[TOC]

Lecture 5

Today

- Last lecture we described central tendency.
- Today we discuss variability or dispersion.
- Range (minimum, maximum, inter-quartile, decile)
- Summed deviations
- Absolute deviations
- Variance
- Standard deviation

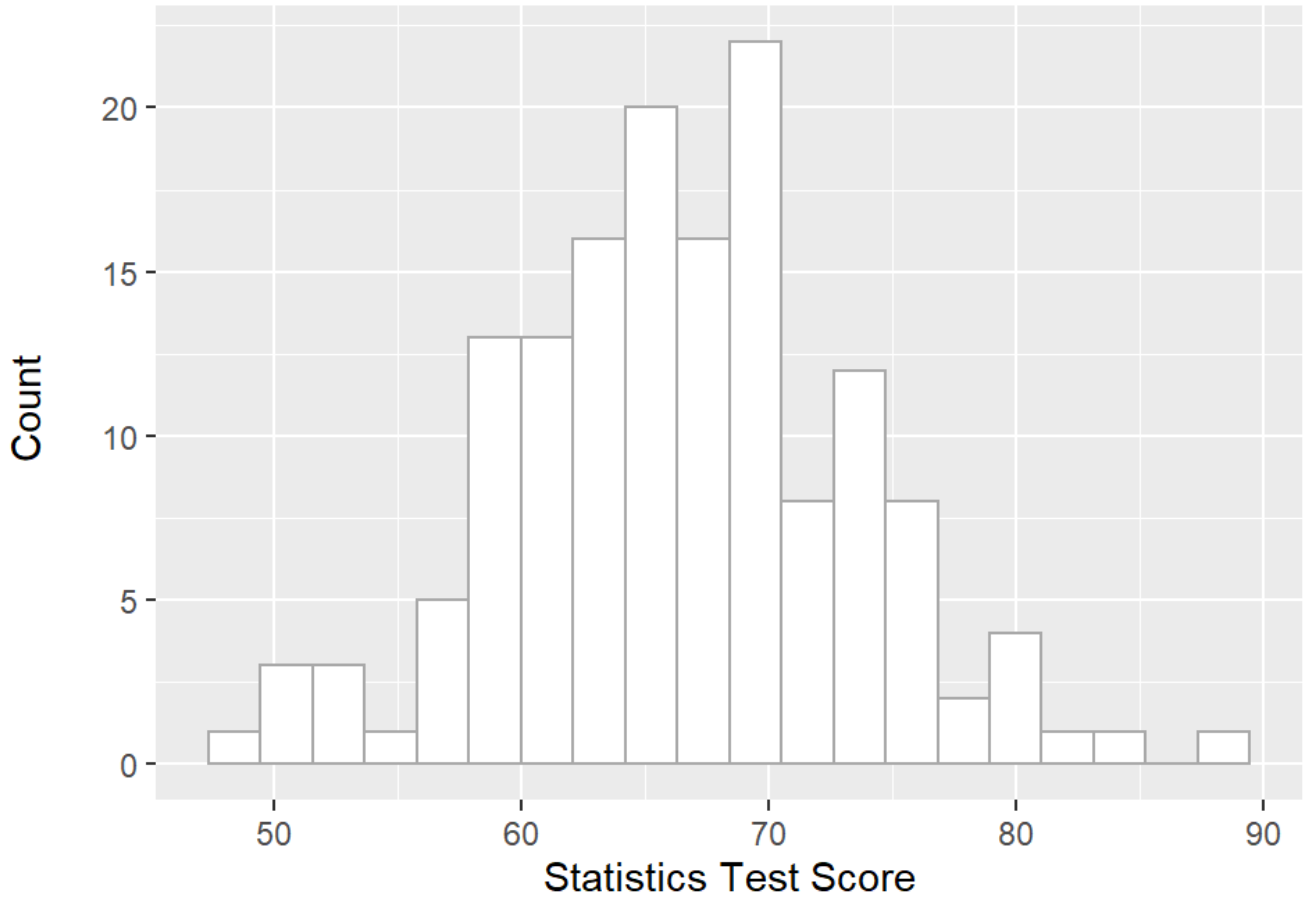
Learning objectives

- Be able to describe the 5 measures of variability.
- Understand to what data we can apply each measure of variability.
- Understand the pro's and con's of each measure.
- Develop an initial understanding of how to calculate in R.

