Week 3: Describing Continuous Data

Data Analysis for Psychology in R 1

Patrick Sturt

Department of Psychology
The University of Edinburgh

Weeks Learning Objectives

1. Understand the appropriate visualization for the distribution of numeric data.
2. Understand methods to calculate the spread for the distribution of numeric data.
3. Understand methods to calculate central tendency for the distribution of numeric data.

Topics for today

- Histograms
- Mean
- Variance and standard deviation

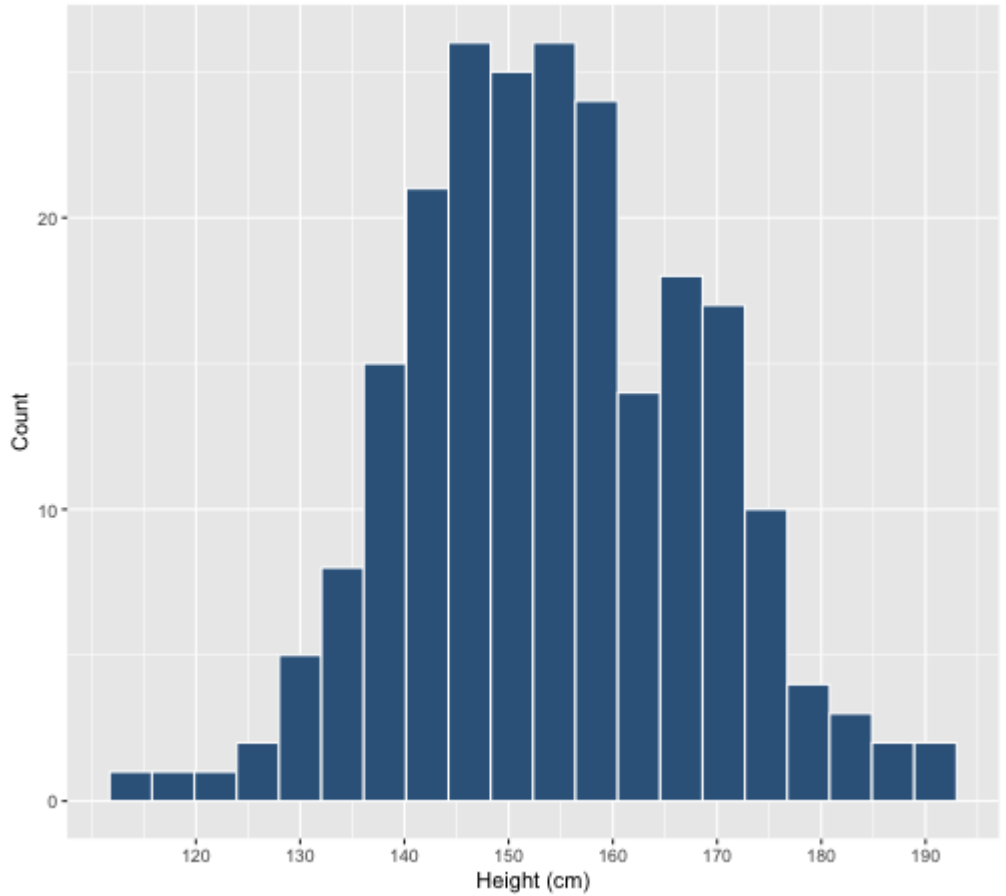
Recap: Continuous data

- Continuous (numeric) data is typically classed as interval or ratio
- That means:
    - The numeric values are meaningful as numbers.
    - We are able to apply mathematical operations to the values.

