Statistics

Data

Data Fundamentals

- **Data**: units of qualitative or quantitative information about persons or objects collected via observation.
  - Note: data is different from information—information resolves uncertainty, while data has the potential to be transformed into information post-analysis.
  - Data as a general concept refers to the fact that some existing information or knowledge can be represented in a form suitable for processing.

Data Types

- Data types have two different general meanings:
  - **Data type (computer science)**: involves the format of data storage and has implications on operations and storage space.
  - **Data type (statistics)**: involves the category of data and has implications on the methods used for analysis.
- There are many data types, with more specific definitions than the following definitions, but for now these are frequently used and adequate for topics covered.

Relevant Statistical Data Types

Category	Type	Description	Example
Numerical	**Interval**	Degree of difference	Temperature °C
	Ratio	Interval + meaningful zero	Height
	Discrete	Count (integers)	Population
Categorical	**Ordinal**	Sortable, discrete	Educational level
	Nominal	Non-sortable, discrete	Movie genre

Population vs. Sample Data

- **Population data** μ: data from *all* members of a group.
- **Sample data** $\hat{\mu}$: data from a *subset* of members of a group (hopefully random).
- Statistical procedures generally are designed for sample or population data; wrong conclusions can be drawn if the distinction is not clear.
  - Note: most data are sample data in practice, as generalization of populations using sample data is usually the goal of statistics.
- **Anecdotes**: a case study of a rare occurrence, or a sample size of only one; insights may be possible, but poor confidence in ability to generalize should be noted.

Data Visualization

- **Data visualization**: a mapping between the original data and graphic elements in order to determine how attributes of interest vary according to the data.
  - The design of the mapping can have a significant effect on information extracted from data, in both beneficial and detrimental ways.
- Data visualization is a core tool of statistics and is generally considered to be a branch descriptive statistics. ↓

Visualization Techniques

- Visualizing data can be an art in and of itself, leading to a wide variety of available techniques, i.e., diagram types, in order to better represent the data.
- The following is a rather shallow list of commonly used techniques; in-depth exploration of data visualization will be pursued in other courses.
- **Bar chart**: a representation of *categorical data* with magnitudes proportional to the values they represent.
  - Displays comparisons among *discrete categories* vs. a measured value.
  - Subcategories can be displayed in clusters within each category, with colors/patterns used to differentiate them.
  - Ordering of the categories (chart shape) do not typically matter, excluding aesthetic reasons.
- **Histogram**: a representation of the *distribution* of numerical data via the use of *binning*.
  - **Binning**: a form *quantization of continuous data*, wherein small intervals (bins) of the data are replaced with a value representative of that interval.
  - The bins are usually specified as consecutive, non-overlapping intervals of a variable; they must be adjacent and are often of equal size.
  - Histograms of *counts* are usually better for *qualitative* inspection of raw data, but can be difficult to compared across datasets.
  - Histograms of *proportion* are usually better for *quantitative* analysis, easier to compare across datasets, but can take extra effort to create.
- **Scatter plot**: a representation of the *relationship between variables*, often two or three (2D/3D graphs).
  - Points can be coded via color, shape, and/or size to display additional variables.
  - Often used to investigate *correlations* between variables.

- **Network graph**: a representation of data as nodes in a network via analysis of *specialization* of the nodes.
  - Used to discover bridges (information brokers) in a network, relative node influence, and outliers via analysis of how the nodes cluster.
  - Node and tie (connection between nodes) size and color can be used to encode additional information about variables in the data.
- **Pie chart**: a representation of one categorical variable via the division of slices in order to illustrate *numerical proportion*.
- **Box plot**: a representation of numerical data via analysis of their quartiles.
  - **Quartiles**: a quantile (division point) of data points into four parts, or quarters.
    - Q_1: the middle number between the smallest minimum and the median of the dataset; 25% of the data lies below this point.
    - Q_2: the median of the data set; 50% of the data lies below this point.
    - Q_3: the middle value between the medium and the maximum of the data set; 75% of the data lies below this point.
  - Often termed box and whisker plot, as the box represents the 50% of the data, and the two whiskers represent the upper and lower 25% of data.
  - Outliers may be plotted as individual points.
  - Useful when examining the *variability of samples* without making any assumptions about underlying statistical distributions.

Descriptive Statistics

Data Normalization and Outliers

Probability Theory

Hypothesis Testing

T-Tests

Confidence Intervals

Correlation

Regression

Statistical Power and Sample Sizes

Clustering and Dimension-Reduction