Data (n=150)

ID	Degree	StatsScore
ID101	Psych	67
ID102	Ling	69
ID103	Phil	70
ID104	Phil	69
ID105	Phil	68
ID106	Phil	70
ID107	Psych	58
ID108	Psych	66
ID109	Ling	66
ID110	Psych	72

Statistic Score Distribution



Range

- The **range** of the data is simply the value between two points.
- We can define these points in different ways.
- Range is generally used to describe the total range (max - min) of the data.

```
ex1 %>%
  select_if(is.numeric) %>%
  summarise(
    Variable = names(.),
    Sample = n(),
    Minimum = min(.),
    Maximum = max(.)) %>%
  mutate(Range = Maximum - Minimum)
```

Variable	Sample	Minimum Maximum	Range
StatsScore	150	48	88

- Problem: Very sensitive to outliers.

Inter-quartile range

- The inter-quartile range (IQR) is the difference between the 1st and 3rd quartile.
  - Rank the data
  - Split data into four equal blocks.
  - Quartiles are the points which divide these blocks.
  - They fall at 25%, 50% and 75% of rank ordered data.
  - IQR is the difference between 25% and 75%

```
iqr <- quantile(ex1$StatsScore)
res <- tibble(
  "0%" = iqr[[1]],
  "25%" = iqr[[2]],
  "50%" = iqr[[3]],
  "75%" = iqr[[4]],
  "100%" = iqr[[5]]
)
```

0%	25%	50%	75%	100%
48	62	66.5	70	88

- So our IQR = 8
- We can calculate various quantiles using the quantile() function, and the IQR directly using the function IQR()
- Problem:
  - Still somewhat sensitive to outliers.
  - Ignores half the data.

Use of range

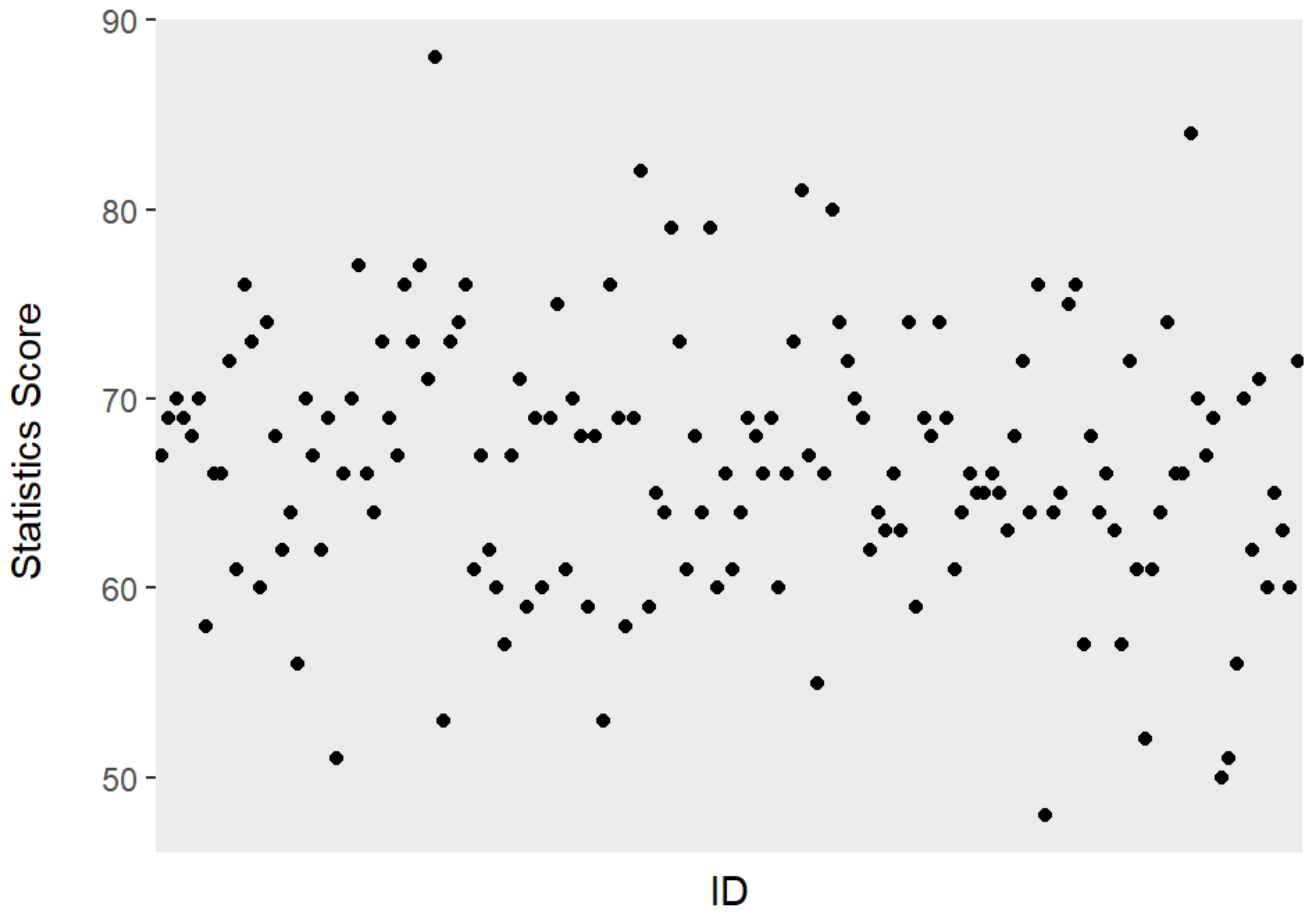
- In principle, we can define a range using any two values.
- Commonly researchers will also use deciles (blocks of 10%) to define a range.
  - E.g. Range between 10% and 90%

```
iqr2 <- quantile(ex1$StatsScore, c(.1, .9))
res2 <- tibble(
  "10%" = iqr2[[1]],
  "80%" = iqr2[[2]],
)
```