Visualization

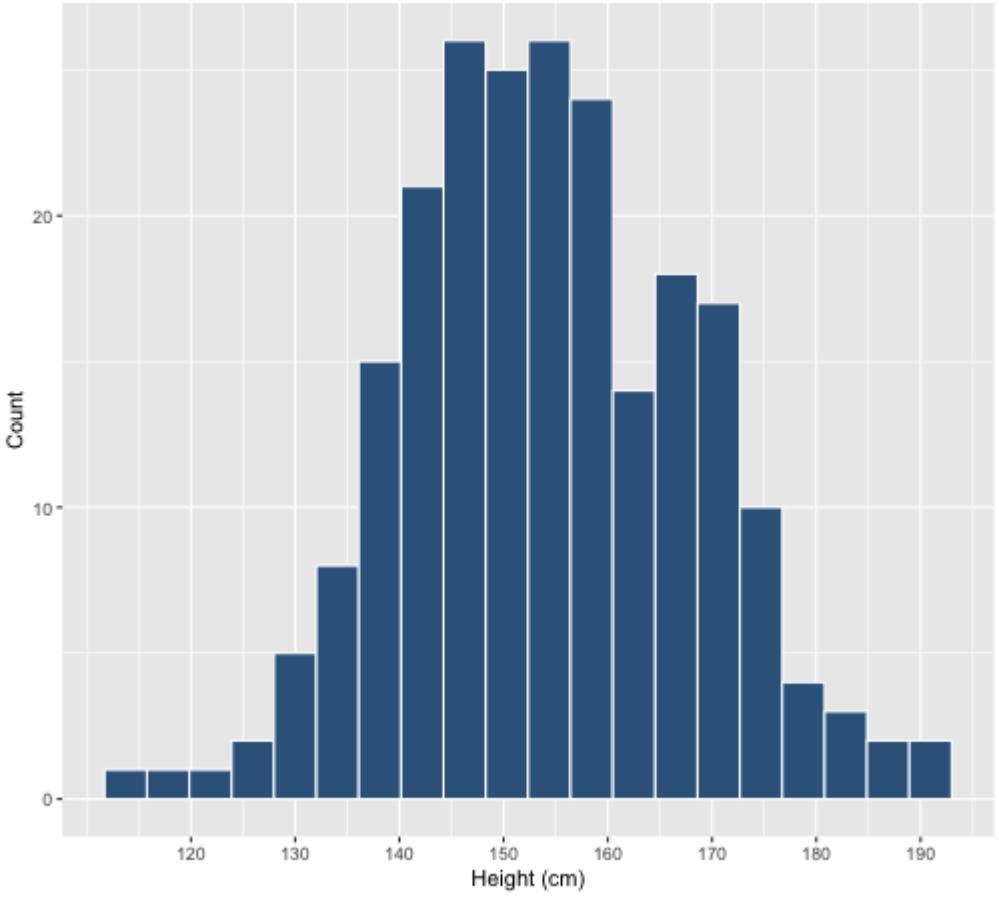
- Last lecture we discussed bar plots for frequency distributions of categorical variables.
- For continuous data, we visualize the distribution using a histogram.

Example histogram on the height of a class



Histogram

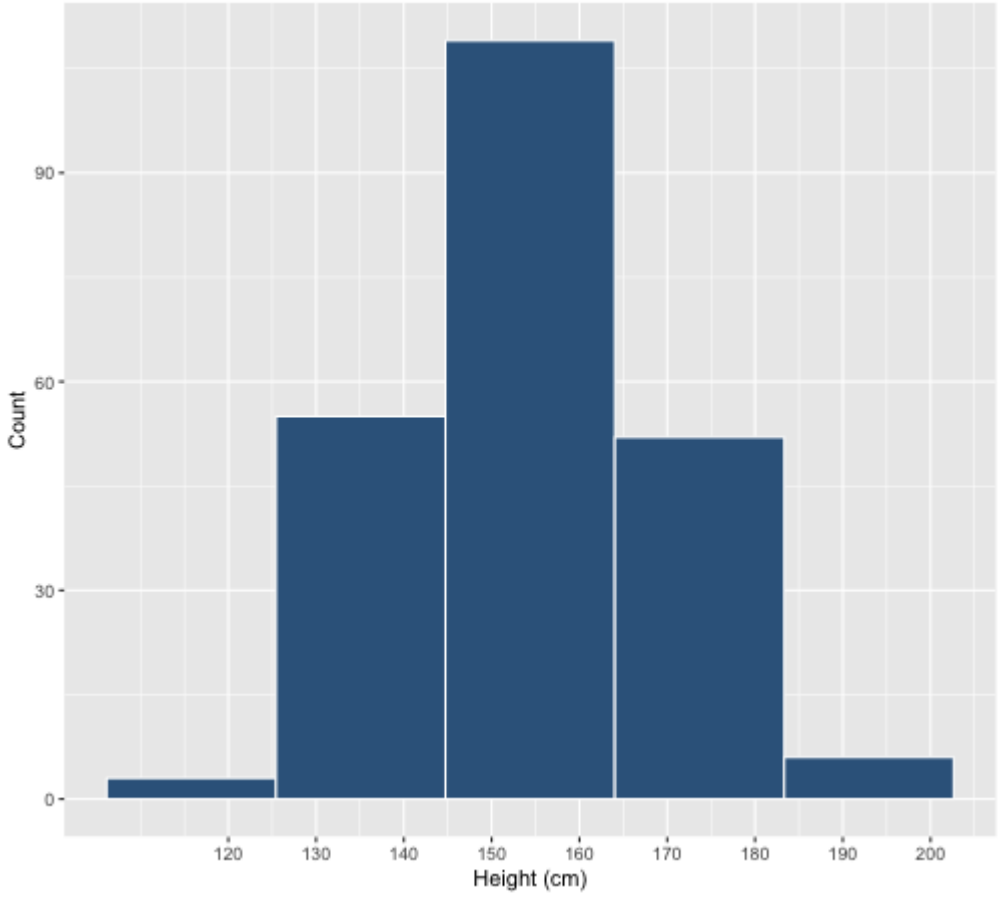
- Properties of a historgram:
  - X-axis: possible values of some variable.
  - Commonly presented in "bins"
  - A bin represents a range of scores (plot can look very different dependent on the bins)
    - Scale = dependent on the form of measurement, here centimetres
  - Y-axis: frequency of a given value or values within "bins"
- Here our data is heights of the class
  - X-Axis values are the possible heights in bins of 4cm.
  - Y=Axis values are the counts of number of students in each bin.

