

Statistics Exercises for the First Course in Medical Statistics

Degree courses in Medicine and Surgery and in Pharmacy

University of Rome "Tor Vergata"

(by Dr Alessia Mammone and Dr Simona Iacobelli)

Descriptive statistics: tables, graphs and summary indicators

Exercise 1.1.

For a sample of 15 students we observe the Time (in minutes, X) usually spent using Facebook per day, and the final grade in the Statistics exam (Y):

Facebook X	Statistics Y
0	25
11	22
17	28
16	30
22	22
17	27
25	26
30	21
27	27
27	28
31	23
35	29
30	30
45	24
60	27

- Compute mean, median and quartiles for both X and Y
- Compute the range and the IQR for both X and Y
- Find the standard deviation and the coefficient of variation for both X and Y. Which variable is more heterogeneous?
- Plot the box-plot for both X and Y

Solution

We work separately for X and Y. In fact all questions regard the distribution of one variable, regardless of the other (this will be called the "marginal" distribution). We will study the relationship between X and Y in exercise 5.5.

- Below we consider the two tables, for X and for Y. First column: ordered values, to identify the median and quartiles. Second column: calculations for the standard deviation, question c (notice: it makes no difference working on the original series or on this one sorted in ascending order)

	Facebook X	X ²
	0	0
	11	121
	16	256
	17	289
	17	289
	22	484
	25	625
	27	729
	27	729
	30	900
	30	900
	31	961
	35	1225
	45	2025
	60	3600
sum	393	13133

Statistics Y	Y ²
21	441
22	484
22	484
23	529
24	576
25	625
26	676
27	729
27	729
27	729
28	784
28	784
29	841
30	900
30	900
389	10211

a)

For Time on Facebook (X)

mean = $393/15 = 26.2$ minutes

median=Q2=27

In fact since $n=15 \rightarrow (n+1)/2=(15+1)/2=8 \rightarrow$ in the 8th position of the ordered list there is the value 27 (minutes)

Q1=17

Since $n=15 \rightarrow (n+1)/4=16/4=4 \rightarrow$ in the 4th position of the ordered list there is 17

Q3=31

Since $n=15 \rightarrow (3*(n+1))/4=3*16/4=12 \rightarrow$ in the 12th position of the ordered list there is 31

For Statistics grade (Y)

mean= $389/15=25.9$

median= Q2=27

In fact since $n=15 \rightarrow (n+1)/2=(15+1)/2=8 \rightarrow$ in the 8th position of the ordered list there is the value 27 (grade)

$$Q1=23$$

Since $n=15 \rightarrow (n+1)/4=16/4=4 \rightarrow$ in the 4th position of the ordered list there is 23

$$Q3=28$$

Since $n=15 \rightarrow (3*(n+1)/4=3*16/4=12 \rightarrow$ in the 12th position of the ordered list there is 28

b)

For Time on Facebook (X)

$$\text{Range} = \text{Max} - \text{min} = 60 - 0 = 60$$

$$\text{IQR} = Q3 - Q1 = 31 - 17 = 14$$

For Statistics grade (Y)

$$\text{Range} = 30 - 21 = 9$$

$$\text{IQR} = Q3 - Q1 = 28 - 23 = 5$$

c) See the calculations in the tables, second column

For Time on Facebook (X)

$$\text{Variance: } (13133/15 - 26.2^2) * 15/14 = (189.0933) * 15/14 = 202.6$$

$$\text{Standard deviation} = 14.23 \text{ minutes}$$

$$\text{VC} = 14.23/26.2 * 100 = 54\%$$

For Statistics grade (Y)

$$\text{Variance: } (10211/15 - 25.9^2) * 15/14$$

$$\text{Standard deviation} = 2.96 \text{ points}$$

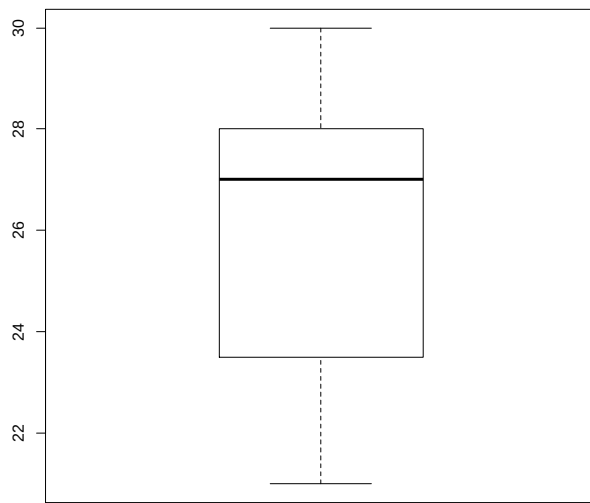
$$\text{VC} = 2.96/25.9 * 100 = 11\%$$

Thus the time spent on FB X is more heterogeneous than the Statistics grade Y

d) Plot the box-plot for both X and Y

The box-plot displays a box delimited by the quartiles Q1 and Q3, with an inner thick line representing the median; let's assume for simplicity that the outer lines go from the Min to the Max (although this is not always the case: most software draw them as a fixed proportion of the standard deviation, and mark with further points the outliers).

Draw by yourself the box-plot for X, the following is for Y:



Grade in Statistics

Exercise 1.2.

Using the data of Exercise 1, Build a frequency distribution both for X and Y choosing a suitable class grouping, for example (choose one for X and one for Y):

Proposed grouping for X

- [0-10), [10-20), [20-30), [30-40), [40-50), [50-60]
- [0-15), [15,30), [30,60]

Proposed grouping for Y

- [18-22), [22-26), [26-30]
- [18-21), [21-24), [24-27), [27-30]
- [18-20), [20-22), [22-24), [24-26), [26-28), [28-30]

Then, for both X and Y:

- a) Plot a histogram using the classes you choose in part a).
- b) Indicate the class representing the mode
- c) Indicate the class including the median
- d) Compute the mean on the basis of the aggregate data (i.e. this table), and compare it to the (exact) mean computed on the original data

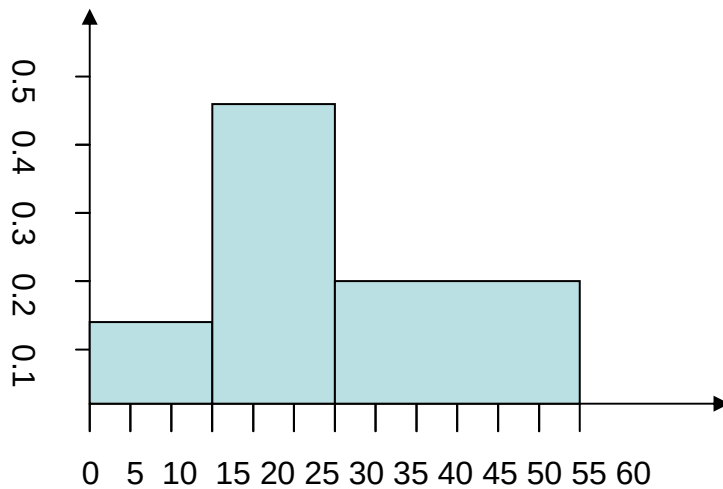
Solution

The solution is proposed for Time spent on Facebook (X) using the second grouping [follow the same procedure for the other cases].

Frequency distribution in classes (notice: an informative table should include: absolute frequency, percentage, cumulative frequency and cumulative percentage – the latter two make sense since a quantitative variable is ordered):

class	n (freq)	p (%)	N (cum)	P (% cum)	width	density	x central value	n·x
[0-15)	2	13.3	2	13.3	15	0.1333	7.5	15.0
[15-30)	7	46.7	9	60.0	15	0.4666	22.5	157.5
[30-60)	6	40.0	15	100.0	30	0.2000	45.0	270.0
tot	15	100.0						442.5

a) The histogram reports on the horizontal axis the classes (contiguous intervals) and on the vertical axis the density:



b) The mode is the class [15-30)

c) The median is included in the class [15-30) [where the cum % reaches 50%]

d) Using the central values as representatives of the classes, the mean is $442.5 / 15 = 29.5$.

The exact value was 26.2. Our approximation overestimates the mean.

Exercise 1.3.

In the following table is reported the frequency distribution of the age of 20 participants in a clinical trial. Compute the mean and the standard deviation.

class	n (freq)	p (%)	N (cum)	P (% cum)
[35-45)	5	25	5	25
[45-55)	2	10	7	35
[55-65)	3	15	10	50
[65-75)	7	35	17	85
[75-85)	3	15	20	100
tot	20			

Solution

We compute a representative value of each class = (upper bound of the class + lower bound of the class)/2. We consider the frequency of the class.

This is the basis also for computing the standard deviation. This is the square root of the variance. For the latter, here we use the first formula, based on the sum of the squares of each error (value-mean), each multiplied by its frequency:

class	n (freq)	p (%)	N (cum)	P (% cum)	x central value	n·x	x-mean	n·(x- mean)^2
[35-45)	5	25	5	25	40	200	-8.5	361.25
[45-55)	2	10	7	35	50	100	1.5	4.50
[55-65)	3	15	10	50	60	180	11.5	396.75
[65-75)	7	35	17	85	70	490	21.5	3235.80
[75-85)	3	15	20	100	80	240	31.5	2976.80
sum	20					970		6975.00

mean for grouped data: $970 / 20 = 48.5$ yrs

variance for grouped data: $6975 / 19 = 367.10$ yrs²

standard deviation for grouped data: $\sqrt{367.10} = 19.16$ yrs

Exercise 1.4.

