuring the share of users every 30 minutes, then 48 measurements are made, times $11 + 29 = 40$ days: 1,920 measurements in total over the period considered. Then, μ_0 is just the uniform average of these 1,920 values.

Since we do not have the physical time to go on this website every 30 minutes, we will only go there once a day, at a time picked at random. That is, our sample will consist of 40 measurements out of the 1,920 made in total.

(Our population is: these total 1,920 measurements.)

Hypotheses:

It is reasonable to think that people went to MegaUpload for movies and TV shows that are anyway not available on Gplay.fr, and hence, that its seizure did not really impact our website.

Of course, a good surprise would be a positive impact, but this good surprise must come with a proof, we cannot take it as our starting point.

H_0 [reasonable]:

The seizure of MegaUpload did not result in an increase of the traffic to Gplay.fr, that is, $\mu_0 = \mu_{ref}$

reference value:

$\mu_{ref} = 0.021\%$

H_1 [what would need to be backed up by strong facts]:

It did result in an increase of our traffic, that is, $\mu_0 > \mu_{ref}$

2. Data collected: $x_1, \dots, x_{40} \in [0,1]$, where x_j denotes the share measured on day j (at a randomly picked hour)

Data summary: $\bar{x}_{40} = 0.023\%$ and $s_{x,40} = 0.005\%$

Question to be answered: Is this sample average of 0.023% significantly larger than the reference value of 0.021%?

Test statistic:

$$t_{40} = \frac{\sqrt{40} (\bar{x}_{40} - 0.021)}{s_{x,40}}$$

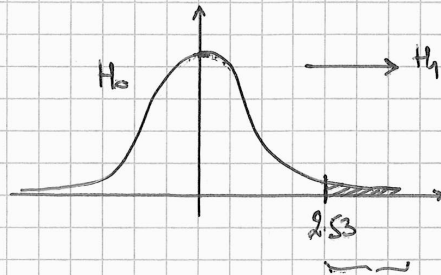
Behavior under H_0 : normal curve

Behavior under H_1 : takes larger values than under H_0

Numerical value on data:

$$\frac{\sqrt{40} (0.023 - 0.021)}{0.005} = 2.53$$

Summary picture:



$$P\text{-value} = 100\% - 99.43\% = 0.57\% \ll 5\%$$

Statistical conclusion:

we (strongly) reject H_0 : these data prove that the Internet traffic to Gplay.fr increased after the seizure of MegaUpload!

! The increase is significant but maybe not so large (= it does exist) (= μ_0 is larger than $\mu_{\text{ref}} = 0.021\%$ but maybe not so much larger!).

Business conclusion(s):

- 1 = Quantify the increase, (via confidence intervals) charge more for online advertisements on Gplay.fr, get more revenue
- 2 = Take legal actions to get many other filesharing websites seized

EXERCISE 5.3 - Public health surveillance.

1. Population of interest: All the kids in Parisian elementary schools (several dozens of thousands?)

Parameter of interest: p_0 , the current asthma prevalence rate among them

(= the current proportion among them suffering from asthma)

Hypotheses to be tested:

This study is a warm-up, we merely want to determine whether we have a case and can alert press. We feel that there is a problem but we need to prove it, or at least, show some red flag. Thus:

H_0 : The situation of Paris w.r.t pediatric asthma is similar to the rest of France, $p_0 = 9.7\%$

H_1 [what we want to prove]: The Parisian air induces asthma in greater proportion than what is nationally observed in France: $p_0 > 9.7\%$

Note: we could also have taken $H_0: p_0 \leq 9.7\%$ but I doubt that any parent would be ready to believe that the Parisian air might be better than in the rest of France... So I prefer my $H_0: p_0 = 9.7\%$

2. Data collected: $x_1, \dots, x_{300} \in \{0,1\}$, where

$x_j = \begin{cases} 1 & \text{if the } j\text{th kid surveyed reports an asthma attack} \\ 0 & \text{otherwise} \end{cases}$

Data Summary: $\bar{x}_{300} = \frac{37}{300} \approx 12.3\%$ of the 300 kids in the sample report an asthma attack in the past six months.

Test statistic: $t_{300} = \sqrt{300} \frac{\bar{x}_{300} - 0.097}{\sqrt{0.097(1-0.097)}}$

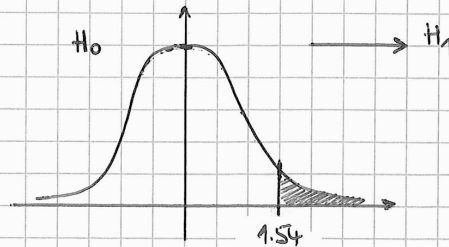
Behavior under H_0 : normal curve

under H_1 : larger values

Numerical value on data:

$$\sqrt{300} \frac{37/300 - 0.097}{\sqrt{0.097(1-0.097)}} = 1.54$$

Summary on a picture:



$$P\text{-value} = 100\% - 93.82\% = 6.18\%$$

Statistical conclusion: a somewhat rather borderline P value!

The best would be to point out that the P-value is close to 5%, hence serious doubts can be raised against H_0 , and further studies are needed.

Business action: Parents have some red flag, they should lobby to get some official, better-quality, study be conducted. They are entitled to alert press given the almost convincing data they gathered.

⚠ You as an individual cannot prove or disprove the detrimental effect of air quality on children's health. You need some official involvement of the health authorities. But your point is to encourage them to start an official study. Your hand-made / DIY study is needed exactly for this encouragement!

3. Which P-value would we have obtained with 38 kids suffering from asthma instead of the 37 ones considered above?

$$\text{Value of } t_{300} \text{ on these new data: } \sqrt{300} \frac{38/300 - 0.097}{\sqrt{0.097(1-0.097)}} \approx 1.74$$

Thus a P-value of $100\% - 95.91\% = 4.09\% < 5\%$

We would reject H_0 , but our business action would remain the same: get some official study conducted!

The aim of this question was to make sure that you would not draw conclusions / make actions that would be completely different between 37 and 38 asthma cases reported.

One additional case should not drastically change the picture.

All in all, this shows how relative / conventional the 5% level is;
this shows that the P-value has to be interpreted in some
cautious way.

(See also the conclusion to the exercise about a controversial
governmental reform.)

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 $p_{ref} = 25\%$.

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.

It's reasonable to think that people have no clue in guessing and hence, are not better or worse than pure chance; the alternative hypothesis will be two-sided as this sentence explains (we are open to people guessing better or worse than pure chance).

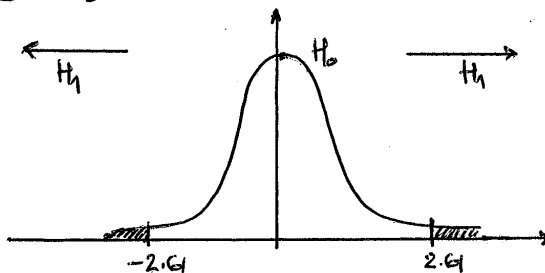
$H_0: p_0 = p_{ref}$, where p_0 is the proportion of people that would guess the correct name \leftrightarrow people guess like at random
 $H_1: p_0 \neq p_{ref} \leftrightarrow$ people guess better or worse than just at random

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.

did not need to be written on your sheet

$$t_{67} = \frac{\bar{x}_{67} - 0.25}{\sqrt{0.25(1-0.25)}} \quad \text{with numerical value} \quad \sqrt{67} \frac{(26/67 - 0.25)}{\sqrt{0.25(1-0.25)}} \approx 2.61$$



$$P\text{-value} = 2(100\% - 99.55\%) = 0.9\% \ll 5\%$$

We strongly reject H_0 .

- ☐ 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 ”).

The data collected from this experiment showed that people can guess names significantly better than if they were just guessing at random. This means that individuals “look like their names”!

(Note: This experiment was successfully conducted with many other individuals to identify.)

NB: no point granted if your answer just mentioned that the ability of guessing was better or worse than a random guess.

$$\bar{x} = \frac{275}{979} \approx 28.1\%$$

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%.



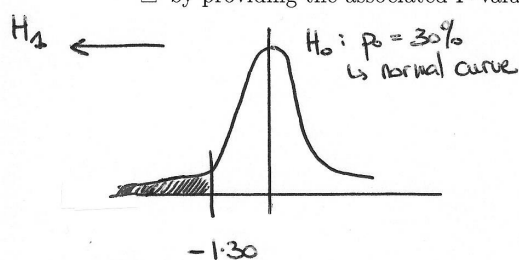
We will first consider two pairs of hypotheses and test each of these pairs, only then we will indicate which pair a given government should choose.

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.



→ Limit case $H_0 : p_0 = 30\%$ instead

$$t = \sqrt{979} \left(\frac{275/979 - 0.3}{\sqrt{0.3(1-0.3)}} \right)$$

$$t \approx -1.30$$

$$P\text{-value} = 100\% - 90.32\% \approx 9.7\% \quad 75\%$$

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

[Beginning]

- ☒ A. The data collected cannot exclude that
- ☐ B. The data collected suggest that
- ☐ C. The data collected show that

☐

[Middle]

- ☒ 1. more than 30%
- ☐ 2. less than 30%

of the population is strongly against the reform under review.

(But not B1)

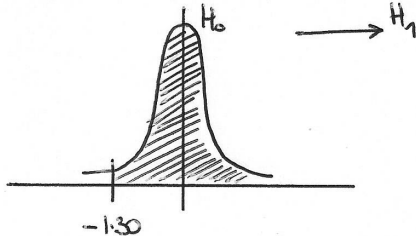
Possibly B2 if you argue that 9.7% is a borderline P-value (if you use a 10% risk level)

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

→ Again, limit case $H_0 : p_0 = 30\%$

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.



Large P-value: 90.32%

(which is not a surprise:
 $\bar{x} = 28.1\%$ is in accordance
 with $H_0 : p_0 \leq 30\%$)

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

- ☐ Letter: **A** Number: **2**

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** governments
- $H_0 : p_0 \leq 30\%$ against $H_1 : p_0 > 30\%$ is for **ideological** governments

- ☐ Provide a brief justification for your choices.

A cautious government would like to prove $H_1 : p_0 < 30\%$
 An ideological government would on the contrary use $H_0 : p_0 \leq 30\%$
 as a starting point (= statement to be challenged)

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)?

these values
 $t = -1.304$
 and $n = 979$
 were useful to
 check your
 answers above!

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

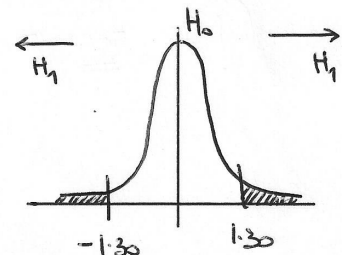
One-Sample Test				
Test Value = 0.3				
	t	df	Sig. (2-tailed)	Mean Difference
Strongly against	-1,304	978	,194	-,019

Reminder:
 Hypotheses cannot
 be set based on
 the data (this, nor
 on P-values)

cf. SPSS does it
 in a 2-tailed
 way:

$\begin{cases} H_0 : p_0 = 30\% \\ H_1 : p_0 \neq 30\% \end{cases}$

P-value
 $= 2 \times (100\% - 90.32\%)$
 $= 2 \times 9.7\% = 19.4\%$



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:

- green
red
orange
- ↔ we are certain that it will be reached: ↔ we reject H_0 and go for $H_1: \mu_0 > 10,000$
 - ↔ we are certain that it will not be reached: ↔ we reject H_0 and go for $H_1: \mu_0 < 10,000$
 - ↔ we do not know yet: ↔ we fail to reject $H_0: \mu_0 = 10,000$
- stick to $H_0: \mu_0 = 10,000$

This of course means that people should walk at least 10,000 steps a day on average (not necessarily every single day)



Design of the underlying test

- ☐ Indicate the parameter of interest μ_0 out of the four following statements:

- 1A. the individual daily numbers of steps made so far = sample data $x_1 \dots x_{49}$
- 1B. the average daily number of steps made so far = sample average \bar{x}_{49}
- 1C. the individual daily numbers of steps (made so far and) to be made in the upcoming months = $x_1 \dots x_{49} x_{50} x_{51} \dots$
- 1D. the average daily number of steps (made so far and) to be made in the upcoming months = μ_0

- ☐ What pair of hypotheses should we consider based on our aim for a color code?

- 2A. $H_0: \mu_0 \geq 10,000$ vs. $H_1: \mu_0 < 10,000$ ← would only result in 2 colors, not 3
- 2B. $H_0: \mu_0 \neq 10,000$ vs. $H_1: \mu_0 = 10,000$ ← We cannot test for this, for mathematical reasons H_0 need to be stated with $>, <, \text{ or } =$, but cannot be stated with \neq
- 2C. $H_0: \mu_0 = 10,000$ vs. $H_1: \mu_0 \neq 10,000$

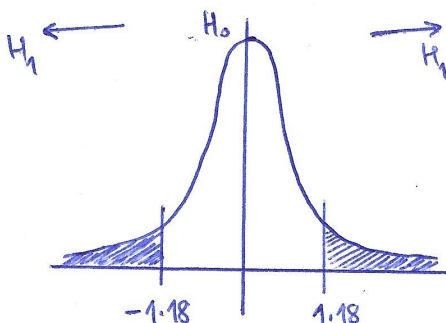
- ☐ Provide a brief justification of your choice, based on our aim for a color code.

