CEU
Data Analysis 1 – Mock exam
2023-10-13

This is a closed book exam. The maximum is 100 percent. You have 90 minutes.
Please write your answer right after the question. You may use as much as space as you'd like.

* Please be very brief and answer the question only. Please do not answer questions that are not asked. Often you will just need a few sentences and some examples to answer.
* If the question asks for a specific answer (yes/no, which one, list advantages/disadvantages etc) then the answer has to contain the answer to the question (yes/no, which one, list advantages/disadvantages etc) in an explicit way.
* It is good practice to give the answer first and the argument next.

Part I: Short questions: 7*10=70 percent

1. (10p)
What are the benefits of a representative sample? Now consider surveying a sample of employees at a large firm. List four selection methods and assess whether each would result in a representative sample.

Answer 1

2. (10p)

Decide if the following sentences are true or false? Give a very short argument of your answer.
a) The amount of water in a swimming pool is a quantitative variable measured on a ratio scale.
b) Ratio variables are also interval variables
c) Flags are variables measured on an ordinal scale
d) Temperature measured on the Celsius scale is a ratio variable
e) The standard deviation of a qualitative variable is the square root of its variance

Answer 2

3. (10p)
What are the main problems of dealing with missing data? Tell an example of how they may be indicated for string, numeric, and binary variables. What options do you have for dealing with missing data? If you made any change in the data, how would you communicate it?

Answer 3

4. (10p)
Please describe the following concepts, You may use formulae but can also describe them verbally.

(a) Statistical dependence
(b) Covariance
(c) Correlation coefficient
(d) Negative and positive correlation between x and y

Answer 4

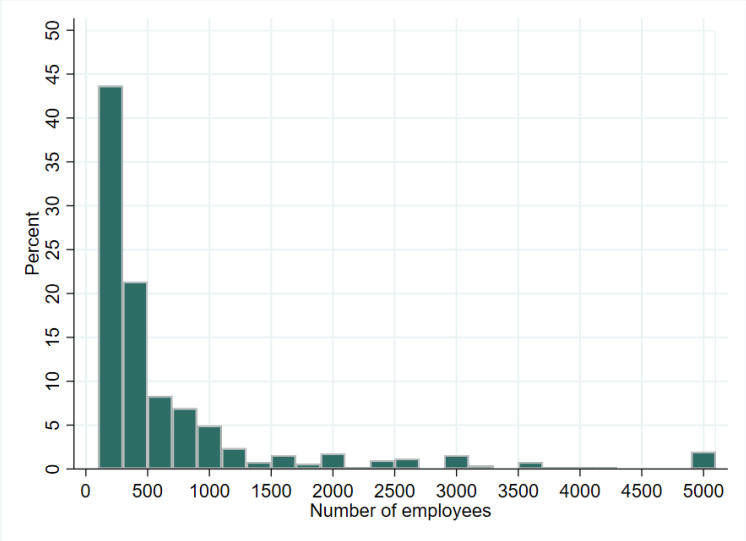
5. (10p)
What is the level of significance or size of a test, and what is the power of a test? Give an example of a test, and explain the concept of size, significance, and power in the context of this example.

Answer 5

6. (10p)
You examine the wages of recent college graduates, and you want to test whether the starting wage of women is the same, on average, as the starting wage of men. Define the statistic you want to test. Define the population for which you can carry out the test if your data is a random sample of college graduates from your country surveyed in 2015. Write down the appropriate null and alternative hypotheses, and describe how you would carry out the test. What would be a false negative in this case? What would be a false positive?

Answer 6

7. (10p)
Consider this for firm size (number of employees) for Brazilian firms in the management survey. The histogram shows firms between 100 and 5000 employees. Your task is to write three bullet points explaining key features of the histogram for a manager.



Answer 7

Part 2: Multiple choice: 6*5=30 percent

Question1 : "You want to collect data on the friendship network of students in your data analysis class from a social media app. 75% of the students are on this social media, and you have full access to data on all users. Which of the following is true about the data you can collect?"

Answer 1: "It will not be a representative sample of all students in the class."

Answer 2: "It will be a representative sample of all students in the class."

Answer 3: "It will be a random sample of all students in the class."

Answer 4: "It will give a good benchmark to the distribution of all students in the class."

Question 2: "When you merge two data tables into one, what's always true of the new data table?"

Answer 1: "It includes variables from both of the original data tables."

Answer 2: "It excludes observations that are in only one of the data tables."

Answer 3: “It excludes variables that are in only one of the data tables.”

Answer 4: "It includes observations that are in neither of the data tables."

Question 3: "A variable with a unimodal distribution has a (substantially) higher mean than median. What does that imply?",

Answer 1: "Its distribution is skewed (such as having a long right tail). "

Answer 2: "The mode is also higher than the median."

Answer 3: "Its distribution is symmetric."

Answer 4: "Its distribution is binomial."

Question 4: "Which of the following variables may be distributed normally?"

Answer 1: "Intelligence test scores of people."

Answer 2: "Sizes of cities in a country."

Answer 3: "Family income in a country."

Answer 4: "Whether it will rain tomorrow."

Question 5: "What's a latent variable?"

Answer 1: "It's a variable that we can define conceptually but can't measure in real-life data."

Answer 2: "It's a variable that is used for conditioning."

Answer 3: "It's a variable that is measured with high validity in the data."

Answer 4: "It's a variable that is measured with high reliability in the data."

Question 6: "Which of the following is true about the t-statistic that you can use to test whether average spending on food delivery is the same (H0) or different (HA) in two groups of people?"

Answer 1: "It measures how large the estimated difference between the two groups is in the data, in units of its SE."

Answer 2: "It measures how large the estimated difference between the two groups in the data, divided by overall average spending."

Answer 3: "It is always between 0 and 1."

[END OF EXAM]