The table on the left below reports the distribution of the final score at the exam of Statistics for the Medicine students (degree course in Italian) of the cohort 2011-2012 (the distribution excludes tests failed and scores rejected; just FYI: 30 out of 45 Score=30 were "30-cum-laude").

The table on the right reports the distribution of the number of attempts = times that each student sat in the exam (only students who passed).

- What is the average score? How many attempts are necessary on average to pass the exam?
- A student has 50% of probability of getting a score equal to or higher than?
- How do you interpret the mode of the distribution of Y?
- Compute the standard deviation of the Score
- Assume that the Score is a continuous variable, and represent its distribution with a graph, using the classes 18|-21, 21|-26, 26|-31

Score (X)	
8	12
19	9
20	7
21	6
22	6
23	15
24	12
25	6
26	12
27	27
28	11
29	3
30	45
	171

Attempts (Y)	
1	83
2	46
3	34
4	8
	171

Solution

Tables with calculations for the answers:

Score (X)	freq	x·freq	cum freq	Error= (x-mean)	Error·freq	Error^2·freq		
18	12	216	12	-7,46199	-89,5439	668,1752		
19	9	171	21	-6,46199	-58,1579	375,8156		
20	7	140	28	-5,46199	-38,2339	208,8332		
21	6	126	34	-4,46199	-26,7719	119,456		
22	6	132	40	-3,46199	-20,7719	71,91218		
23	15	345	55	-2,46199	-36,9298	90,9208		
24	12	288	67	-1,46199	-17,5439	25,64892		
25	6	150	73	-0,46199	-2,77193	1,280599		
26	12	312	85	0,538012	6,45614	3,473479		
27	27	729	112	1,538012	41,52632	63,86796		
28	11	308	123	2,538012	27,91813	70,85654		
29	3	87	126	3,538012	10,61404	37,55258		
30	45	1350	171	4,538012	204,2105	926,7098		
	171	4354		-19,0058	0	2664,503		

Attempts (Y)	freq	x*freq	% freq
1	114	114	49%
2	64	128	28%
3	39	117	17%
4	12	48	5%
5	2	10	1%
	231	417	100%

a) Average score: $\text{Mean}(X) = 4354 / 171 = 25.46$

Average number of attempts: $\text{Mean}(Y) = 417 / 231 = 1.8$

b) The answer is found by computing the median of X. Median: the modality with rank = $(171+1)/2=86$. It is the modality "27". Thus, there is 50% probability of getting 27 or more.

c) Mode=1 (with frequency 114). Most of the students (49%) pass the exam after only 1 attempt.

d) For the variance of X:

We compute the sum of squared errors (SSE) and divide by $(n-1)$.

The logic is: we want a sort of average of the distance between each observation and their mean. Each distance $(x-\text{mean})$ is called "error". The usual average would require to sum

all errors (because we have a frequency table, each error is multiplied by its frequency, to get the total) and divide by n .

Now, by definition of the mean, the simple sum of all errors equals 0 [you can check your calculations using this property, as we do in the above table]. So we make a different type of mean, called "quadratic mean": we first square the errors, and then we sum. To get an unbiased estimator of the variance [topic seen in Statistical Inference] we divide by $(n-1)$ instead of n .

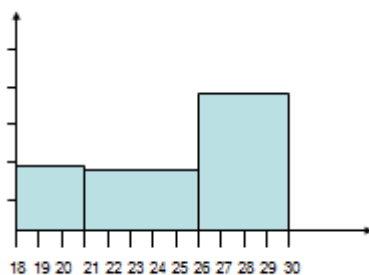
$$\text{variance} = 2664 / 170 = 15.673$$

$$\text{st.dev.} = \sqrt{15.673} = 3.959$$

e) The appropriate graph to represent this distribution in classes (the variable is treated as continuous) is an histogram

Distribution of the Score in classes:

Score	freq	width	density
18 -21	28	3	9.3
21 -26	45	5	9.0
26 -31	98	5	19.6
	171		



Exercise 1.5.

The mean and the standard deviation of the score at the exam of Statistics for the past 2 cohorts of students was:

- cohort 2011-2012 (Italian; $n=171$): mean=25.5, std=3.959
- cohort 2012-2013 (English; $n=16$): mean=26.4, std=4.11

Compute the overall average and compare the variability.

Solution

The overall mean is a weighted average:

$$\text{Mean} = [(25.2 \cdot 171) + (26.4 \cdot 16)] / (171+16) = 4782.9 / 187 = 25.58$$

The variability is compared by looking at the coefficient of variation:

$$\text{CV for cohort 2011-2012: } 25.5 / 3.959 = 16\%$$

$$\text{CV for cohort 2011-2012: } 26.6 / 4.3 = 16\%$$

Probability: basic rules, conditional probability and Bayes' formula**Exercise 2.1.**

There are two urns containing coloured balls. The first urn contains 50 red balls and 50 blue balls. The second urn contains 30 red balls and 70 blue balls. One of the two urns is randomly chosen (both urns have equal probability of being chosen) and then a ball is drawn at random from that urn.

- (a) What is the probability to draw a red ball?
- (b) If a red ball is drawn, what is the probability that it comes from the first urn?

Solution

$$P(U1)=P(U2)=1/2$$

$$P(\text{Red}|U1)=1/2$$

$$P(\text{Red}|U2)=3/10$$

$$\text{a) } P(\text{Red})=P(\text{Red}|U1) \cdot P(U1)+P(\text{Red}|U2) \cdot P(U2)=0.5 \cdot 0.5+0.3 \cdot 0.5=0.4$$

$$\text{b) } P(U1|\text{Red}) = P(U1 \& \text{Red})/P(\text{Red}) = P(\text{Red}|U1) \cdot P(U1)/P(\text{Red}) = 0.5 \cdot 0.5/0.4 = 0.625$$

(Notice that the latter is an application of the Bayes formula)

Exercise 2.1-b A similar exercise is:

Two hospitals use the same innovative surgical technique for a certain intervention. In the first hospital the probability of successful intervention is 50%. In the second hospital this probability is only 30%. A patient can be admitted to any of the two hospitals with equal probability.

- (a) What is the probability that the patient has a successful intervention?
- (b) If you know a person who had a successful intervention, what is the probability that he/she was admitted to the first hospital?

In fact: Each hospital is a Urn. Red ball = successful intervention.

Exercise 2.2.

There are two urns: the first urn contains 3 red balls, 3 blue balls, and 1 black ball. The second urn contains 1 red ball, 2 blue balls, and 1 black ball.

We choose the first urn with probability $2/3$ or the second with probability $1/3$. Then from the chosen urn we randomly draw one ball (that is, all balls in the urn have equal probabilities).

Let $B1$ be the event that I choose the first urn, and $B2$ denote the events that I choose the second, and let A be the event that I draw a red ball.

- (a) Compute $P(A|B1)$
- (b) Compute $P(A \cap B1)$
- (c) Compute $P(A)$

Solution

$$\text{(a) } P(A|B1) = 3/7$$

$$\text{(b) } P(A \cap B1) = P(A|B1) \cdot P(B1) = 3/7 \cdot 2/3 = 2/7$$

$$\text{(c) } P(A) = P(A|B1) \cdot P(B1) + P(A|B2) \cdot P(B2) = 2/7 + 1/4 \cdot 1/3 = 31/84$$

Exercise 2.3.

It is estimated that for the U.S. adult population as a whole, 55 percent are above ideal weight, 20 percent have high blood pressure, and 60 percent either are above ideal weight or have high blood pressure.

- (a) What percentage of the population is above ideal weight and have high blood pressure?
- (b) Find the conditional probability that a randomly chosen adult of the U.S. population:
 - (i) Has high blood pressure given that she or he is above ideal weight
 - (ii) Is above ideal weight given that he or she has high blood pressure
- (c) Let A be the event that a randomly chosen member of the population is above his or her ideal weight, and let B be the event that this person has high blood pressure. Are A and B independent events?

Solution

$$P(\text{Overweight})=P(O)=0.55$$

$$P(\text{High Blood Pressure})=P(P)=0.2$$

$$P(\text{Overweight or High Blood Pressure})=P(O \cup P)=0.6$$

$$(a) \quad P(O \cap P) = P(O) + P(P) - P(O \cup P) = 0.55 + 0.2 - 0.6 = 0.15$$

$$(b) \quad i) \quad P(P|O) = P(O \cap P) / P(O) = 0.15 / 0.55 = 0.273$$

$$ii) \quad P(O|P) = P(O \cap P) / P(P) = 0.15 / 0.2 = 0.75$$

$$(c) \quad \text{Pressure independent Overweight if and only if } P(P \cap O) = P(O) \cdot P(P)$$

$$P(O \cap P) = 0.15$$

$$P(O) \cdot P(P) = 0.55 \cdot 0.2 = 0.11$$

$$0.15 \neq 0.11 \text{ so they are not independent}$$

Exercise 2.4.

