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 generally considered to be a branch descriptive statistics↓; more techniques will be covered in that chapter.

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 data sets.
    - Histograms of proportion are usually better for quantitative analysis, as they are typically easier to compare across data sets, 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 data set; 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.
        - **Interquartile range IQR**: the box, i.e., the difference between upper and lower quartiles; $IQR = Q_3 - Q_1$.
    - Outliers may be plotted as individual points.
    - Useful when examining the variability of samples without making any assumptions about underlying statistical distributions.

Descriptive Statistics

Descriptive Statistics Fundamentals

Descriptive vs. Inferential Statistics

- **Descriptive statistics**: the processes of using and analyzing summary statistics that quantitatively describes or summarizes features of a collection of information.
  - Methods/measures of descriptive statistics:
    - Distribution shape↓
    - Mean, median, mode↓
    - Variance↓
    - Kurtosis, skew↓
  - No relation to population.
  - No generalization to other data sets.
  - Concerned only with properties of observed data.
- **Inferential statistics**: the process data analysis to deduce properties of an underlying probability distribution.
  - Methods/measures of inferential statistics:
    - P-value↓
    - Hypothesis testing↓
    - T/F/χ^2 value↓
    - Confidence intervals↓
    - And essentially all of applied statistics.
  - Assumes that the observed data set is sampled from a larger population.
  - Entire purpose is to generalize/relate features to other data sets.

Accuracy, Precision, Resolution

- **Accuracy**: the relationship between the measurement and the actual truth.
  - Inversely related to bias; colloquially interchangeable with accuracy.
- **Precision**: the certainty of each measurement.
  - Inversely related to variance↓
- **Resolution**: the number of data points per unit measurement (e.g., time, space, individual, etc).

Data Distributions

- The shape of data distributions are functions of probability theory↓; a more in-depth explanation will be covered later, but for now coverage of common distribution types might be useful.
- There is one major distinction of distributions based on data types↑, either discrete or continuous.
- **Discrete distribution**:
  - Deals with events that occur in countable sample spaces; contains finite number of outcomes.
  - Summation of values can be done to estimate probability of an interval.
  - Expressed with graphs, piece-wise functions, or tables.
  - Expected values might not be achievable.
  - Common examples:
    - Bernoulli : a model for the set of possible outcomes of any single binary experiment.
    - Binomial : a sequence of n independent Bernoulli experiments; a basis for the binomial test.
    - Uniform : a known, finite number of values are equally likely to be observed.
    - Poisson : a sequence of independent events over a specified interval with a known constant mean rate.
- **Continuous distribution**:
  - Deals with events that occur in a continuous sample space; contains infinitely many consecutive values.
  - Summation of values in order to determine probability of interval not possible; integrals used instead.
  - Expressed with continuous functions or graphs.
  - Common examples:
    - Normal : used to represent real-valued random variables who are not known; very common.
    - Chi-Squared : the sum of squares of k independent standard normal random variables.
    - Lognormal : distribution of a random variable whose logarithm is normally distributed.
    - Student's T : estimations of the mean using mall sample sizes with unknown standard deviations.
- Wikipedia's list of probability distributions

Descriptive Techniques

Measures of Central Tendency

- **Mean** $\overline{x}$: the sum of all measurements x_i divided by the number n of observations in the data set x, i.e.,

$$\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$$

  - Suitable for roughly normally distributed data of continuous data types.
- **Median** $\operatorname{med}(x)$: the middle value of the data, i.e.,

$$x_i, \quad i = \frac{n+1}{2}$$

  - Suitable for unimodal distributions of continuous data types.
  - Odd number of observations with no distinct middle value are usually defined as the mean of the two middle values.
- **Mode**: most common value.
  - Suitable for any discrete distribution, usually used for nominal data types.

