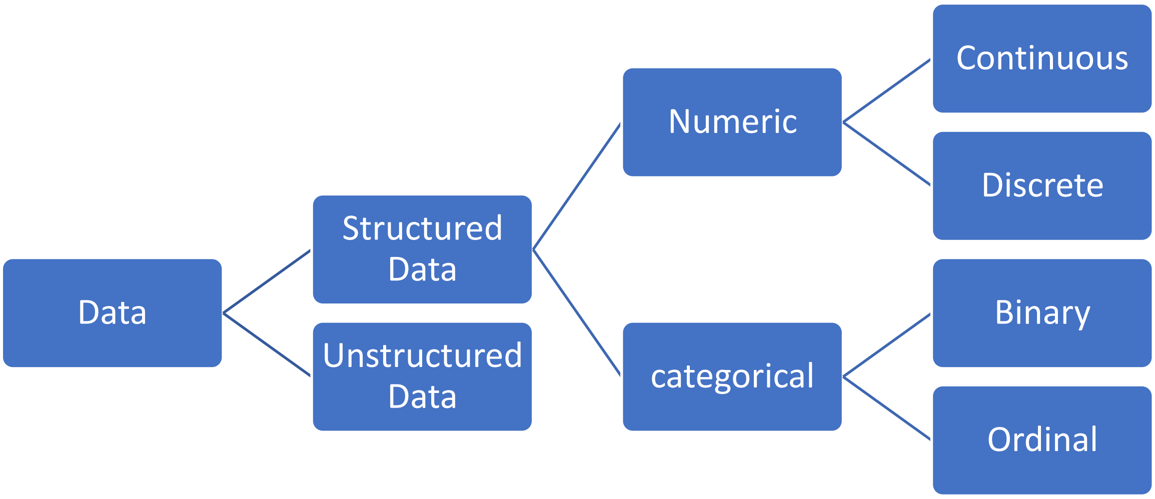
Statistics and Probability



- **Numeric Data** that are expressed on a numeric scale.
  - **Continuous Data** that can take on any value in an interval. (Synonyms: interval, float, numeric)
  - **Discrete Data** that can take on only integer values, such as counts. (Synonyms: integer, count)
- **Categorical Data** that can take on only a specific set of values representing a set of possible categories. (Synonyms: enums, enumerated, factors, nominal)
  - **Binary** A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (Synonyms: dichotomous, logical, indicator, boolean)
  - **Ordinal** Categorical data that has an explicit ordering. (Synonym: ordered factor)

Rectangular Data

Rectangular data is the general term for a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables).

Key Terms for Rectangular Data

- **Data frame** Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.
- **Feature** A **column** within a table is commonly referred to as a feature. (Synonyms: attribute, input, predictor, variable)
- **Outcome** Many data science projects involve predicting an outcome often a yes/no outcome. The features are sometimes used to predict the outcome in an experiment or a study. (Synonyms: dependent variable, response, target, output)
- **Records** A **row** within a table is commonly referred to as a record. (Synonyms: case, example, instance, observation, pattern, sample)

Estimates of Location

Estimates of location, also known as measures of central tendency, provide a single value that characterizes the **center point or typical value** of a dataset. It gives an estimate of where most of our data is located.

Mean

The **sum of all values divided by the number of values**. The mean is much more sensitive to the data which is the disadvantages in some cases. (Synonym: average)

Formula

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Weighted mean

The sum of all values times a weight divided by the sum of the weights. you calculate by **multiplying each data value x_i by a user-specified weight w_i** and **dividing their sum by the sum of the weights.** (Synonym: weighted average)

Formula

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed mean

The **average of all values** after **dropping** a fixed number of **extreme values**. For example, in international diving the top score and bottom score from five judges are dropped, and the final score is the average of the scores from the three remaining judges. This makes it difficult for a single judge to manipulate the score, perhaps to favour their country's contestant. (Synonym: truncated mean)

Formula

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

There are two main motivations for using a weighted mean:

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight.
- The data collected does not equally represent the different groups that we are interested in measuring.