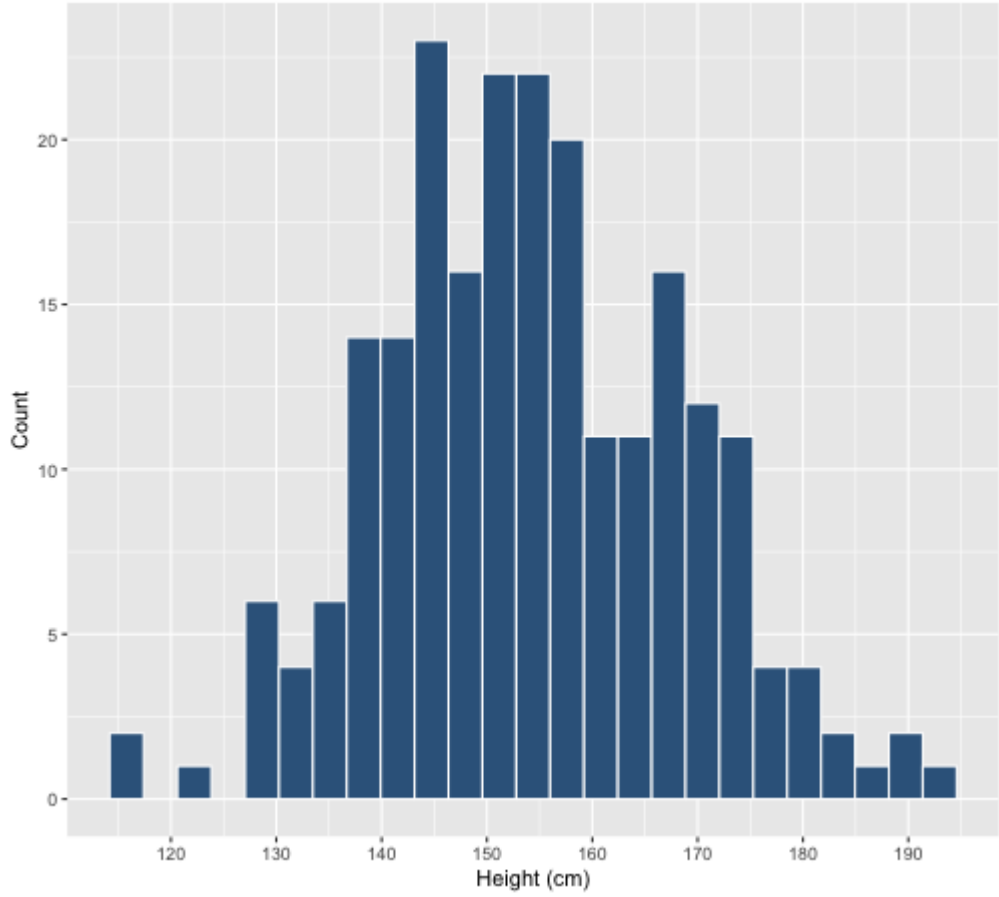


Pause for thought?

Why have we used bins for ranges of values and not individual values?