Red/Green will be for the cases when we prove something, i.e. when we are able to reject $H_0: \mu_0 = 10,000$ and go either for $H_1: \mu_0 > 10,000$ (Green) or $H_1: \mu_0 < 10,000$ (Red); otherwise, as long as we stick to H_0 , we don't know anything for sure, we just cannot exclude that $\mu_0 = 10,000$

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.



$$t = \sqrt{49} \frac{10,532 - 10,000}{3,154} \approx 1.18$$

$$P\text{-value} = 2 \times (100\% - 88.10\%) = 23.8\% > 5\%$$

We fail to reject H_0

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

Reminder:
SPSS uses a
Student's distribution
(with 48 degrees of
freedom) to compute the
P-value

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.

Value of the test statistic
= the same as we had

P-value for a 2-tailed test (ie, with $H_1: \mu_0 \neq 10,000$)
is slightly different from what
we computed because SPSS relies on a
Student's distribution and
does not approximate
it by a
normal distribution

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

→ Orange

as we failed
to reject $H_0: \mu_0 = 10,000$

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.)

[Similar picture]

$$t = \sqrt{115} \left(\frac{10,452 - 10,000}{2,356} \right) = 2.05$$

$$P\text{-value} = 2 \times (100\% - 97.98\%) = 4.04\% < 5\%$$

We reject H_0 and go for $H_1: \mu_0 \neq 10,000$
Given the sample average $\bar{x}_{30} = 10,452 > 10,000$
we thus proved that $\mu_0 > 10,000$

Third data set

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

→ Green color

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

Green is ruled out, but both Orange and Red are possible.

→ Green is ruled out: impossible to prove that $\mu_0 > 10,000$ based on $\bar{x}_{62} = 9,759 < 10,000$

→ Orange and Red are possible: the question is whether $\bar{x}_{62} = 9,759$ is significantly smaller than 10,000; to determine this, we would need the standard deviation $s_{x,62}$.
If we had it, we could compute $t = \sqrt{62} \frac{9,759 - 10,000}{s_{x,62}}$

Two-sample tests (Comparing two populations)

Elementary exercises	page 76
Advanced exercises	
Separate or pooled marketing campaign?	page 80
Comparing prices in two local supermarkets	page 84
Gender pay gap, revisited	page 89
Alcohol consumption during the POWs at HEC, first version	page 91
Cash in the wallet, by country	page 94
Speedy self-assessment	page 95
The effect of touch, re-worked	page 96
Choosing between two gifts	page 98
Lead levels in children's blood	page 100

Elementary Exercise 6.1

The exercise about independent data, as two unrelated samples (one of 10 left-handed persons, and another one of 10 right-handed persons) are considered.

- (1) The reasonable viewpoint is that, a priori, there should not be much of a difference between left-handed and right-handed persons (otherwise, an evolutionary argument à la Darwin would have led to the extinction of one of these two categories).

↳ Thus H_0 is the equality of the average reaction times, $\mu_0^R = \mu_0^L$.

Now, as far as H_1 is concerned, the statement clearly explains that there are two competing phenomena: better brain connections for left-handed people vs. objects of daily life typically designed for right-handed people. So, we have no expectation as to which category could be favored... Therefore, we are open to any deviation and consider the two-sided alternative $H_1: \mu_0^R \neq \mu_0^L$.

- (2) The respective sample average reaction times equal 81.392 s for right-handed people and 94.958 s for left-handed people.

Are these two values significantly different?

We read the second table:

- First, the standard deviations of the reaction times are not significantly different (P-value of 20.7%)

- Therefore, we will read the first line of this table

	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	
Reaction time (sec.)	1,711	,207	-2,488	18	,023	-13,5659	5,4620	-25,0200	-2,1117
			-2,488	14,076	,026	-13,5659	5,4620	-25,2533	-1,8785

- Finally, the P-value associated to the two-sided test of $H_0: \mu_0^L = \mu_0^R$ vs. $H_1: \mu_0^L \neq \mu_0^R$ is then seen to be equal to 2.3%, which is $< 5\%$

Statistical conclusion: We reject H_0 . These data show that the average reaction times are different.

Now, we *must* go one step further and say who is faster than whom: given the sample values, we infer that $\mu_0^R < \mu_0^L$, i.e. that right-handed people have faster reaction time on average.

Elementary Exercise 6.2

The exercise is about independent data, as two unrelated samples are simultaneously studied.

- (1) The two gifts seem indistinguishable, it's reasonable to start with the assumption H_0 that they have the same impact.

As for H_1 , we are ready to be convinced that either of them has more impact than the other, we have no clue in advance regarding which of them could be more efficient.

- (2) We denote by x_1, \dots, x_{300} and y_1, \dots, y_{300} the facts whether an order was placed ($x_j = 1$ and $y_k = 1$) or not ($x_j = 0$ and $y_k = 0$).

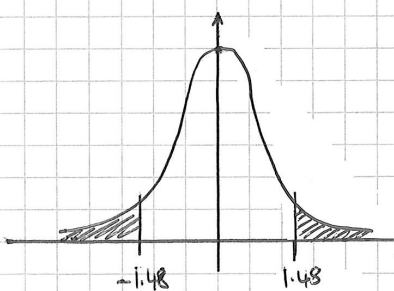
Test statistic
$$t_{300, 300} = \frac{1}{\sqrt{1/300 + 1/300}} \left(\frac{\bar{x}_{300} - \bar{y}_{300}}{\sqrt{\bar{x}_{300}(1-\bar{x}_{300})}} \right)$$

where $\bar{x}_{300} = \text{pooled estimation}$

Behavior under H_0 : normal curve
under H_1 : larger or smaller values than under H_0 (large positive or large negative values)

Numerical values on data:
$$\bar{x}_{300} = \frac{125 + 143}{300 + 300} = \frac{268}{600} \approx 44.7\%$$

and
$$t_{300, 300} = \frac{1}{\sqrt{2/300}} \left(\frac{125/300 - 143/300}{\sqrt{0.447(1-0.447)}} \right)$$



P-value = $2 \times (100\% - 93.06\%) = 13.88\%$
 $> 5\%$

We fail to reject H_0 , that is:

These data do not bring to light any superiority of one gift over the other.

Corresponding business action: if one of the two considered gifts is way cheaper than the other one, pick the former!
If not, let the customer choose!

Elementary Exercise 6.3

(1)

This is paired data as the same 200 customers are followed. We measure 2 values x_j and y_j for customer j , namely,

- the amount x_j ordered in the first trimester (with no gift)
- the amount y_j ordered in the second trimester (with a gift)

We will study the differences $y_j - x_j \stackrel{\text{not}}{=} z_j$ in the amounts ordered.

We got a sample average increase of $\bar{z}_{200} = 1.58 \text{ €}$.

The corresponding parameter of interest is Δ_0 , a population average increase: the quantity by which the amounts of the orders placed would increase in the presence of a gift, where the average is computed over all the members of the online buying club.

(2) We take as hypotheses:

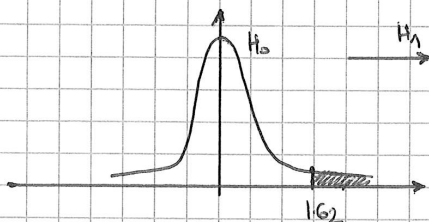
prudent / risky
because
the gift costs
money!

- H_0 [prudent]: The small gift has no impact, $\Delta = 0$
- H_1 [risky]: The small gift has a positive impact on sales, $\Delta > 0$

Test statistic:
$$t_{200} = \sqrt{200} \frac{\bar{z}_{200} - 0}{s_{z,200}}$$

Behavior under H_0 : normal curve
under H_1 : larger values than under H_0

Numerical value:
$$\sqrt{200} \frac{1.58}{13.8} \approx 1.62$$



P-value = $100\% - 94.74\% = 5.26\% \approx 5\%$

↳ a borderline value...

but as this is a business problem (and not a critical issue), we may well declare that these data lead to the rejection of H_0 , i.e., that they show a positive impact of the small gift.

↳ we should offer the small gift to everyone purchasing for more than 30 €.

Elementary Exercise 6.4

Answering students' complaints

This exercise is about independent data:

There are two populations, 1. the students taking the quiz first
2. the students taking the quiz second

Each of these populations is composed of thousands of students as they gather students over several years.

The independent samples considered here have sizes 40 and 36, respectively.

The parameters of interest are $\mu_0^{(1)}$ and $\mu_0^{(2)}$, the average grades in each group.

We test $H_0: \mu_0^{(1)} = \mu_0^{(2)}$ the communication during the break has no effect
 $H_1: \mu_0^{(1)} \neq \mu_0^{(2)}$ this communication has an effect, either a detrimental or a positive effect

what the studies claim

what the students think

Outcome of the test:

P-value 48.5%, no significant difference
between the samples averages
 $\bar{x}_{40} = 14.3$ and $\bar{y}_{36} = 14.847$

Statistical conclusion:

Data do not rule out that communication during the break has no effect

(but as usual we haven't proved that fact).

Instructors' action:

So far so good, a single statement does a reasonable job but continue monitoring the differences and collect further evidence

→ If for 10 years in a row no significant difference can be proved and if the rankings of \bar{x} and \bar{y} are not always the same (sometimes a non-significant $\bar{x} > \bar{y}$, sometimes a non-significant $\bar{x} < \bar{y}$); then we may have a proof.

Chapter Two-sample tests:

SEPARATE OR POOLED?
MARKETING CAMPAIGN?

Ante-scriptum:

My business situation is a bit unclear.
Let us assume that we want to determine whether men and women share similar tastes when it comes to the advertisement of fitness machines, or not. If so, then our marketing department will work on one single campaign (but maybe with a third, different, picture); otherwise, it will build two separate campaigns.

↳ The two pictures considered here are only used as a test to assess whether the preferences vary or are similar between men and women.

Question 1:

Type of data:

independent samples + proportions

Populations:

the men and the women practising in fitness centers; by extension, we hope to study all men and women interested in using fitness machines.

We pick one picture as a reference picture (left or right: say, right).

↳ Parameters of interest:

p_m , the proportion of these men preferring the right picture

and p_f , the proportion of these women preferring the right picture.

Data available:

x_1, \dots, x_{164}

where $x_j = \begin{cases} 1 & \text{if the } j\text{-th man interviewed preferred the right picture} \\ 0 & \text{otherwise} \end{cases}$

y_1, \dots, y_{108}

where

$y_j = \begin{cases} 1 & \text{if the } j\text{-th woman interviewed preferred the right picture} \\ 0 & \text{otherwise} \end{cases}$

Hypotheses to be tested:

↳ We need to decide them before digging into the data

H_0 [prudent option]: same preferences among men and women, the design of a single, common, campaign is enough.

H_1 [costly option]: different preferences, two separate campaigns would need to be created.

Put differently:

$$\begin{cases} H_0: p_m = p_f \\ H_1: p_m \neq p_f \end{cases}$$

Data summary: $\bar{x}_{164} = \frac{89}{164} = 54.3\%$ of the men in the sample prefer the right picture.
 $\bar{y}_{105} = \frac{54}{105} = 51.4\%$ of the women in the sample prefer the right picture.

Pooled (grouped) sample proportion: $\bar{xy}_{269} = \frac{89+54}{164+105} = 53.1\%$

the global proportion in favor of the right picture when putting men and women in the same bag

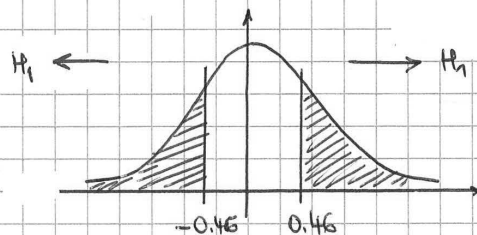
↳ Rephrasing of our question:

Are the sample average preferences 54.3% and 51.4% significantly different or not?

Test worked out:

- Test statistic: $t_{164, 105} = \frac{1}{\sqrt{1/164 + 1/105}} \frac{\bar{x}_{164} - \bar{y}_{105}}{\sqrt{\bar{xy}_{269}(1 - \bar{xy}_{269})}}$
- Behavior under H_0 : normal curve
- Behavior under H_1 : takes larger or smaller values
- Value on our data: ≈ 0.46

Picture:



$$P\text{-value} = 2 \times (1 - 0.6772) \approx 64.6\%$$

We fail to reject H_0 :

→ These data cannot exclude that men and women share similar preferences when it comes to advertising fitness machines.

→ Business conclusion: We have no ground for designing two separate campaigns, so we will pay for the design of a single common campaign.