Median

The median is the **middle number on a sorted list** of the data. If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves. Median is robust (Synonym: 50th percentile)

Weighted median

Instead of the middle number, the weighted median is a value such that the **sum of the weights is equal for the lower and upper halves of the sorted list**. Like the median, the weighted median is robust to outliers.

Other Estimates of Locations

- **Percentile** The value such that **P percent** of the data lies below. (Synonym: quantile)
- **Robust Not sensitive** to extreme values. (Synonym: resistant)
- **Outlier** A data value that is **very different** from most of the **data**. (Synonym: extreme value)

Estimates of Variability

Location is just one dimension in summarizing a feature. A second dimension, variability, also referred to as *dispersion*, **measures whether the data values are tightly clustered or spread out**. At the heart of statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.

Key Terms for Variability

- **Metrics Deviations** The difference between the observed values and the estimate of location. (Synonyms: errors, residuals)
- **Variance** The sum of squared deviations from the mean divided by n – 1 where n is the number of data values. (Synonym mean-squared-error)
  Formula:

  $$\text{Variance} = s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

  If you use the intuitive denominator of n in the variance formula, you will underestimate the true value of the variance and the standard deviation in the population. This is referred to as a biased estimate. However, if you divide by n – 1 instead of n, the variance becomes an unbiased estimate.
- **Standard deviation** The square root of the variance.
  Formula:

  $$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

  The variance and standard deviation are especially sensitive to outliers since they are based on the squared deviations.
- **Mean absolute deviation** The mean of the absolute values of the deviations from the mean. (Synonyms: l1-norm, Manhattan norm)
  Formula:

  $$\text{Mean absolute deviation} = \frac{\sum_{i=1}^{n}\left|x_i - \bar{x}\right|}{n}$$

  where $\bar{x}$ is the sample mean.
- **Median absolute deviation from the median** The median of the absolute values of the deviations from the median.

  Formula:

  $$\text{Median absolute deviation} = \text{Median}\left(\left|x_1 - m\right|, \left|x_2 - m\right|, \ldots, \left|x_N - m\right|\right)$$

  where m is the median.
- **Range** The difference between the largest and the smallest value in a data set.
- **Order statistics** Metrics based on the data values sorted from smallest to biggest. (Synonym: ranks)
- **Percentile** The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more. (Synonym: quantile)
- **Interquartile range** The difference between the 75th percentile and the 25th percentile. (Synonym: IQR)

Exploring the Data Distribution

It useful to explore how the data is distributed overall.

- **Boxplot** A plot as a quick way to visualize the distribution of data. (Synonym: box and whiskers plot). Boxplots are a simple way to visually compare the distributions of a numeric variable grouped according to a categorical variable.
- **Frequency table** A tally of the count of numeric data values that fall into a set of intervals (bins).
- **Histogram** A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis. While visually similar, bar charts should not be confused with histograms.
- **Density plot** A smoothed version of the histogram, often based on a kernel density estimate.

Exploring Binary and Categorical Data

- **Mode** The most commonly occurring category or value in a data set.
- **Expected value** When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence. "Expected value," which is a form of weighted mean, in which the weights are probabilities.

  The expected value is calculated as follows:

  1. Multiply each outcome by its probability of occurrence.
  2. Sum these values.
- **Bar charts** The frequency or proportion for each category plotted as bars.
- **Pie charts** The frequency or proportion for each category plotted as wedges in a pie.
- **Probability** The probability that an event will happen is the proportion of times it will occur if the situation could be repeated over and over, countless times.

Correlation

Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated. The correlation coefficient is sensitive to outliers in the data

- **Correlation coefficient**

  A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1). The correlation coefficient, which gives an esti mate of the correlation between two variables that always lies on the same scale. To compute Pearson's correlation coefficient, we multiply deviations from the mean for variable 1 times those for variable 2, and divide by the product of the standard deviations:

  $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

  The correlation coefficient always lies between +1 (perfect positive correlation) and –1 (perfect negative correlation); 0 indicates no correlation.
- **Correlation matrix** A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.
- **Scatter plot** A plot in which the x-axis is the value of one variable, and the y-axis the value of another. The standard way to visualize the relationship between two measured data variables is with a scatterplot. **Scatter plots** are fine when there is a relatively small number of data values.