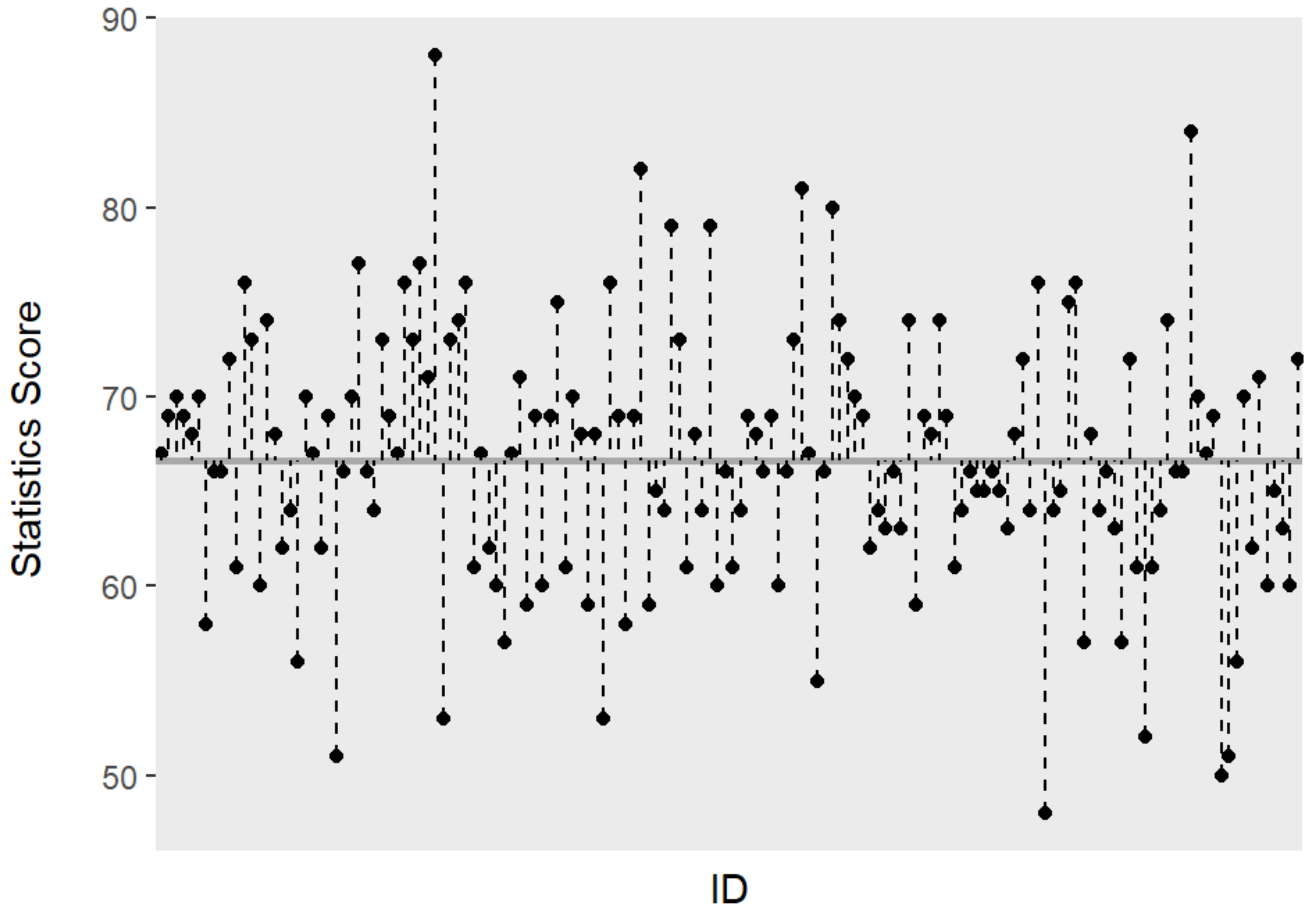
10%	80%
58.9	75.1

- So the range across 80% of the ranked data is 16.2.
- Ranges rely on **rank-ordered data** .
- Hence they are most often used alongside the median as a measure of central tendency.

