THE UNIVERSITY of EDINBURGH

Week 2: Measurement and Distributions

Univariate Statistics and Methodology using R

USMR Team

Department of Psychology
The University of Edinburgh

Today's Key Topics

- Histograms & Density Plots
- The Normal Distribution
- Populations & Samples
- Central Limit Theorem

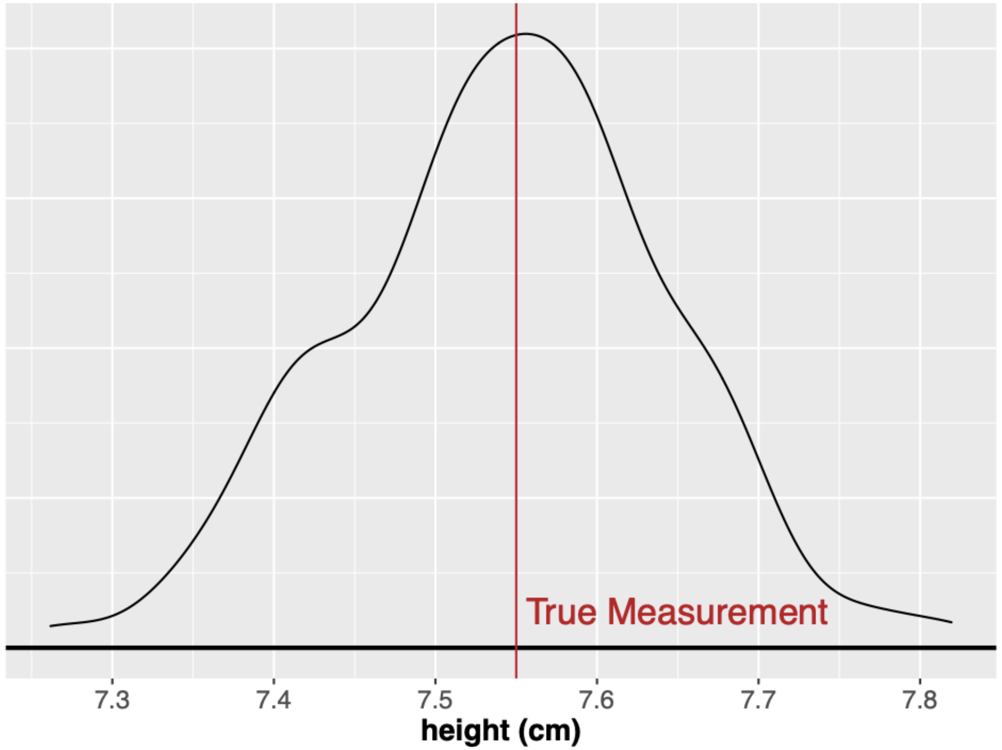
Part 1

Measurement

- When we measure something, we attempt to identify its **true measurement**, or the **ground truth**
- The problem is that we don't have any way of measuring accurately enough
  - Our measurements are likely to be close to the truth
  - They will likely vary if we take multiple measures
- Let's run a quick experiment:

Measurement

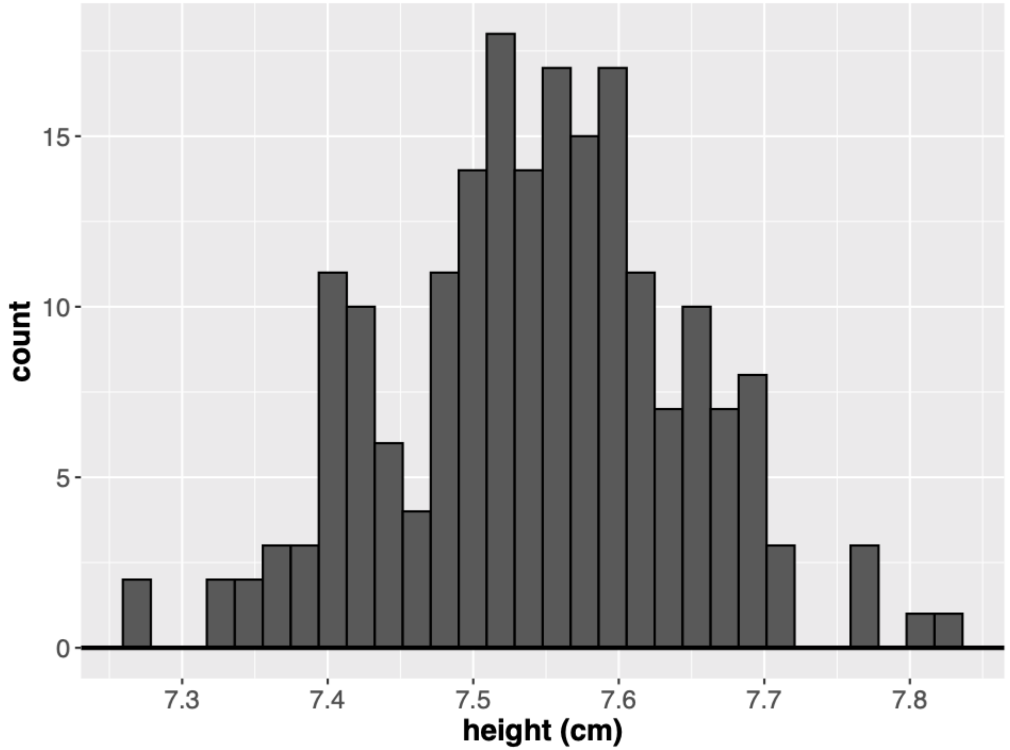
We might expect values close to the true measurement to be more frequent if we take multiple measurements:



