DATA 100: Vitamin 3 Solutions

February 18, 2019

1 Ordinal Data

The best way to plot ordinal data would likely be:

☐ A histogram

☐ A density plot

☑ A bar chart

☐ Using a color scale

Explanation: Histograms and density plots can only be used to visualize quantitati variables. A color scale is typically used to visualize discrete and continuous quantitative variables, but in some situations may be used to illustrate the scale of an ordinal variable. A bar chart is therefore the best way to plot ordinal data of the method listed.

2 Quantitative Data

When trying to understand the distribution of scores on an assignment in a large class in which many different scores are assigned, which of the following plots would be most appropriate?

☐ A pie chart

☑ A density plot

☐ A bar chart

☐ A scatter diagram

Explanation: Beyond the problems already associated with this plotting method, a pie chart would require the data to be aggregated (i.e. number of students in some grade interval), leading to a loss of potentially important information. A bar chart should only be used for qualitative data or discrete quantitative data. A scatter diagram is used to visualize the relationship between a minimum of two variables. A density plot (i.e. KDE) is the most appropriate option since it visualizes the distribution of continuous variables with minimal loss of information.

3 Comparison

Which of the following kinds of plot would be useful for comparing the distribution of scores for second-year and third-year students?

☑ A violin plot

☑ A box plot

□ A scatter diagram

□ A lollipop chart

Explanation: Violin plots and box plots can be used to present the score distributions of both groups in a single plot. Having the distributions side-by-side greatly enhances the ability to compare the second-year and third-year students. Lollipop charts would also allow a side-by-side comparison of the groups, but requires the data to be aggregated prior to plotting, leading to a loss of information. A scatter diagram is only appropriate for the visualization of a minimum of two quantitative variables.

4 Scatter Diagrams

Which of the following are uses of a scatter plot matrix?

☑ Compare every pair of quantitative variables in a dataset

☑ Identify associations between pairs of quantitative variables

☑ Visualize more than two quantitative variables at a time

Explanation: Scatter diagrams are used to compare a minimum of two quantitative variables at the same time. These plots are often used to understand the association between these variables, and can be used to illustrate multiple pairs of quantitative variables in a data set. (**Note:** Given a data set with n quantitative variables, $\binom{n}{2}$ scatter diagrams illustrating the relationship between pairs of these variables can be created. Thus, for large values of n, this method of studying the associations between variables is inefficient. Can you think of a visualization that would be useful in this situation?).

5 Overplotting

Which of the following are instances of overplotting?

- ☑ Multiple points in a scatter diagram have the same position, making it difficult to determine how many elements appear at that position.
- □ Too many dimensions of the data are included in the same figure, making it difficult to determine the associations between different variables.
- ☑ Histograms for several different variables are overlaid in one figure, making it difficult to determine the height of each histogram.

Explanation: Overplotting consists of layering many observations or variables in a plot such that the visualization is too densely packed with information to be interpretable. Answers 1 and 3 are clear examples of overplotting. Answer 2 does not represent a situation of overplotting since the observations are not layered over one another. A remedy to the problem described in answer 1 may be to "jitter" the observations in the scatter diagram (i.e. slightly perturb each point in a random direction of the cartesian plane) or to plot a random subset of the data. In the case of answer 3, each histogram could be placed in an individual plot, or the opacity and colors of the histograms can be modified such that comparisons are easier.