CEU
Data Analysis 1 – Mock exam
2024-10-22

This is a closed book exam. The maximum is 40 points (100 percent). You have 45 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: 28p (70 percent)

1. (8p)

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) Temperature measured on the Celsius scale is a ratio variable
d) The standard deviation of a qualitative variable is the square root of its variance

Answer 1

a)True - it is measured in liters, and both the difference and the relative value are meaningful.
b) True – If the ratio is meaningful so is the difference
c) False – It is interval, relative degree (twice as cold) does not make sense
d) False – only works for quantitative variables

2. (7p)
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 2

There are two main problems, firstly that missing data means fewer observations that can be used for the analysis. Secondly, missing data may introduce selection into our sample.
Missing string variables maybe indicated with the empty string or "NA", missing numeric variables with a . or a value outside of the range of the variable. Binary variables often use the value 9 for missing values.
The two options are dropping the observations with missing value or imputing their values.
Communicating changes in data is essential. I'd use a flag, a binary variable: 1 when a value is changed, 0 otherwise. I'd also take a note on a README file.

3. 7p
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 3

H_0: $\overline{w}_m - \overline{w}_f = 0$
H_1: $\overline{w}_m - \overline{w}_f \neq 0$

where $\overline{w}_m and\ \overline{w}_f$ are average wage of males and females in the group. To control for other effects, we may select graduates only from one field of study at a time (only MSc in Finance or only BSc in Electrical Engineering, etc.) for both male and female graduates. These comparisons then can be repeated and tested for other subgroups as well.

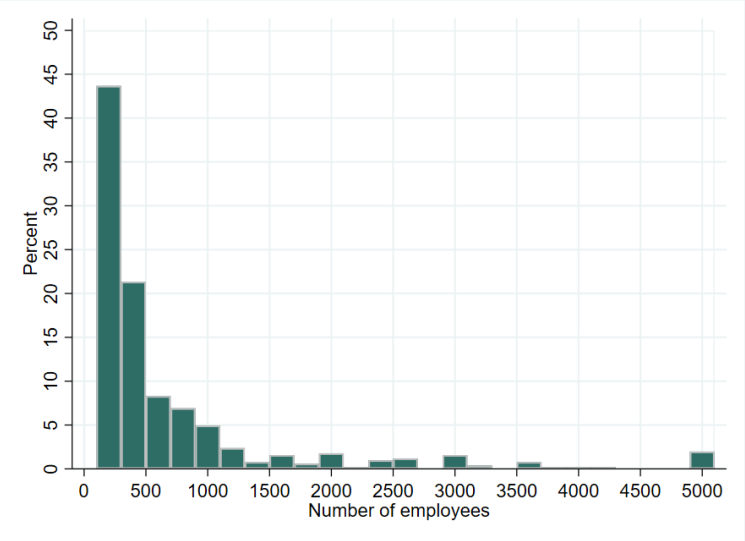
We calculate the t-stat:
$t = \frac{\overline{w}_m - \overline{w}_f}{SE(\overline{w}_m - \overline{w}_f)}$

If we take a level of significance of 5 percent, then, in case of a large sample, a t-value above 2 would prompt us to reject H_0 and say that male and female graduates have different starting wages.

False positive: we reject H_0 although the wages are essentially the same for males and females, on average.
False negative: we do not reject H_0 although the wages are different for males and females, on average.

4. (6p)

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 4

- It has a power-law distribution.
- Almost two-third of Brazilian firms have less than 500 hundred employees, and appr 43 percent has less than 200 employees.
- Having more than a thousand employees is very rare in Brazil. This is especially interesting given the size of Brazil's population, which would imply that there is room for utilizing economies of scale and which may lead to a larger percentage of companies with many employees.
- We need to put these numbers in context though. Additional examples from large developed economies (US, Germany, Japan) and large emerging economies (China, India, Russia) would help interpreting Brazil's results better.

Part 2: Multiple choice: 4*3=12p (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: "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 3: "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 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."

[END OF EXAM]