HSE Introductory Probability and Statistics

Home assignment 1
22.02.2025

Problem 1. An environmental group conducted a study to determine whether animals in a certain region were ingesting food containing unhealthy levels of lead. A biologist classified lead levels greater than 7.5 parts per million (ppm) as unhealthy. The lead levels of a random sample of animals in the region were measured and recorded. The data were as follows:

2.7 2.8 2.8 4.0 4.1 4.6 4.8 5.1 5.3 5.3 5.4 5.8 6.0 6.1 6.3 6.4 6.7 6.8 6.9 7.1 7.5 7.5 7.7 8.1 8.2 8.4 8.8 8.8 9.1 9.1 9.2 9.3 9.4 9.5 10.1

(a) What proportion of animals in the sample had lead levels that are classified by the biologist as unhealthy?
(b) Find the mean and median lead level that animals in the sample had.
(c) Find the standard deviation and the interquartile range of lead level that animals in the sample had.
(d) Are there any outliers in the sample?
(e) Represent the data by a box plot.
(f) Represent the data by a histogram.

Problem 2. File "Hw-1-crypto.xlsx" contains data on top 50 cryptocurrencies. Analyse the market capitalisation of presented coins.

(a) Represent the data by a histogram.
(b) Compute the descriptive statistics: mean, median, standard deviation, IQR.
(c) Represent the data by a box plot.

Problem 3. Continue analysing top 50 cryptocurrencies presented at file "Hw-1-crypto.xlsx". Notice the column ''Type" which represents the type of validation which coin uses. As you can see ''Proof of Stake (PoS)" type is dominant. Analyse PoS coins against non-PoS ones. In case when type is unknown (blank cell) you need to exclude the corresponding coin from analysis.

(a) Draw a parallel histograms that allow to compare the distribution of market capitalisation across PoS and non-PoS cryptocurrencies.
(b) Represent the same data by parallel boxplots. Show your work to compute the necessary descriptive statistics.
(c) Compare the datasets (PoS vs non-PoS) based on location of centers; spread; shapes and special features.

Problem 4. A sample of 200 third year students was selected. The gender of each student was recorded, and each student was asked the following questions:

1. Have you ever had a job?
2. If you answered yes to the previous question, was your job part-time in the summer only? Are you currently working?
3. Are you currently working?

The responses are summarized in the table below.

Job Experience	Male	Female	Total
Never had a job	18	16	34
Had a part-time job during summer only. Currently not working.	53	55	108
Had a part-time job but not during summer only. Currently not working.	17	11	28
Currently working	24	6	30
Total	112	88	200

(a) Construct a graphical display that represents the association between gender and job experience for the students in the sample. Which is better to use in this situation: relative of absolute frequencies?

(b) Write a few sentences summarizing what the display in part (a) reveals about the association between gender and job experience for the students in the sample (datasets comparison).

Problem 5. A statistician collected a sample of 10 elements and computed descriptive statistics. The value of the mean he obtained was 80 and the value of the standard deviation 150.

Later it turned out that there was an error in the data and the value of one element was recorded as 500 instead of the true value 50.

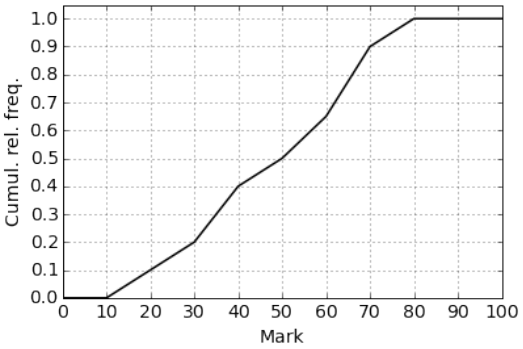
(a) What will be the mean if it is computed for the correct data?

(b) What will be the standard deviation if it is computed for the correct data?

Hint: you may find useful the following formula for sample variance $s^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2\right)$.

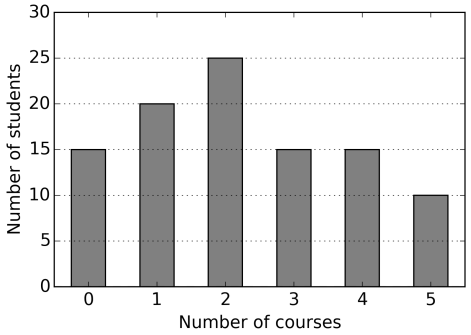
Additional problems

Problem 6. The cumulative relative frequency plot below represents marks that students of some university got for an exam.



(a) What proportion of students got a mark between 40 and 60 (inclusively)?
(b) It is required to get at least 30 to pass the exam. What proportion of students passed the exam?
(c) What was the median mark?
(d) The graph is flat between 80 and 100. What does that mean?

Problem 7. The graph below shows how many additional courses are taken by students in a sample.



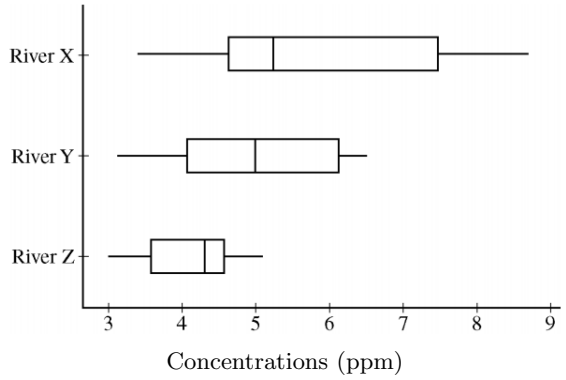
(a) Find the mean and the standard deviation of the number of additional courses taken by the students.
(b) Draw a box plot for the number of additional courses taken by the students.

Problem 8. The data below contain the results (in minutes) of 1 mile amateur swimming competition in two groups of athletes: 18-30 years old and 31-49 years old.

18–30: 21, 25, 31, 34, 36, 39, 42, 44, 44, 45, 52, 55
31–49: 24, 29, 33, 34, 37, 42, 43, 44, 48

(a) Find the medians, lower quartiles, and upper quartiles in each group.
(b) Are there any outliers in each data set? Represent each data set by a box plot.

Problem 9. The graph below displays the concentrations of some pesticide in three rivers.



(a) Compare the distributions of the concentration of the pesticide among the three rivers.

(b) The data for River X were obtained by sampling the water in 20 different locations. The concentrations at those locations were as follows:

3.4	4.0	5.6	3.7	8.0	5.5	5.3	4.2	4.3	7.3
8.6	5.1	8.7	4.6	7.5	5.3	8.2	4.7	4.8	4.6

Construct a histogram that displays the concentrations for River X.

(c) Describe a characteristic of the distribution in River X that can be seen in the histogram but cannot be seen in the boxplot.