(Though there are still limits to our precision)

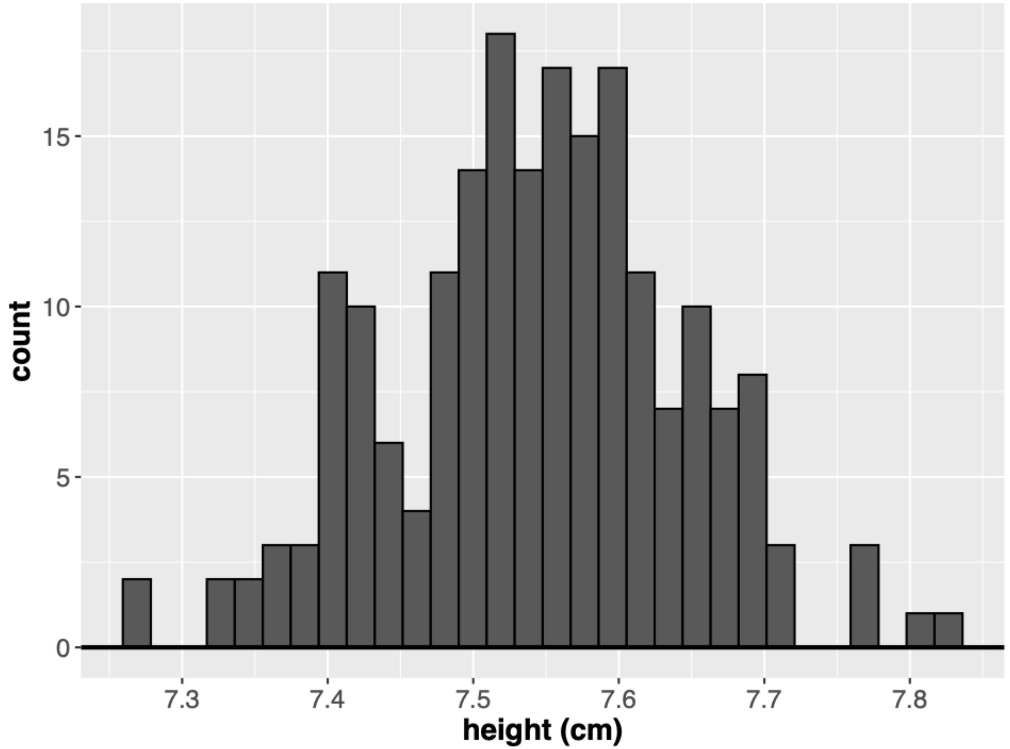
Histograms

Considering this principle, it might be useful to create a histogram of all measurements taken.



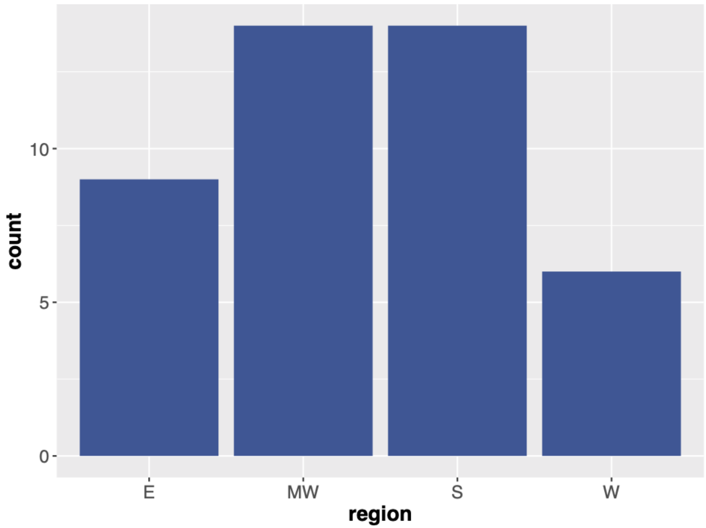
Histograms

Considering this principle, it might be useful to create a histogram of all measurements taken.