Measures of Dispersion

- **Dispersion**: the measure of how distributed, or deviated, data are around a central value.
- **Variance** σ^2, s^2: the primary measure of dispersion, or more explicitly, the expectation of the squared deviation of a random variable from its mean, i.e.,

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

  - Suitable for any distribution; better for normally distributed data.
  - Mean centering, i.e., $(x_i - \overline{x})$, is done to capture the dispersion around the average, but not the magnitude of the values themselves.
  - The sum of a mean-centered data set would be zero, thus it is squared.
    - **Mean absolute difference (MAD)**: when the absolute value of mean-centered data is taken instead of the square value.
    - MAD is more robust to outliers, but further from Euclidean distance and less commonly used.
  - Division by $n-1$ is used for sample variance, as often sample sizes can be small and are considered empirical quantities; n^{-1} is used for population variance (a theoretical quantity).
  - **Standard Deviation** σ: simply the square root of variance, $\sqrt{\sigma^2}$

Statistical Moments

- **Moments**: a quantitative measure related to shape of a functions graph; relates to physics and statistics.
  - Regarding probability distributions, the general formula can be defined as:

$$m_k = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^k$$

  - Increments of k define particular moments, i.e.,
    - First moment $k = 1$: expected value, or mean↑.
    - Second moment $k = 2$: central moment, or variance↑.
    - Third moment $k = 3$: dispersion asymmetry, or skewness.
    - Fourth moment $k = 4$: tail "thickness," or kurtosis.
    - Further moments are possible, but useful applications are less common.
  - **Skewness**: a measure of asymmetry of a probability distribution of a real-valued random variable about its mean.
    - Can be positive, zero, negative, or undefined.
    - Negative skew: an indication that the tail is on the left.
    - Zero skew: an indication that tails balance out; can be true for both asymmetric and symmetric distributions depending on kurtosis.
    - Positive skew: an indication that the tail is on the right.
  - **Kurtosis**: a measure of the thickness/curvature of the tail of a probability distribution is; an indication of deviation/outliers.
    - Univariate normal distributions have a kurtosis of 3, leading to a common basis.
    - Platykurtic < 3: a term for low kurtosis, indicating that a lesser degree of deviations or outliers is observed.
    - Leptokurtic > 3: a term for high kurtosis, indicating that a greater degree of deviations or outliers is observed.
    - **Excess kurtosis**: kurtosis minus 3, often colloquially termed as kurtosis; an indication a greater degree outliers compared to a normal distribution.

Visualizations Revisited

- **Q-Q (quantile-quantile) plot**: a graphical method for comparing two probability distributions by plotting their quantiles against each other.
  - **Quantile**: cut points dividing the range of probability distributions into continuous intervals with equal probabilities, e.g.,
    - Percentiles: 0–100
    - Quartiles: 0–4
    - Quantiles: 0–x
  - The points of similar distributions will lie approximately on the line $y = x$;
  - However, other linear relations are possible, meaning points may not necessarily lie on the line $y = x$.
  - Provides a mean for comparing location, scale, and skewness of similarities of differences in two distributions.
- **Histogram bin number** k: the is no "best" number of bins, different bin sizes can reveal different features of the data, but there are several methods of determining k;
  - Determination via suggested bin width h:
$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil$$
  - Sturges' formula: derived from binomial distribution; assumes approximately normal distribution:
$$k = \lceil \log_2(n) \rceil$$
  - Freedman-Diaconis' rule: method of determining h using interquartile range (IQR); often method of choice:
$$h = 2\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$
  - Arbitrary ≈ 42: often intuitive guesses are sufficient and yield useable results:
- **Violin plot**: similar to a box plot, but rotated with addition of a kernel density plot on each side.
  - **Kernel density plot**: essentially a smoothing estimation based on finite data samples.
  - Statistical and IQR moments can be conveniently shown, sometimes with asymmetric comparisons of similar data sets (rather than a mirrored version).

Data Normalization and Outliers

Hypothesis Testing

T-Tests

Confidence Intervals

Correlation

Regression

Statistical Power and Sample Sizes

Clustering and Dimension-Reduction