DATA 100: Vitamin 5 Solutions

September 25, 2019

1 What to do with missing values?

Real world data is messy; observations are often missing values for some variables. Which of the following approaches may be reasonable for dealing with this missingness?

☑ Drop the observations with missing values

☑ Replace missing values with an average value (mean imputation)

☑ Replace missing values with random values (hot deck imputation)

□ Ignore the missing values, they won't affect our analyses anyways

Explanation: The first three options may be reasonable approaches to handling the missing values in data set. This will depend on the cause of the missingness. If the values are missing completely at random, then these options are reasonable. However, it is almost never wise to ignore these missing values, as they will affect downstream analyses. In fact, many statistical methods cannot be applied to data with missing values.

2 Missingness

Fill in the blank: When the missing values in a dataset are ____, it means that the missingness is not correlated with other variables.

□ few

☑ missing at random

□ missing not at random

□ plentiful

Explanation: We say that the missing values are missing at random when their missingness is not correlated with any other variable.

3 Keys

The following definition corresponds to which kind of key? "The column or sets of columns in a table that determine the values in the remaining columns."

□ A foreign key

□ The key to success

☑ A primary key

Explanation: This is the definition of a primary key.

4 Plotting Qualitative Variables

Which of the following plots can be used to depict a single qualitative variable?

□ histograms

□ box plots

☑ bar plots

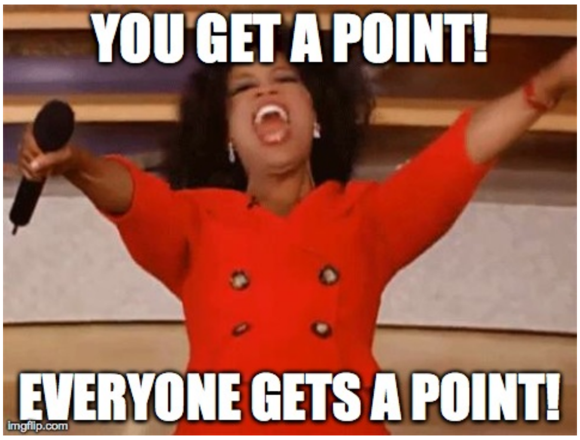
☑ dot plots

□ scatter plots

Explanation: Histograms and box plots are used to depict the distributions of quantitative variables, and therefore cannot be used when only considering qualitative variables. Scatter plots are used to illustrate pairs of quatitative variables. Of the options listed, only bar plots and dot plots are well suited for presenting a lone qualitative variable.

5 Data transformations

This question is out of scope!



Non-linear relationships between variables of interest are often transformed into linear relationships. Why?

- ☑ If the transformation is simple, the relationship will be easier to interpret
- □ The relationship will always be easier to interpret, regardless of the transformation used
- ☑ Models are easily fit to linear relationships
- □ Models can only be fit to linear relationships