Consider a given population, a certain disease D and a certain symptom S. We know that 85% of the population members who have the disease D do have the symptom S, while the remaining 15% do not. Suppose that 95% of those who do not have the disease D do not have the symptom S, while 5% do have the symptom S (due to some other cause). Suppose that 10% of the given population have the disease.

- (a) What is the probability that a random person chosen from the population has the symptom?
- (b) Given that a random person was sampled, and he has the symptom, what is the probability that he or she has disease?
- (c) What is the probability that a random person does not have the disease and does not have the symptom?

Solution

$$P(D)=0.1 \rightarrow P(D^c)=0.9$$

$$P(S|D)=0.85 \rightarrow P(S^c|D)=0.15$$

$$P(S^c|D^c)=0.95 \rightarrow P(S|D^c)=0.05$$

$$a) \quad P(S) = P(S|D) \cdot P(D) + P(S|D^c) \cdot P(D^c) = 0.85 \cdot 0.1 + 0.05 \cdot 0.9 = 0.13$$

$$b) \quad P(D|S) = P(D \cap S) / P(S) = P(S|D) \cdot P(D) / P(S) = 0.85 \cdot 0.1 / 0.13 = 0.65$$

$$c) \quad P(D^c \text{ \& } S^c) = P(S^c|D^c) \cdot P(D^c) = 0.95 \cdot 0.9 = 0.855$$

Exercise 2.5

Consider the results of the exams in Statistics for the cohort 2011-2012 reported in Exercise 1.4. Additionally, be aware that in total 228 students tried the exam. Assume that the distribution of scores for those who pass remains the same during the academic year 2013-2014 for $n=9$ students of the Medicine course in English.

- a) What is the probability of passing the exam?
- b) If you pass, what is the probability that you get 30 (included 30-cum-laude)?
- c) If you pass, what is the probability that you get a score ≥ 27 ?
- d) What is the probability that you pass the exam with score ≥ 27 ?
- e) How many students are expected to pass the exam?
- f) What is the probability that all students pass the exam?
- g) What is the probability that no student passes the exam?

Solution

- a) $P(\text{pass}) = \text{favourable cases} / \text{possible cases} = 171 / 228 = 0.75$
- b) $P(\text{Score} = 30 \mid \text{pass}) = 45 / 171 = 0.26$ (this is the frequency returned by the table)
- c) $P(\text{Score} \geq 27 \mid \text{pass}) \approx 0.5$: in fact, 27 was the median. More precisely:
 $= (27+11+3+45) / 171 = 0.5029$ (again using the table; we could have used the cumulative frequency to find the numerator more quickly: $171-85 = 86$)
- d) Notice that this is NOT the same question as above:
 $P(\text{pass} \ \& \ \text{Score} \geq 27) = P(\text{pass}) \cdot P(\text{Score} \geq 27 \mid \text{pass}) = 0.75 \cdot 0.5029 = 0.377$
- e) $P(\text{pass}) \cdot 9 = 6.75$
- f) $P(\text{pass} \ \& \ \text{pass} \ \& \ \dots \ \text{pass}) = P(\text{pass})^9 = 0.75^9 = 0.075$
- g) $P(\text{Not-pass} \ \& \ \text{Not-pass} \ \& \ \dots \ \text{Not-pass}) = P(\text{Not-pass})^9 = (1-0.75)^9 < 0.0001$

Or, we use the Binomial distribution, with parameters $N=9$ and $p=0.75$, computing respectively:

$P(X=9)$

$P(X=0)$

For example, the probability that only 1 student passes is:

$$P(X=1) = \binom{9}{1} 0.75^1 0.25^{9-1} = 0.000103$$

(The binomial coefficient is $\frac{9!}{1!8!} = 9$)

Exercise 2.6

According to a school teacher, when one of his students makes all homeworks assigned, he/she will get the maximum score in 95% of the cases; instead, students who do not make the homeworks properly, have only 45% probability of getting the maximum score. The teacher believes (say, with probability 90%) that his student Tom did not do the

homeworks before the last test; however, Tom got the maximum score. What is the probability that Tom cheated at the test? (i.e. Tom did not do the homeworks)

Solution

The question is: $\Pr(\text{no homework} \mid \text{max score})$. The solution is provided by the Bayes formula.

prior: $\Pr(\text{no homeworks})=0.9$

likelihood: $\Pr(\text{max score} \mid \text{homeworks}) = 0.95$ $\Pr(\text{max score} \mid \text{no homeworks}) = 0.45$

$\Pr(\text{no homework} \mid \text{max score}) = 0.9 \cdot 0.45 / (0.9 \cdot 0.45 + 0.1 \cdot 0.95) = 0.81$

Exercise 2.6-b A similar exercise is:

There's a leak in an apartment, the owner is pretty sure (say, with probability 90%) that the problem originates in his neighbour's apartment, in this case he can be reimbursed of all the expenses to repair the leak and re-painting. The plumber states that the probability of having the leak if the problem is in the next apartment is only 45%, while there is a 95% probability of having that leak when the problem is in the same apartment. What is the probability that the owner will be reimbursed?

In fact, problem in next apartment = no homework, leak = max grade.

Probability: random variables; Binomial, Poisson, Normal**Exercise 3.1**

Consider a coin and assign to Head and Tail respectively the values 0 and 1. Suppose that $P(1)=3/4$, and $P(0)=1/4$. This coin is tossed 3 times independently. Let us define the random variable X = the average result of the 3 tosses.

Write down the distribution of X , that is, the list of all possible values it can take along with their probabilities. Then compute the expectation

** More advanced: compute the variance.*

Solution

$$P\{X=0\} = P\{0,0,0\} = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{64}$$

$$P\left\{X=\frac{1}{3}\right\} = P\{0,1,0\} + P\{0,0,1\} + P\{1,0,0\} = 3 \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{9}{64}$$

$$P\left\{X=\frac{2}{3}\right\} = P\{1,1,0\} + P\{1,0,1\} + P\{0,1,1\} = 3 \times \frac{3}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{27}{64}$$

$$P\{X=1\} = P\{1,1,1\} = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$$

(Check: the sum of all these probability must be =1)

Expected value:

$$E(X) = \sum x_i p_i = 0 \cdot \frac{1}{64} + \frac{1}{3} \cdot \frac{9}{64} + \frac{2}{3} \cdot \frac{27}{64} + 1 \cdot \frac{27}{64} = 0.75$$

** Variance:*

$$E(X^2) = \sum x_i^2 p_i = 0^2 \cdot \frac{1}{64} + \frac{1^2}{3} \cdot \frac{9}{64} + \frac{2^2}{3} \cdot \frac{27}{64} + 1^2 \cdot \frac{27}{64} = 0.625$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = 0.625 - 0.75^2 = 0.0625$$

Exercise 3.2.

Consider Exercise 3.1. Given the same hypotheses, let us define now the random variable Y = number of 1's obtained in three tosses of the coin. Which is the distribution of Y ? and the expected value?

** More advanced: Which is the relationship between the random variable X of Exercise 3.1 and Y of Exercise 3.2? Establish it and compute the expected value and variance of X using those of Y .*

Solution

We could do calculations similarly to what was done in Exercise 3.1; since the probabilities of Head and Tail are the same, we can actually re-use the computations already done in 3.1 (approach 1); a quick alternative is to realize that we can use a known probability distribution (approach 2).

1) Notice that:

$$Y=0 \leftrightarrow X=0$$

$$Y=1 \leftrightarrow X=1/3$$

$$Y=2 \leftrightarrow X=2/3$$

$$Y=3 \leftrightarrow X=1$$

Thus:

$$\Pr(Y=0)=1/64$$

$$\Pr(Y=1)=9/64$$

$$\Pr(Y=2)=27/64$$

$$\Pr(Y=3)=27/64$$

$$E(Y) = 0 \cdot 1/64 + 1 \cdot 9/64 + 2 \cdot 27/64 + 3 \cdot 27/64 = 2.25$$

2) Y follows a Binomial distribution with n=3 is (number of trials) and p=3/4 (the probability of success, Tail i.e. value=1)

(Check that you obtain the same probabilities for all possible values of Y: 0, 1, 2, 3)

Mean and the variance* of Y are found applying the formulas that hold for the Binomial(n,p):

$$E[Y] = np = 3 \cdot \frac{3}{4} = \frac{9}{4}$$

$$VAR[Y] = np(1-p) = 3 \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{9}{16}$$

* (more advanced)

The relationship between the random variable X of Exercise 1 and Y of Exercise 2 is:

$$X = \frac{Y}{3}$$

It's a linear transform, thus we can e.g. compute E(X) and Var(X) using the following properties of the expected value and of the variance:

$$E[cX] = c \cdot E[X]$$

$$VAR[cX] = c^2 \cdot VAR[Y]$$

Exercise 3.3

Usually (90% of the cases) patients undergoing an intervention in day-hospital actually go home after the intervention, but in 10% of the cases they need to be admitted in hospital. Today there are 3 interventions planned, but only 1 bed available. Compute the probability that today beds will not suffice.

