Visualizing Variable Distributions

Reading

Sections 2.1, 2.2

Notes

For every variable we talk about the variable's **distribution**, which means a description of what values the variable takes, and how frequently it takes those values. Variables are visualized differently depending on their type.

- Visualizing Categorical Variables

  Frequency Table A table showing each possible value, along with its *frequency* i.e. the count of its occurencies. One can also include *relative frequencies*.

  Pie Chart A circular shape is divided in parts proportional to the relative frequency of each value. Good for showing relation of each part to total.

  Bar Chart A rectangular bar for each value, whose height is proportional to the frequency. Good for comparing frequencies of values to each other.

  Pareto Chart A bar chart where the values have been ordered from most frequent to least frequent.

- Visualizing Scalar Variables

  Summaries Numerical summaries can give us some limited but easy-to-work-with information. Frequency tables turn out to be too unwieldy in this case.

  Histogram Values are broken into equally spaced intervals. Draw one bar per interval whose height is proportional to the frequency of values in that range.

  Stem-Leaf Plot Useful for certain types of values. Use first 1-2 digits for "stem", then add one value via its "leaf" on the correct stem row.

  Box-plot Visual representation of the "five number summary" that we will talk about later.

  Density plot A continuous line that describes the data a bit like a histogram, only more precisely. We will not be using them in this course, but they are out there and are useful.

- When visualizing scalar variables, there is some terminology we use and patterns we look for:

  Modes A mode refers to a distinct section of the data that "stands out" as a spike in the graph. It need not be a single value, more of a tendency for values to concentrate around that point. A graph with a single mode is called **unimodal**, one with two modes is called **bimodal**. When multiple modes are present, they become the main characteristic of the dataset.

Tails A term referring to the two ends of the data. A long tail indicates that the data on that side is spread out and "goes on for a while".

Skewed Left A skewed left distribution is one where the left tail is longer. This represents a concentration of data (higher bars) to the right.

Skewed Right A skewed right distribution is one where the right tail is longer. This represents a concentration of data (higher bars) to the left.

Symmetric In a symmetric distribution both tails seem to be about the same. If you look in the middle of a symmetric distribution, then the two sides around the middle should be (close to) mirror images of each other.

Outliers Any values that seem to deviate from the overall pattern are called outliers. Some times these are simply values that are too far from the rest. But some times they can be outliers for other reasons.