Variation around the mean



!



Sum of deviations

- We could just add up the amount by which each observation differs from the mean.
- This is called the **sum of deviations**.

$$S = \sum_{N}^{i=1}(x_i - \bar{x})$$

x_i

individual observations

$\bar{x}$

mean of x

Calculation

ID	StatsScore	Mean	Deviance
ID102	69	67.5	1.5
ID103	70	67.5	2.5

ID	StatsScore	Mean	Deviance
ID104	68	67.5	0.5
ID106	70	67.5	2.5
ID107	58	67.5	-9.5
ID108	66	67.5	-1.5
ID109	66	67.5	-1.5
ID110	72	67.5	4.5

Problem: Sum of deviations

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(StatsScore - mean(StatsScore)),2)
  )
```

Variable	Sum Deviation
Statistics	Test Score 0

- The positive and negative values cancel.
- This means S will always be 0 when calculated around the mean.

Sum of absolute deviations

- Suppose we just ignored the direction of the difference, and just considered the magnitude of the difference?
- This is the sum of absolute deviations.
- Here the `||` are read as absolute, which means that we remove the sign.

ID	StatsScore	Mean	Deviance	Absolute
ID101	67	67.5	-0.5	0.5
ID102	69	67.5	1.5	1.5
ID103	70	67.5	2.5	2.5
ID104	69	67.5	1.5	1.5

ID	StatsScore	Mean	Deviance	Absolute
ID105	68	67.5	0.5	0.5
ID106	70	67.5	2.5	2.5
ID107	58	67.5	-9.5	9.5
ID108	66	67.5	-1.5	1.5
ID109	66	67.5	-1.5	1.5
ID110	72	67.5	4.5	4.5

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(StatsScore - mean(StatsScore)),2),
    "Abs Sum Deviation" =  round(sum(abs(StatsScore - mean(StatsScore))),2)
  )
```

Variable	Sum Deviation	Abs Sum Deviation
Statistics Test Score	0	796

- Problem:
  - As sample gets bigger, the sum of deviations continues to grow.

Mean sum of absolute deviations

- To resolve, we could scale it by sample size.
- Thus we get the **mean sum of absolute deviations**.

$$\bar{S}_{abc} = \frac{\sum_{N}^{i=1} |(x_i - \bar{x})|}{N}$$

- Tells us how far on average all points are from the centre of the values.

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(StatsScore - mean(StatsScore)),2),
    "Abs Sum" =  round(sum(abs(StatsScore - mean(StatsScore))),2),
    "Mean Abs Sum" = round(sum(abs(StatsScore - mean(StatsScore)))/length(StatsS
  )
```

Variable	Sum Deviation	Abs Sum Mean	Abs Sum
Statistics Test Score	0	796	5.31

- Divide the absolute sum by N.

Variance

- Instead of using absolute values to deal with negatives, we could instead square the differences.
- This is called the variance .

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

- Variance is the mean squared deviation from the mean.

$$\sigma^2$$

= variance (Greek letter lower case sigma)

ID	StatsScore	Mean	Deviance	Deviance_sq
ID101	67	67.5	-0.5	0.25
ID102	69	67.5	1.5	2.25
ID103	70	67.5	2.5	6.25
ID104	69	67.5	1.5	.25
ID105	68	67.5	0.5	0.25
ID106	70	67.5	2.5	6.25
ID107	58	67.5	-9.5	90.25
ID108	66	67.5	-1.5	2.25
ID109	66	67.5	-1.5	2.25
ID110	72	67.5	4.5	20.25

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(StatsScore - mean(StatsScore)),2),
    "Abs Sum" =  round(sum(abs(StatsScore - mean(StatsScore))),2),
    "Mean Abs Sum" = round(sum(abs(StatsScore - mean(StatsScore)))/length(Stats
```

```
    Variance = round((sum((StatsScore - mean(StatsScore))^2))/length(StatsScore)
  )
```

Variable	Sum Deviation	Abs Sum	Mean Abs Sum	Variance
Statistics Test Score	0	796	5.31	47.18

- Problem:
- Our units here are not quite right.
- Variance is the mean squared deviation from the mean.