Post-scriptum:

Of course, this exercise is a stylized version of what is done in qualitative and quantitative marketing!

Question 2:

[Not so related... but interesting! It's the simplest situation of a goodness of fit. In Chapter 7, we will review more complex such situations.]

I write: «not so related» because despite the intentionally misleading formulation, all boils down to a one-sample test: we compare to the reference proportion $p_{ref} = 51.4\%$.

Population: a single population, those who attend fitness centers

Parameter: p_0 , the proportion of women in this population

Reference value: $p_{ref} = 51.4\%$, the proportion of women in the global population

Hypotheses:

Here we have no agenda but our choice is led by mathematical constraints. By design, the tests we have studied pick

and $H_0: p_0 = p_{ref}$
 $H_1: p_0 \neq p_{ref}$

Now, note that $p_0 = p_{ref}$ corresponds to the fact that men and women go to the gym in equal proportions,

while $p_0 > p_{ref}$ corresponds to women going more often* than men to the gym

and $p_0 < p_{ref}$ is the "less often" case.

* Meaning: the proportion of men going to the gym among all men is larger than the corresponding proportion for women

Data: z_1, \dots, z_{269} where $z_j = \begin{cases} 1 & \text{if the } j\text{-th interviewee was a woman,} \\ 0 & \text{if he was a man.} \end{cases}$

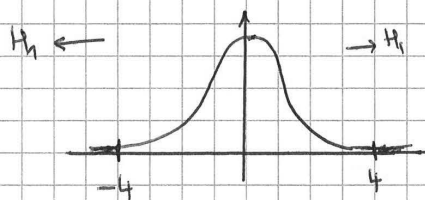
Data summary: $\bar{z}_{269} = \frac{105}{269} = 39.0\%$ of the persons in the sample were women.

↳ Is this sample proportion of 39.0% significantly different

from the 51.4% reference proportion?

Test statistic :
$$\sqrt{269} \frac{\bar{z}_{269} - 0.514}{\sqrt{0.514(1-0.514)}}$$

- H_0 behavior : normal curve
- H_1 behavior : takes larger or smaller values
- Value on the data : ≈ -4



P-value : almost null

We strongly reject H_0 ; given the sample value, we see that women are under-represented in this sample and thus** in the studied population:

These data show that men and women do not go to the gym in equal proportions : men go much more !

** assuming that our sample is representative (in particular, that the fitness centers selected are representative...)

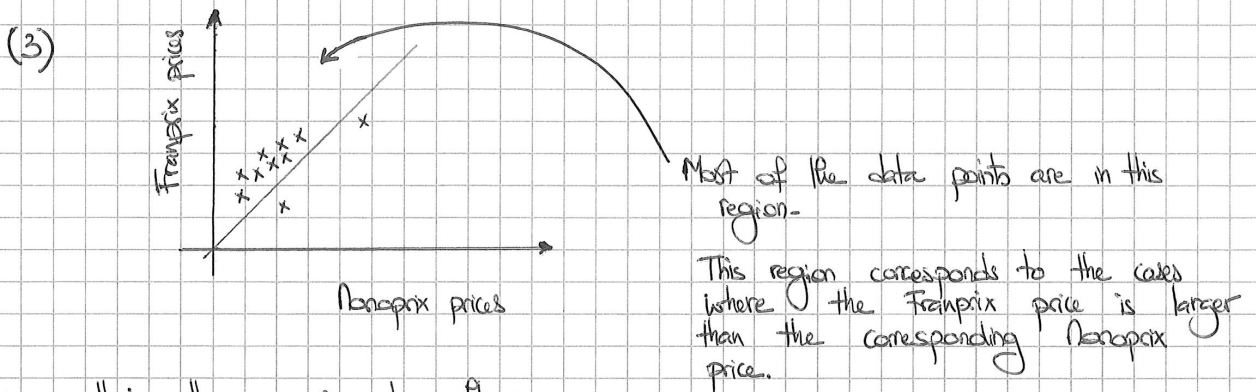
Comparing prices in two local supermarkets.

(1) We deal with paired data: for each sample unit (= each product, in this case), we made two measurements (= the prices in the two supermarkets considered).

(2) It is reasonable to think that at economic equilibrium, if both supermarkets exist and do well, then they should have the same average prices. This will form H_0 .

There is no context or background suggesting that one of them could be cheaper than the other one. Hence we will take a two-sided hypothesis H_1 .

Summary: H_0 [reasonable]: Same average prices at Franprix and Monoprix
 H_1 [no context]: Different average prices



Hence, all in all, we suspect, after data collection, that Franprix could be more expensive.

(4) Population = All products sold in (both) these supermarkets

Sample = 51 such products (picked somewhat at random, but trying to figure out a typical grocery list)

Variable of interest (variable « difference ») = The difference in prices, computed as Franprix price - Monoprix price

Statistics				
	Franprix	Monoprix	Difference	LogRatio
N Valid	51	51	51	51
Missing	0	0	0	0
Mean	2,9741	2,7667	,2075	,1214
Median	2,3200	2,1600	,2500	,17093
Std. Deviation	2,37978	2,34611	,53437	
Minimum	1,11	1,03	-2,00	
Maximum	17,63	16,74	1,07	

(how do we know that? See the « Statistics » table, line « Mean »)

Parameter of interest = The average difference in prices μ_0 , where the said average is computed over all the common products (not just the 51 of the sample)

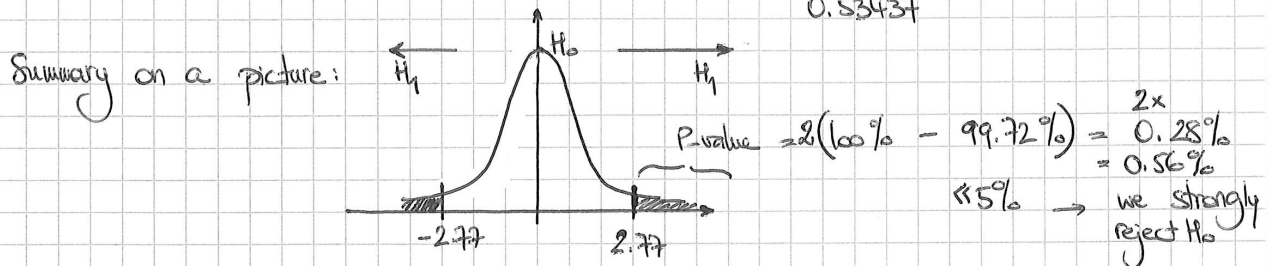
Data collected = $z_1 \dots z_{51}$ where z_j denotes the difference in prices for the j -th product.

Data summary = $\begin{cases} \bar{z}_{51} = 0.2075 \text{ €} \\ s_{z,51} = 0.53437 \text{ €} \end{cases}$
(see the "Statistics" table)

(5) * We resort to the test statistic $t_{51} = \sqrt{51} \frac{\bar{z}_{51} - 0}{s_{z,51}}$

Behavior under H_0 : normal curve
under H_1 : larger or smaller values than under H_0

Numerical value on data: $\sqrt{51} \frac{0.2075}{0.53437} = 2.77$



Statistical conclusion in plain words: These data show that one of the two supermarkets is cheaper than the other, namely (looking at the sample statistics), Nonpax.

Business conclusion: Go to Nonpax: it's cheaper but also fancier!
(At least if you can: Franprix' opening hours are probably wider...)

* Comparison to SPSS output: we read the same numerical value for our test statistic (2.772) but a slightly different P-value: 0.8%

The reason behind that is that we computed the P-value based on the standard normal distribution while SPSS uses Student's distribution with 50 degrees of freedom.

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

(6)

Our well-chosen figure is given by a confidence interval: an underestimate of the difference in the average prices.

With high confidence, this difference μ_0 is larger than:

$$\bar{z}_{51} - 1.65 \frac{s_{\bar{z}_{51}}}{\sqrt{51}} = 0.2075 - 1.65 \frac{0.53437}{\sqrt{51}} = 0.084035 \geq 0.08 \text{ €}$$

↳ These data show that Monoprix is cheaper than Franprix (by at least 8 cents per product on average).

... Except that "8 cents per product" is not easily understood, hence Question 7 of this exercise (beware! it's a difficult question). ...

Question 7: To be skipped in class (it's a difficult question).

Let us first look at what «LogRatio» corresponds to: we read

Vodka:	17.63	16.74	0.89	0.05
Chocolat:	2.73	2.47	0.26	0.10

↑
Fransprix
prices

↑
Monoprix
prices

↑
Arithmetic
differences in
prices; eg:

$$\ln \left(\frac{\text{Fremprix price}}{\text{Nemprix price}} \right)$$

e.g.,

$$0.89 = 17.63 - 16.74$$

$$0.26' = 2.73 - 2.47$$

$$Q_{OS} = \ln \left(\frac{17.63}{16.74} \right)$$

$$0.10 = \ln\left(\frac{2.73}{2.47}\right)$$

We therefore study the averages of $z_j = \ln(\bar{F}_j / M_j)$

where F_j is the Fraprix price of product j

p_j is the Nonoprix price of product j

Put differently:

$$F_j = M_j \times \exp(x_j)$$

Thus :

The $\frac{\exp(x_j) - 1}{\text{product}}$ is how much (in %) more expensive or cheaper at Temprix .

Now, what is the associated population parameter?

We denote by γ_0 the arithmetic average over all products of $\ln(F/M)$

Then $e^{\bar{f}_0}$ is the geometric average over all products of F/M

The right notion of average variation in % is $e^{\sigma} - 1$

We test H_0 : bike prices on average, $\mu_0 = 0$

H_1 : one supermarket is more expensive on average, $\neq 0$

The second line of the "One sample T-test table" indicates an (almost) null P value. It is obtained by using the test

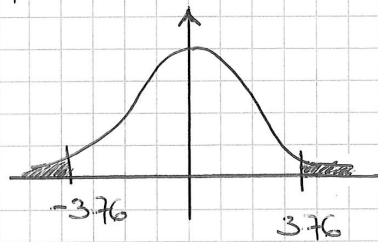
Statistic

$$t_{51} = \sqrt{51} \frac{\bar{x}_{51}}{s_{x,51}}$$

with value on data (LogRatio) : $\sqrt{51} \frac{0.0901}{0.17093} = 3.76$

with H_0 behavior: normal-curve

H_1 behavior: smaller or larger values



$$P\text{-value} = 2 \times (1 - 0.999892) \approx 0.02\%$$

is almost null

We reject H_0 and conclude to a significant difference in prices.

We also see that γ_0 is estimated by:

- the sample mean: $\bar{x}_{51} = 0.0901$

- the confidence interval: $[0.0420, 0.1382]$

So that the average variation $e^{\gamma_0} - 1$ is estimated by:

- the single value $\exp(0.0901) - 1 \approx 9.4\%$

- the interval $[\exp(0.0420) - 1, \exp(0.1382) - 1] = [4.3\%, 14.9\%]$

Hence our general-public conclusion:

The data show a significant difference in average prices between the two supermarkets: the Franprix prices are higher than the Monoprix prices, by about 9.4%. (To be more precise, we can tell that with high confidence, this average difference in prices lies between +4.3% and +14.9%.)

Exercise "Gender pay gap, revisited"

We consider again the same notation as in Chapter 3.

SPSS tests $H_0: \mu_o^{\sigma} = \mu_o^{\text{f}}$ vs. $H_1: \mu_o^{\sigma} \neq \mu_o^{\text{f}}$
and gets a P-value of 0.1% as indicated below:

Group Statistics				
Gender	N	Mean	Std. Deviation	Std. Error Mean
Monthly net salary Men	179	3431,46	3895,437	291,159
Monthly net salary 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 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

(2) this we read the second line of the table

(1) 0.4% < 5%
↳ variances of the salaries are different by gender

(4) for later purpose, this is the numerical value of the test statistic

(3) in this second line the most important quantity is this P-value: 0.1%

But

in our case, we have some heavy glass-ceiling-effect assumption that we would actually like to prove/bring to light and hence we would rather consider

H_0 : Same average salaries between men and women, $\mu_o^{\sigma} = \mu_o^{\text{f}}$

H_1 [what we want to prove]: Women earn less on average, $\mu_o^{\sigma} > \mu_o^{\text{f}}$

SPSS measures the differences in the men-women direction (see the positive values for "Mean Difference" in the table).

The test statistic at stake here is of the form $\frac{\bar{x}_{179} - \bar{y}_{147}}{\text{normalization}}$, where the x_j are the men data and the y_j the women data.

Its H_0 behavior is close to a normal curve, its H_1 behavior is that it takes larger values than under H_0 (see: $\mu_o^{\sigma} > \mu_o^{\text{f}}$ under H_1 and $\bar{x}_{179} \approx \mu_o^{\sigma}$ while $\bar{y}_{147} \approx \mu_o^{\text{f}}$).