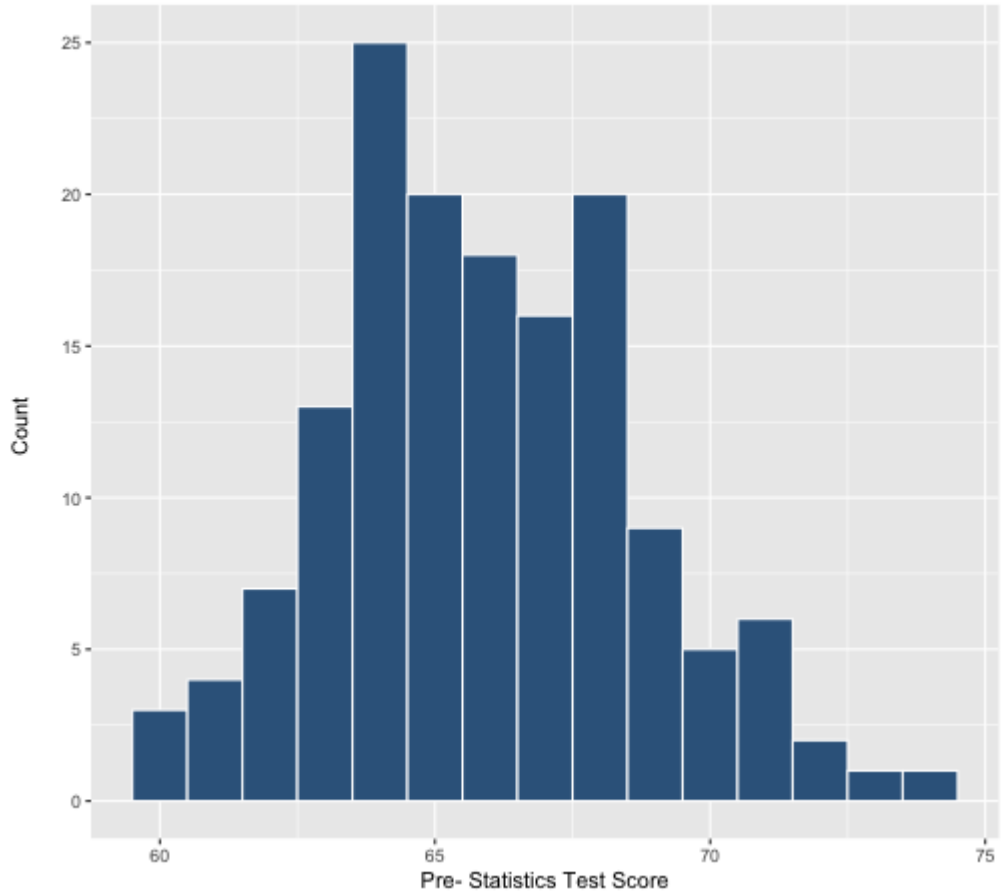
Impact of bins





Stats summer school example: test score

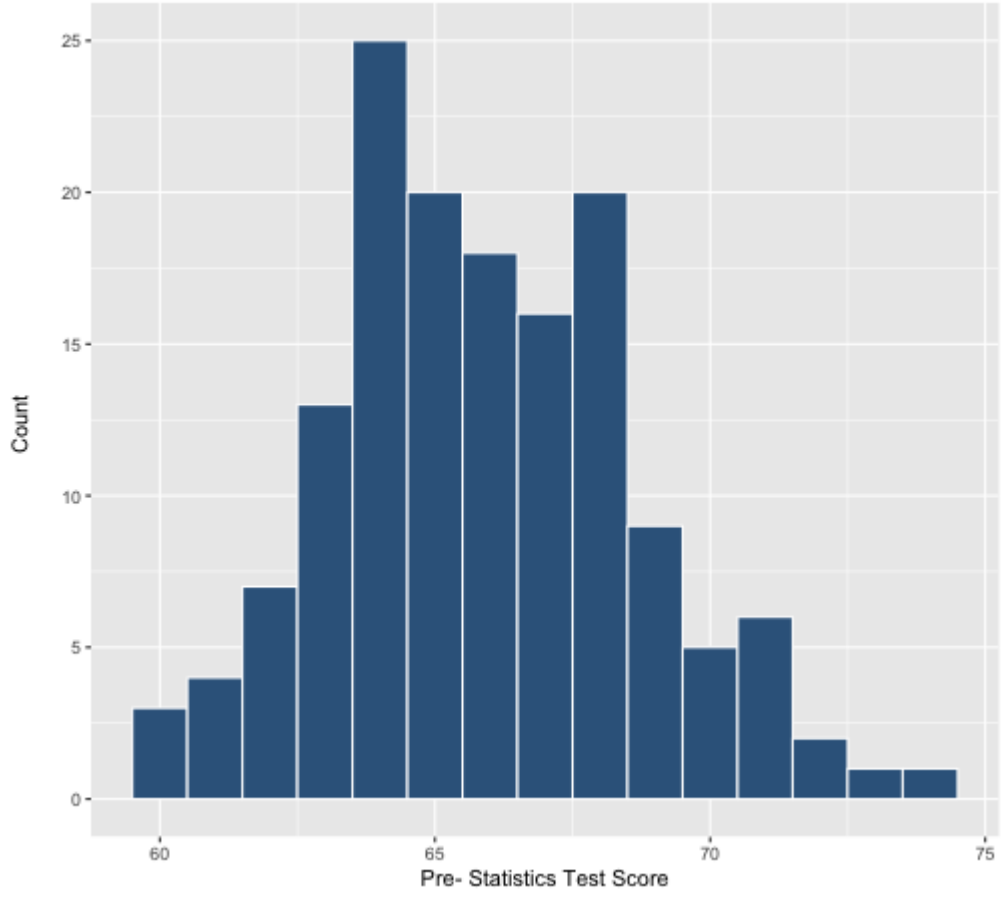
ID	Degree	Year	Score1	Score2
ID101	Psych	2	71	74
ID102	Ling	2	65	72
ID103	Ling	2	64	72
ID104	Phil	1	69	74
ID105	Ling	3	62	69
ID106	Ling	1	68	72
ID107	Phil	3	66	75
ID108	Psych	1	64	71
ID109	Psych	3	65	73
ID110	Ling	1	64	72



Stats summer school example: test score

```
ex1 %>%
  ggplot(., aes(x=Score1)) +
  geom_histogram(bins = 15,
                 color = "white",
                 fill = "steelblue4")+
  xlab("Pre- Statistics Test Score") +
  ylab("Count \n")
```

- New bits of code:
  - `geom_histogram` is used to make histograms
  - `bins` is the number of columns we want
  - `color` provides the colour for the outline of the column
  - `fill` provides the main colour



Central Tendency: Mean

- Last lecture we looked at the mode and median.
- Both can be used for continuous data, but the optimal measure is the **arithmetic mean.**
- **Mean:** is the sum of all values, divided by the total number of observations.
    - I.e. this is the average as most people think about the average.

