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 quartiles (the box) and data min/max points (whiskers) in order to illustrate overall distribution of data and its potential outliers (often denoted by **).
- **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:

- $x_i,\ i = \dfrac{n+1}{2}$
- 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.

Measures of Disperion

▷ **Dispersion**: also called variability, scatter, or spread; a single number that describes how dispersed the data is around the central tendency.

▷ Main measures of dispersion: standard deviation and variance.

▷ Variance: indicates dispersion around the mean.

- Formula: $\sigma^2 = \dfrac{1}{n-1} \sum (x_i - \bar{x})^2$
- Suitable for any distribution.
- Works best with numerical data, or ordinal data with a mean.
- Taking the absolute value instead of the square of the mean difference results in the *mean absolute difference (MAD)*.
- Squaring emphasizes large values; better for optimization; closer to euclidean distance; is the second "moment"; better link to least-squares regression; and more.
- MAd is robust to outliers, though less commonly used.
- Dividing by $N-1$ is for sample variance, while N is for population.

▷ Standard deviation: simply the square root of variance.

▷ Knowing the standard deivation gives you variance and vice versa. Variance is more useful mathematically, while standard deviation has convenience of being expressed in units of the original variable.

▷ There other related measures such as *Fano factor* and *Coefficient of variation*, which are normalized measures of variability. Sensible only for datasets with positive values.

- *Fano factor*: $F = \frac{\sigma^2}{\mu}$; variance divided by the mean.
- *Coefficient of variation*: $CV = \frac{\sigma}{\mu}$; standard deviation divided by the mean.

Interquartile Range and QQ Plots

- Each half of the data made by the median can be divided further by taking the median again, resulting in 3 boundary points, or quartiles
- Quartile 1 is the "left"; quatile 2 is the middle, or "global median", and quartile 3 is the right.
- **Interquartile range (IQR)**: the range between quartile 1 and 2 that represnets 50% of the data.
- *Revisiting box plots*: IQR is represented by the box of the plot.
- **QQ plots**: aka quantile-quantile plots; a diagnostic scatter plot that compares two probability distributions by plotting their quantiles against each other in order to determin if it comes from a normal distribution.