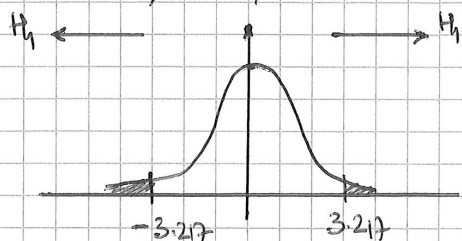
Its numerical value on the data is 3.217 (see the SPSS output).

Thus, we have the following summaries as far as the computation of P-values is concerned:

What SPSS does

$$H_0: \mu_{\text{M}} = \mu_{\text{F}}$$

$$H_1: \mu_{\text{M}} \neq \mu_{\text{F}}$$

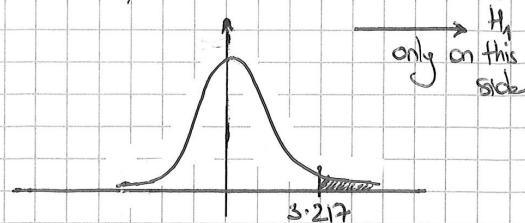


P-value = 0.1%
(as read in the table)

What we do

$$H_0: \mu_{\text{M}} = \mu_{\text{F}}$$

$$H_1: \mu_{\text{M}} > \mu_{\text{F}}$$



Thus P-value here
equals $0.1\% / 2$
 $= 0.05\%$

↳ Strong rejection of H_0

These data strongly show that women earn less than men on average.

But (see Chapter 3) this is not, per se, a proof of the existence of a gender pay gap, as the effect of gender here is not isolated out from the other factors (education level, full-time vs. part-time work, etc.).

Exercise "Alcohol consumption during the Pows at HEC Paris"

(1) We actually had an agenda:

In that good old time, my colleague and I were young professors (we just had joined in September 2017) and did not dare insist that the students be present — of course, that dramatically changed meanwhile as we got more experience and more self-confidence, not mentioning the fact that over the years, we could clearly bring to light a correlation between absence rate and bad grades. Anyway.

In that good old time, our assumption was that:

- At 8am we only get sober students who went to bed reasonably early; thus, we get many absentees (all those who drank too much and can't get up) and the students actually attending the classes behaved the night before;
- At 10am, you could have slept longer and thus could be attending while partying and binge drinking the night before; the absence rate was typically much lower.

Hence, denoting by μ_0^{8am} and μ_0^{10am} the respective average alcohol consumptions over the weeks for each group, we had in mind to prove that $\mu_0^{10am} > \mu_0^{8am}$, and hence wanted to test:

$$H_0: \mu_0^{10am} = \mu_0^{8am} \\ H_1 \text{ [what we want to prove]}: \mu_0^{10am} > \mu_0^{8am}$$

⚠ Note that SPSS uses in its outputs $H_0: \mu_0^{10am} = \mu_0^{8am}$ vs. $H_1: \mu_0^{10am} \neq \mu_0^{8am}$

(2) The second column reports the number of alcohol glasses drunk (and in that good old time, there was not only beer but also stronger alcohols like vodka).

Thus, values like 18.7 (!) — or even 10 — are highly unplausible.

⚠ We were surprised that students indicate non-integer values for the number of glasses drunk ... They of course were right (!) doing so!

(5) Two independent samples:

We have two different populations (the 8am students that are present and the 10am students that are present) and we measure the same quantity (number of glasses drunk) for each sample member.

Fyi: these were 31 and 51

(4) In the second treatment of the data, we eliminated the numbers of glasses that were > 20 (!). This only suppressed 2 data points, in the 10am group:

Sample sizes	Before suppression	After suppression
8am sample	23	23
10am sample	31	29

SPSS measures the differences as: 8am - 10am
The test statistic thus:

- behaves approximately as a normal curve under H_0
 - takes smaller values under H_1 than under H_0
- and it has a -1.7 value on the data.

Thus the P-value output by SPSS should be divided by 2, as the summary pictures on the next page.

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

Sample data are such that the 10am group drank more than the 8am group; but is this gap significant?

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
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	-7.2084 .6311

We read the second line

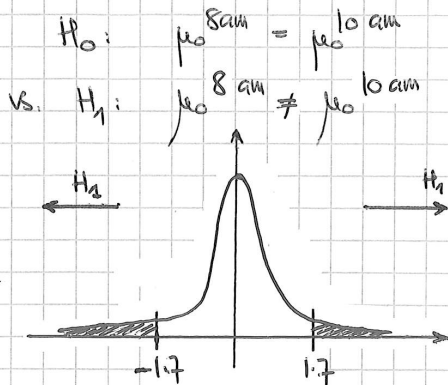
Case of RAW DATA

Variances are thus different

P-value equals 9.8%
for $H_0: \mu_{8am} = \mu_{10am}$
vs. $H_1: \mu_{8am} \neq \mu_{10am}$
(2-sided test with SPSS)

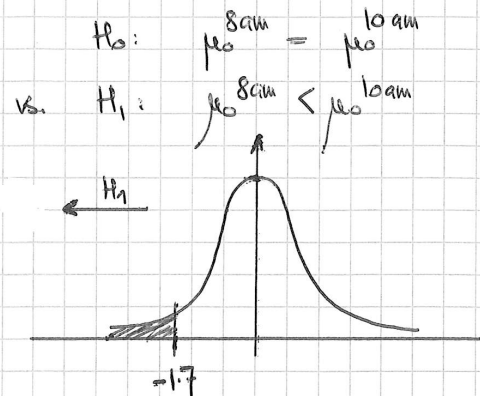
Numerical value of the test statistic

What SPSS does



P-value (read in the SPSS output):
9.8%

What we do



Thus: P-value of 4.9%
in this case

Case of (partially) CLEANED DATA:

Same methodology, we get a P-value of $36\% / 2 = 18\%$

Group Statistics				
Group	N	Mean	Std. Deviation	Std. Error Mean
8am	23	4.448	3.0598	.6380
10am	29	5.441	4.3697	.8114

On the cleaned sample data also did the 10 am group drink more than the 8 am group.

		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

We read the first line

P-value 13.8%
this variances are not significantly different

P-value equals 36% for $H_0: \mu_{8am} = \mu_{10am}$
vs. $H_1: \mu_{8am} \neq \mu_{10am}$

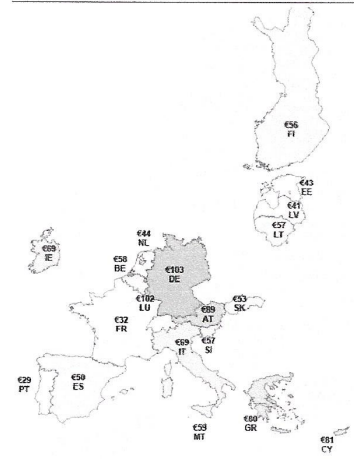
(5) We of course trust more the cleaned data (how can I one drink more than 20 glasses and not end up at hospital? and be present in class the day after?). But even the raw data was not very conclusive (we had a borderline P-value of 4.9%).

All in all, we fail to reject H_0 , which was a surprise to us: These data do not suggest a significant higher alcohol consumption for the students present at 10 am compared to the students present at 8 am. \rightarrow 8 am students cannot be declared more boring!

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.

Chart 20
Average amount of cash in wallet



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.

! We use independent samples but we are not testing for independence, we are testing equality of (population) means.

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.	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
Cash in wallet	Equal variances assumed	136,018	,000	1881	,000	-13,9438750	1,209231121	Lower	Upper
	Equal variances not assumed								
				1787,518	P-value: ,000	-13,9438750	1,166962653	-16,2326295	-11,6551205

We read this line →

Shows that variances are different

2-sided test

SPSS tests equality in a 2-sided way

↑ effect size to be read here

- □ 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 ?

H_0 : Germans and Austrians carry the same average amount of cash in their wallets
 H_1 : Germans and Austrians carry different such average amounts

P-value almost null: we (strongly) reject H_0

- □ Write a statistical conclusion quantifying the effect size. (Again, do not forget *important adjectives*.)

Data show that Germans and Austrians carry different average amounts of cash in their wallets, namely (looking at sample data) Germans carry more cash on average.

Effect size: With high confidence, they carry at least 11€ and at most 17€ more than Austrians.

Alternative numbers: - at least 11.65€ and at most 16.24€ more
- at least $13.94 - 1.65 \times 1.167 \geq 12$ € more

! You didn't need to provide the parameters of interest nor the mathematical formulations of H_0, H_1

The parameters of interest are, by the way:
 P_{company} = the success rate with the company's pitch
 P_{salesman} = the success rate with the salesman's pitch

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.

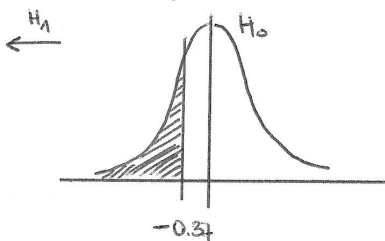
H_0 [prudent]: the company's pitch is as efficient or more efficient than his own pitch (ie, $P_{\text{company}} \geq P_{\text{salesman}}$)

H_1 [risky]: His own pitch is (strictly) superior (ie, $P_{\text{company}} < P_{\text{salesman}}$)

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.

Note: we are measuring the difference in the day 1 - day 2



$$t = \frac{64/534 - 67/526}{\sqrt{(1/534 + 1/526) \cdot 0.124 \cdot (1 - 0.124)}} \approx -0.37$$

as the pooled estimate of success rate equals

$$\frac{64 + 67}{534 + 526} \approx 12.4\%$$

$$\begin{aligned} \text{P-value} &= 100\% - 64.43\% \\ &= 35.57\% \gg 5\% \end{aligned}$$

→ We fail to reject H_0

☐ 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:

- A These data show that
- B These data suggest that
- C These data do not bring to light that
- D These data cannot exclude that

- 1 both pitches work equally well
- 2 the company's pitch is more effective
- 3 the salesman's personal pitch is more effective
- 4 the salesman's personal pitch is less effective

Equivalently nice answers:

→ 2
→ 4

☐ Draw a strategic conclusion: which pitch should he use the next day?

As he is prudent he should thus stick to the company's pitch for now.

Answer → 1 is not as nice though I counted it as correct despite all.

Alternative correct answer: He could continue alternating the two pitches till he can rank them; for now they seem to have a somewhat comparable performance.

Alternative incorrect statements: (too affirmative!)
 - Both pitches have the same performance so he can pick whichever he prefers
 - His pitch is slightly better so he should use it

Exact same statement
as in a previous quiz!

But this time, with the better methodology
we will be able to conclude to an impact.

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.



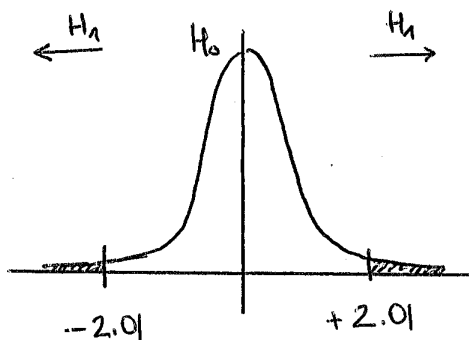
We want to illustrate this fact by performing the following experiment. We consider two similar stores (e.g., two Hollister stores) and ask the sellers of the first store to avoid any physical contact with the customers, while the ones of the second store are asked to lightly touch the customers' arm. 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.



$$\text{Pooled estimate: } \bar{p}_{240} = \frac{12 + 23}{120 + 120} \approx 14.6\%$$

$$t = \frac{12/120 - 23/120}{\sqrt{(1/120 + 1/120) \cdot 0.146(1 - 0.146)}} = -2.01$$

$$P\text{-value} = 2 \times (100\% - 97.78\%) = 4.44\% < 5\%$$

↳ Reject H_0

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]

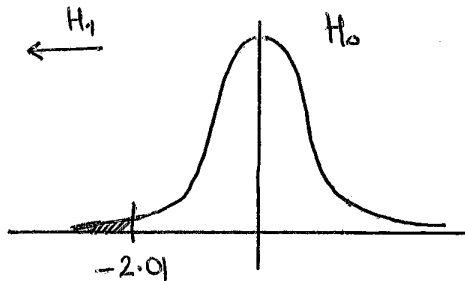
- ~~1. purchase rates are different with and without a light contact~~
- ~~2. purchase rates are similar with and without a light contact~~
- ☒ 3. the purchase rate increases with a light contact
- ~~4. the purchase rate decreases with a light contact~~

So, we conclude that purchase rates are significantly different, namely, looking at the sample data, that the purchase rate with a light contact is larger than the one without a 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.