Solution

We have n=3 trials (interventions) where the probability of "success" (value=1) i.e. requiring hospitalization is p=0.1. The random variable of interest is the total number of hospitalizations needed, Y = sum of the values=1. The question is: what is the probability that Y>1?

We have learned in Exercise 3.1-2 how to compute the distribution probability of a similar random variable, but we have also learned that we can alternatively use $Y \sim \text{Bi}(3, 0.1)$.

$\Pr(Y>1) = \Pr(Y=2 \text{ or } Y=3) = \Pr(Y=2) + \Pr(Y=3)$. Alternative:
 $\Pr(Y>1) = 1 - \Pr(Y=0 \text{ or } Y=1) = 1 - \Pr(Y=0) + \Pr(Y=1)$
 $= 0.028$

Exercise 3.4

Consider the same situation as in Exercise 3.3, but with probability of hospitalization equal to 0.01 and 300 interventions planned; still - incredibly - only 1 bed available. Compute the probability that beds will not suffice in this new situation.

Solution

We can again use a $\text{Bi}(n=300, p=0.01)$

$$\Pr(Y>1) = 1 - \Pr(Y=0 \text{ or } Y=1) = 1 - \Pr(Y=0) + \Pr(Y=1) = 1 - 0.049 - 0.149 = 0.802$$

Since n is large and p is small, we can also use the Poisson approximation to the Binomial: $Y \sim \text{Po}(\lambda = \text{mean} = 3 \cdot 0.01 = 3)$

$$\Pr(Y>1) = 1 - \Pr(Y=0 \text{ or } Y=1) = 1 - \Pr(Y=0) + \Pr(Y=1) = 1 - 0.050 - 0.149 = 0.801$$

Exercise 3.5

Let Z be a standard Normal random variable. Compute the probability of the following intervals: $(-\infty, 2)$; $(-\infty, 2.1)$; $(-\infty, -2.1)$; $(-2.18, +\infty)$; $(0, 2.21)$; $(-2.21, 2.21)$; $(-1, 2.18)$

Solution

$$P(-\infty \leq Z \leq 2) = 0.977$$

$$P(-\infty \leq Z \leq 2.1) = 0.982$$

$$P(-\infty \leq Z \leq -2.1) = P(2.1 \leq Z \leq \infty) = 1 - P(-\infty \leq Z \leq 2.1) = 1 - 0.982 = 0.0178$$

$$P(-2.18 \leq Z \leq \infty) = P(-\infty \leq Z \leq 2.18) = 0.985$$

$$P(0 \leq Z \leq 2.21) = P(-\infty \leq Z \leq 2.21) - P(-\infty \leq Z \leq 0) = 0.986 - 0.5 = 0.486$$

$$\begin{aligned}
 P(-2.21 \leq Z \leq 2.21) &= P(-\infty \leq Z \leq 2.21) - P(-\infty \leq Z \leq -2.21) = \\
 &= P(-\infty \leq Z \leq 2.21) - [1 - P(-\infty \leq Z \leq 2.21)] = 0.986 - [1 - 0.986] = 0.9728
 \end{aligned}$$

$$\text{Or in another way: } P(-2.21 \leq Z \leq 2.21) = 2 \cdot P(0 \leq Z \leq 2.21)$$

$$\begin{aligned}
 P(-1 \leq Z \leq 2.18) &= P(-\infty \leq Z \leq 2.18) - P(-\infty \leq Z \leq -1) = P(-\infty \leq Z \leq 2.18) - [1 - P(-\infty \leq Z \leq 1)] \\
 &= 0.985 - [1 - 0.841] = 0.826
 \end{aligned}$$

Exercise 3.6

Consider the Score X obtained by the students of Statistics seen in Exercise 1.4 and 2.5, and assume it is a continuous variable with Normal distribution (we know it isn't!) with mean and standard deviation as observed, i.e. $\mu=25.46$ and $\sigma=3.959$.

- If you pass, what is the probability that you get a score ≥ 27 ? (compare to Exercise 2.5)
- If you pass, what is the probability that you get a score ≤ 21 ?

Solution

a) We compute $\Pr(X > 27)$ as follows:

$$\begin{aligned}
 &\frac{27 - 25.5}{3.959} = 0.379 \\
 &\text{Standardize } x=27: z = \frac{27 - 25.5}{3.959} = 0.379 \\
 &\text{Area until 0.38 (from the table of the Normal): } 0.648
 \end{aligned}$$

$\Pr(X > 27) = \Pr(Z > 0.379) = 1 - 0.648 = 0.352$
 (can you say why this is lower than with the previous calculations?)

b) For $\Pr(X < 21)$:

Standardize $x=21$: $z = \frac{21 - 25.5}{3.959} = -1.137$
 Area until 1.14 (from the table of the Normal): 0.873

$\Pr(X < 21) = \Pr(Z < -1.137) = \Pr(Z > 1.137) = 1 - 0.873 = 0.127$

Exercise 3.7

The number of bottles of shampoo sold monthly by a certain drug store is a Normal random variable with mean 212 and standard deviation 40. Find the probability that the next month's shampoo sales will be:

- a) Greater than 200
- b) Less than 250
- c) Greater than 200 but less than 250

Solution

$$a) \Pr(X > 200) = \Pr\left(Z > \frac{200 - 212}{40}\right) = \Pr(Z > -0.3) = \Pr(Z < 0.3) = 0.618$$

$$b) \Pr(X < 250) = \Pr\left(Z < \frac{250 - 212}{40}\right) = \Pr(Z < 0.95) = 0.829$$

$$c) \Pr(200 < X < 250) = \Pr(-0.3 < Z < 0.95) = \Pr(Z < 0.95) - \Pr(Z < -0.3) = 0.829 - (1 - 0.618) = 0.447$$

Exercise 3.8

Consider exactly the same situation as in Exercise 3.4. This time you have 5 beds available. Compute the probability that beds will not suffice in this third situation.

Solution

In principle, we could proceed computing:

$$\Pr(Y > 5) = \Pr(Y=6) + \Pr(Y=7) + \dots + \Pr(Y=300) = 1 - \Pr(Y=0) - \Pr(Y=1) - \dots - \Pr(Y=5)$$

but this is quite boring to do with a pocket calculator! It is then useful to know that another approximation to the Binomial is with the Normal:

$$Y \sim N(\text{mean} = 3 \cdot 0.01 = 3, \text{std} = \sqrt{3 \cdot 0.01 \cdot 0.99}) = \sqrt{2.97} = 1.72$$

Now for $\Pr(Y > 5)$ we can proceed quickly, standardizing this value and looking up in the table the corresponding probability. The only new thing is to apply a **continuity correction**, to account for the fact that the random variable Y is discrete; proceed as follows:

compute $\Pr(Y = i)$ as $\Pr\{i - 0.5 \leq X \leq i + 0.5\}$

when you need $\Pr(Y > y)$ standardize $y - 0.05$

when you need $\text{pr}(Y < y)$ standardize $y+0.05$
when you need $\text{pr}(y_1 < Y < y_2)$ standardize $y_1-0.05$ and $y_2+0.5$

$y=5 \rightarrow$ standardize $5-0.05=4.5 \rightarrow z = (4.5-3) / 1.72 = 1.163 \rightarrow \Phi(z) = 0.877$
 $\text{Pr}(Y > 5) = 1 - 0.877 = 0.123$

Exercise 3.9

Suppose that 46% of the population favours a particular candidate to Mayor of the town. If a random sample of 200 citizens is chosen, what is the probability that at least 100 of them favour this candidate?

Solution

Indicating with X the number of persons who favour the candidate, then X is a Binomial random variable with parameters $n = 200$ and $p = 0.46$. The desired probability is $P\{X \geq 100\}$. It is definitely inconvenient to use the formula of the Binomial. Instead, we can use the Normal approximation, assuming:

$X \sim N(\text{mean}=200 \cdot 0.46=92, \text{var}=200 \cdot 0.46 \cdot (1-0.46)=49.68 \text{ ie. std}=7.05)$

Since the Binomial is a discrete and the Normal is a continuous random variable, it is better to apply a continuity correction (see Exercise 3.8). We will thus compute:

$100 \rightarrow 100-0.5 = 99.5 \rightarrow z=(99.5 - 92)/7.05 = 1.06 \rightarrow \Phi(1.06) = 0.855$
 $P(X > 100) = P(Z > 1.06) = 1-0.855 = 0.145$

Statistical inference: distribution of the sample mean; confidence intervals and hypothesis testing on the mean and the proportion in one sample**Exercise 4.1**

The blood cholesterol level of a population of workers has mean 202 and standard deviation 14. A sample of 36 workers is selected. Compute the probability that the sample mean of their blood cholesterol level will lie between 198 and 206. Repeat the exercise for a sample size equal to 64, and explain the change.