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

- $\bar{x}$ = estimate of mean of variable x
- x_i = individual values of x
- n = sample size

Hand calculation

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

- Our data:

$$x = [10, 40, 30, 25, 15, 6]$$

- Worked calculation

$$\frac{\sum_{i=1}^{n} (10 + 40 + 30 + 25 + 15 + 6)}{6} = \frac{126}{6} = 21$$

Arithmetic Mean: Test score

Following hand-calculation in R

```
sum(ex1$Score1)/length(ex1$Score1)
```

```
## [1] 65.9
```

Short way in R

```
mean(ex1$Score1)
```

```
## [1] 65.9
```

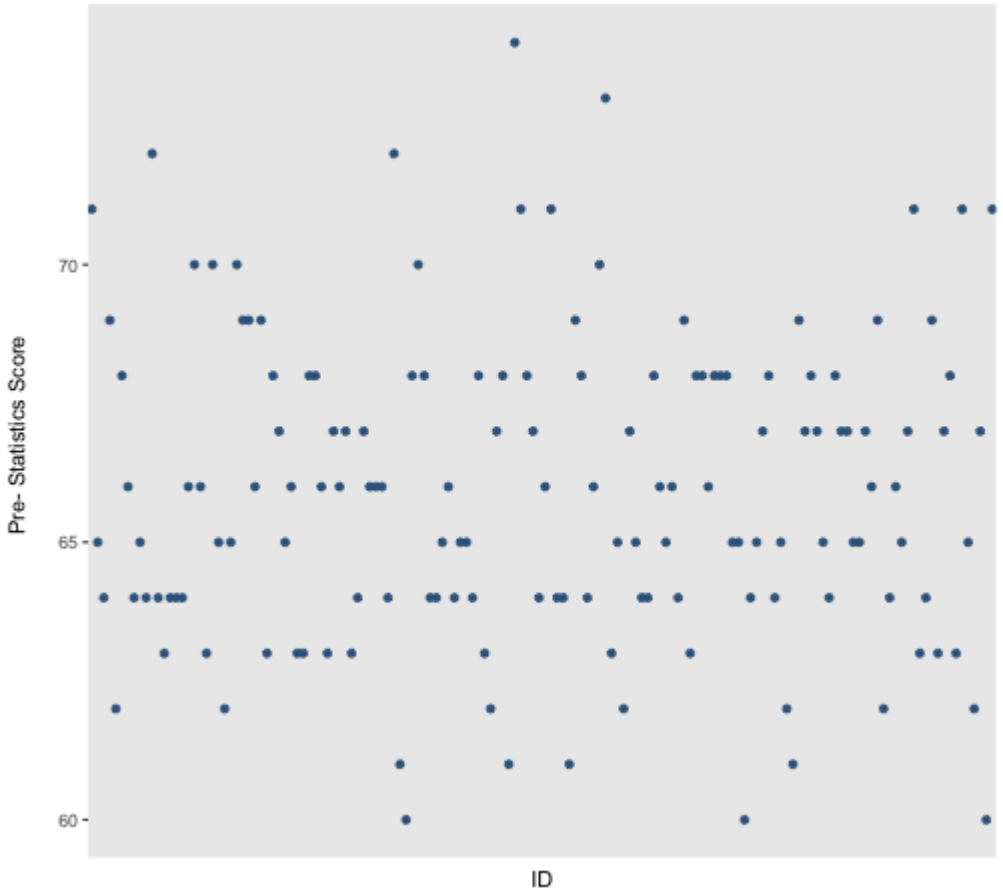
Working with `tidyverse`

```
ex1 %>%
    summarise(
        mean = mean(Score1)
    )
```

