COMPLEMENTARY EXERCISES WITH DESCRIPTIVE STATISTICS

EX 1

Given the following series of data on Gender and Height for 8 patients, fill in two frequency tables one for each Variable, according to the model below. Then add a graph. Then create a contingency table end describe the relation between Gender and Height using appropriate statistical summaries.

id	Height, cm	Gender: 1=M, 2=F
1	165	M
2	157	F
3	168	F
4	178	M
5	171	F
6	182	M
7	182	M
8	153	F

modality	Absolute freq.	Percent freq.	Cumulative freq.

EX 2

186 patients were given a therapy for a certain disease; 122 had therapy A, the other 64 had therapy B.

In group A, there were 37 responders (patients who had benefit from the therapy). In group B, there were 32 responders. Which was the best treatment? How can we measure the advantage with this treatment?

Among responders, how many had treatment B?

EX 3

Compute the mean and the median class for the following distribution of the number of nurses in 23 medical institutes.

nurses	n
1 – 10	6
11 – 20	13
21 – 40	4
	23

EX 4

We know the values of haemoglobin for 6 patients before and after a course of chemotherapy: we wish to compute the mean reduction. What is the relation between the latter and the means of the values "before" and "after"?

before	after
13.0	9.4
12.8	11.5
11.0	11.5
13.2	13.1
12.5	10.2
11.9	12.0

EX 5

A certain treatment is used in two different centres, A and B; patients in centre A were 25 and were on average 54 years old; patients treated in centre B were 62 and had mean age equal to 58 years. What is the overall mean among all patients who got the treatment?

EX 6

Pregnant women (within month 4) who are being followed-up by a nutritionist had weights (kg) equal to: 64.3; 65.2; 70.0; 54.5; 58.8; 81.5; 61.0; 62.0. What was the mean? and the median? Do data suggest a strong skewness of the distribution of the Weight?

EX 7

The following data regard 10 male adults; we consider age, value of FEV1 (Forced Espiratory volume in 1 second) and diastolic pressure. Compute median and standard deviation of the three variables. Then tell what is the variable with higher variation, using an appropriate measure for comparisons.

Age	FEV1	pressure
25	2.5	85
32	1.8	71
28	1.5	92
21	2.5	80
33	4.5	87
33	2.1	83
34	3.4	70
24	1.2	101
41	2.8	90
26	3.9	83

EX 8

The Age quartiles in a sample of participants in a clinical trial were respectively 27, 41 e 59.

a) This means that:

- 1 out of 4 was younger than … years
- 1 out of 4 was older than … years
- 2 out of 4 were between … and … years old
- half of them was more than … years old

b) Additionally we know that mean and standard deviation were respectively equal to 42 and 12. Can we say whether the distribution was approximately Normal or not?

c) Which index of position is appropriate to give a synthetic description of the distribution?

EX 9

Consiedr the 6 patients with values of haemoglobin before and after chemotherapy of EX 4. We computed the means: resepctively 12.40 and 11.28 - and thus 1.12 for the reduction.

Now compute the standard deviation for the variable Before, After and for the Reduction (Before - After): does the linearity property hold?

EX 10

The weight distribution of a sample of adults with physical inabilities is approximately Normal, with mean 72 and standard deviation 8. Find an interval of values around the mean such that:

a) includes 95% of the observed values

b) includes almost all observed values (and thus coincides with the range, min-max)

c) includes 50% of the observed values

SOLUTIONS

EX 1

Variable Gender:

modality	Absolute freq.	Percent freq.	Cumulative freq.
M	4	50%	
F	4	50%	
tot	8	100%	

* For Gender we DON'T compute cumulative frequencies as it is a non-ordered qualitative variable. We will compute the cumulative frequencies for Height, which is quantitative (and thus necessarily ordered) and continuous.