Exploring Two or More Variables

Familiar estimators like mean and variance look at variables one at a time (univariate analysis). Correlation analysis is an important method that compares two variables (bivariate analysis). At more than two variables (multivariate analysis).

- **Contingency table** A tally of counts between two or more categorical variables.
- **Hexagonal binning** A plot of two numeric variables with the records binned into hexagons.
- **Contour plot** A plot showing the density of two numeric variables like a topographical map.
- **Violin plot** Similar to a boxplot but showing the density estimate.
- **Heat maps, hexagonal binning, and contour plots** all give a visual representation of a two-dimensional density. In this way, they are natural analogs to histograms and density plots.

Random Sampling and Sample Bias

- **Random sampling** Drawing elements into a sample at random. It is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw.
- **Sample** A subset from a larger data set.
- **Population** The larger data set or idea of a data set.
- **Stratified sampling** Dividing the population into strata and randomly sampling from each strata.
- **Stratum** (pl., strata) A homogeneous subgroup of a population with common characteristics.
- **Simple random sample** The sample that results from random sampling without stratifying the population.
- **Bias** Systematic error.
- **Sample bias** A sample that misrepresents the population.

Size Versus Quality: When Does Size Matter?

Data quality often matters more than data quantity when making an estimate or a model based on a sample. Data quality in data science involves completeness, consistency of format, cleanliness, and accuracy of individual data points. In the era of big data, it is sometimes surprising that smaller is better. Time and effort spent on random sampling not only reduces bias but also allows greater attention to data exploration and data quality.

Sample Mean Versus Population Mean

The symbol $\bar{x}$ (pronounced “x-bar”) is used to represent the mean of a sample from a population, whereas μ is used to represent the mean of a population.

- **Selection bias** Bias resulting from the way in which observations are selected.
- **Data snooping** Extensive hunting through data in search of something interesting.

Standard Error

The standard error is a single metric that sums up the variability in the sampling distribution for a statistic.

Formula:

$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

As the sample size increases, the standard error decreases. The relationship between standard error and sample size is sometimes referred to as the square root of n rule: to reduce the standard error by a factor of 2, the sample size must be increased by a factor of 4.

Standard Deviation Versus Standard Error

Standard deviation (which measures the variability of individual data points) with standard error (which measures the variability of a sample metric).

The Bootstrap

One easy and effective way to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample. This procedure is called the bootstrap, and it does not necessarily involve any assumptions about the data or the sample statistic being normally distributed.

Conceptually, you can imagine the bootstrap as replicating the original sample thousands or millions of times so that you have a hypothetical population that embodies all the knowledge from your original sample (it's just larger). You can then draw samples from this hypothetical population for the purpose of estimating a sampling distribution.

In practice, it is not necessary to actually replicate the sample a huge number of times. We simply replace each observation after each draw; that is, we sample with replacement. In this way we effectively create an infinite population in which the probability of an element being drawn remains unchanged from draw to draw. The algorithm for a bootstrap resampling of the mean, for a sample of size n, is as follows:

1. Draw a sample value, record it, and then replace it.
2. Repeat n times.
3. Record the mean of the n resampled values.
4. Repeat steps 1–3 R times.
5. Use the R results to:
    a. Calculate their standard deviation (this estimates sample mean standard error).
    b. Produce a histogram or boxplot.
    c. Find a confidence interval.

R, the number of iterations of the bootstrap.

The bootstrap does not compensate for a small sample size; it does not create new data, nor does it fill in holes in an existing data set. It merely informs us about how lots of additional samples would behave when drawn from a population like our original sample.

Resampling Versus Bootstrapping

Sometimes the term resampling is used synonymously with the term bootstrapping, as just outlined. More often, the term resampling also includes permutation procedures, where multiple samples are combined and the sampling may be done without replacement. In any case, the term bootstrap always implies sampling with replacement from an observed data set.