[As t measures performance in the without contact - with a contact direction]

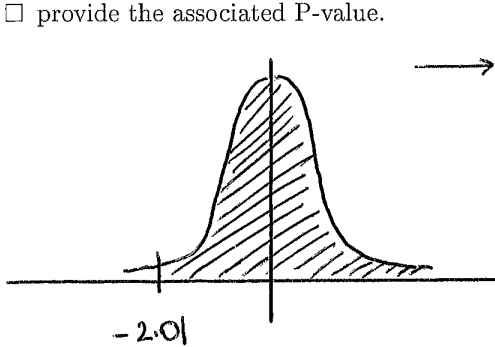
$$P\text{-value} = 2.22\%$$

(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.



$$P\text{-value} = 97.78\% \gg 5\%$$

(Not a surprise: this test is equivalent to $H_0 : p_0 \leq q_0$ vs. $H_1 : p_0 > q_0$ and H_0 is consistent with the data do not contradict H_0)

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:

- ☐ First conclusion possible Letter: **A** Number: **2**
 - ☐ Second conclusion possible Letter: **D** Number: **4**
- Cannot exclude H_0 Fails to prove H_1

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 **1**
- ☐ Shopkeeper: Pair **2**

(as is neutral)
(has an agenda: wants to determine whether to implement a light-touch policy for the sellers)

} same justifications as in the previous quiz

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. Its 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 μ_0^1 and μ_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

variances are significantly different

- ☐ 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 ?

$$\begin{cases} H_0: \mu_0^1 = \mu_0^2 \\ H_1: \mu_0^1 \neq \mu_0^2 \end{cases}$$

P-value = 15% → we fail to reject 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?

Conclusion: - These data cannot exclude that the two gifts would lead to the same average amount of orders.

[or] - These data do not bring to light any superiority of one gift over the other as far as the average amounts of orders are concerned.

Which gift should be chosen?

- The treatment above (consisting of comparing the means) does not indicate *per se* which gift to choose.
- Other dimensions / criteria may help to pick a gift (their prices if different; also, note that gift #1 is safer in some sense as it leads to a smaller variance in the obtained amounts ordered).

These tests complement the one performed on the previous page. We test whether each gift is effective in leading to larger amounts of orders, by comparing to the reference value 165. They look at the same data but with a different perspective.

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					
	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

We define $\mu_{ref} = 165$

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					
	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.)

We answer for Group 1 only, answers are similar for Group 2.

Tested: $H_0: \mu_0^1 = \mu_{ref}$
 $H_1: \mu_0^1 \neq \mu_{ref}$

What we would want to test: $H_0: \mu_0^1 = \mu_{ref}$
 $H_1: \mu_0^1 > \mu_{ref}$

- ☐ 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?

In these cases (cf. positive numerical values of the test statistic), they are given by the P-values read above divided by 2, that is:

Group 1: $50.7\% / 2 = 25.35\% \rightarrow$ we fail to reject H_0 / Group 2: $2.5\% / 2 = 1.5\% \rightarrow$ we reject H_0

- ☐ 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?

Superiority? Yes and no:

- Yes: The real point of this study is not to compare gifts but pick one; we could prove that the second gift was effective but could not prove this (nor exclude this) for the first gift; however, if we need to pick one gift, then let's pick the second one, it's a safe option.

- No: we did not prove that the second gift is superior to the first gift: we proved $\mu_0^2 > \mu_{ref}$ but still don't know whether $\mu_0^1 > \mu_{ref}$ or $\mu_0^1 = \mu_{ref}$; it could still well be that $\mu_0^1 = \mu_0^2$, we did not exclude nor prove this yet.

Contradiction / Confirmation? None of them: merely yet another non-conclusive statement on the equality of μ_0^1 and μ_0^2

Why do you think this study was conducted in the first place? Because there was some background that (surprisingly enough) children whose parents work in lead-related industries suffer from lead poisoning as well...

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.)

Paired samples: for each location, two measurements are made (one on a control child and one on a child whose parents work in a lead-related industry)

Parameter of interest: the average difference Δ in lead levels in blood between all children whose parents work in lead-related industries and all other children.

- ☐ State relevant hypotheses to be tested. Explain with few words in brackets why you picked them.

H_0 [reasonable] $\Delta = 0$,

only parents are exposed to lead and they do not communicate that exposure to their children

H_1 [what we want to prove]

$\Delta > 0$,

parental exposure to lead at work has an effect on their children's own blood lead level

H_0 is "reasonable" because why/how could children be contaminated through their parents? Sample data is summarized here: It seems odd at first.

Descriptive Statistics

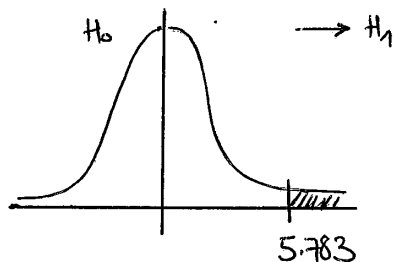
	N	Minimum	Maximum	Mean	Std. Deviation
Exposed	33	10	73	31,85	14,407
Control	33	7	25	15,88	4,540
Difference	33	-9	60	15,97	15,864
Valid N (listwise)	33				

- ☐ 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.

$$t_{33} = \frac{15.97}{15.864} \sqrt{33} \approx 5.783$$



P-value is very small, actually almost null (value 5.783 is out of the table provided, and remember from other exercises done in class that P-value are already almost null when the test statistic is larger than 4)

- ☐ Write a statistical conclusion (in plain words, that are understandable by a layman).

These data strongly show that children whose parents work in lead-related industries suffer from this and get higher lead levels in blood (on average).

- ☐ 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?

keyword "Paired"

→ [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

keyword "Paired"

→ [Paired] Samples Test

		Paired Differences					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	,000	

Yes, these values are the same as the ones we calculated {
(the P-value here is the double of our P-value as SPSS conducts a two-sided test, but as both are almost null, we cannot read any difference for one...)

↑ Numerical value of the test statistic

↑ P-value

Note: The study then needs to quantify Δ to see how detrimental the job of the parents is to the health of the children.
Here, we get the confidence interval 10.3 - 21.6 mg/dl on the average difference in lead levels in blood.

χ^2 -tests of independence and of goodness of fit

Elementary exercises	page 104
Advanced exercises	
Reading habits per socio-professional category	page 107
The sinking of the Titanic	page 109
Hair color by gender	page 110
Call centers	page 112
“Junior entreprise” campaign at HEC Paris	page 114
Mendel's experiment	page 115
Alcohol consumption during the POWs at HEC, second version	page 117
M&M colors	page 118
Satisfaction survey	page 120
Births by month	page 122

Elementary exercises.

Exercise 1 (χ^2 -test of independence)

(1) This expected count equals

$$\begin{array}{ccc} 29 & \times & 69.0\% \approx 20.0 \\ \uparrow & & \uparrow \\ \text{total number} & & \text{overall} \\ \text{of customers for} & & \text{satisfaction} \\ \text{sale representative \#1} & & \text{rate } (= 80/116, \\ & & \text{where 116 customers} \\ & & \text{in total, including 80 satisfied ones}) \end{array}$$

(2) We check whether the validity conditions are met:

- Total sample size $n = 116 \geq 30$
- No expected count smaller than 5 (see table note)

(3) We read a P-value of 56.3% $\gg 5\%$, we fail to reject H_0 .

(4) That is, no sale representative stands out, their satisfaction rates (respectively: 72.4%, 76.7%, 65.5% and 66.7%) are NOT significantly different... contrary to what we might have had the intuition of!

Exercise 2.

Instructor 1

(1) $50 \times 15\% = 7.5$

\uparrow \uparrow
 total target
 number of proportion
 grads given of A

(2) Yes, the validity conditions are met in both cases:

- Total sample size $n = 50 \geq 30$ and $n = 49 \geq 30$
- No expected count smaller than 5 (see table note)

(3) P-value = 48.4% \gg 5%

We fail to reject H_0

(4) No significant discrepancy to the target distribution can be found, thus we could declare the grading performed consistent with the school's policy.

Instructor 2

$49 \times 15\% \approx 7.4$

(similar explanation)

P-value = 4.3% $<$ 5%

We should reject H_0
(though it is a borderline value)

The grading performed is not consistent with the school's policy. It should be revised.

The biggest issue (compare observed counts to expected counts) is that there are too many C's and too few D-E-F. An easy

patch would be to convert some C's into D's (2 of them might be enough? → to know for sure we would then re-run the test).

Exercise 3:

χ^2 -test version of "Answering students' complaints"

- (1) Hypotheses
- H_0 [homogeneity] Same performance profiles in the two groups / The distributions of grades obtained in the two groups are the same
- H_1 [lack of homogeneity] Different performance profiles / Different grade distributions

- (2) 1pm / 18-20 Observed count: 8 vs.
- Expected count: $84 \approx \frac{16}{76} \times 40$
- $\underbrace{\frac{16}{76}}_{\text{global proportion of 18-20}} \times \underbrace{40}_{\text{size of 1pm group}}$

- (3) Yes because sample size $n=76 \geq 30$ and all expected counts are ≥ 5 (see table note)

- (4) P-value = 62.0% \gg 5% \rightarrow Fail to reject H_0

- (5) The performance profiles are not significantly different between the two groups, at least for this quiz.
- Some actions to take as for the other version of this exercise: Instructors should repeat that test year after year and continue collecting evidence. For now, a single quiz statement is fine.

Reading habits per socio-professional category

(1) Test: χ^2 test of homogeneity / independence

H_0 [homogeneity / independence]:

Reading habits do not vary by the socio-professional category

or:

The socio-professional category has no influence on the reading habits

vs.

H_1 [lack of homogeneity / some dependence]:

Reading habits vary by the socio-professional category
(depend on)

or:

The socio-professional category has an influence on the reading habits

(2) Output 1/2: The test is NOT validly worked out because one category has an expected count smaller than 5, namely: Farmers / 1 book per month, or more

It would be awkward to merge the farmers with any other socio-professional category, so we simply drop them.

↳ The consequence will be that our study will be mute about the reading habits of farmers.

This suppression of farmers leads to the second output:

Output 2/2:

This second test is now validly worked out as:

- the sample size $n = 1,484$ is ≥ 30
- all expected counts are larger than 5 (see the table note)

(3) Output 1/2: The reading of the P-value is not applicable, as the test is not validly applied.
We cannot exploit this output

Output 2/2:

We read an almost null P-value, we strongly reject H_0 .

(4) The reading habits differ among the various socio-professional categories.

↳ We need to explain why categories read more than average

(To do so, we compare expected and observed counts) as the average less than average

Craftsmen & co: as the average? or just slightly less than average
 Company executives & co: (much) more than average
 Associate/intermediate professions: more than average
 Employees: almost exactly as average
 Workers: (much, much) less than average
 Other non-economically active: (slightly) more than average
 Retired: mixed picture but overall they read less than average (see the 2nd and 3rd columns)
 because it's a mixed category, with various backgrounds!

eg, long novels, essays, etc.

(5) What should a publisher do based on this data?

Already a competitive market {
 - Continue publishing books for those who love them of course!
 - Identify a segment where a strong progression in sales could be made: eg, publish plain books with simple but thoughtful stories, targeted to workers and retired people. Start with a qualitative panel to understand why they don't read (yet) and what they would like to read (if any) and why they can't find it yet.

That's an empty slot to fill!

(6) Company executives & co. / 1 book per month, or more

$$\hookrightarrow 129 \times \frac{233}{1,484} \approx 20.3$$

\uparrow \uparrow
 size of the company executives category in the sample global proportion of people reading at least 1 book per month

Exercise "The sinking of the Titanic"

(1) Test: χ^2 test of independence

H_0 [independence / homogeneity]: Same { death / survival } rate among all categories of passengers

that is: all Titanic passengers were equal before death

vs. H_1 [dependence / lack of homogeneity]: Some categories of passengers fared better than others

(2) The test is validly worked out:

- Total sample size $n = 2201 \geq 30$
- No expected count smaller than 5 (see table note)

(3) We read an almost null P-value

(4) H_0 is strongly rejected given these data

↳ These data show that some categories of passengers fared better than others, i.e., that passengers were not equal before death:

compare observed counts to expected counts → Namely, first-class and to a lesser extent second-class passengers enjoyed a survival rate larger than the global survival rate. Crew members and third-class passengers suffered from a larger death rate than the global death rate.

Why?

1. Because the crew members sacrificed themselves
2. Because it was not "women and kids first" to get a seat in the rescue boats but rather whoever came first, and as the rescue boats are close to the deck, where first-class passengers are hosted (while third-class passengers travel in the bowels of the ship...), "whoever comes first" was given mostly by these upper-class / upper-ship passengers.

(5) N/A

(6) Expected count for first-class survivors:

$$\frac{325}{\text{total number of 1st-class passengers}} \times \frac{711}{2201} \approx 105.0$$

global survival rate

Hair color by gender.

1. Test: χ^2 of independence

H_0 [independence / homogeneity]: Same hair color distribution, regardless of the gender

(or: Hair color does not depend on the gender

or: The gender does not influence the hair color.)

H_1 [dependence / lack of homogeneity]: Hair color distributions vary by gender.

2. The test is validly worked out:

- Total sample size $n = 3883 \geq 30$

- No expected count less than 5 (see table note)

3. We read a P-value of 3.3%

4. P-value $< 5\%$ \rightarrow We reject H_0

These data show that some hair colors are more frequent among women than men, or vice versa: which ones? We should indicate them but it is not so easy to read. See the subsidiary question below.

