Contents

1 Data

Data Basics

▷ Frequent types of data in statistics:

- ○ **Interval**: numeric scale with meaningful intervals, e.g. temperature in celsius.

- ○ **Ratio**: numeric but with a meaningful zero, e.g. height.

- ○ **Discrete**: numeric with with no arbitrary precision, e.g. population.

- ○ **Ordinal**: sortable and discrete, e.g. education level.

- ○ **Nominal**: non-sortable and discrete, e.g. genre.

▷ **Sample data**: Data from *some* members of a group.

▷ **Population data**: Data from *all* members of a group.

▷ Sample population sometimes uses hat notation, e.g. $\hat{\beta}$, $\hat{\sigma}$, or other slight ambiguities. Sample data is used more often than population in statistics.

Visualizing Data

▷ **Bar plots**: used to represnet categorical (nominal and ordinal) and discrete numerical data.

▷ **Box plots**: collection of a data that is split into separate quatiles in order to illustrate overall distribution of data and its potential outliers.

▷ **Histograms**: similar to bar plots, but with binned continuous data on the x-axis. Shape and order is meaningful.

- ○ Histograms of **counts**:

  - – Often more meaningful interpretation of raw data.

  - – Difficult to compare across datasets.

  - – Does not need to sum up to 1.

  - – Usually better for qualitative inspection.

- ○ Histograms of **proportion**:

  - – Can be more difficult to relate to raw data.

  - – Easier to compare across datasets.

  - – Illustrates proportion of dataset.

  - – Usually better for quantitative analysis.

▷ Translating from counts to proportions: $bin_i = 100\,(bin_i\,/\,sum(bins))$

▷ **Pie charts**: representation of nominal, ordinal, or discrete data that must sum up to 1.

2   Descriptive Statistics

Descriptive vs. Inferential

▷