

Descriptive Statistics

Pablo E. Gutierrez-Fonseca

Fall 2024

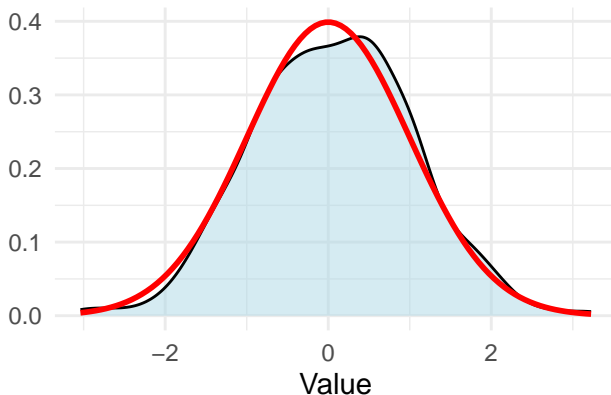


Why describe data?

- Determine if our sample reflects the population of interest.
- Identify outliers.
- Obtain metrics necessary for inferential tests.
- Understand the distribution of our data values – test for normality.
- Identify the type of statistical test to run.

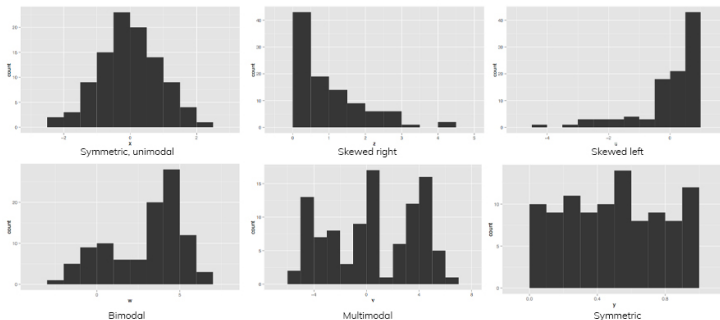
Data description and visualization

- We can examine our data and run statistical tests to see if the distribution approximates a normal curve.
- Typically, we start by visualizing our data.

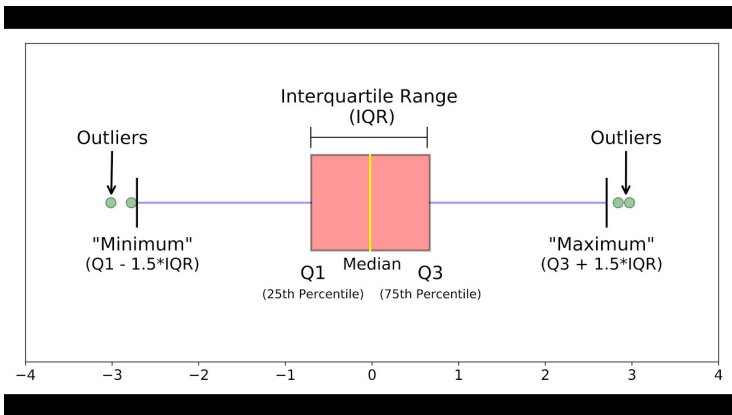


Histogram basic

- Continuous data are most commonly visualized using Histograms.



Box and Whisker Basics



Metrics to Describe data distribution.

- Data and their associated distributions can be described in four primary way:
 - ▶ Central Tendency (mean, median, mode)
 - ▶ Variability (standard deviation, variance, quantiles)
 - ▶ Skew
 - ▶ Kurtosis (Peakedness)

Central tendency

- Mean (Sum of scores/ N)
 - ▶ Most often used measure of central tendency.
 - ▶ Works well with normal and relatively normal curves.
- Median (50th Percentile)
 - ▶ No formula. Rank order observations then find the middle.
 - ▶ The second most used measure of central tendency.
 - ▶ Works best with highly skewed populations.
- Mode (Most Frequent Score)
 - ▶ Least used measure of central tendency.
 - ▶ Works best for highly irregular and multimodal distributions.

Central tendency: Mean

- Sample mean is the measure of central tendency that best represents the population mean.
- Mean is **very** sensitive to extreme scores that can “skew” or distort findings.

Central tendency: Median

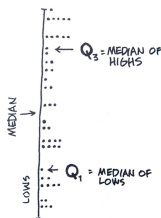
- Percentiles are used to define the percent of cases equal to and below a certain point on a distribution.
 - ▶ The median **is the 50th percentile** half of all observations fall at or below this value.
- But lots of other percentiles are also important.

A little about Percentiles

- Quartiles are a common percentile used to represent the value below which.
 - ▶ 25% (Q_1 or first quartile)
 - ▶ 75% (Q_3 or third quartile)

HERE'S THE RECIPE:

- 1)** PUT THE DATA IN NUMERICAL ORDER.
- 2)** DIVIDE THE DATA INTO TWO EQUAL HIGH AND LOW GROUPS AT THE MEDIAN. (IF THE MEDIAN IS A DATA POINT, INCLUDE IT IN BOTH THE HIGH AND LOW GROUPS.)
- 3)** FIND THE MEDIAN OF THE LOW GROUP. THIS IS CALLED THE FIRST QUARTILE, OR Q_1 .
- 4)** THE MEDIAN OF THE HIGH GROUP IS THE THIRD QUARTILE, OR Q_3 .

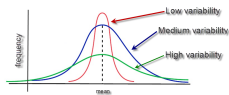


When to use What

- Use the **Mode** when the data are categorical:
 - ▶ **Mode**: is the value that occurs most frequently in your data.
 - ▶ This is because having the same value occur for measurements with many significant digits is highly unlikely.
- Use the **Median** when you have extreme scores:
 - ▶ **Median**: is simply the value that falls in the middle of all your data.
- Use the **Mean** the rest of the time.



Variability



Variability: Standard Deviation

- Standard Deviation measures how spread out the numbers in a dataset are around the mean.
- The sample standard deviation s is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Variability

- Variance measures the average of the squared differences from the mean, indicating how spread out the data points are.

The variance σ^2 is calculated as:

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

Variability: Range

- Range is the difference between the largest and smallest values in a dataset, providing a measure of the spread or dispersion of the data.
- The range is calculated as:

$$\text{Range} = \max(x) - \min(x)$$

Percentiles are useful for spread too

- You can use percentiles to get a feel for how spread out the data is and where most of your observations are contained:
 - ▶ Inter-quartile range (IQR) = $Q3 - Q1$

Identifying outliers

- An outlier is an observation that lies outside the overall pattern of a distribution (Moore and McCabe 1999).
- Usually, the presence of an outlier indicates some sort of problem. (e.g. an error in measurement or sample selection).
- But they may also be an indicator of novel data or identification of unique and exciting observations.

Identifying outliers

- The first and third quantiles (Q1 and Q3) are often calculated to identify outliers.
- One method for systematically identifying outliers uses:
 - ▶ $Q1 - (1.5 \times \text{the inter-quartile range})$
 - ▶ $Q3 + (1.5 \times \text{the inter-quartile range})$
- Others identify outliers as any values below the 0.5th or above the 99.5th percentile.

When to use What

- Use the **Standard deviation (SD)** in most cases.
 - ▶ SD quantifies how far, on average, each observation is from the mean.
 - ▶ The larger the SD, the more highly variable your data.
- Use range (R) when describing predictive models.
 - ▶ R is simply the maximum minus the minimum value in your data set
 - ▶ R is important when modeling or making predictions, since your algorithms are valid only over the range of values used to calibrate your predictive model
- Use the IQR to identify and test potential outliers in your data.

Skewness

- Skewness: This metric quantifies how balanced (symmetrical) your distribution curve is.