Standard deviation

- What about a measure of variation in the same units as the mean/variable?
- The **standard deviation**.
- The standard deviation is the square root of the variance.
  - This fixes our unit/scaling problem.

$$\sigma = \sqrt{\frac{sum_{N}^{i=1}(x_i - barx)^2}{N}}$$

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(StatsScore - mean(StatsScore)),2),
    "Abs Sum" =  round(sum(abs(StatsScore - mean(StatsScore))),2),
    "Mean Abs Sum" = round(mean(abs(StatsScore - mean(StatsScore))),2),
    Variance = round((sum((StatsScore - mean(StatsScore))^2))/length(StatsScore)
    SD = round(sqrt((sum((StatsScore - mean(StatsScore))^2))/length(StatsScore))
  )
```

Variable	Sum Deviation	Abs Sum	Mean Abs Sum	Variance	SD
Statistics Test Score	0	796	5.31	47.18	6.87

Easier options

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(StatsScore - mean(StatsScore)),2),
    "Abs Sum" =  round(sum(abs(StatsScore - mean(StatsScore))),2),
    "Mean Abs Sum" = round(mean(abs(StatsScore - mean(StatsScore))),2),
```

```
    Variance = round(var(StatsScore),2),
    SD = round(sd(StatsScore),2)
  )
```

Which measure should we use?

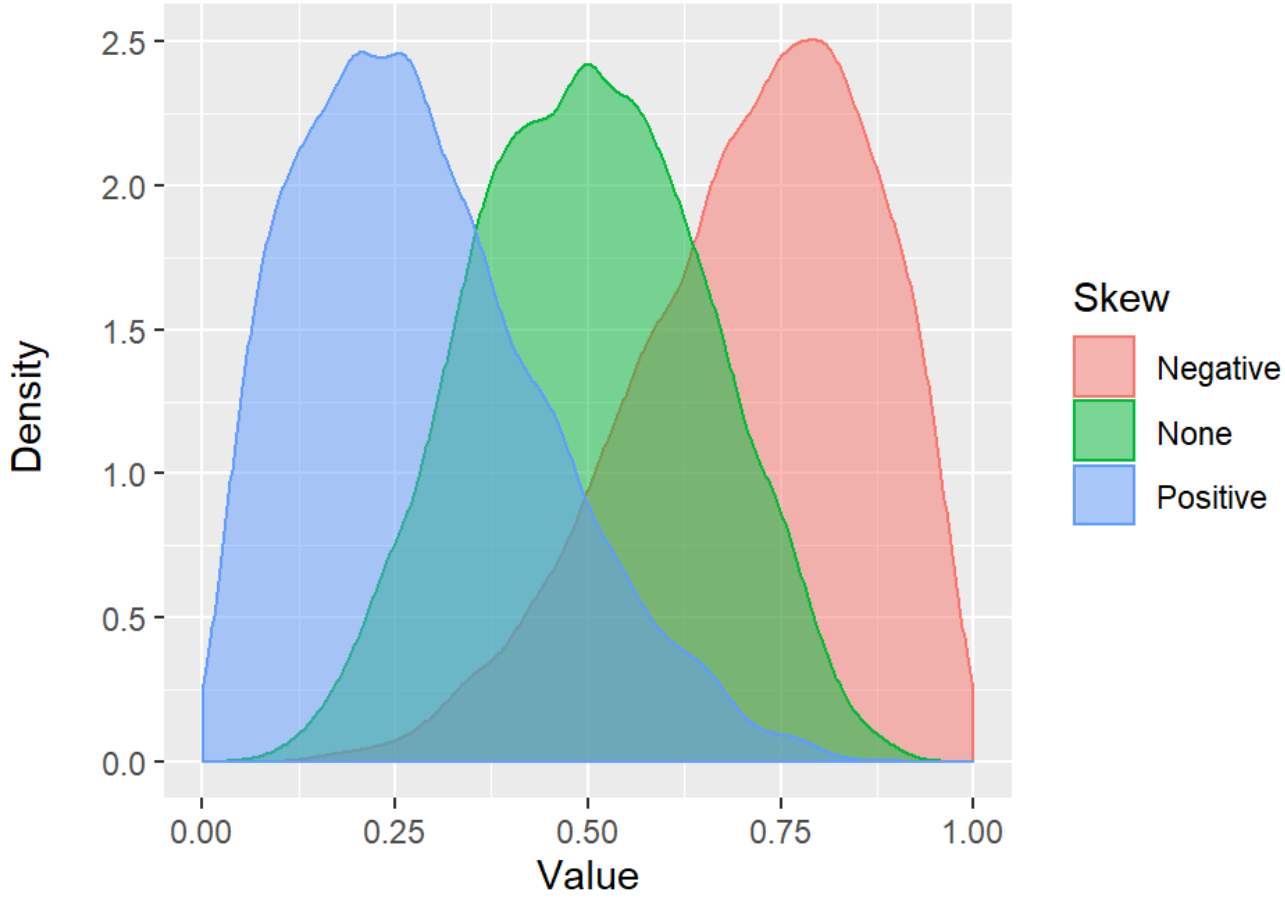
Variable Type	Central Tendency	Dispersion
Categorical (Nominal)	Mode	Frequency Table
Categorical (Ordered)	Mode/Median	Range
Continuous	Mean (any in fact)	Variance & Standard Deviation
Count	Mode (mean)	Range (Variance & SD)