5. N/A (no business conclusion; at least, Fisher did not conduct this study with a business implication in mind!)

6. E.g., expected count for red-haired women:

$$99.2 = \frac{216}{3883} \times 1783 \quad (= \text{global proportion of red-haired persons} \times \text{total number of women})$$

To be compared to the observed value for this category: 97

8 VERY DIFFICULT

7. Subsidiary question: which hair color drives the results?

The value of the test statistic is computed as:

$$D = \sum \frac{(N_{obs} - N_{exp})^2}{N_{exp}}$$

$$= \frac{(592 - 614.4)^2}{614.4} + \frac{(544 - 521.6)^2}{521.6} + \dots + \frac{(14 - 23.0)^2}{23.0}$$

\uparrow for blond/men equals ≈ 0.82
 \uparrow for blond women equals ≈ 0.96
 \uparrow for jet-black/women equals ≈ 3.52

We have, for the values of $\frac{(N_{obs} - N_{exp})^2}{N_{exp}}$:

	Men	Women
Blond	0.82	0.96
Red	0.04	0.05
Chestnut	0.68	0.80
Brown	0.30	0.36
Jet-black	3.00	3.52

⏟

Sum of all these numbers ≈ 10.467 as we read in the SPSS output

⇒ Two cells drive the results: jet-black/men & jet-black women.

Actually, without the jet-black color, there would be no significant difference in the distribution of the other hair colors!

Exercise "Call centers"

Ante-scriptum:

We consider here a whole family of suitable distributions for the waiting times, of the form

a fraction x of the customers waits ≤ 2 min
 a fraction y $2 \text{ min} < \dots \leq 5 \text{ min}$
 where $x \geq 50\%$
 and $x+y \geq 90\%$

Many (x, y) pairs are possible, eg: $x = 60\%, y = 30\%$
 $x = 70\%, y = 25\%$
 etc.

The one when people wait the longest is: $x = 50\%, y = 40\%$

but don't forget that it's just the least favorable distribution among all acceptable distributions.
 this will be our reference distribution (= limit distribution)

After this lengthy introduction, we answer the questions in order:

(1) Test: χ^2 test of goodness of fit

H_0 [conformity]:

The waiting times of the customers follow the limit distribution:

50% of them are ≤ 2 min
 40% are $2 \text{ min} < \dots \leq 5 \text{ min}$
 10% are $> 5 \text{ min}$

for once, the hypotheses are stated in a technical though accurate way

H_1 [non-conformity]:

The waiting times follow some other distribution (more favorable or less favorable: both alternatives have to be considered)

→ Softer statements would be: (but not accurate enough, as we shall see below)

H_0 : Customers wait as prescribed

H_1 : Customers do not (they wait less or more)

(2) The two tests are validly worked out:

- Respective sample sizes of 111 and 104 phone calls, which are ≥ 30 as required
- No expected counts less than 5 (see the two table notes)

(3)

P values :

Call Center #1:

77.4%,

we fail to reject H_0

Call Center #2:

2.2%,

we reject H_0

(4) & (5)

Statistical + business conclusions:

Call Center #1:

We cannot exclude that data abides by the prescribed distribution for the waiting times. The head of the Internet-access provider cannot raise any claim against the performance of this call center.

Call Center #2:

Here we reject H_0 but if we dig into the data, we see that it's because customers waited less than what is prescribed by the limit distribution: there are more customers than expected in the two ≤ 5 min categories and fewer in the 75 min category. We hence reach the same conclusion as for the other center: no claim can be raised.

(6)

E.g., expected counts in the 75 min category:

Call center #1:

$$111 \times 10\% = 11.1$$

(vs. observed count of 9)

Call center #2:

$$104 \times 10\% = 10.4$$

(vs. observed count of 2)

↑
total
sample
size

↑
frequency of
this category
under the
prescribed distribution.

"Junior enterprise" campaign at HEC Paris.

(1) Test: χ^2 -test of goodness of fit

H_0 [conformity]:

The sample is representative, i.e. the distribution of the 200 sample members by gender and program matches the overall such distribution (over the entire 2,983 students)

H_1 [non-conformity]:

The sample is not representative, its distribution by gender and program differs from the overall such distribution.

(2) Raw data:

The test is NOT validly worked out because some categories have expected counts smaller than 5, namely:

PhD/man — 3.0
PhD/woman — 2.6

An easy way out is to merge these two categories into a greater PhD/man or woman category, which will be associated with an expected count of 5.6. This leads to the second application of the test principle:

After some treatment:

This second test is now validly worked out as:

- the sample size $n=200$ is ≥ 30
- all expected counts are larger than 5 (see the table note)

(3) Raw data:

The reading of the P-value is not applicable, as the test is not validly applied. We cannot exploit this output.

After some treatment:

P-value of 66.3%, we fail to reject H_0 .

(4) The sample does not exhibit any obvious/significant bias as to its gender/program distribution.

(5) This Junior Enterprise list should point out in its study that it gathered a quality sample and thus, that the results of its study about cultural activities are highly reliable.

(6) Expected count for MIM/Woman: $200 \times 31.3\% = 62.6$
↑ ↑
sample size overall frequency

Mendel's experiment.

1. Test: χ^2 of goodness of fit

H_0 [conformity]: Random transmission of alleles to children, resulting in a 2nd-generation distribution of phenotypes equal to

$9/16$,	$3/16$,	$3/16$,	$1/16$
Y+R	G+R	Y+W	G+W

H_1 [non-conformity]: Some other, less or non-random, transmission resulting in another 2nd-generation distribution of phenotypes

2. The test is validly worked out:

- Total sample size $n = 556 \geq 30$
- No expected count less than 5 (see table note)

3. We read a P-value of 92.5%, much larger than 5%.

4. We fail to reject H_0 . These data are compatible with the theory of random transmission.

(BTW: As this theory of random transmission was deemed compatible with the many experiments Mendel performed, it was finally, and only then, held true.)

5. In his case: perform many other experiments of the same kind (but on different species, etc.) and publish the results, so that other scientists can read, evaluate and validate them (scientific results are always and have always been peer-reviewed before disclosed to the general public).

6. E.g., the expected count for Green+Round:

$$104.3 = 556 \times \frac{3}{16}$$

↑
sample size

↑
frequency for this class
given by the reference distribution

To be compared to the observed value 108.

7. Subsidiary (and very difficult) question:

Fisher performed ex-post (several decades later) better statistical treatments of Mendel's data (the χ^2 tests did not exist in his time!) and systematically obtained P-values larger than 90%, ... which is implausible!

Indeed, if H_0 is true, then the χ^2 test statistic follows some χ^2 distribution and the P-value is uniformly distributed between 0% and 100% (ie, Mendel should only have got P-values larger than 50% in a fraction 50% of his experiments, and in only 10% of the experiments should he have got P-values larger than 90%....).

What happened? Mendel cheated and always multiplied his sample size by 10 or 20, adjusting the observed counts at random.

E.g., he would cultivate 56 pea plants, get the observed counts 31 - 11 - 10 - 3 and make up the numbers we read in the table:

556 plants (instead of 56)

315 - 108 - 101 - 32 as observed counts (instead of 31 - 11 - 10 - 3)

But it was for a good cause! No one was ready to believe in a random transmission, he needed convincing data...

⚠ This exercise is not at all about participation rates vs. drinking behaviors...

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 ^a	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.

Some mentioned "drinking patterns" or "drinking profiles", which are very good answers as well.

- What do we compare here, given that we are not comparing average numbers of glasses? Carefully state the corresponding hypotheses.

We compare distributions of numbers of glasses (to do so, we group data in 3 categories).

H_0 [homogeneity]: Same drinking behavior among 8am and 10am students present in class.

vs. H_1 [lack of homogeneity] Different drinking behaviors between 8am and 10am students present in class.

- Which is the complete name of the test worked out here? Can the outcome of the test be validly exploited? Explain.

χ^2 test of homogeneity / independence

Yes, validly applied: total sample size $n = 52 \geq 30$ & no expected count less than 5 (see table note)

- 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).

P-value 31.1% → we fail to reject H_0 :

These data cannot exclude that students present in class at 8am or at 10am have a similar drinking behavior: our statistical analysis does not

- What calculations led to the expected count of 6.6 in the top-right cell? To which observed value should it be compared?

expected count
 $6.6 = \frac{23}{52} \times 15$
 ↑ ↑
 # 8am students global proportion of 6+ glasses

vs. observed count 5

bring to light any significant difference in the drinking behaviors.

in short: averages → distributions

Note: "Data are compatible with the Tennessee factory distribution" is equivalent to "We cannot exclude that data are distributed according to the Tennessee factory distribution"

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.

He tested in this order:

1. [Uniform distribution of colors]

2. M&Ms color distribution, c. 2008



3.1. M&Ms color distribution, c. 2017



3.2.



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!).

The data collected show that the M&Ms currently available at SAS are not uniformly distributed in colors, nor do they follow the color distribution used in 2008, nor the one currently output by the NJ factory, but the sample data are compatible with being produced by the Tennessee factory.

[Since the candies have to come from somewhere, we deduce that they must come from the Tennessee factory]

Shorter alternative sentence: Data are incompatible with a uniform distribution of colors, with the 2008 distribution and the NJ factory but they are Tennessee factory distribution -

Note:

- All χ^2 -tests of goodness of fit are validly applied, as:
- sample size equals $n = 712 > 30$
 - all expected counts are > 5 (they are actually all even > 88)

Uniform color distribution

2017 New Jersey factory distribution

cf.: all expected counts are equal

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)

P-value 1,4% $< 5\%$
(Strong) rejection of H_0

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)

P-value almost null
(Strong) rejection of H_0

2008 website-posted distribution

2017 Tennessee factory distribution

$= 712 \times 24\%$

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)

P-value 3,8% $< 5\%$
Rejection of H_0

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)

P-value 38,8% $> 5\%$
We fail to reject H_0

We obtained the assignment by looking at the expected counts (the observed counts do not vary and correspond to Wicklin's sample): one expected count per table is enough, e.g., for Green.

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%

! Keep the context and the aim (does satisfaction vary?) when you answer this question

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.

The two sets of circled triplets (22.1% - 35.1% - 33.3% for highly satisfied students and 26.3% - 18.6% - 14.6% for highly unsatisfied students) might suggest that pre-MiM students are less satisfied than the two other categories of students (and among these, that S.M. students are even more satisfied than MiM students).

To know for sure and get a proof, we of course must run a test! (namely a χ^2 -test of independence)

Next you perform, on second thoughts, the treatment reproduced on the next page.

Note: Too many students compared the modes (= the most frequent values):
 1 for Pre-MiM students (with empirical frequency 26.3%)
 5 for 1st year MiM (35.1%)
 5 for S.M. (33.3%)
 but this doesn't tell whether satisfaction varies...

		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 ^a	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~~

which can be rephrased as:
 H_0 : homogeneity of satisfaction profiles
 H_1 : lack of homogeneity thereof

□ Is the test validly applied? Recall all elements to check to that end.

Yes: Sample size $n = 288 \geq 30$

+ All expected counts ≥ 5 (the smallest expected count equals 11.2)

[I didn't want to write the word "homogeneity", I wanted just to introduce it]

□ Does the table above prove that satisfaction varies among the three subpopulations considered? Explain.

No, on the contrary! P-value = 18.5%, we fail to reject the H_0 hypothesis of similar satisfaction profiles

↳ These data do not bring to light any obvious / significant differences in satisfaction profiles between the three groups of students considered.

Put differently: the differences we spotted in the table, even though they looked important to us (e.g., 22.1% vs. 35.1%), were not significant!

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 ^a
df	11
Asymp. Sig.	,495

a. 0 cells (0,0%) have expected counts less than 5.

the expected counts show that the reference distribution is a uniform distribution.

These data do not suggest any evidence against an even distribution of births among the 12 months of a year.

Which is the complete name of the test worked out here? χ^2 test of goodness of fit
Can the outcome of the test be validly exploited? Explain.

Yes, as: total sample size $n = 8,015 \geq 30$
& no expected count less than 5 (see table note)

- ☐ 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 ?

P-value 49.5%, data fail to reject H_0

H_0 [conformity]: Births occur uniformly over the year
 H_1 [non-conformity]: Births occur with some seasonality

- ☐ State a statistical conclusion (in plain words, that are understandable by a layman).