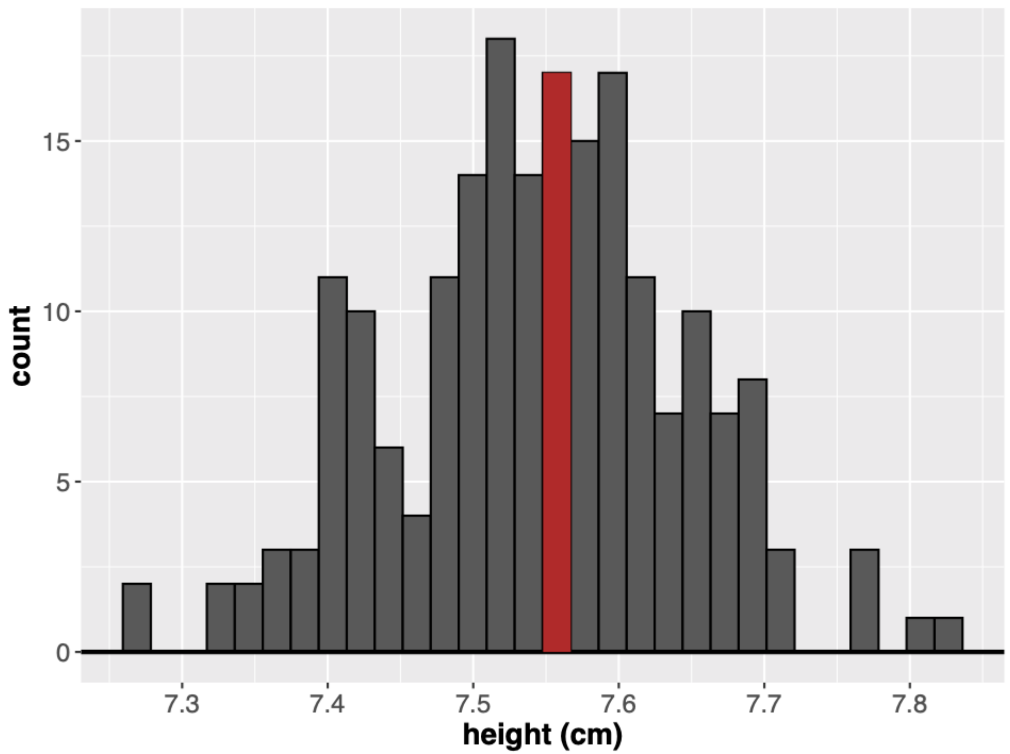
Note difference from a bar chart:

- Histogram represents **continuous** data
- Bar Chart represents **categorical** data



Histograms

Considering this principle, it might be useful to create a histogram of all measurements taken.



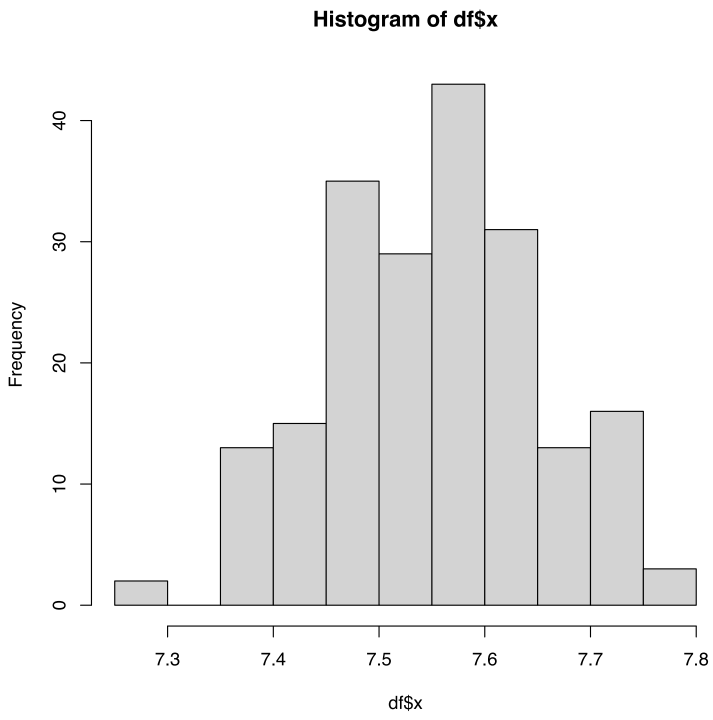
- We know that there are 17 measurements around 7.55
- Strictly, between 7.548 & 7.567

Histograms in R

```r
df <- data.frame(x=rnorm(200, mean = 7.55, sd = 0.1))
head(df)
```

```
##       x
## 1 7.609
## 2 7.614
## 3 7.471
## 4 7.578
## 5 7.491
## 6 7.524
```

```r
hist(df$x)
```