Solution

Indicate with $\bar{X}$ the average blood cholesterol level in the sample. It follows from the central limit theorem that $\bar{X}$ is approximately normal with mean $\mu = 202$ and standard deviation $\sigma = 14 / \sqrt{36} = 2.33$ (1.75 for $n=64$). Thus:

$P(198 < \bar{X} < 206) = P(z_1 < Z < z_2)$ being z_1 and z_2 the values obtained by standardization of 198 and 206 (on the right, the calculations for $n=64$):

$$198 \rightarrow (198-202)/2.33 = -1.71 \quad (-2.29)$$

$$206 \rightarrow (206-202)/2.33 = +1.71 \quad (2.29)$$

$$\Phi(1.71) = 0.956 \quad (0.989)$$

$$P(198 < \bar{X} < 206) = P(z_1 < Z < z_2) = 2 \cdot P(0 < Z < 1.71) = 2 \cdot (0.956 - 0.5) = 0.912 \quad (0.978)$$

With larger sample size, the variability of the distribution of the sample mean is reduced; thus the tails $\bar{X} < 198$ and $\bar{X} > 206$ have less probability.

Exercise 4.2

This exercise helps getting familiar with the quantiles that we use in the inferential procedures that involve using the Normal distribution. Use the table of the standard Normal to find the value(s) z (with precision equal to two decimal places) such that:

- a) $P(Z > z) = 0.05$
- b) $P(Z < -z) = 0.05$
- c) $P(Z > z) = 0.025$
- d) $P(z_1 < Z < z_2) = 0.95$
- e) $P(Z > z) = 0.01$
- f) $P(Z > z) = 0.005$
- g) $P(z_1 < Z < z_2) = 0.99$

Solution

- a) $P(Z > \underline{1.64}) = 0.05$
- b) $P(Z < \underline{-1.64}) = 0.05$
- c) $P(Z > \underline{1.96}) = 0.025$
- d) $P(\underline{-1.96} < Z < \underline{1.96}) = 0.95$
- e) $P(Z > \underline{2.33}) = 0.01$
- f) $P(Z > \underline{2.58}) = 0.005$
- g) $P(\underline{-2.58} < Z < \underline{2.58}) = 0.99$

Exercise 4.3

To check compliance with national regulatory, the city board of education needs to estimate the proportion π of women among all secondary school teachers. If there are 518 females in a random sample of 1000 teachers:

- Construct the confidence interval estimate for π at 95%, 90% and 99% confidence level.
- Before doing this, what do you expect: which confidence interval should be the largest one?
- How do you interpret the 95% CI?
- For example, if the regulatory fixes the target proportion of female teachers to be at least equal to 50%, what does the 95% CI tell about the compliance to this requirement?

Solution

The point estimate for the percent π is $p=0.518$. The various CI will be constructed multiplying different quantiles of the Standard Normal distribution by the standard deviation of the sample distribution, which for the case of inference on a proportion is estimated as $\text{Sqrt}(p \cdot (1-p)/n)=0.0158$.

Definition of a CI at $1 - \alpha$ confidence level for a proportion π , based on the sample estimate p :

$$\Pr\left(p - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

[The random quantity in this expression is the sample estimate p , while the parameter π is a fixed, although unknown, quantity. In other terms, repeated sampling will generate each time a different couple of extreme points of the confidence interval, such that they will include the fixed value π in $(1-\alpha)\%$ of the times]

The quantiles $z_{\alpha/2}$ are respectively:

- i) For the 95% CI: $1-\alpha=0.95$ ($\alpha=0.05$) so 1.96
- ii) For the 90% CI: $1-\alpha=0.90$ ($\alpha=0.10$) so 1.64
- iii) For the 99% CI: $1-\alpha=0.99$ ($\alpha=0.01$) so 2.58

Thus the CI are:

- i) (0.487, 0.549)
- ii) (0.492, 0.544)
- iii) (0.477, 0.559)

The confidence level expresses the “strength of believe” that the interval (our estimation) will include the true value (parameter of the population); a larger confidence level corresponds to a stronger trust in our interval, but the obvious ‘cost’ is that we will have a less precise estimate – i.e. a wider interval. So the largest CI is the one at 99% level.

The 95% CI = (0.487, 0.549) means that we estimate that the true percent of women among teachers in our population is between 48.7% and 54.9%. Thus, despite we might be fairly compliant to the regulation, with possibly almost 55% female teachers, it is yet possible that we have a slight gap to fill-in (49% is a possible value).

Exercise 4.3-b. A similar exercise is:

The most popular drug against menstrual pain induces complete disappearance of pain within 30 minutes from administration in 50% of the women. A pharmaceutical company invested money to produce a drug with larger efficacy. When the drug is tested on 1,000 women, 518 report the target outcome (pain disappeared within 30 min). Did the company achieve the objective? Answer by producing the 95% CI for the proportion π of pain disappearance.

We get (see above) that the interval is between 48.7% and 54.9. So the company is very close to be satisfied, but is not yet sure that the new drug gives more than 50% success.

Notice that this problem can be formulated as an hypothesis test, with $H_0: \pi=0.5$ vs. $H_1: \pi \neq 0.5$ (we would actually think of a one-sided test, $H_1: \pi > 0.5$; then instead of a too-large conventional $\alpha=5\%$ we should reduce α to 2.5%. This is equivalent to test against the two-sided H_1 at the total α level of 5%).

Due to the relationship between a 2-sided test at α level and the $(1-\alpha)\%$ confidence interval, we do not need to develop the calculations for the test: the null hypothesis cannot be rejected at 5% level, since the value 0.5 of the null hypothesis belongs to the 95% CI.

Exercise 4.4

Continue ex. 4.3-b: as an exercise, proceed anyhow to the test, and:

- compute the p-value
- compute the upper threshold of the rejection region on the original scale (instead of the standardized scale)

Solution

$H_0: \pi=0.5$ vs. $H_1: \pi \neq 0.5$

Test statistic, on the original scale: $p=518/1000=0.518$

Standard Error: $\text{Sqrt}(0.518 \cdot (1-0.518)/1000)=0.0158$ [*]

Test statistic, standardized: $z=(0.518-0.5)/0.0158=1.14$

$\Phi(1.14)=0.873$

p-value= $2 \cdot (1-0.873)=0.254$

Thresholds of the rejection region, on the standardized scale: ± 1.96 [our $z=1.14$ falls in the acceptance region]

Thresholds of the rejection region, on the original scale: $0.5 \pm 1.96 \cdot 0.0158 = 0.469; 0.531$ [we got 0.518, which falls in the acceptance region; in order to reject H_0 , we should have had $p > 0.513$, i.e. more than 513 successes out of 1,000 women tested]

[*] (Advanced)

Actually, under the null hypothesis, $\pi=0.5$ and thus we could assume that the standard deviation is $\text{Sqrt}(0.5 \cdot (1-0.5)/1000)$. The approach we follow, i.e. using the estimated percentage p instead of the value of π under the null hypothesis, is due to the fact that we have a very large sample and that we approximate the distribution of X with a Normal.

You could repeat the exercise using the standard deviation under H_0 .

Exercise 4.5

Medical students undergo a test with score varying between 0 and 100. It is known that the standard deviation of test scores is equal to 11.3, but the mean is unknown. A random sample of 81 students had a sample mean score of 74.6. Compute the 90% confidence interval estimate for the mean μ that represents the average score of all medical students.

Solution

Data: $n=81$ $\bar{x} = 74.6$ (sample value) $\sigma = 11.3$ (known population value)

From the central limit theorem - i.e. by general property of the random variable sample mean, its standard deviation is:

Standard error = $11.3 / \sqrt{81} = 1.256$

Having fixed the confidence level $1-\alpha = 0.90$, $z_{\alpha/2}=1.64$

Thus the 90% confidence interval for μ is (72.54, 76.66)

$\Pr(74.6 - 1.64 \cdot 1.256 \leq \mu \leq 74.6 + 1.64 \cdot 1.256) = 0.90$

Exercise 4.6

Traffic authorities claim that the duration of the red light of the town traffic lights is Normally distributed with mean equal to 30 seconds and standard deviation equal to 1.4 seconds. To test this feature a sample of 40 traffic lights's durations was checked, and the observed average duration was 32.2 seconds.

Can we conclude at the 5% level of significance that the authorities are incorrect?

What about using instead a 1% level of significance?

Solution

X =duration of traffic lights (measured in seconds). This random variable has a Normal distribution with mean μ unknown and standard deviation $\sigma=1.4$ (assumed to be known). About the mean, we want to compare the two hypotheses:

$H_0: \mu=\mu_0=30$ $H_1: \mu \neq 30$

So, under H_0 , $X \sim \text{Normal}(30, 1.4)$, and if we draw a sample of $n=40$ traffic lights' durations, and we take their sample average $\bar{X}$, this is a random variable with Normal distribution with mean 30 and standard deviation $1.4/\sqrt{40}=0.221$.

In our observed sample, we got an average $\bar{x}=32.2$. The test statistics (standardized) is thus:

$z=(32.2-30)/0.221=9.9$

We immediately realize that our observations tell that the authorities have completely wrong values! This is in fact much further than the values expected under the null hypothesis. The thresholds at 5% level are ± 1.96 . The p-value is in practice $=0$ (in publications, you might read $p<0.0001$). So we definitely conclude that the authorities are incorrect about the mean duration.