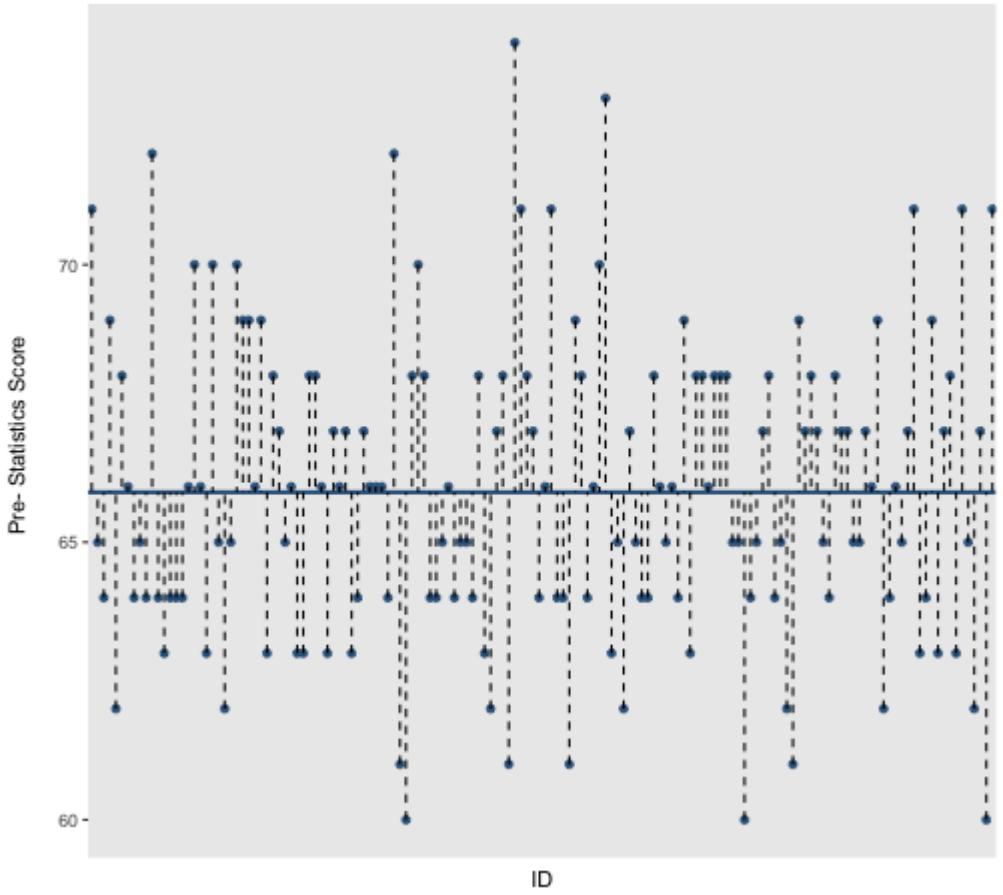
```
## # A tibble: 1 × 1
##    mean
##   <dbl>
## 1  65.9
```

- We will work with `tidyverse` and summarise as we can build up summary tables for our data sets.

Variation around the mean



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.**

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

- x_i = individual observations
- $\bar{x}$ = mean of x

Calculation: First 10 rows

ID	Score1	Score	Mean	Deviance
ID101	71	71	65.8	5.2
ID102	65	65	65.8	-0.8
ID103	64	64	65.8	-1.8
ID104	69	69	65.8	3.2
ID105	62	62	65.8	-3.8
ID106	68	68	65.8	2.2
ID107	66	66	65.8	0.2
ID108	64	64	65.8	-1.8
ID109	65	65	65.8	-0.8
ID110	64	64	65.8	-1.8

Problem: Sum of deviations

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(Score1 - mean(Score1)),2)
  )
```

Variable	Sum Deviation
Statistics Test Score	0

- Uh oh! The positive and negative values cancel.
- That means the sum of deviations from the mean will always be 0.

Variance

- In order to remove the effect of sign, we can square each of the deviations.
- This is called the ***variance*** .

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

- Variance is the average squared deviation from the mean.
- s^2 = sample variance

Calculation

ID	Score1	Score	Mean	Deviance	Deviance_sq
ID101	71	71	65.8	5.2	27.04
ID102	65	65	65.8	-0.8	0.64
ID103	64	64	65.8	-1.8	3.24
ID104	69	69	65.8	3.2	10.24
ID105	62	62	65.8	-3.8	14.44
ID106	68	68	65.8	2.2	4.84
ID107	66	66	65.8	0.2	0.04
ID108	64	64	65.8	-1.8	3.24
ID109	65	65	65.8	-0.8	0.64
ID110	64	64	65.8	-1.8	3.24

Variance

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(Score1 - mean(Score1)),2),
    Variance = round((sum((Score1 - mean(Score1))^2))/(length(Score1)-1),2)
  )
```

Variable	Sum Deviation	Variance
Statistics Test Score	0	7.7

- 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.
    - Taking the square root undoes (or fixes) the squaring of deviations that we did to get variance.

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

Standard deviation

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    Variance = round(sum((Score1 - mean(Score1))^2)/(length(Score1)-1),2),
    SD = round(sqrt(sum((Score1 - mean(Score1))^2)/(length(Score1)-1)),2)
  )
```

```
## # A tibble: 1 × 3
##   Variable              Variance    SD
##   <chr>                    <dbl> <dbl>
## 1 Statistics Test Score      7.7  2.78
```

Standard deviation

- Easier R calculation

```
ex1 %>%
  summarise(
    Variable = "Statistics Test Score",
    "Sum Deviation" = round(sum(Score1 - mean(Score1)),2),
    Variance = round(var(Score1),2),
    SD = round(sd(Score1),2)
  )
```

