Chapter I: Univariate Descriptive Statistics

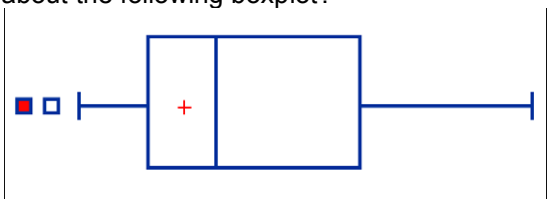
PROBLEMS

Proposed Problems

1. Show that if all data of one variable are multiplied for a constant $k > 0$, the mean and the standard deviation change by a factor k as well.
2. A journal makes a study that shows that the Spanish people spends 400€ per person in the average during Christmas holidays. On the contrary, the economists say that the distribution of the expenses per person relative to any product has a unimodal distribution skewed to the right. What do people think about the average expense of the journal' study? Will they think that the news is exaggerated? Or will they think that the study has underestimated the average expense?

   SOLUCTION:

   Spanish people will think that the estimated average expense is higher than in the reality
3. We have n manufacturing products, among which d are faulty and $n - d$ are acceptable. Let us give to any product a variable x that takes value 1 if the product is acceptable and 0 if the product is faulty. Show that $\bar{x}$ is equal to the proportion of acceptable products (i.e. the relative frequency of the acceptable products)
4. Which of the following sentences is correct?
   a. If a dataset har the CA>0 its histogram shows a skewness to the right
   b. If an histogram shows a skewness to the right than we have CA>0
5. Determine if the following sentences are false or true relative to a whatever dataset and justify your answer
   a. It is possible that most of the data is on the left of the mean
   b. It is impossible that most of the data is of the right of the mean
   c. In presence of skewness to the right most of the data are on the left of the median
   d. When the histogram shows more than one mode it means that there are outlines
   e. In a dataset with mean 100, maximum 300 and minimum value 0, we know that there will be atypical values (outlines).
   f. If we move one value we can increase the variance as much as we want
   g. Moving one values we can increase the mean as much as we want
6. What would you say about the following boxplot?



7. An analyst proposes the following dispersion measure for a set of data $x_1, \cdots, x_n$

$$D = \frac{\sum_{i=1}^{n} x - \bar{x}}{n}.$$

   Show that $D = 0$ for any dataset and therefore if is a useless measure.
8. One machine produced 1837554 identical articles. 80% of such articles does not show any defect, 10% of them has 1 defect, 7% of them has 2 defects and the rest has 3 defects. How many defects per article has this machine produced in average?

   SOLUCTION: $\bar{x} = 0.33\ defect/article$
9. A production process has two production lines: line **A** and line **B** who works independently each to the other (different machines, different workers, etc.) An analyst takes notes at the end of each production line of the number of faults per article. In the note s/he writes in two columns the faults that each article contains. First s/he writes about 50 articles produced by line **A** and later he writes about 50 articles produced by line **B**.

#	Line **A**	Line **B**
1	2	0
2	1	3
⋮	⋮	⋮
50	1	1

After inspecting 50 articles for each line s/he computes a bivariant frequency table reproduced in the following picture

Faults product of production Line **A**

Faults product of production Line **B**

Faults product of production Line A	0	1	2	3	Row Total
0	10	4	1	2	17
	20,00%	8,00%	2,00%	4,00%	34,00%
1	6	4	2	1	13
	12,00%	8,00%	4,00%	2,00%	26,00%
2	4	2	1	2	9
	8,00%	4,00%	2,00%	4,00%	18,00%
3	4	3	4	0	11
	8,00%	6,00%	8,00%	0,00%	22,00%
Column Total	24	13	8	5	50
	48,00%	26,00%	16,00%	10,00%	100,00%

What conclusion could you deduce about?

Resolved Problems

10. Surveying 300 students, we get that 10% of them smokes and 40% of them are women. Writes down a absolute frequency table of these 300 students for the variables Does/Does NOT smoke and Male/Female knowing that the number of men who smokes is equal to the number of women that smokes.

SOLUCTION:

	Female	Male	Row Total
Does smoke	15	15	30
Does NOT smoke	105	165	270
Column Total	120	180	300

10. Given the previous table you are asked to answer the following questions
    a. Write the distribution of the relative joint frequencies
       (check that they sum up to 1)
    b. Write the relative marginal distribution of the variable Does/Does NOT smoke
       (check that they sum up to 1)
    c. Write the absolute frequency distribution of the variable Does/Does NOT smoke conditioned to the fact that the individuals are all women
       (check that they sum up to 120)
    d. Write the relative frequency distribution of the variable Does/Does NOT smoke conditioned to the fact that the individuals are all women
       (check that they sum up to 1)
    e. Which group does smoke the most (male or female)?

    SOLUCTION:

    a. Dividing all values by 300 we get

	Female	Male	Row Total
Does smoke	0.05	0.05	0.10
Does NOT smoke	0.35	0.55	0.90
Column Total	0.40	0.60	1.00

b. The relative marginal distribution of the variable Does/Does NOT smoke is given by

Does smoke	0.10
Does NOT smoke	0.90
Column Total	1.00

c. The absolute frequency distribution of the variable Does/Does NOT smoke conditioned to the fact that the individuals are all women is given by

	Female
Does smoke	15
Does NOT smoke	105
Column Total	120

d. The relative frequency distribution of the variable Does/Does NOT smoke conditioned to the fact that the individuals are all women is given by

	Female
Does smoke	0.125
Does NOT smoke	0.875
Column Total	1.000

By the previous table we see that the 12.5% of women smokes. If we do compute the same frequencies for the men as before we get

Male	Absolute	Relative
Does smoke	15	0.083
Does NOT smoke	165	0.917
Column Total	180	1.000

Therefore, even if the number of men that smoke is equal to the number of women that smoke, speaking about percentage the men's one is smaller. Only the 8.3% of the men smokes, while among the women 12.5% of them smokes.