Introduction to summary statistics

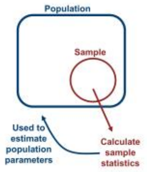
Location, spread, and shape

Estimating data values

We usually care about the **parameters** of a **population** (all the individuals of interest), but we can't measure them all.

We therefore measure the values for a subset and calculate the **statistics** of a **sample**.

Our sample statistics are **estimates** of the population parameters.


Population
Sample
Used to estimate population parameters
Calculate sample statistics

Summarizing a sample or population

Populations and samples usually have too many values to make sense of every single one of them.

We typically want to know their ***basic properties*** anyway, not every single value.

e.g., to see whether a disease raises diastolic pressure, do we really need to look at all the numbers individually, or does the average accomplish the goal?

Baseline diastolic pressures:
125,144,119,115,131,125,136,131,117,120,115,110,119,126,113,134,128

Disease diastolic pressures:
122,141,138,141,143,127,132,137,135,141,129,131,130,127,125,127,118

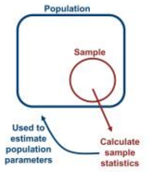
Baseline diastolic pressure average:
124

Disease diastolic pressure average:
132

Fundamental properties of data sets

What are the basic properties of data sets that we typically want to know?

- Location: what is the typical value?
- Spread: how variable are the values?
- Shape: how does the distribution compare to others with similar locations and spreads?


Population
Sample
Used to estimate population parameters
Calculate sample statistics

Statistics of location

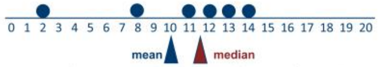
Question: what is the typical value ... the *average* ?

4 different averages:
- **Mean**: sum of values, divided by the number of values.
- **Median**: value in the center, 50% on each side.
- **Mid-range**: halfway between smallest and largest.
- **Mode**: Most frequent or common value.

First two commonly used, last two rarely used.

Statistics of location

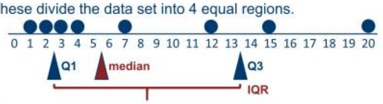
- **Mean**: sum of values, divided by the number of values.
- **Median**: value in the center, 50% on each side.



Mean uses the exact positions, like a "center of mass."
Median divides data into regions, upper and lower.
Both provide a location the values are arranged around.

Quartiles

These divide the data set into 4 equal regions.

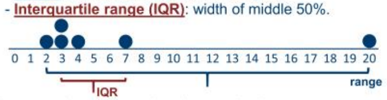


Q1: **first quartile**, placed at 25% spot. (25/75)
Q2: median, placed at 50% spot. (50/50)
Q3: **third quartile**, placed at 75% spot. (75/50)

IQR = Q3 - Q1

Statistics of spread

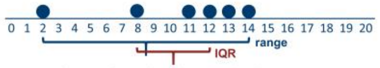
- **Range**: distance from smallest to largest.
- **Interquartile range (IQR)**: width of middle 50%.



The range is not robust, it only uses 2 values.
The IQR is robust, uses quartiles.

Statistics of spread

- **Range**: distance from smallest to largest.
- **Interquartile range (IQR)**: width of middle 50%.



The range is not robust, it only uses 2 values.
The IQR is robust, uses quartiles.

Statistics of spread (i.e., variability or dispersion)

Question: how variable are the values ?

6 different spreads:
- **Range**: distance from smallest to largest.
- **Interquartile range (IQR)**: width of middle 50%.
- **Sum of squares (SS)**: sum of squared differences from the mean.
- **Variance (Var)**: Mean of the SS values.
- **Standard deviation (SD)**: square root of the variance.
- **Coefficient of variation**: SD relative to the mean.

Statistics of location

How robust (consistent or resistant to randomness) are these values?

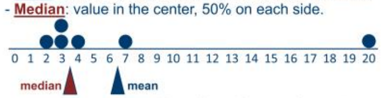
Presence or absence of outliers (rare extreme values):
- Mean: not robust.
- Median: **robust**
- Mid-range: not robust
- Mode: **robust**

When calculated from repeated samples from a population:
- Mean: **robust**
- Median: **robust**
- Mid-range: not robust
- Mode: not robust

Statistics of location

Beware of "the average ..."