These data cannot exclude that births now occur uniformly over the year and that their seasonality disappeared. At least, these data do not exhibit any significant seasonality.

expected count $667.9 = 8,015 \times \frac{1}{12}$

vs.

observed count: 667

reference distribution is: $(\frac{1}{12}, \dots, \frac{1}{12})$

Alternative wording:



Simple linear regression

Elementary exercises	page 124
Advanced exercises	
The historical example of regression	page 126
An example of a spurious correlation	page 128
Some French politics: demonstrations	page 130
Prices of ski passes	page 132

Elementary exercise on "Growth of orange trees"

1. (a) Existence of a significant linear regression:
The indicated P value (to be read in the 1st column of the middle table or equivalently in the 1st column & 1st line of the third table) is almost null.
Hence the slope coefficient is significantly different from 0.
- (b) We read $r^2 = 83.5\%$, which is very good, truly excellent actually.
 \hookrightarrow 83.5% of the (variations in the) circumference(s) are explained by the (variations in the) age(s).
 Note: the remaining 16.5% are due to genetic factors, environment (though all these trees are grown up side by side, minor differences still exist in environments, etc).
- (c) Circumference (in mm) = 17.4
 $+ 0.107 \times \text{Age (in days)}$
 $+ \text{Residual term}$
 (with standard deviation: 23.738)
- (d) Well... the intercept here would be the circumference at day 0... while at day 0, the tree is just a seed.
 We would expect a 0 circumference.
 Now, it turns out that our 17.4 here is NOT significantly different from 0 (P value 52% for the test of its nullity, where H_0 is the nullity and H_1 the non-nullity).
- Slope coefficient:
 - of course significantly different from 0 (see question (a)!)
 - quantifies the growth rate

2. Approximate growth rate: 0.107 millimeters per day (= estimated coefficient)

Conversion into a more palatable number: $365 \times 0.107 / 10 \approx 3.9$ cm per year

More precise answer: with high confidence, the growth rate equals $0.107 \pm 2 \times 0.008$ mm per day

That is, we provide a confidence interval on the coefficient

The same conversion would indicate: 3.9 ± 0.6 cm per year

3. (a) Point estimate of this circumference:

$$17.4 + 0.107 \times 852 \approx 108.6 \text{ mm} \\ \approx 10.9 \text{ cm}$$

(b) Calculation: $108.6 \pm 2 \times 23.738 / \sqrt{35}$
 $= 108.6 \pm 8.03$

35
data points,
see column
in the
middle table

rounded off to $10.9 \pm 0.9 \text{ cm}$

↳ With high confidence, the average circumference of 2-year-and-4-month old trees should lie in the interval $10.9 \pm 0.9 \text{ cm}$

(c) Calculation: $108.6 \pm 2 \times 23.738$
 $= 108.6 \pm 47.476$

rounded off to $10.9 \pm 4.8 \text{ cm}$

or even $11 \pm 5 \text{ cm} = [6 - 16 \text{ cm}]$

↳ Most (95%) of the 2-year-and-4-month-old trees should have circumferences within the 6-16 cm range.

4. Solid lines are $\pm 47.5 \text{ mm}$ away from the regression line, hence they correspond to the lower and upper bounds of prediction intervals, as in Question 3(c).

No outliers: all data points are within their prediction intervals, all are compatible with the model.

Exercise "The historical example of regression"

1. Dependent variable: sons' heights \rightarrow y-axis
 to be statistically explained by the
 Independent (explanatory) variable: fathers' heights \rightarrow x-axis

We should thus read the Regression output #2.

2. - Existence of a significant linear regression: Yes, see the almost null P-value in the middle table, last column.
- Quality of this relation: $r^2 = 25.1\%$, good!
 \hookrightarrow 25.1% of the (variations of the) sons' heights are explained by the (variations of the) fathers' heights.
 Where do the other almost 75% lie?
 Perhaps, 25% of the sons' heights could be explained by mothers' heights, which would make 50% for genetic factors... and thus, it would remain 50% of the variable of interest to be explained by environmental factors.

- Writing of the relation:
- (*) Sons' heights (in cm) = 86.072
 $+ 0.514 \times$ Fathers' heights (in cm)
 $+ \text{Residual term}$
 (with standard deviation: 6.1889)

- Interpretation: both 86.072 and 0.514 are significantly different from 0 and need to be interpreted...
 \hookrightarrow However we can find no easy such interpretation!
 Hence the hint given by the next question.

3. We know thanks to the theory that the regression line goes through the point
 $\bar{x} = 171.925$
 $\bar{y} = 174.458$
- That is (as can be checked by direct computations as well):
- (**) $174.458 = 86.072 + 0.514 \times 171.925$

We subtract (**) from (*) and get:

$$\begin{aligned} \text{Sons' heights (in cm)} - 174.58 &= 0.514 \times (\text{Fathers' heights (in cm)} - 171.925) \\ &+ \text{Residual term (with standard deviation: 6.189)} \end{aligned}$$

That is,

$$\begin{aligned} \text{Differences of sons' heights to sons' average height (in cm)} &= 0.514 \times \text{Differences of fathers' heights to fathers' average height (in cm)} \\ &+ \text{Residual term (with standard deviation: 6.189)} \end{aligned}$$

↳ Differences to the average are reduced (on average) by a factor of $0.514 < 1$ → we indeed observe a regression toward the average height.
Gallien was right!

But wait, just a sanity check:

how confident are we that the slope coefficient 0.514 is (significantly) smaller than 1?

↳ Its 95% confidence interval $0.461 - 0.567$ does not contain 1.

Exercise "An example of a spurious correlation"

(1) Existence of a significant linear regression:

Yes, cf. the almost null P-value in the middle table, last column

Quality of this relation:

excellent - $r^2 = 98.4\%$

↳ Over the considered period, 98.4% of the (variations in the) mental-disease rate are explained by the (variations of the) TV-equipment rates.

Writing of the relation:

$$\begin{aligned} \text{Mental-disease rate (in \%)} &= 4.552 \\ &+ 0.222 \times \text{TV-equipment rate (in \%)} \\ &+ \text{residual term} \\ &\quad (\text{with standard deviation: } 0.728) \end{aligned}$$

Interpretation of the coefficients:

really?!

4.552 is significantly $\neq 0$, thus has to be interpreted:
 = some baseline value for the mental-disease rate?
 ie, the rate that existed before TV was created?

0.222 quantifies how much TV gets people mentally ill?
 + 10% TV equipment rate \Rightarrow + 2.22% in the mental-disease rate

Outliers:

we can see from the scatterplot that there are no outliers...

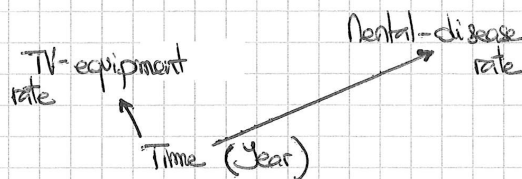
(2)

Be cautious with the interpretations!

We can only say here that we observe a strong linear association between the two variables over the considered period.

It does not mean that the explanatory variable is the "true" source of variation of the dependent variable.

Here, we probably rather have the following causality triangle:



where \uparrow denotes some causal linear relation.

Indeed, reading the "Explanation: Part 2/2" outputs, we see that

TV-equipment rate / Time (Year) is a very LINEAR relation, see: $r^2 = 97.2\%$

Mental-disease rate / Time (Year) is a very LINEAR relation, see: $r^2 = 96.4\%$

So that by transitivity,

Mental-disease rate / TV-equipment rate is also a very LINEAR relation, see: $r^2 = 95.4\%$

Causality explanations are just that over the considered period (which was probably hand-picked to fool you...),

science, industry & medicine made huge progress

↳ the mental disorders are just better diagnosed (they probably don't happen more often than before)

↳ the time period corresponded to TV sets becoming a mass consumption good.

→ Time is our third, latent variable, explaining in a causal way the two initial variables.

(3) We will briefly comment on all this material in class.

Exercice on "Some French politics: demonstrations."

Question 1

We take the number of demonstrators as counted by organizers as the dependent variable, and the number of demonstrators as counted by the police as the independent variable.

1. Existence of a significant linear regression: yes, see the almost null P-value in table 2.

2. Quality of this relation: very good, $r^2 = 82.2\%$ (meaning that 82.2% of the variations in one count is explained by the other count's variations)

where do the remaining 17.8% lie? I guess that there is a variability on the type of organizers: unions, political parties, "ordinary" citizens, etc.

3. Writing the relation:

$$\begin{aligned} \text{Number of demonstrators} \\ \text{as counted by organizers} \\ \text{(in thousands)} &= 358.197 \\ &+ 2.254 \times \text{Number of demonstrators} \\ &\text{as counted by the police} \\ &\text{(in thousands)} \\ &+ \text{residual term (with standard} \\ &\text{deviation 413.4445)} \end{aligned}$$

4. Interpretation:

- The constant factor 358.197 is significantly larger than 0 (see the P-value 2.4% for the test of nullity of the intercept coefficient).
- The scale factor 2.254 is significantly larger than 1 (see its confidence interval).

So yes, definitely, organizers and police disagree on the numbers!

(They would have agreed if both the constant term would not have been significantly different from 0 and the slope coefficient would not have been significantly different from 1).

Question 2:

See my interpretation in question 1.

Question 3:

- * With high confidence, when the police reports 500,000 demonstrators, the organizers report, on average,

$$(358.197 + 2.254 \times 500) \pm 2 \times \frac{413.4445}{\sqrt{29}}$$

$$= 1485.197 \pm 153.550$$

Thousands of demonstrators, that is, after rounding: they report 1,480,000 \pm 160,000 demonstrators on average.

- * 95% of the demonstrations for which the police reports 500,000 demonstrators include, in the organizers' eyes,

$$1485.197 \pm 2 \times 413.4445 = 1485.197 \pm 826.89$$

thousands,

that is, after rounding:

$$1,480,000 \pm 830,000 \text{ demonstrators.}$$

"Prices of ski passes"

(1) Existence of a significant linear regression:

Yes, cf. the almost null P -value in the middle table, last column

Quality of this relation:

very good — $r^2 = 57.1\%$

↳ 57.1% of the (variations in the) ski pass prices are explained by the (variations in the) sizes of the ski areas.

Writing of the relation:

$$\begin{aligned} \text{Price of the weekly ski pass (in €)} &= 92.366 \\ &+ 0.434 \times \text{Size of the ski area (in km)} \\ &+ \text{Residual term} \\ &\quad (\text{with standard deviation: } 20.594) \end{aligned}$$

Interpretation of the coefficients:

[We try to imagine some interpretations, that would then need to be confirmed by more extensive analyses...]

92.366 is significantly $\neq 0$, this needs to be interpreted:

it corresponds to some fixed costs, eg:

- cable cars (to go from the main village to bottom of the main ski runs)
- buses connecting remote villages to the central one
- advertisement for the ski resort, etc.

0.434 (significantly $\neq 0$) measures the average extra costs per additional kilometer in the ski area; eg, an additional ski run often requires an additional ski tow or ski chairlift (which needs to be built and maintained + needs to be operated by employees).

(2) Let us compute the prediction interval for a ski resort with size 250 km:

Our model indicates that 95% of these ski resorts have pass prices included in the interval

$$\begin{aligned} &92.366 + 0.434 \times 250 \pm 2 \times 20.594 \\ &= 200.866 \pm 41.188 \\ &\approx [159 - 243] \end{aligned}$$

and the ski pass price observed for Serre-Chevrier (251 €) does indeed NOT belong to this interval.

Multiple linear regression

Excerpts of past exams

Wage discrimination?

page 134

Modeling life expectancy

page 138

Wage discrimination?

- (1) The top-left histogram pair illustrates that men's salaries cover a wider spectrum than women's ones (basically: men's median salary \approx women's maximal salaries!).

* but *

The total experiences for men and women share similar behaviors (same for Years junior / Years senior), so it is unclear whether the salary differences are due to discrimination or to these differences in experiences.

↳ We are not measuring differences all things being equal, there might be some confusion factors.

- (2) We read the last line of scatterplots and try to assess where data points would lie the closest to the regression line: probably for Salary / Total experience.

- (3) Statistically valid models: all but Salary / Skill

1 point
0.5 point
Economic validity: whether the relation makes sense, i.e. whether the increases or decreases of the dependent variable as functions of the independent variable make intuitive sense.

1.5 points
All models are economically valid: it is intuitive that more experience (especially in higher levels) leads to larger salaries (models: Salary / Total experience, Salary / Years junior and Salary / Years senior have positive slopes). For Salary / Age, it is probably not age per se, but its positive correlation with experience that is at stake and explains the positive coefficient. Now, unfortunately, Salary / Gender comes with a negative slope coefficient, which (given that women are coded by 1) shows that women earn less on average... hence the feeling of wage discrimination.

- (4) Best model \leftrightarrow largest r^2 \leftrightarrow Total experience

0.5 point
1 point
Writing:
$$\text{Salary (in \$)} = 30,816.788 + 924.169 \times \text{Total experience (in years)} + \text{Residual term (with standard deviation 4677)}$$

Interpretation: (Let us take 0 as baseline value for Total experience.)

1.5 points
 $30,816.788 \$$: average annual starting salary
 $924.169 \$$: average annual additive increase of salary for each year worked in the company

(5) Automatic method 1: Backward selection → we iteratively simplify the model (by removing the least individually significant variable at each step), as long as needed (as long as at least one variable is still not individually significant).

