DATA1001 Topic 2 Flashcards

Data and Graphical Summaries

Abyan Majid

Question	**My answer**
What type of data is a histogram used to represent? And for what purpose is it best suited for?	1 quantitative variable. It's particularly used to highlight the percentage of subjects/data in different class intervals.
What is the whole area of a histogram?	100%
What is the horizontal scale in a histogram?	Class intervals
What does the area of a block in a class interval in a histogram represent?	The percentage of subjects/data in that particular class interval
What does the height of a histogram represent?	Crowding - the percentage of subjects/data per horizontal unit
Why do we use a density scale in a histogram?	To standardize the y-axis thereby removing the dependence on absolute frequency counts, which makes it more reliable for comparing the shapes of the class intervals.

Given two class intervals A: 0 to 15, and B: 15 to 30, to which interval will a data that is exactly 15 be classified? Then, write A and B using the interval notation based on the endpoint convention!	It will be classified into class interval B, by the endpoint convention "left-closed, right-open"
	A = [0, 15)
	B = [15, 30)

Recite the formula for the height of the bin, which involves percentage of subjects in the class interval and the length of the class interval!	$Height\ of\ bin\ = \frac{\%\ of\ subjects}{length\ of\ bin}$
Why shouldn't you use too many bins in a histogram? What do you reckon is an acceptable range for the number of bins?	Having too many bins makes your histogram cluttered. As a rule of thumb, 10-15 bins is preferable.
What type of data is a simple box plot used to represent? And for what purpose is it best suited for?	1 quantitative variable. It is particularly useful for visualizing mean, spread, and outliers.
How does a comparative box plot	A comparative box plot splits a quantitative variable by a

differ from a simple box plot?	qualitative variable into subgroups such that you get multiple boxes and whiskers in the plot that are comparable to one another.
What type of data does a scatter plot represent? And for what purpose is it best suited for?	2 quantitative variables. It is particularly useful for visualizing the relationship or trend between 2 quantitative variables and modeling.
What is data?	Data is information about the subjects being studied
What is initial data analysis (IDA) and what is its purpose?	Initial data analysis refers to the first look at data with the purpose of getting you familiarized about the data's main qualities such that you can get an idea of whether the data is appropriate or is sufficient for investigating your research questions.
What are the steps involved in initial data analysis (IDA)?	- Checking the background and credibility of the data - Exploring the data, its structure and variables, to get an understanding of all information it provides for your analyses. - Data wrangling (cleaning data, renaming variables, removing invalid entries, etc) - Making numerical/graphical summaries where necessary (to further enhance understanding of the data)
What is a variable?	Variable is a measurement that describes an attribute of a subject
Identify the two different types of qualitative variables, and two different types of quantitative variables!	Qualitative variables: ordinal (ordered), nominal (unordered) Quantitative variables: discrete, continuous
What type of data does a simple bar plot represent? And what is it particularly useful for?	1 qualitative variable. It's particularly useful for visualizing the frequencies/distribution of a qualitative variable with respect to the y-axis.
What type of data does a stacked bar plot and side-by-side bar plot represent? And what are they particularly useful for?	2 qualitative variables. It's particularly useful for comparing the distribution of two qualitative variables with respect to the y-axis.
What is big data?	Data that commonly has immensely high dimensionality. That is, it has extremely many variables, and it is often the case that the number of variables exceeds the number of subjects.