For Gender, an appropriate graph is with columns (or bars), i.e. with two separate rectangles one for each modality, M and F, with height proportional to the percentage. It is in general correct to use a vertical axis going from 0% to 100% in order to avoid a distorted perception of the importance of the frequencies.

For Height, we consider a division in classes. Let us assume we know the minimum (140 cm) and the maximum value (200 cm) *(notice: a different choice of the extremes as well as of the number and width of the classses will lead to results slightly different from the following)*:

modality	Absolute freq.	Percent freq.	Cumulative freq.	Class width	Frequency density*
140 - | 160	2	25%	25%	20	= 2/20=0.1
160 -| 170	2	25%	50%	10	= 2/10=0.2
170 -| 200	4	50%	100%	30	= 4/30=0.13
tot	8	100%			

The appropriate graphical representation is the histogram, drawn on a cartesian diagram, putting the classes on the horizontal axis, and drawing for each class a rectangle with height equal to the frequency density of the class, so that the area of the rectangle equals the frequency of the class. (Notice the difference with the column chart used for qualitative variables: rectangles here are contiguous and their area, not their height, is equal to the frequency)

Contingency table:

	Height			
Gender	140 - | 160	160 -| 170	170 -| 200	Tot
M	0	1	3	4
F	2	1	1	4
tot	2	2	4	8

To describe the relation between Gender and Height, we can compute separately for calcoliamo M and F the percentages for each age class: these are also called 'row profiles' or conditional distributions of Height (conditional on Gender, i.e. "restricted to" a specific gender):

	Height		
Gender	140 - | 160	160 -| 170	170 -| 200
M	0%	25%	75%
F	50%	25%	25%

This table suggests that M are taller than F. Let us observe also that for Males the Mode is the class 170 -| 200, while for Females the Mode is 140 - | 160.

EX 2

We fill-in a table with the data given (grey cells) and we complete the missing cells:

	Response		
Treatment	no	yes	Tot
A	85	37	122
B	32	32	64
tot	117	69	186

(e.g. 85=122-37, 69=37+32. It is a good exercise to repeat what we read in the table, e.g. "there were 122 patients in group A, 37 of them responded to treatment. In total, 189 patients were treated, and 69 of them responded")

Now we compute the appropriate percentages:

	Response		
Treatment	no	yes	Tot
A	85/122=69.7%	37/122=30.3%	100%
B	32/64=50.0%	32/64=50.0%	100%
tot	117/186=62.9%	69/186=37.1%	100%

The best treatment seems to be treatment B. It is intuitive, and it will be seen in the course, that to compare the percentages in two groups it is useful to compute the ratio. This measure of comparison of percentages or probabilities is called Risk Ratio (*see Probability. The comparison of numbers via a ratio is illustrated in the Appendix I, among Prerequisites*):

RR=50/30.3=1.65

Thus B has a response percentage superior by 65% with respect to treatment A.

Among responders, the percent of those who got treatment B was:

32/69=46,4%

This percentage is obtained by considering the column profile, or in other terms the distribution of Treatment conditional on Response equal to Yes.

EX 3

The variable Number of Nurses observed in a sample of 23 medical institutes (statistical units) is a quantitative discrete variable, which we can treat as if it was continuous since it has many modalities (the numbers from 1 to 40); in fact the distribution is described by frequencies in classes. (Let us remark that usually the classes should be contiguous while here there are gaps in between: this is due to the discrete nature of the variable. This fact has no consequences in our exercise but it could be annoying when making a graphical representation. For example, in order to make a representation using an histogram, e.g. the class 1-10 should be considered as 1 |- 11; if we adopt this convention in our exercise, the mid-values are different from the ones used below)

To compute the mean, we chose a representative value for each class: we take the value in the middle, computed as (lower limit + upper limit)/2. The total amount of nurses for each class is then found as this mid-value times the frequency. The mean is the total amount among the classes divided by the total sample size (the number of statistical units) which is 23.