- Depends on the level of measurement.

A few extra bits?

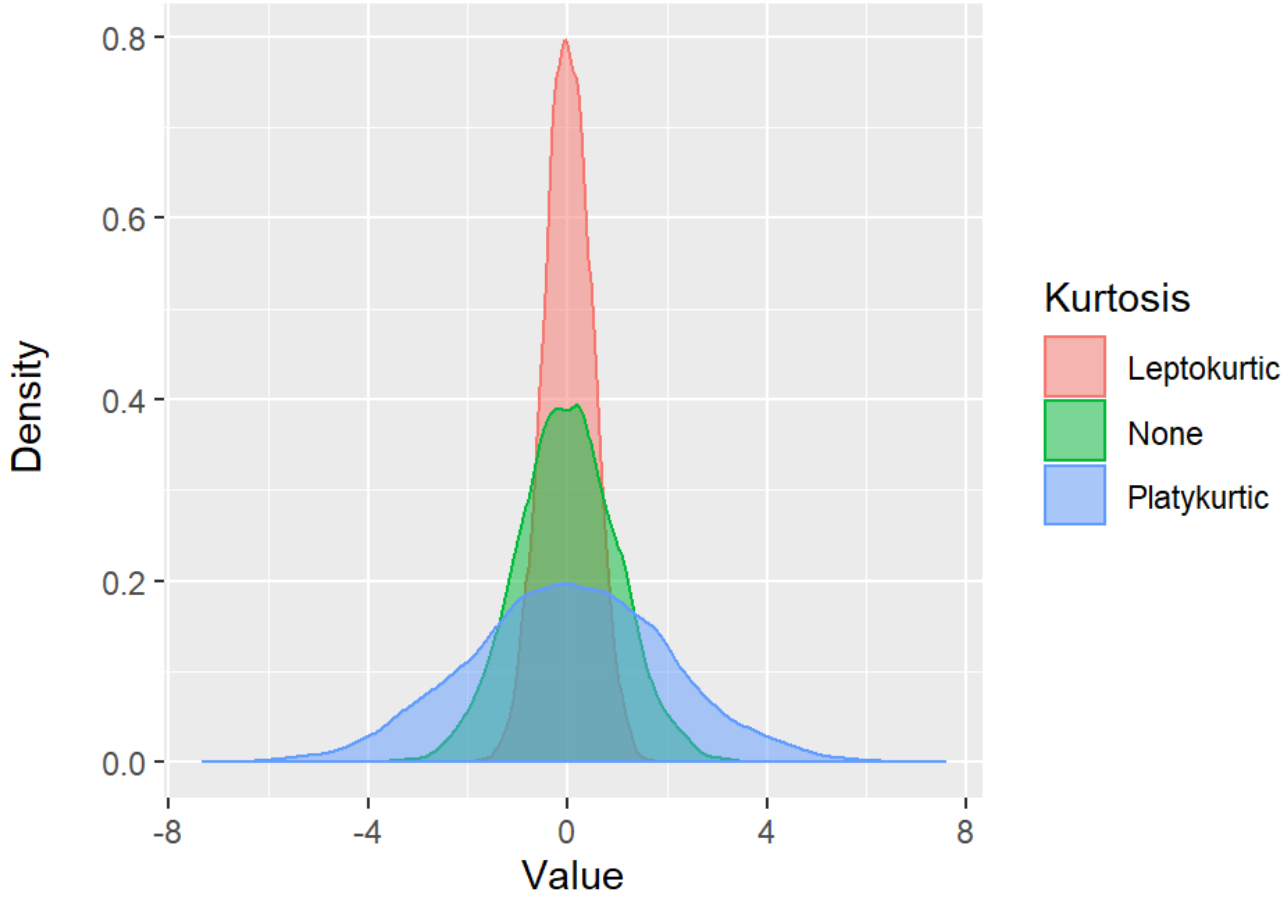
- You may come across the mathematical language of moments.
- Moments describe the shape of a set of points
- Mean
- Variance
- Skew
- Kurtosis