Now, even if we wish to answer to the same question being more cautious before saying that the traffic authorities "lie", and thus we choose a smaller significance level, equal to 0.01, our evidence that the true mean duration is longer than 0 seconds is so strong

(standardized sample mean=9.9) that we still reject the null hypothesis. In fact, with $\alpha=0.01$, the thresholds are the quantiles ± 2.58 , our $z=9.9$ is yet external.

Exercise 4.7

A certain disease is known to have a prevalence equal to 1%. However, in a random sample of $n=400$ members of the population, 11 people were found to have the disease. Test the hypothesis H_1 that the true prevalence π in this population is different than 0.01 against the null hypothesis that the prevalence is $\pi = \pi_0 = 0.01$.

Solution

$H_0: \pi = 0.01$ vs. $H_1: \pi \neq 0.01$

sample value (point estimate of π): $p = 11/400 = 0.0275$ [almost three times higher than what is reported in the literature: worth checking if this could be due to chance, or it is an evidence that something is ongoing in that population, such that there is more risk of that disease (or, if the disease rapidly evolves towards death, that instead in this population the diseased people survive longer, so that the prevalence is higher)]

standardized test statistics:

$$z = \frac{0.0275 - 0.01}{\sqrt{\frac{0.0275 \cdot 0.9725}{40}}} = \frac{0.0175}{0.0082} = 2.14$$

[See ex. 4.5, the note on the formula of the standard deviation ("Advanced")]

this value is significant at 5% level (it is higher than 1.96) but not at 1% level (it is lower than 2.58). more precisely, the p-value is $= 2 \cdot (1 - \Phi(2.14)) = 2 \cdot (1 - 0.984) = 0.032$

So there is good evidence, but not a very strong one, that this population is different from what reported in the literature.

Exercise 4.8

Using the data of exercise 1.4 on the score at the exam of Statistics for the cohort 2011-2012 (degree course in Italian) compute a confidence interval for the mean score (at 95% confidence level). Explain in words how we interpret the result. Then test the hypothesis H_1 that the mean score is different than 26, with a 2-sided test at 5% significance level.

Solution

We have already computed the sample mean and standard deviation (or, we compute them again):

Mean = 25.5 St.dev. = 3.959

The 95% CI is:

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(25.5 - 1.96 \cdot \frac{3.959}{\sqrt{171}}, 25.5 + 1.96 \cdot \frac{3.959}{\sqrt{171}} \right) = (24.9, 26.1)$$

It means that we expect an average score between 24.9 and 26.1 (and our estimation is wrong – in the sense that the true average score is lower than 24.9 or larger than 26.1 - in 5% of the cases).

Notice that it does not mean that all students will get a score between 24.9 and 26.1!! this range is valid for the mean score in a population of students.

We now don't need to compute a T statistic and check with the rejection region or p-value: since the 95% CI includes $\mu_0=26$, we know that the null hypothesis $H_0: \mu=26$ is accepted at 5% level when compared to the 2-sided alternative $H_1: \mu \neq 26$.

Anyhow, we WILL proceed with the test T to verify this conclusion:

$$t = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{25.5 - 26}{3.959/\sqrt{171}} = -1.65$$

This value falls in the acceptance rejection region, delimited (for $\alpha=0.05$) by -1.96 and 1.96.

We also compute the p-value:

$$p = 2 \cdot (1 - \Phi(1.65)) = 2 \cdot (1 - 0.951) = 0.098$$

thus it is confirmed that the sample value 25.5 does not differ significantly from the null value 26.

Statistical inference & the analysis of associations: hypothesis testing to compare the mean and the proportion in two samples and the linear association of two continuous variables

Exercise 5.1

An official ministry office is willing to check whether woman receive lower salaries than men, while being at the same level of experience, expertise, etc. To this purpose they set up a survey among a well-defined population of people with a history of regular employment during the last 10 years, similar experience, similar education etc. These are the results (wages in K euros/year, gross salaries):

Men (group 1): $n_1=72$, mean $\bar{x}_1=12.2$, std = 1.1

Women (group 2): $n_2=55$, mean $\bar{x}_2 = 10.8$, std=0.9

Assume that the variances of the two populations of men and women are equal, and set up a test, writing the null and the alternative hypotheses, computing the p-value, and drawing a conclusion.

Solution

We could set up the test as a one-sided one, focusing on the main alternative hypothesis that men have higher salaries than women - this is in fact a well-known social problem, so, even without seeing the results from the sample, we would have excluded a-priori that the difference between men and women was in the other direction.

Similarly, in some clinical trials that use a placebo as a control treatment, it is usually possible to test one-sided that the experimental arm using the active drug will have larger efficacy (and also larger toxicity) than the placebo arm.

So we could test:

$H_0: \mu_1=\mu_2$ vs. $H_1: \mu_1>\mu_2$

For this test, we should use a small alpha, e.g. 0.025 or smaller.

Alternatively, and equivalently, we could test the following hypotheses, at alpha level 0.05:

$H_0: \mu_1=\mu_2$ vs. $H_1: \mu_1\neq\mu_2$

The test statistic needs the following estimate of the common (unknown) value of the standard deviation of the population:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = s = \sqrt{\frac{(72 - 1)1.1^2 + (55 - 1)0.9^2}{72 + 55 - 2}} = 1.0184$$

then:

$$z = \frac{12.2 - 10.8}{1.0184 \sqrt{\frac{1}{72} + \frac{1}{55}}} = 7.68$$

The p-value associated with the test statistic $z=7.68$ is approximately equal to 0, thus in practice lower than any α level that we could choose; so we always reject the null hypothesis, and conclude that there is a very strong evidence that women' wages are (on average) lower than men'.

Exercise 5.2

A public health office detailed the outcome of a prevention project, where 260 elderly people were advised to have a vaccine against flu. A total of 184 agreed to have the vaccine, while the other 76 declined. At the end of the flu season the following outcomes were collected:

	Vaccine	No vaccine
Got flu	10	6
No Flu	174	70

Do the data provide evidence that the people receiving the vaccine had a different chance of contracting the flu from those not receiving the vaccine? Compare these probabilities by an appropriate measure of the effect of "exposure" to vaccine, then test the hypothesis of absence of relation at 5% significance level.

** More advanced (if introduced during the course: repeat the test and compute the p-value using a T-test for two proportions).*

Solution

The probability of getting the flu in the two groups were:

	Vaccine	No vaccine	Tot
Got flu	10	6	16
No Flu	174	70	244
Tot	184	76	260

$$P(V) = 10/184 = 0.054$$

$$P(NV) = 6/76 = 0.079$$

$$\text{Risk Ratio} = 0.054/0.079 = 0.69$$

So the elderly people who got the vaccine had 30% less probability of getting the flu with respect to the elderly people who did not receive the vaccine.

The next step is checking whether this is due to chance (H_0 : no association between vaccine administration and occurrence of flu) or it is statistically significant (H_1 : there is an association, the prob. of flu is different according to administration or not of vaccine).

$$H_0: \pi_V = \pi_{NV} \quad \text{vs.} \quad H_1: \pi_V \neq \pi_{NV}$$

In absence of any confounding factor, if we can reject H_0 in favour of H_1 , we could interpret the result in causal terms, ie claim that the vaccine is effective in preventing the flu.

To test for the difference between the two proportions, we can use the X^2 test. We need the table of the absolute frequencies that we would expect under the null hypothesis:

	Vaccine	No vaccine	Tot
Got flu	$184 \cdot 16 / 260 = 11.32$	$16 \cdot 76 / 260 = 4.68$	16
No Flu	$184 \cdot 244 / 260 = 172.68$	$244 \cdot 76 / 260 = 71.32$	244
Tot	184	76	260

(notice that the totals remain the same as in the original table; this should always happen, if not, and it is not a small discrepancy possibly due to rounding, the calculations are wrong)

The test statistic is the sum of the terms $(\text{observed} - \text{expected})^2 / \text{expected}$:

$$\chi^2 = \frac{(10 - 11.32)^2}{11.32} + \frac{(6 - 4.68)^2}{4.68} + \frac{(174 - 172.68)^2}{172.68} + \frac{(70 - 71.32)^2}{71.32} = 0.564$$

The threshold of the region region (or critical value) is determined by the degrees of freedom, a characteristic of the table, here $(r-1) \cdot (c-1) = (2-1) \cdot (2-1) = 1$, and by the significance level; for $\alpha=0.05$, it is 3.84.

Since our χ^2 is < 3.84 , we do not reject the null hypothesis at the level of significance 0.05.

* *T-test for two proportions:*

We compute an overall estimate of the probability of flu, and use it to compute a standard deviation for the difference between the two proportions:

$$\text{overall } p = 16 / 260 = 0.062$$

(notice: it is found also as a weighted average of the subgroup probabilities:

$$p = \frac{0.054 \cdot 184 + 0.079 \cdot 76}{184 + 76})$$

The test statistic is:

$$z = \frac{0.054 - 0.079}{\sqrt{0.062 \cdot (1 - 0.062) \left(\frac{1}{184} + \frac{1}{76} \right)}} = -0.76$$

which lies in the acceptance region delimited by ± 1.96 (the two tests should return the same conclusion).

