# 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

- **Descriptive**:
  - The point is to obtain individual numbers that describe a dataset.
  - Mean, median, mode, variance, kurtosis, skew, distribution, spectrum.
  - No relation to population; no generalization to other datasets of groups.
- **Inferential**:
  - Use features of sample data set to make generalizations about a population.
  - P-value, T/F/chi-square value.
  - Confidence intervals.
  - Hypothesis testing.

### Accuracy, Precision, Resolution

- **Accuracy**: the relationship between measurement and the actual truth. Inversely related to bias.
- **Precision**: the certainty of each measurement. Inversely related to variance.
- **Resolution**: the number of data points per unit measurement.

### Data Distribution

- **Data Distribution**: a function that lists values or intervals of data, and how often each value occurs.
- Common distributions include power-law, gaussian (bell curve), t, F, and Chi-squared.
- Most statistical procedures are based on assumptions about distributions.
- Data distributions provide insights into nature and often used to model physical and biological systems.

### Measures of Central Tendency

- **Central tendency**: the center of typical value for a probability distribution.
- Common measures of central tendency: mean, median, mode.
- Mean, aka average or arithmetic mean:

- Formula: $\bar{x} = n^{-1} \sum x_i$.
- Alternate notations for mean: $\mu$, $\mu_x$.
- The mean is most suitable for normally distributed interval and ratio data.
- Discrete and ordinal data can be useful, but must be carefully interpreted.

▷ Median:

- $x_i,\ i = \frac{n+1}{2}$
- Most suitable for unimodal distributed interval and ratio data.

▷ Mode: the most common value that is suitable for any distribution and data type, though mostly used for nominal.