Skew



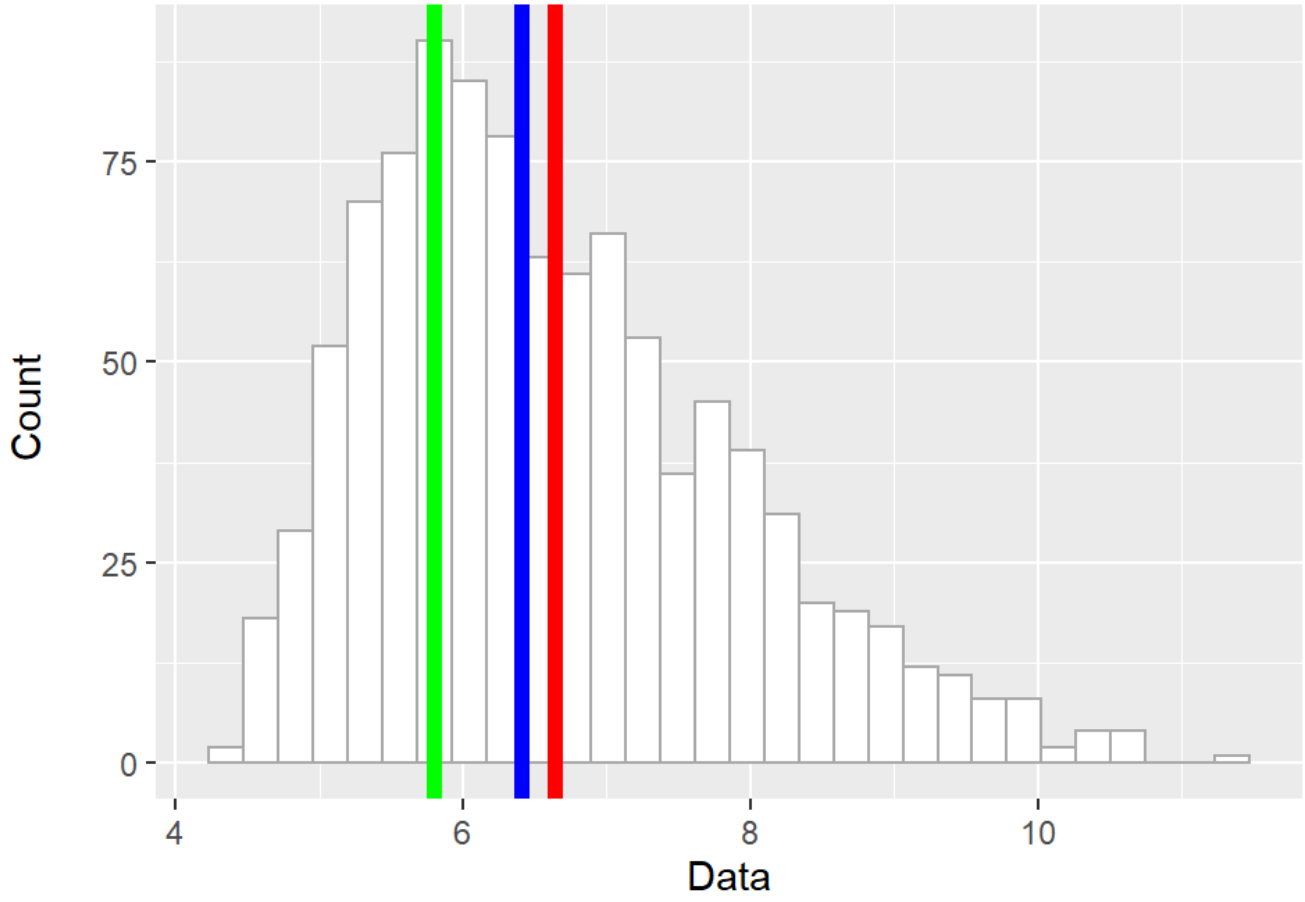
- Is a measure of asymmetry of a distribution.

Kurtosis



- Kurtosis is a measure of the flatness of the peak and the fatness of the tails of the distribution.

Do they matter?