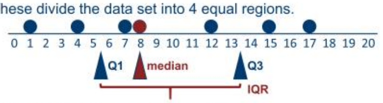
- **Mean**: sum of values, divided by the number of values.
- **Median**: value in the center, 50% on each side.



Mean uses the exact positions, like a "center of mass."
Median divides data into regions, upper and lower.
Both provide a location the values are arranged around.

Quartiles

These divide the data set into 4 equal regions.



Q1: **first quartile**, placed at 25% spot. (25/75)
Q2: median, placed at 50% spot. (50/50)
Q3: **third quartile**, placed at 75% spot. (75/50)

IQR = Q3 - Q1

Calculating quartiles

Simple method. This method ignores the shape of the distribution and just focuses on the values.

- Step 1. Arrange the n values in order, smallest to largest.
- Step 2a. *If n is odd*: median is the center value (include this value in both halves during step 3).
- Step 2b. *If n is even*: median is the mean of the two middle values (do not include this value when doing step 3).
- Step 3. Repeat steps 2a and 2b for each half of the data set.
  - Q1 is the median of the small half of data.
  - Q3 is the median of the large half of data.

Calculating quartiles

Other methods exist because the definition is arbitrary. If the exact quartiles are important, make sure you know which is being used.

Alternative 1. Same as previous, but don't include median to calculate Q1 and Q3 for data sets with an odd number of values.

Alternative 2. Calculate Q1 and Q3 using a weighted average of the data points if data set has an odd number of values.

For 4n+1 values: Q1 is 0.25 nth value + 0.75 (n+1)th value
Q3 is 0.75 (3n+1)th value + 0.25 (3n+2)th value
For 4n+3 values: Q1 is 0.75 (n+1)th value + 0.25 (n+2)th value
Q3 is 0.25 (n+2)th value + 0.75 (n+3)th value

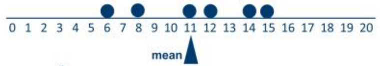
Statistics of spread (i.e., variability or dispersion)

- **Sum of squares (SS)**: sum of squared differences from the mean.
- **Variance (Var)**: Mean of the SS values.
- **Standard deviation (SD)**: square root of the variance.
- **Coefficient of variation**: SD relative to the mean.

Medians and quartiles are used for describing data sets, but these are used in statistical tests (due to math property of sums of squares and variances).

Statistics of spread (i.e., variability or dispersion)

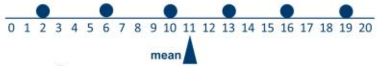
- **Sum of squares (SS)**: sum of squared differences from the mean.



$$SS = \sum_{i=1}^{6}(x_i - \bar{x})^2 = 5^2 + 3^2 + 0^2 + 1^2 + 3^2 + 4^2 = 60$$

Statistics of spread (i.e., variability or dispersion)

- **Sum of squares (SS)**: sum of squared differences from the mean.



$$SS = \sum_{i=1}^{6}(x_i - \bar{x})^2 = 9^2 + 5^2 + 1^2 + 2^2 + 5^2 + 8^2 = 200$$

Something to be aware of

$$\sum(x_i - \bar{x})^2 = \sum x_i{}^2 - \frac{(\sum x_i)^2}{n}$$

There are two different equations that can be used to calculate the sum of squares. One is easy to understand, the other faster to compute. e.g., data = 3,4,6,7 ($\bar{x}$ = 5)

$$SS = \sum(x_i - \bar{x})^2 = (3-5)^2 + (4-5)^2 + (6-5)^2 + (7-5)^2 = 10$$

$$SS = \sum x_i{}^2 - \frac{(\sum x_i)^2}{n} = 3^2 + 4^2 + 6^2 + 7^2 - \frac{(3+4+6+7)^2}{4} = 10$$

The second was used for years so is still in textbooks.

The coefficient of variation (i.e., CV)

Always positive.
Not robust to outliers (since based on SS).

Population

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

Sample

$$CV = \frac{s}{\bar{x}} \times 100$$

Puts variation into context. Used as descriptive statistic.
Not super common, rarely used in statistical tests.