```
## # A tibble: 1 × 4
##   Variable              `Sum Deviation` Variance    SD
##   <chr>                           <dbl>    <dbl> <dbl>
## 1 Statistics Test Score               0      7.7  2.78
```

Population vs. Sample statistics

Population Variance

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

Sample Variance

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

Population SD

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

Sample SD

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

- NOTE: R defaults to sample values.
- μ = population mean; $\bar{x}$ = sample mean
- σ = population standard deviation; s = sample standard deviation
- N = population size; n = sample size

Summary of last 2 lectures

Measure	Strength	Weakness
Mode	Actually occurs in our data	Not algebraically calculable
	Unaffected by extreme values	Probably does not exist for true continuous data (think reaction time)
Median	No assumptions about interval value of data	Not relatable to measures of dispersion (see next week)
	Unaffected by extreme values	
Mean	Algebraically tractable	Sensitive to extreme values
	Related to measures of dispersion (see next week)	Assumes data are interval or better
		Possible no case in your data takes the value of the mean

Which measure should we use?

Variable Type	Central Tendency
Categorical (Nominal)	Mode
Categorical (Ordered)	Mode/Median
Continuous	Mean (any in fact)
Count	Mode (mean)

- Depends on the level of measurement.

Which measure should we use?

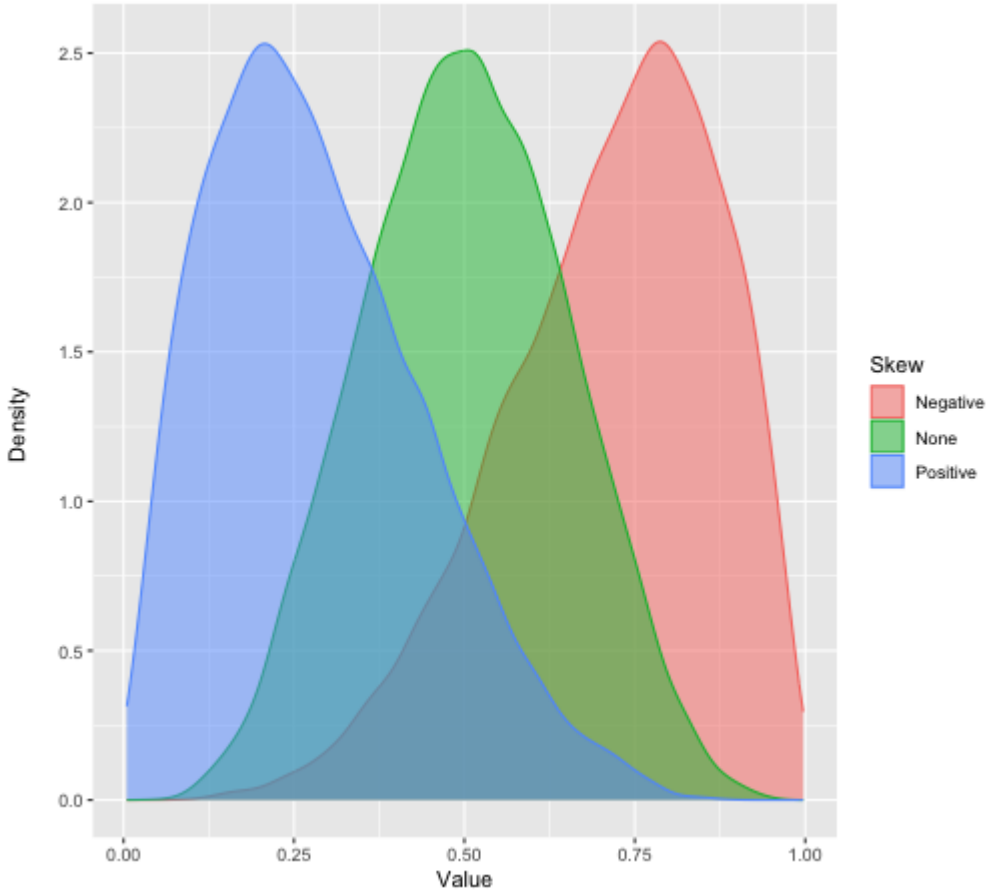
Variable Type	Central Tendency	Dispersion
Categorical (Nominal)	Mode	Frequency Table
Categorical (Ordered)	Mode/Median	Range
Continuous	Mean (any in fact)	Variance & Standard Deviation
Count	Mode (mean)	Range (Variance & SD)

- Depends on the level of measurement.

A few extra bits?

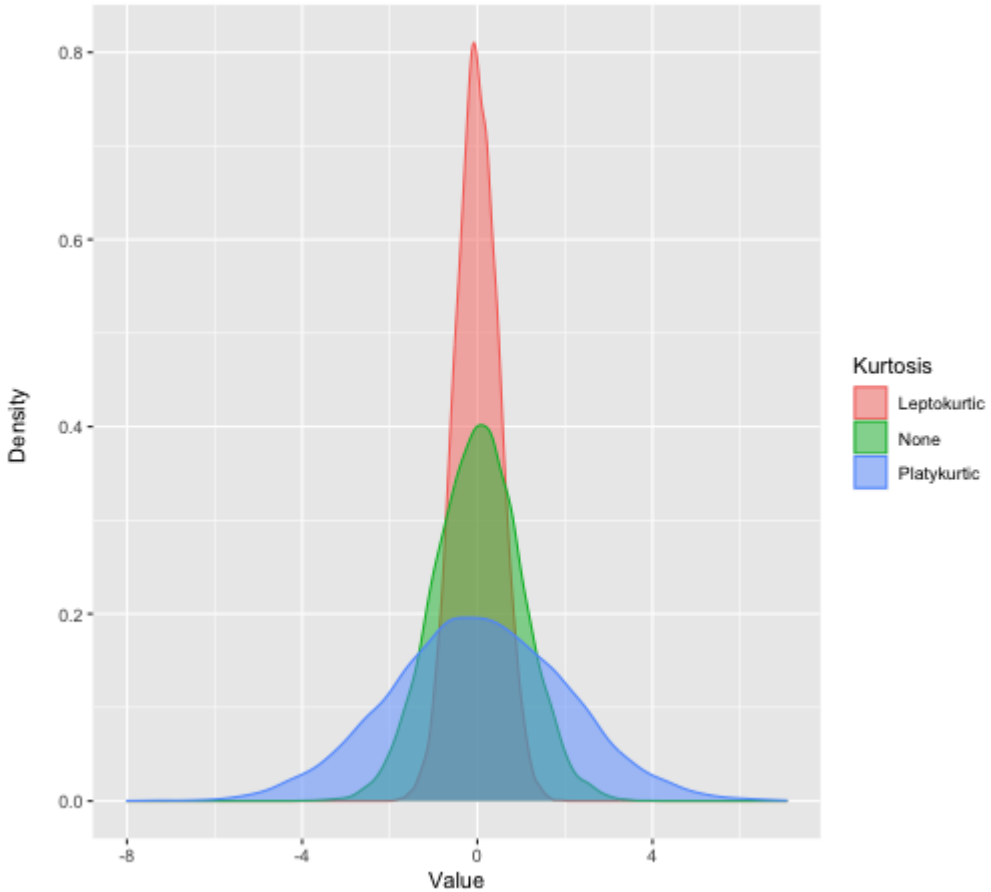
- You may come across the mathematical language of *moments.*
- Moments describe the shape of a set of points
    - Mean
    - Variance
    - Skew
    - Kurtosis

Skew



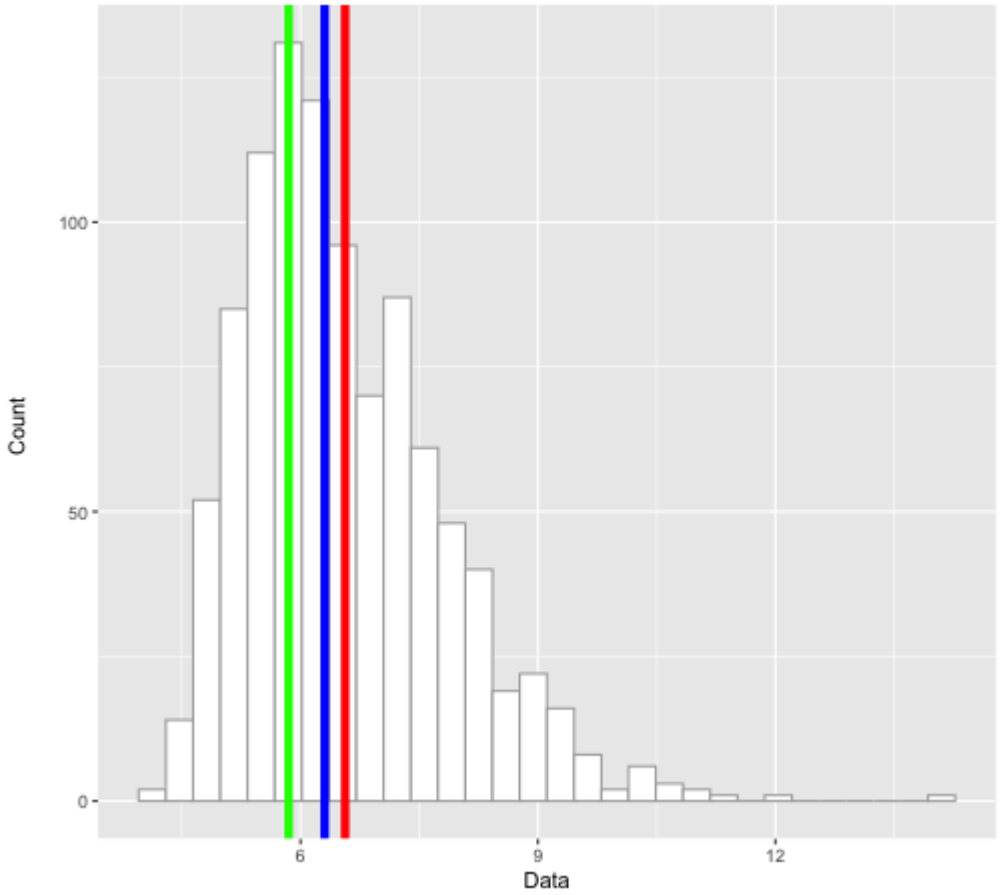
- Is a measure of asymmetry of a distribution.

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.

Summary of today

- Continuous variables are...
    - Visualized with histogram
    - summarised with mean and standard deviation
- We can describe the shape of the distribution with skew and kurtosis

Next tasks

- Next week, we will look at describing relationships.
- This week:
    - Complete your lab
    - Come to office hours
    - No quiz this week (week 3). Assessed quiz on categorical data next week (week 4)