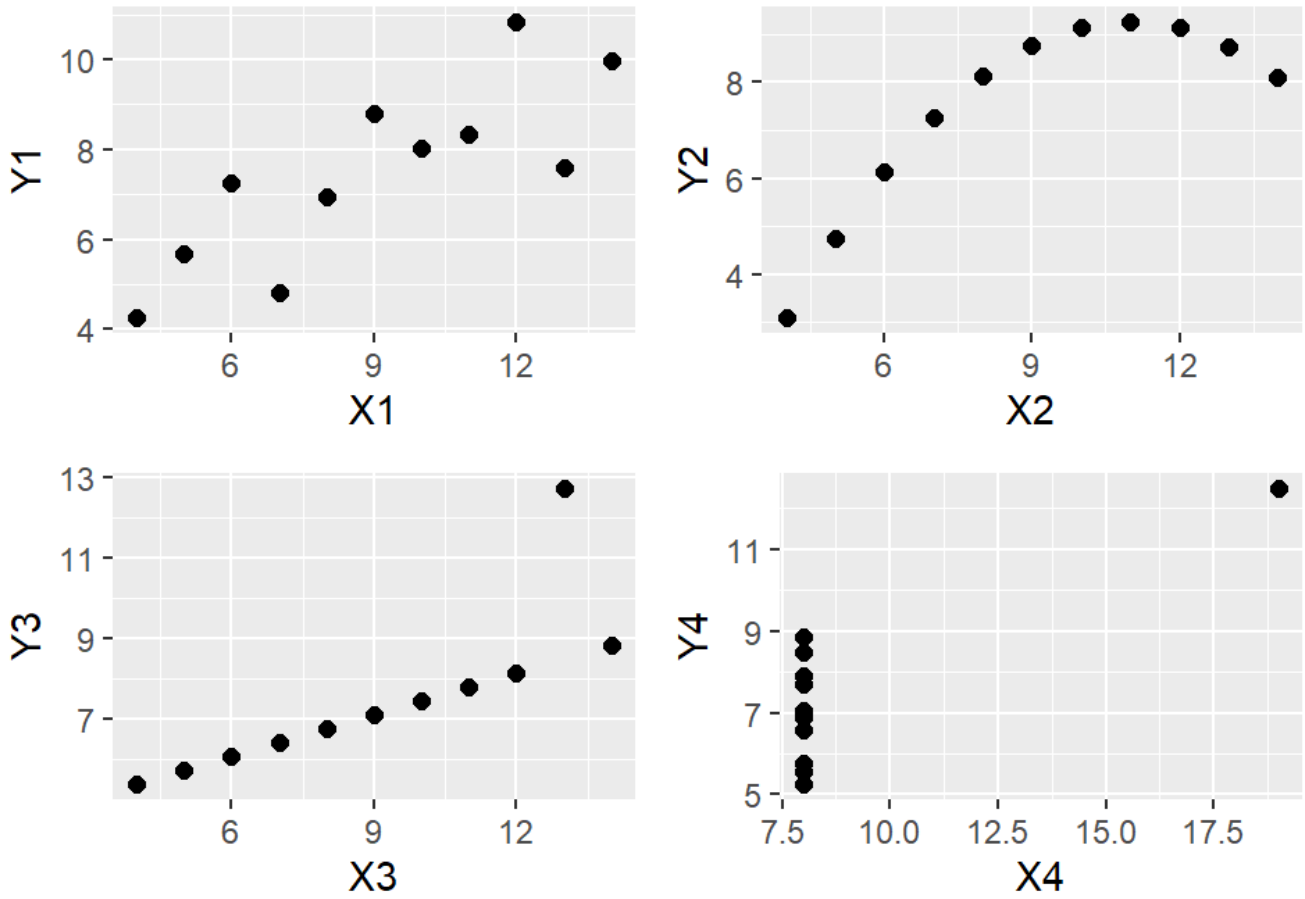


- It can make a difference in how we describe data.
- Both skew and kurtosis impact the normality of the distribution of the data.

Descriptive statistics can deceive

set	mean	sd
x1	9.0	3.32
x2	9.0	3.32
x3	9.0	3.32
x4	9.0	3.32
y1	7.5	2.03
y2	7.5	2.03
y3	7.5	2.03
y4	7.5	2.03

Always visualize data



Tasks for this week...

1. Finish any existing tasks from lab 4.
2. Reading: Linked at the top of lab 5.
3. Quiz 5: Central tendency

- **This quiz counts**
- Live now (as of Monday at 09:00).
- Closes Sunday at 17:00
-

Recommendations of the week

Podcast: Ineos159 Preview documentary Recipe: Roasted Veggie Lasagna Book: Stoner, John Williams Thing to do: Portobello Beach Coffee/food Place: Mary's Milk Bar