To identify which is the median class, or in other terms the class that contains the median, we use the cumulative frequencies.

nurses	n	value x_i	$x_i \cdot n_i$	Cumulative Freq
1 – 10	6	5.5	33.0	6
11 – 20	13	15.5	201.5	19
21 – 40	4	30.5	122.0	23
	23		356.5	

Mean=365.5 / 23 = 15.5

Median: modality that occupies the rank 12. Looking at the column of cumulative frequencies, it belongs to the class 11-20 (in fact, the first class includes only the first 6 units; as another example, the modality that occupies the 20th rank belongs to the class 21-40)

EX 4

The Reduction is the difference between the value Before and the value After; in some case a reduction is negative, this happens since the variable X (haemoglobin) has actually increased.

We can compute the reduction for each of the six patients (statistical units) and then compute their average. Another way to compute the mean reduction is by using the property of LINEARITY: given the variables X and Z, if we apply a linear transform such as

Y=aX+bZ

it is always true that mean(Y)=a·mean(X)+b·mean(Z).

In our exercise, a=1 and b=(-1), and thus the mean of the difference is the difference of the means:

mean(Before-After)= mean(Before)-mean(After). All computations are in the table below.

Notice: the demonstration of such properties will not be a subject of the course tests, as in general the theoretical results are not part of the knowledge required to the Students. This exercise is illustrated as a complement to the classes.

Another exercise in this document uses again the property of linearity, in the version: $mean(a+bx) = a+b\bar{x}$. *Think of a situation when it could be useful to compute the mean of a+bX when only a, b and mean(X) are known.*

before	after	reduction

	13.0	9.4	3.6
	12.8	11.5	1.3
	11.0	11.5	-0.5
	13.2	13.1	0.1
	12.5	10.2	2.3
	11.9	12.0	-0.1
Sum	74.4	67.7	6.7
sum/6	12.4	11.28333	1.116667
		12.4-11.3=	1.116667

EX 5

We need to compute a weighted average, i.e. the average of two means (54 and 58) weighted by the size of the two groups (25 and 62).

overall Mean = (54·25 + 58·62) / (25+62) = 4946 / 87 = 56.85

EX 6

We can sort the observed values and identify the values that occupies the position (rank) 4 and 5 (since we have 8 statistical units - pregnant women).

A slightly different way of illustrating this same procedure is assigning to each value the corresponding rank in the following table (in other terms, we avoid to write down the values sorted, but we need to sort to assign the rank!):

value x_1	rank r_i
64.3	5
65.2	6
70	7
54.5	1
58.8	2
81.5	8
61	3
62	4

Sum of values = 517.3 → Mean = 517.3 / 8 = 64.66

Central values (look at ranks 4 and 5): 62 e 64.3 → Median = (62 + 64.3) / 2 = 63.15

Since Mean and Median are not very far from each other, the data don't suggest that the variable Weight has a strongly skewed distribution.

EX 7

We have 3 quantitative continuous variables. Mean (arithmetic mean) and standard deviation provide a synthesis of position and variability. The mean is computed as the sum of all values divided by 10 (n=10 sample size). Let us apply the "fast" formula for computing the standard deviation. Computations are reported in the table.

To compare variability of these three variables it is NOT sufficient to look at standard deviations, that by the way are expressed in different units of measurement and describe variables with

different nature! We must express variability in relative terms with respect to the mean, using the coefficient of variation. The variable with highest variability is FEV1, 4 times more variable than Pressure and 2 times more variable than Age (notice that FEV1 had apparently the smallest value for the standard deviation ...)