The p-value is found from the table of the Normal distribution, as double the area of the tail: $2 \cdot (1 - \Phi(0.76)) = 2 \cdot (1 - 0.776) = 0.448$

Exercise 5.3

A survey investigates on the preferences of Tor Vergata students in terms of computer operating system and type of mobile telephone. The following table represents the results collected on a random sample of size 500.

	Smartphone	Mobile Phone	Tot
MAC OS	102	52	154
WINDOWS	238	40	278
LINUX	52	16	68
Tot	392	108	500

Are the two variables independent at the level of significance $\alpha=0.05$? What about $\alpha=0.01$? and how do you interpret your answers?

Solution

H0: The two categorical variables are independent (i.e., there is no relationship, or no association, between them)

H1: The two categorical variables are dependent (i.e., there is a relationship, or association, between them)

We use a chi-squared test, whose test statistic is the sum of the terms $(\text{observed} - \text{expected})^2 / \text{expected}$, and the expected frequencies are:

Table of the expected frequencies:

	Smartphone	Mobile Phone	Tot
MAC OS	$392 \cdot 154 / 500 = 120.7$	$154 \cdot 108 / 500 = 33.3$	154
WINDOWS	$392 \cdot 278 / 500 = 218.0$	$278 \cdot 108 / 500 = 60.0$	278
LINUX	$392 \cdot 68 / 500 = 53.3$	$108 \cdot 68 / 500 = 14.7$	68
Tot	392.0	108.0	500

Terms of the sum:

2.91	10.55
1.84	6.69
0.03	0.12

$$\chi^2 = 22.15$$

The degrees of freedom are $(r-1) \cdot (c-1) = (3-1) \cdot (2-1) = 2$.

The critical value for 2 df and $\alpha=5\%$ is 5.991, while for $\alpha=1\%$ it is 9.21. Thus our chi-squared is highly significant, and allows us to reject the null hypothesis both at 5% level and at 1% level (the latter test is a "less cautious" one, i.e. requires stronger evidence to reject H0).

We can conclude that the two variables are dependent or associated. This means that the proportion of use of the three operative systems is different depending on the type of device (smartphone or traditional mobile phone) used. Just for descriptive purposes, let's compute for example the percent of use of each system within the groups:

	Smartphone	Mobile Phone
MAC OS	$102/392 = 26\%$	$52/108 = 48\%$
WINDOWS	$238/392 = 61\%$	$40/108 = 37\%$
LINUX	$52/392 = 13\%$	$16/108 = 15\%$
Tot	100%	100%

Exercise 5.4

Take the mean and the standard deviation of the score at the exam of Statistics for the 2 cohorts of students seen in exercise 1.5 (and before, in exercise 1.4 and 4.8):

- cohort 2011-2012 (Italian; n=171): mean=25.5, std=3.959
- cohort 2012-2013 (English; n=16): mean=26.4, std=4.11

Do we have evidence to conclude that there is a difference between the two cohorts? Test the hypothesis both computing the p-value and using the rejection region at 5% significance level.

Solution

Indicating with A the “Italian” students and with B the “English” students, we will compare the null and alternative hypotheses on the difference of their mean score:

$$H_0: \delta = \mu_A - \mu_B = 0 \quad \text{vs} \quad H_1: \delta = \mu_A - \mu_B \neq 0$$

We will use the t-test, thus computing:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(171 - 1)25.5^2 + (16 - 1)26.4^2}{171 + 16 - 2}} = 3.97$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{25.5 - 26.4}{3.97 \sqrt{\frac{1}{171} + \frac{1}{16}}} = -0.87$$

This is clearly a small (standardized) difference, very close to 0 – and definitely included in the acceptance region at 5% significance level (with threshold ± 1.96).

To compute the p-value for assessing the significance of the observed difference, we look up in the table of the standard Normal the value corresponding to $z=0.87$: the area is 0.808; thus the area in one tail is $1-0.808=0.192$, and the total probability of the tails (2-sided test) is $p=0.384$.

Thus from the data we have we cannot reject the hypothesis that there is no difference between the “Italian” and “English” cohorts.

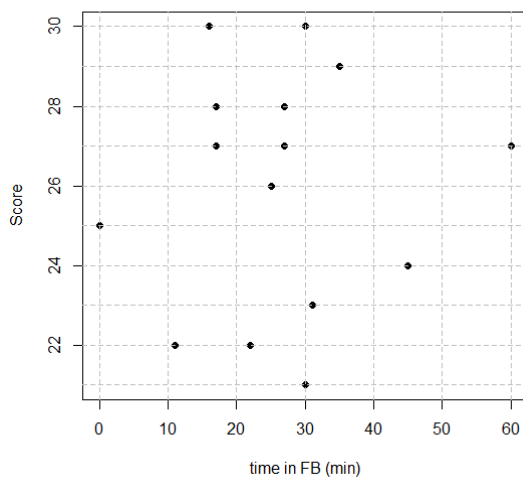
Exercise 5.5

Represent and analyze the (linear) relationship between time spent using Facebook and the grades of Statistics exam seen in exercise 1.1, in particular assessing if the time spent on Facebook affects the score in Statistics (discuss the results).

Solution

- To represent the association between the two quantitative variables, we use a scatter plot (important: we must use the initial table and not the ones where the two variables were sorted for the marginal analysis).
- To assess the degree of linear association, we can compute the correlation coefficient.
- To assess the impact of FB on the score, we can compute the equation of the line that interpolates the points, and test the slope.

Notice that, considering the latter purpose of our analysis, the scatter-plot is more informative if we put FB time on the horizontal axis and the Statistics score on the vertical axis.



FB X	Stat Y	Xi-mean(X)	Yi-mean(Y)	(Xi-m(X))*(Yi-m(Y))
0	25	0-26.2=-26.2	25-25.9=-0.9	(-26.2)*(-0.9)=23.6
11	22	11-26.2=-15.2	22-25.9=-3.9	(-15.2)*(-3.9)=59.3
17	28	17-26.2=-9.2	28-25.9=2.1	(-9.2)*(2.1)=-19.3
16	30	-10.2	4.1	-41.8
22	22	-4.2	-3.9	16.4
17	27	-9.2	1.1	-10.1
25	26	-1.2	0.1	-0.1
30	21	3.8	-4.9	-18.6
27	27	0.8	1.1	0.9
27	28	0.8	2.1	1.7
31	23	4.8	-2.9	-13.9
35	29	8.8	3.1	27.3
30	30	3.8	4.1	15.6
45	24	18.8	-1.9	-35.7
60	27	33.8	1.1	37.2

$$\text{Cov}(X,Y)=42.2/15=2.64$$

(notice: we can see in this exercise how much rounding the calculations can affect the results; here, by using 25.9 instead of the correct mean 25.9333 for Y, implies a large error, which we can detect by making the sum of the "errors" $Y_i - \bar{Y}$, which is not =0 as it should be by construction, but is equal to 0.5)

index of linear relation:

$$r(X,Y)=2.64/(14.23 \cdot 2.96)=0.06$$

Comment: the linear relationship between the score in the exam of statistics and the time spent on Facebook is very weak.

Slope of the regression line:

$$b=2.64/(14.23)^2=0.013$$

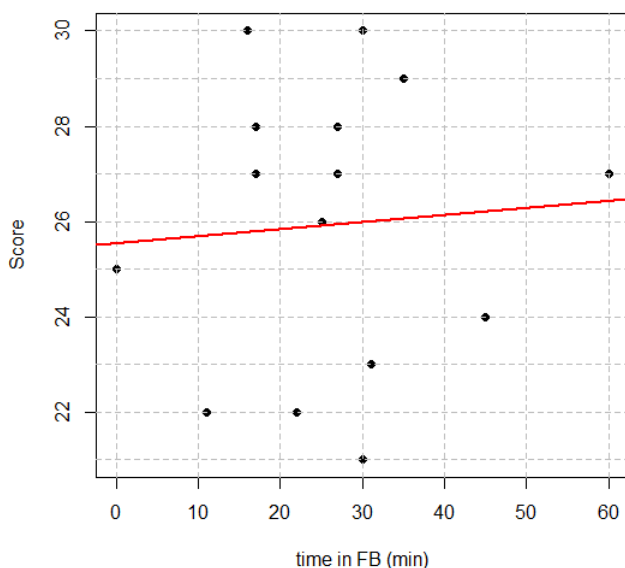
This means that when the time on FB (X) increases by 1 minute, the score increases by 0.01. so we need an increase of 100 minutes to detect an increase of 1 point in the score.

Intercept of the regression line:

$$a=25.9-0.013 \cdot 26.2 = 25.6$$

Thus when a student spends every day 1 hour in FB, he/she can expect to have a score given by:

$$y(x=60)=25.6+0.013 \cdot 60= 26.3$$



(here we plot the exact regression line, computed with no rounding errors: $\beta=0.01488$, $\text{intercept}=25.54353$)

This regression line is useful if we can "believe" that there is indeed a relationship such that when time in FB increases, the score also increases, although, as we saw, little. We thus need to test the significance of the slope.

The test statistic is the estimated slope $b=0.013$, properly standardized.