The standard deviation (i.e., σ or s)

Always positive.
Not robust to outliers (since based on SS).

Population: σ

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

Sample: s

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

Standard deviation has same units as the original data, unlike the variance.

The variance (i.e., σ^2 or s^2)

Always positive.
Not robust to outliers (since based on SS).

Population: σ^2

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

Sample: s^2

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

The estimate of a population variance using a sample often underestimates, therefore the different denominator is used to create an *unbiased* estimate.

Statistics of spread (i.e., variability or dispersion)

- **Sum of squares (SS)**: sum of squared differences from the mean.

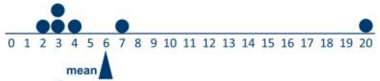
This is the foundation of the variance, standard deviation, and coefficient of variation.

Extremely prone to being inflated by outliers. ☹

However, the variance (next statistic) has a useful property: for two *independent* data sets, the sum of their variances is the same as the variance of the combined data set.

Statistics of spread (i.e., variability or dispersion)

- **Sum of squares (SS)**: sum of squared differences from the mean.



$$SS = \sum_{i=1}^{6}(x_i - \bar{x})^2 = 4^2 + 3^2 + 3^2 + 2^2 + 1^2 + 14^2 = 235$$

Statistics of shape (relative to the normal)

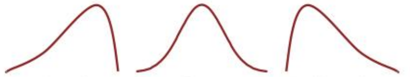
The normal distribution (i.e., bell curve) is a baseline shape for comparison.

2 (or 3) different shape statistics:
- **Skewness**: measures asymmetry.
- **Kurtosis**: measures peakedness (really measures tail thickness).
- **Excess Kurtosis**: Kurtosis -3.

Skewness

Measures asymmetry.

$$Skewness = \frac{\left(\frac{\sum(x_i - \bar{x})^3}{n}\right)}{\sigma^3}$$



Negative value
"skewed left"
A tail to the left.

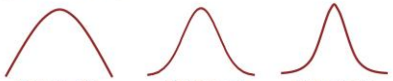
Zero
Balanced
Symmetric

Positive value
"skewed right"
A tail to the right.

Kurtosis

Measures peakedness (thickness of tails)

$$Kurtosis = \frac{\left(\frac{\sum(x_i - \bar{x})^4}{n}\right)}{\sigma^4}$$



Kurtosis < 3
Excess kurtosis<0
Platykurtic

Kurtosis = 3
Excess kurtosis=0
Mesokurtic

Kurtosis >3
Excess kurtosis>0
Leptokurtic

Estimating skewness and kurtosis

The previous formulas are population formulas.
Estimates of the population skewness or excess kurtosis from sample data can be biased, these are better:

$$Skewness = \frac{\left(\frac{\sum(x_i - \bar{x})^3}{n}\right)}{s^3} \quad or \quad \frac{\sqrt{n(n-1)}}{n-2}\frac{\left(\frac{\sum(x_i - \bar{x})^3}{n}\right)}{\sigma^3}$$

2nd equation is more common

$$\begin{matrix}Excess\\Kurtosis\end{matrix} = \frac{(n+1)n}{(n-1)(n-2)}\frac{\left(\frac{\sum(x_i - \bar{x})^4}{n}\right)}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Application of skewness and kurtosis

The skewness and kurtosis are rarely studied for their own sake.

They are usually calculated to see if the distribution is normal (which we usually want).

$$Skewness = \frac{\left(\frac{\sum(x_i - \bar{x})^3}{n}\right)}{\sigma^3} = 0?$$

$$Kurtosis = \frac{\left(\frac{\sum(x_i - \bar{x})^4}{n}\right)}{\sigma^4} = 3?$$

$$Excess\ Kurtosis = Kurtosis - 3 = 0?$$

Statistics of location, spread, and shape

Location	mean	DESCRIPTIVE	TESTING
	median	DESCRIPTIVE	
	mid-range		
	mode		
Spread	Range	descriptive	
	IQR	descriptive	
	SS		
	Var		TESTING
	SD	DESCRIPTIVE	
	CV	descriptive	
Shape	Skewness		testing pre-req
	kurtosis		testing pre-req

StatsExamples.com