id	Age	FEV1	pressure	Age^2	FEV1^2	pressure^2
1	25	2.5	85	625	6.25	7225
2	32	1.8	71	1024	3.24	5041
3	28	1.5	92	784	2.25	8464
4	21	2.5	80	441	6.25	6400
5	33	4.5	87	1089	20.25	7569
6	33	2.1	83	1089	4.41	6889
7	34	3.4	70	1156	11.56	4900
8	24	1.2	101	576	1.44	10201
9	41	2.8	90	1681	7.84	8100
10	26	3.9	83	676	15.21	6889
Sum	297	26.2	842	9141	78.70	71678
sum/10	29.7	2.62	84.2	914.1	7.87	7167.8
			Variance	35.57	1.12	86.84
			st. dev.	5.96	1.06	9.32
			cv	20%	40%	11%

EX 8

Point a):

1 out of 4 was younger than 27 years: this is the definition of first quartile, ¼=25% of observed values was lower than Q1=27

1 out of 4 was older than 59 years: similarly, this is the definition of third quartile, ¾=75% of observations was lower than Q3, and the other 25% was larger than Q3=59

2 out of 4 are …. We can claim: "between 0 years and the median 41 years", but also "between Q1 and Q3" and also "between the median 41 and the maximum age" (but we don0t know the latter).

Half of them was more than 41 years old: this comes from the definition of the Median.

b) First we notice that the mean is 42 and it is very close to the median, in fact their distance (equal to 1) is small (1/12) compared to the standard deviation. Thus the observed distribution is rather symmetric. But the Normal is not the only kind of symmetric distribution; we can go further by looking at the quartiles. In a Normal curve the first and third quartile should be at a specific distance from the mean, given by 0.67 times the standard deviation, thus $0.67 \cdot 12 = 8$. Thus if the observed distribution was approximately Normal the first and third quartiles were expected to be 34 e 50. Our observed quartiles are instead 27 and 59, rather distant from the ones of a Normal distribution with the same mean and standard deviation. Thus, our distribution is NOT Normal-shaped; it was symmetric but not bell-shaped; it could have been a distribution with very high tails and few observations in the middle, possibly a distribution with two modes..

c) Given what we just remarked, neither the mean nor the median are good indexes to describe the distribution; if it was bi-modal, we should use the two modes, and if we could recognize the presence of two subpopulations, we should use the means or the medians of these subpopulations..

EX 9

The computation of the standard deviations for Before and After is let for the Student; the results are 0.822912 and 1.313646 respectively. For the Reduction, using the "fast" formula:

	Before	After	Reduction	Reduct^2
	13	9.4	3.6	12.96
	12.8	11.5	1.3	1.69
	11	11.5	-0.5	0.25
	13.2	13.1	0.1	0.01
	12.5	10.2	2.3	5.29
	11.9	12	-0.1	0.01
Sum	74.4	67.7	6.7	20.21
Sum/6	12.4	11.2833	1.11667	3.36833

The variance is:

$$\text{var} = \left(3.36833 - (1.11667^2\right) \cdot \frac{6}{6-1} = 2.545666$$

and the standard deviation is its squared root: 1.595514

Thus for the standard deviation linearity does not hold - this is because its calculation requires computing the power 2 of the values and the squared root, and these opearations do not follow linearity: $(a+bx)^2 \neq a^2 + bx^2$

We now give the answer to the question of EX 4: using the linearity for the mean is useful for example when the values are transformed into another unit of measurements of the kind $y = a + bx$. *For example, a mean temperature is expressed in Fahrenheit degrees and we want to express it in Celsius degrees. This won't be possible for the standard deviation!*

EX 10

we use the properties of each Normal distribution.

In an uinterval given by mean ± 2·st.dev. we have about 95% of the values (more precisely, we should use 1.96 as a factor instead of 2) . This answers to question a) . similarly for question b) we shall compute the interval with radius 3·st.dev. which includes 99.7% of values:

a) $72 \pm 2 \cdot 8 = (56,88)$

b) $72 \pm 3 \cdot 8 = (48,96)$

For the last point, let us remark that an interval around the mean=median which includes 50% of observations is by definition delimited by the first and third quartile (Q1,Q3), thus we can compute the limits as:

c) $72 \pm 0.67 \cdot 8 = (66.64,77.36)$