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

1. Selecting the first x percent of employees by their last names in alphabetical order. Not representative as people with different names may belong to different groups of society.
2. Selecting employees the first x percent of employees by the time they have spent at the company. Not representative as employees with long employment history may be very different in terms of loyalty, and as such, personality than those with short employment history.
3. Adding a random id to each employee and select the first x percent based on that id. It is likely a representative sample.
4. Group employees by their departments. Select random employees from each department so that the share of the departments in the sample is the same as / similar to the share of departments to all employees. This will likely result in a representative sample.

General rule for the answer: you'd better make sure that randomness is present in your sampling. Many selecting methods can be listed at this question, but most of them will be not really representative if it is tied to a non-random factor, such as name, employee id (which is likely to be of a low nominal value in case of employees with longer employment history), etc.

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

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 – Flags are binary indicators
d) False – It is interval, relative degree (twice as cold) does not make sense
e) False – only works for quantitative variables

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

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.

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

a) The conditional distributions of one variable (y) are not the same when conditional on different values of the other variable (x).
b) It is a measure of mean-dependence, or of joint variability of two variables. Alternatively: a directional relationship between the values of two variables.
c) It is derived from covariance by dividing it by the standard deviations of the two variables (x and y). One important attribute of covariance is that it is always between -1 and +1.
d) Positive correlation: observations with larger-than-average x values tend to have larger-than-average y values as well.
Negative correlation: observations with larger-than-average x values tend to have lower-than-average y values and vice versa.

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

Size of a test: the probability of a false positive decision.
Level of significance: the maximum probability of a false positive decision that we tolerate.
Power of a test: the probability of avoiding a false negative (accepting a null-hypothesis which is in fact not true).

Example: measure the difference of online and offline prices of some products. Calculate the price differences and test whether this difference is significantly different from zero. The null-hypothesis (which we want to reject), is H_0: $p_{offline} - p_{online} = 0$.
Alternative hypothesis H1: $p_{offline} - p_{online} \neq 0$.
The p-value of our test is 0.04. In this case to **size of the test** (probability of rejecting a true null hypothesis, false positive) is 4 percent. If we take 5 percent as the **level of significance**, we reject H_0. With the size of the test being only 4 percent, we can hope for a **high power of the test**, which is more likely if the sample size is larger, or further away the true value is from the null-hypothesis.

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

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

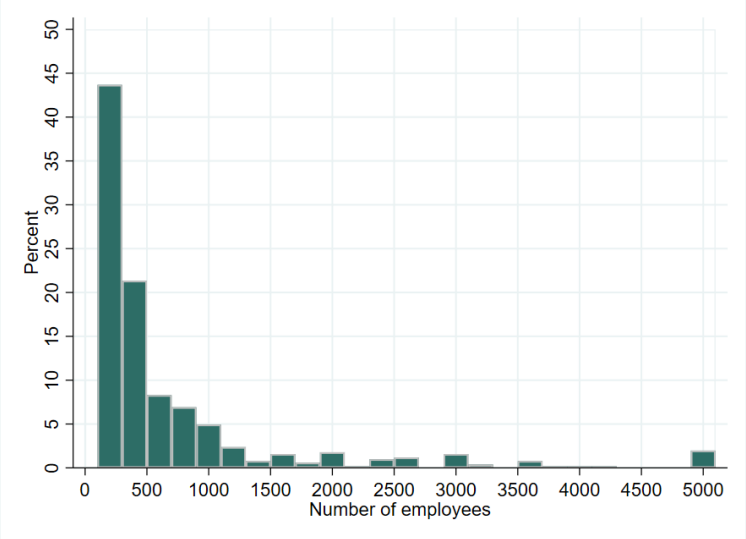
where $\bar{w}_m and\ \bar{w}_f$ are average wage of males and females in the group. To control for other effects, we may select graduates only from one field of study at a time (only MSc in Finance or only BSc in Electrical Engineering, etc.) for both male and female graduates. These comparisons then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

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

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.

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

- 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: 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]