

Contents

Data

Data Basics	2
Visualizing Data	2
Visualizing Data: Revisited (Post-Chapter 2)	3

Describing Statistics

Descriptive vs. Inferential	4
Accuracy, Precision, Resolution.	4
Data Distribution.	4
Measures of Central Tendency	4
Measures of Dispersion.	5
Interquartile Range and QQ Plots	6
Statistical Moments	6

Data Normalization

Z-Score Normalization.	8
Min-Max Scaling	8
Outliers	8

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 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 represent **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: $bin_i = 100 (bin_i / \text{sum}(bins))$
- ▷ **Pie charts**: representation of nominal, ordinal, or discrete data that must sum up to 1.

Visualizing Data: Revisited (Post-Chapter 2)

- ▷ Determining **number of bins** for histograms:
 - Number of bins (k) can be specified directly or calculated from width of bins h :
 - $k = \left\lceil \frac{\Delta x}{h} \right\rceil$
 - $\lceil \cdot \rceil$ represents the ceiling, or rounding up to nearest int.
 - **Sturges** guideline: $k = \lceil \log 2(n) \rceil + 1$
 - Advantage: relates to the data count.
 - **Freedman-Diaconis**: $h = 2 \frac{IQR}{\sqrt[3]{n}}$
 - Advantage: relates to both the data count and data spread.
 - Arbitrary choices, or other methods, often are decent enough and easier to implement, though Freedman-Diaconis is usually the best choice.
- ▷ **Violin plots**: a **rotated** and **mirrored** histogram.
 - IQR, mean, median, and distribution can all be easily represented.
 - Usually symmetric, but can compare two similar datasets asymmetrically.
 - Swarm plots are similar, except you can see individual data points.

2 Describing 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 or 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: μ , μ_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.

Measures of Dispersion

- ▷ **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: $\sigma^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
 - 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 deviation 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"; quartile 2 is the middle, or "global median", and quartile 3 is the right.
- ▷ **Interquartile range (IQR)**: the range between quartile 1 and 2 that represents 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 determine if it comes from a normal distribution.

Statistical Moments

- ▷ Unstandardized statistical moments:
 - General formula: $m_k = n^{-1} \sum (x_i - \bar{x})^k$
 - *First moment*: the **mean**, with a k value of 1.
 - *Second moment*: the **variance**, with k value of 2.
 - Further moments are increments of k .
- ▷ Standardized statistical moments:
 - Third and fourth moments are standardized with additional variance terms, $(n\sigma^k)^{-1}$ instead of just n .
 - **Skewness**: the *third moment*; dispersion asymmetry around the mean.
 - **positive skew**; the range of outliers pulled to right.
 - **negative skew**; the range of outliers pulled to the left.
 - **Kurtosis**: the *fourth moment*; the length of the distribution from the mean, heavy-tailed or light-tailed, relative to the normal distribution.
 - Data with **low kurtosis** have light tails, or **lack of outliers**.
 - Data with **high kurtosis** have heavy tails, or **more outliers**.

- ▷ There are further moments, but generally lack significance.
- ▷ **Shannon entropy**: entropy related to information processing that represents the average level of information/uncertainty inherent in a variable possible outcomes.
 - Surprising events convey more information.
 - Formula: $H = - \sum p(x_i) \log_2(p(x_i))$
 - x = data values, p = probability.
 - Used for nominal, ordinal, or discrete data.
 - Interval or ratio data must be converted to discrete by binning; entropy is affected by bin width and number.
 - **High entropy** means the dataset has **high variability**.
 - **Low entropy** means most values repeat and offer **redundant** information.
 - Entropy is related to variance, though it is **nonlinear** and makes **no assumptions** about distribution.
 - \log_2 entropy gives **bit** based units.
 - In entropy across datasets of consists units gives **nat** based units.

3 Data Normalization

Z-Score Normalization

- ▷ Problem: comparing two things of unrelated units.
- ▷ Solution: normalize both measurements to a unit-less scale.
- ▷ **Z-score**: or *standard score*: the **number of standard deviations** of a value from the mean.
 - Formula: $z_i = \frac{x_i - \bar{x}}{\sigma_x}$
 - Mean center: subtract the average from each value.
 - Variance-normalize: divide by the standard deviation.
 - Shifts and stretches, but doesn't change shape.
 - The distributions should be **roughly Gaussian**.

Min-Max Scaling

- ▷ Scale data into any arbitrary range, usually 0–1.
 - 0–1 scaled is often called unity-normalized.
- ▷ No relative information is lost.
- ▷ Formula: $\tilde{x}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
- ▷ Scale to arbitrary range: $x^* = a + \tilde{x}(b - a)$

Outliers

- ▷ Names for outlier: anomaly, extreme (deviant) data, non-representative data, noise, etc.
- ▷ Multi-dimensional data can make it very hard to determine outliers.
- ▷ **Leverage**: measure of how far away an independent variable value differs from other variables.
- ▷ Dealing with outliers:
 - Identify and remove outliers, assuming they are invalid.
 - Leave outliers in but use *other methods* that reduce negative impact of the outliers, assuming they are unusual but valid.
- ▷ Outliers must be investigated; don't ignore or remove without thought.
- ▷ Using z-score method:

- Convert data to z-score.
- Set a standard deviation threshold (usually 3), then remove outliers.
- Can iterate over same steps after removing an outlier.
- ▷ Modified z-score method for non-normal distributions:
 - Formula: $M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$
 - MAD: median absolute deviation (vs mean absolute deviation)
 - $MAD = \text{median}(|x_i - \tilde{x}|)$
 - $\tilde{x} = \text{median}(x)$
 - 0.6745 is the standard deviation units for 3rd quartile.