Automatic method 2: Forward selection → we iteratively enrich the model by adding one variable at each step (that is individually significant in the extended model), as long as we can.

Both methods recommend the same model:
Salary / Total experience, Years senior, Years junior, Skill, Gender

(6) Cf. scatterplot of Total Experience / Age (top-left corner of the matrix of scatterplots in Fig. 3): the two variables are linearly linked, it is redundant to consider Age when one already considers Total Experience.

(7) Statistical validity: OK, from a global and from an individual point of view; cf. all P-values < 5%.
Quality: $r^2 = 87.8\%$, which is very good!

Writing: $\text{Salary (in \$)} = 34,252.779 + 461.729 \times \text{Total experience (in years)} + 401.043 \times \text{Years juniors (in years)} + 751.522 \times \text{Years senior (in years)} - 1,776.308 \times \text{Gender (0 for 0, 1 for 1)} + 4,657.344 \times \text{Skill (1 if exists, 0 otherwise)} + \text{Residual term (with standard deviation: 3,510)}$

Economic validity: clear, cf. same signs for the coefficients as already discussed in Question 3.

(8) Yes, we are now studying salaries all other things being equal ("ceteris paribus"), and thus we can isolate the effect of gender and get a (beginning of a) proof of gender discrimination.

Impact (cf. confidence interval): women earn at least 357.447 \$ less than men all other things being equal.

However, given that annual salaries (cf. histograms) are typically larger than 30,000 \$, this salary difference that we can minimally prove:
- is significant (ie, it does exist)
- but is small (less than 1% of the quantities at hand).

2 points

"small effect size" in the statistical jargon

(9) We recover this (significant but) small-impact salary difference when comparing the r^2 of the two models: when dropping the Gender variable (which BTR was the least individually significant), the r^2 only drops from 87.8% to 86.9% (and the standard deviation only increases from $\pm 3,540$ \$ to $\pm 3,602$ \$).

(10) This woman: $\left\{ \begin{array}{l} \text{observed salary: } 32,250 \\ \text{Total experience: } 5 \text{ years} \\ \text{No highly valued skill} \end{array} \right.$ (and 0 for Years junior, Years senior)

The model of Question (7) predicts an individual salary, given these characteristics, lying approximately in the interval:

$$= \frac{34,282.779}{34,816.116} \pm \frac{461.729 \times 5 - 1,775.308}{7,080} \pm 2 \times 3,540$$

The model of Question (9) predicts

$$= \frac{33,520.114}{35,826.514} \pm \frac{461.080 \times 5}{7,204} \pm 2 \times 3,602$$

In both cases, the observed value 32,250 lies in the prediction interval, hence the observed salary is compatible with each of the two models.

(11) We use that $\text{Total experience} = \text{Total experience men} + \text{Total experience women}$

to transform the second relation

$$\text{Salary (in \$)} = 31,499,900 + 706.472 \times \text{Total experience (in years)} + 208.252 \times \text{Total experience men (in years)} + \text{Residual term (with standard deviation 4,553)}$$

into the first relation (via the fact that $914.724 = 706.472 + 208.252$)

$$\text{Salary (in \$)} = 31,499,900 + 706.472 \times \text{Total experience women (in years)} + 914.724 \times \text{Total experience men (in years)} + \text{Residual term (with standard deviation 4,553)}$$

Actually, 706.472 is the baseline (= for women) average annual increase in salary per additional year of experience, 208.252 is the additional such increase men get on top of this baseline (so all in all, they get on average $+ (706.472 + 208.252) = + 914.724$ \$ per additional year)

(12)

The studied regression is not statistically valid, as the Total experience (women) variable is not individually significant. It means that the salary increases between genders are not significantly different (all other things being equal / ceteris paribus).

Note: the +166.220 is not significantly positive but you may still be surprised to get a non-significantly positive (rather than a non-significantly negative) number, especially given the previous question.

So, we have no good argument for discrimination in salary increases based on these data.

If however one wanted to discuss that topic, one would have to check whether there would not be a glass-ceiling effect preventing women from accessing the higher (junior / senior) levels in the first place.

Here we are measuring salaries conditional on the years spent in these higher levels, but we are not studying equal access to these levels.

Thus, based on these data, we are only left with the additive difference already discussed in Question (8) and which probably corresponds to the starting salary. It is proven to exist but is small comparatively to the order of magnitude of the problem (see our detailed answer therein).

Modeling life expectancy.

1. [0.25 point]

a. Developed countries ↔ High consumption
Developing countries ↔ Average consumption
Muslim countries ↔ Very low consumption

[0.25 point]

b. IQ \approx 100 ↔ Western countries
IQ much lower than 100 ↔ Africa, Middle-East, South America, Asia

↳ Clearly indicates that IQ is a very biased measure of intelligence! Measures "Westernness"...

2. Scatterplots:

[0.25 point]

Top-left matrix, last line: Points "follow the line" on LifeExp / LnGDP while at least half of the points are very far from the line on LifeExp / GDP. We can even guess some logarithmic arrangement of the points on LifeExp / GDP: compare with LnGDP / GDP same trend!

Numerically:

[0.25 point]

Compare the regression outputs for LifeExp / LnGDP and LifeExp / GDP: both are statistically and economically valid (see next question) and we compare the r^2 :

LifeExp / GDP: $r^2 = 50.7\%$
LifeExp / LnGDP: $r^2 = 71.6\%$

↳ Both criteria indicate that LifeExp / LnGDP is a superior model to LifeExp / GDP.

3. We study LifeExp / LnGDP (Simple regression #2)

[0.5 point]

Statistically valid: yes, see P-value in the 2nd table: almost null, $< 5\%$.
Economically valid: yes, the positive coefficient indicates that the higher the LnGDP (or the GDP), the longer the life expectancy, which makes sense: the country is often more developed, with a better healthcare system, etc.

[0.5 point]

Writing the model:

$$\begin{aligned} \text{Life expectancy (in years)} &= 56.054 \\ &+ 6.718 \times \text{LnGDP} \\ &\quad (\text{GDP in K\$ / capita}) \\ &+ \text{residual term} \\ &\quad (\text{with standard dev: } 5.6) \end{aligned}$$

[0.5 point]

Interpretation:

* 6.718: $\text{LnGDP} + 0.69 = \ln(2)$ if $\text{GDP} \times 2$
Thus: LifeExp $+ 0.69 \times 6.718 \approx 4.6$ years when GDP doubles

* SG. 054: $\ln GDP$ is typically ≥ 0 (though occasionally < 0) so that we can take 56 years as a baseline value for life expectancy (in poor underdeveloped countries).

4. [0.5 point] All of them are! (see the P-values $< 5\%$ in the ANOVA tables.)

5. [0.5 point] Alcohol and tobacco are bad for one's health. They should have a detrimental effect on life expectancy, at least from an individual viewpoint. We expect negative coefficients.

But we read positive coefficients!
So, we are puzzled.

[0.5 point] Actually (see the matrix of scatter plots on page 11) Alcohol & Tobacco consumptions increase with the per capita GDP. They measure in some sense (see also question 1a: up to Muslim countries) how developed is a country.

How developed is a country (the GDP / $\ln GDP$) is a latent variable positively correlated with LifeExp and Alcohol / Tobacco. It explains why we read this surprising positive coefficient.

6. [0.5 point] * Statistical interest: measured by r^2 . The best models would be

LifeExp / $\ln GDP$: $r^2 = 71.6\%$
LifeExp / IQ: $r^2 = 72.1\%$
LifeExp / $\ln IQ$: $r^2 = 73.3\%$

while all other models have $r^2 \leq 50\%$

* Economic meaning: IQ variable very questionable (see question 1b).

* Thus, best trade-off: LifeExp / $\ln GDP$.

7. [0.5 point] We compared LifeExp / Var and LifeExp / $\ln Var$

* Var = GDP \rightarrow $\ln GDP$ is better than GDP (see question 2)

same method as in question 2 (scatterplots of page 10 + computing r^2)

* Var = Alcohol \rightarrow	Alcohol is better than $\ln Alcohol$
* Var = Tobacco \rightarrow	$\ln Tobacco$
* Var = Democr \rightarrow	Democr

but for IQ vs Ln IQ: same performance or almost for both models, so we prefer to stick to the nominal variable IQ for simplicity.

8. [0.5 point] We read the outputs of page 17.

Model on 5 variables:

- globally valid (at least one variable is useful), see P-value < 5% in the ANOVA table
- but issue with the marginal validity of one variable, namely, LnTobacco (P-value = 23.4%).

The 5-variable model is unnecessarily complex, the incremental contribution of the LnTobacco is not significant → we drop this variable and get the 4-variable model in the second part of page 17.

9. [0.5 point] Statistical validity: global & marginal same walked as in the previous question

Economic validity: → Positive coefficient for LnGDP as already discussed in question 3

→ Negative coefficient for Alcohol as expected in question 5

→ For IQ & Democr: why not... The model says that countries with smarter people and more democratic countries have higher life expectancies.

[0.5 point]

$$\begin{aligned} \text{Life expectancy (years)} &= 22.067 \\ &- 0.361 \times \text{Alcohol (liters/adult)} \\ &+ 0.461 \times \text{IQ (no unit)} \\ &+ 0.752 \times \text{Democr (no unit)} \\ &+ 3.107 \times \text{LnGDP (GDP in K$/capita)} \\ &+ \text{residual term (with std: 4.4)} \end{aligned}$$

[0.5 point]

Interpretation:

alternative baseline value for a developing country at the end of this correction

- Baseline value for a western country

$$\begin{aligned} 22.067 &- 0.361 \times 10 \\ &+ 0.461 \times 100 \\ &+ 0.752 \times 7 \\ &+ 3.107 \times 3 \end{aligned}$$

$$= 79$$

- While the coefficients -0.361, 0.461, 0.752 and 3.107 measure how the per country life expectancy evolves as

each of the variable increases or decreases :

$+ 1 \text{ liter/adult of alcohol} \rightarrow - 0.361 \text{ years of life expectancy}$
 $+ 10 \text{ IQ number} \rightarrow + 4.6 \text{ years}$
 $+ 1 \text{ Democr grade} \rightarrow + 0.752 \text{ years}$
 $\text{GDP} \times 2 \rightarrow + 3.107 \times \ln(2) \approx 2.15 \text{ years}$

10. [0.5 point]

This is the forward selection method:

- we start with the best individual variable (from a statistical viewpoint)
- as long as we can, we add a variable at each step in the model, that should be marginally valid in the extended model; if several such additions are possible, we pick the one leading to the best model (ie, with largest r^2).

It recommends the same model as the one studied in the previous question.

11. [0.5 point]

Life expectancy observed in France: 81 years

Point estimate given by the model of question 9:

$$\begin{aligned}
 & 22.067 \\
 & - 0.361 \times 12.48 \\
 & + 0.461 \times 98 \\
 & + 0.752 \times 7.77 \\
 & + 3.107 \times 3.529 \\
 & \approx 79.5
 \end{aligned}$$

[1 point]

Compatibility: yes, the difference between the observed and predicted value is $\leq 2 \times \text{standard deviation}$.

Put differently, the prediction interval for France was $79.5 \pm 2 \times 4.4 = 79.5 \pm 8.8$ years and 81 lies in this interval.

12. [0.5 point]

a) & b)

IQ is not a nice variable, see question 1b., so we tried to drop it but obtained a model not statistically valid (because of an issue of marginal validity for Alcohol).

Thus, we further dropped Alcohol and got the second alternative model Life Exp / $\ln \text{GDP}$ & Democr, which is satisfactory both from a statistical and economic viewpoint.

[0.5 point]

c)

To me, we should only compare the following models:

question 9	Life Exp / LnGDP, Democr, Alcohol, IQ	$r^2 = 82.8\%$
question 12 b)	Life Exp / LnGDP, Democr	$r^2 = 72.9\%$
question 6	Life Exp / Ln GDP	$r^2 = 71.6\%$

Is the 10% difference in r^2 worth adding 3 variables in the model, two of which (Democr, IQ) are very difficult to grasp?

↑
the r^2_{adj} are very close to these r^2

↳ I don't think so. So I would stick to Life Exp / GDP!

Question 9, alternative interpretation:

Taking a developing country as a baseline: [baseline IQ: 70
Democr.: 4 in this case]

$$\begin{aligned} \text{Life expectancy (years)} &= 57.345 \\ &- 0.361 \times \text{Alcohol (liters/adult)} \\ &+ 0.461 \times (\text{IQ} - 70) \quad (\text{no unit}) \\ &+ 0.752 \times (\text{Democr} - 4) \quad (\text{no unit}) \\ &+ 3.107 \times \text{Ln GDP} \quad (\text{GDP in K\$/capita}) \\ &+ \text{residual term (with standard deviation: 4.4)} \end{aligned}$$

where we got the 57.345 by writing $57.345 = 22.067 + 0.461 \times 70 + 0.752 \times 4$