For this, we first need to compute (with a slight adjustment at the denominator) the variance of the "residuals", which are the differences between the observed Y_i and the expected $Y(x_i)$ i.e. the value for Y predicted on the line given x_i :

X	Y	Y(x)	y-Y(x) (residual)	residual^2
0	25	25.54	-0.54	0.29
11	22	25.71	-3.71	13.73
17	28	25.80	2.20	4.86
16	30	25.78	4.22	17.81
22	22	25.87	-3.87	14.98
17	27	25.80	1.20	1.45
25	26	25.92	0.08	0.01
30	21	25.99	-4.99	24.90
27	27	25.95	1.05	1.11
27	28	25.95	2.05	4.22
31	23	26.01	-3.01	9.03
35	29	26.07	2.93	8.61
30	30	25.99	4.01	16.08
45	24	26.22	-2.22	4.91
60	27	26.44	0.56	0.31
tot				122.30

$$s_{RES}^2 = \frac{\sum (y_i - Y(x_i))^2}{n-2} = 122.30 / 13 = 9.41$$

Then we compute the standard deviation of b , as:

$$SE(b) = \frac{s_{RES}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s_{RES}}{std(x)\sqrt{n}} = \frac{\sqrt{9.41}}{14.23\sqrt{15}} = 0.056$$

Thus, test statistic = $0.013/0.056 = 0.270$. This is inside the acceptance region at 5% level, delimited (as usual) by ± 1.96 . We can also compute the p-value = $2 \cdot (1 - \Phi(0.056)) = 2 \cdot (1 - 0.520) = 0.96$.

So the association is highly NON significant, and we can conclude that according to our survey there is no association between time spent in FB and score in Statistics.

Exercise 5.6

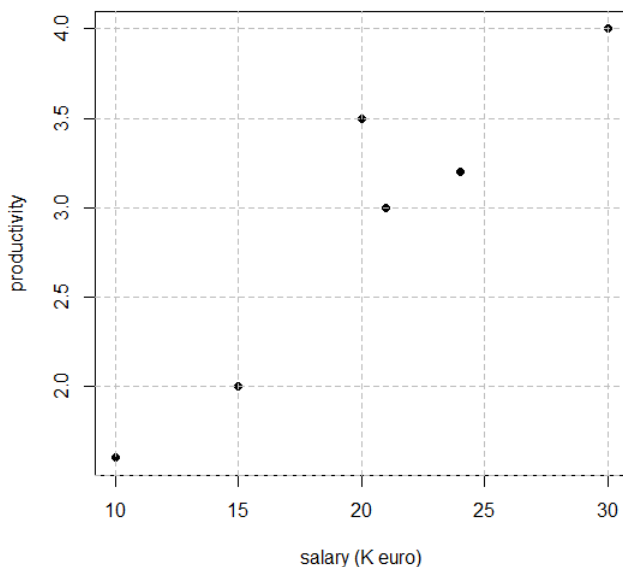
The data in the following table show the annual salary (in K euro) X and a measure Y of the productivity of a sample of employees of a company. Some computations are done already in extra columns.

Salary X	Productivity Y	$X_i - m(X)$	$Y_i - m(Y)$	$(X_i - m(X)) \cdot (Y_i - m(Y))$
10.0	1.6	-10.0	-1.3	12.8
15.0	2.0	-5.0	-0.9	4.4
20.0	3.5	0.0	0.6	0.0
21.0	3.0	1.0	0.1	0.1
24.0	3.2	4.0	0.3	1.3
30.0	4.0	10.0	1.1	11.2

Represent the data in a suitable form to show a possible association, then measure it (by the linear correlation coefficient) and find an equation to represent it; furthermore, assess its significance by computing the p-value. Finally, compute the expected productivity of an employee who earns 25 thousand euros.

Solution

Scatter-plot of the observations: each point has coordinates equal to the couples of X and Y observed.



sample means and standard deviations:

$$m(x) = 20 \quad \text{Std}(X) = 6.96$$

$$m(y) = 2.88 \quad \text{Std}(Y) = 0.91$$

covariance and correlation coefficient:

$$\text{cov}(X, Y) = (12.8 + \dots + 11.2)/6 = 29.8/6 = 4.97$$

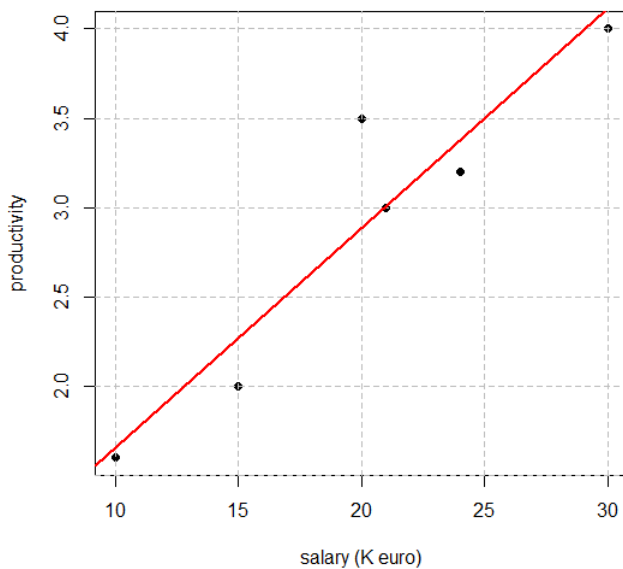
$$r(X, Y) = 4.97 / (6.96 \cdot 0.91) = 0.94$$

The linear correlation is almost perfect between X and Y, since r approaches to 1, meaning that there is a strong positive linear relationship between the productivity and the salary.

The line that interpolates the data in the population $Y = \alpha + \beta X$ thus has a positive slope and represents well the data*; it is found as:

$$\text{slope: } b = 4.97 / 6.96^2 = 0.103$$

$$\text{intercept: } a = 2.88 - 0.103 \cdot 20 = 0.83$$



(graph obtained using the precise estimation: $b=0.1231$, $a=0.4205$)

* advanced: the squared of the correlation coefficient measures the "goodness of fit", i.e. how good is the representation of the observed points using the regression line; it varies from 0=very bad to 1=perfect. This index is also defined in other, more complex regression methods, and although it loses the relationship with the simple correlation coefficient, it is still called R^2 .

We have $R^2=0.88$: a very good representation, as it is evident from the graph.

Let's see if we have sufficient power to exclude that this linear relation is due to chance, which implies that in another sample we would not find it again (it seems rather unlikely that our data are significant, since we have a very small sample size; however, because the linear relation is so strong, i.e. relevant, we could find significance).

We test:

H_0 : no linear association, $\beta=0$ versus H_1 : presence of a linear relation, $\beta \neq 0$

To standardize the test statistics $b=0.103$ we need first to compute the residuals and their estimated variance, and then the standard deviation of b .

X	Y	$Y(x)$	$\text{res}=y-Y(x)$	res^2
10	1.6	1.857	-0.257	0.066
15	2	2.370	-0.370	0.137
20	3.5	2.883	0.617	0.380
21	3	2.986	0.014	0.000
24	3.2	3.294	-0.094	0.009
30	4	3.909	0.091	0.008
tot				0.600

SSRes=0.150
Std(b)=0.022

Standardized test statistics: $t=4.51$

So the p-value is <0.0001 : the association is highly significant, i.e. the data support the hypothesis that in general with higher salary the productivity is also high.

On this basis, the expected productivity of an employee earning $X=25$ is:
 $Y = 0.83 + 0.103 \cdot 25 = 3.405$

Exercise 5.7

Continues exercise 5.6. The executive director of the company does not like the conclusion that if you pay the employees better, they are more productive. He/she has the following data regarding productivity and salaries in relation to other factors, observed in a larger group of employees: could you suggest him/her some additional analysis to support a position against a salary increase, and find other solutions to increase the productivity?

Impact of Salary on Productivity based on a regression line	$b=0.12$	p-value <0.0001
Salary according to gender	Salary M: average=22 Salary F: average=19	p-value=0.01
Productivity according to gender	Productivity M: average=2.7 Productivity F: average=2.9	p-value=0.31
Salary according to experience (measured in years)	Salary Exp <5 yrs: average=18 Salary Exp ≥ 5 yrs: average=24	p-value=0.002
Productivity according to experience (measured in years)	Productivity Exp <5 yrs: average=2.4 Productivity Exp ≥ 5 yrs: average=3.2	p-value=0.012
Salary according to training (measured in hours/year)	Salary <20 h: average=19.1 Salary ≥ 20 h: average=21.3	p-value=0.010
Productivity according to training (measured in hours/year)	Productivity <20 h: average=2.5 Productivity ≥ 20 h: average=3.0	p-value=0.030

Solution

It should be checked whether the relation between higher salaries and higher productivity holds given the experience; in fact, experience could be a confounding factor, being longer experience associated to both higher salaries and higher productivity. Similarly, also training could be a confounder. Gender is not a possible confounder, since it is not associated to productivity.

After checking these assumptions, it could be possible to increase the productivity (not just by increasing the salaries, but also) by giving more training and making sure that the more experienced employees remain in the staff.