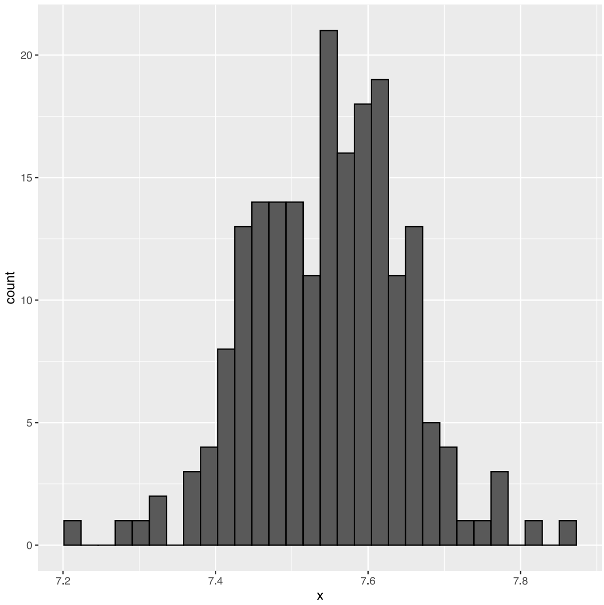


Histograms in R

```
df <- data.frame(x=rnorm(200, mean = 7.55, sd = 0.1))
head(df)
```

```
##        x
## 1 7.780
## 2 7.461
## 3 7.514
## 4 7.638
## 5 7.476
## 6 7.583
```

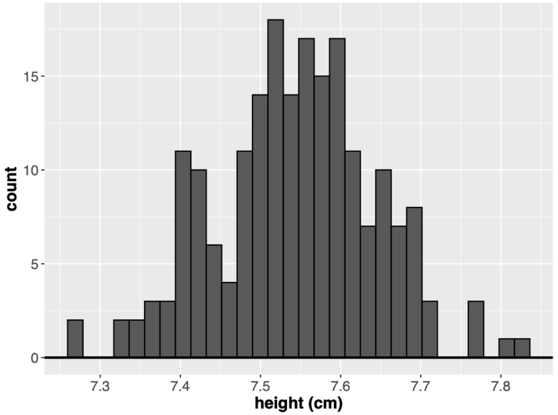
```
library(ggplot2)
ggplot(df, aes(x)) +
  geom_histogram(colour='black')
```



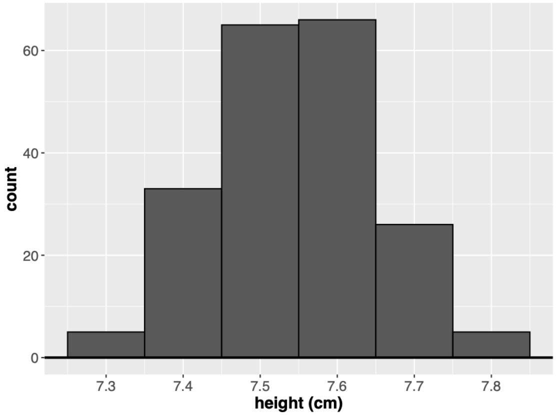

```
stat_bin() using bins = 30. Pick better value with
binwidth.
```

Histograms in R

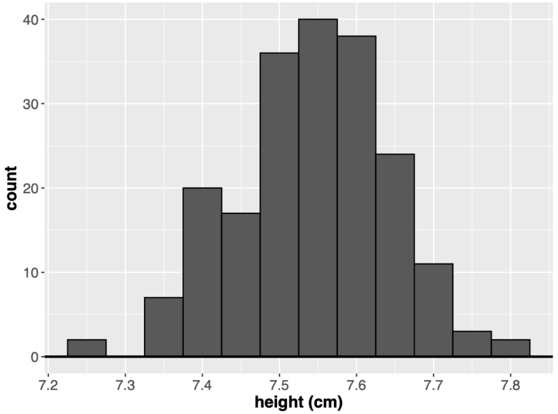
Note that the bin width of the histogram matters. Every figure below displays the same data.



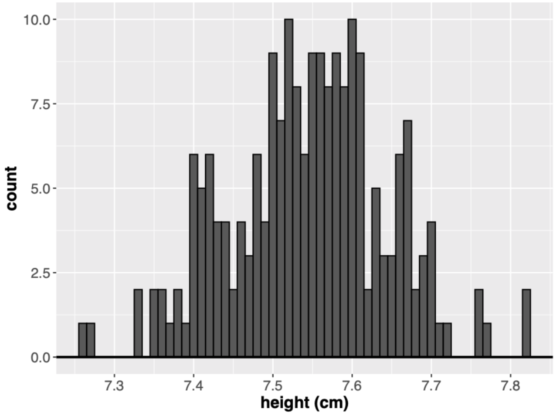
ggplot default



binwidth = .1



binwidth = .05



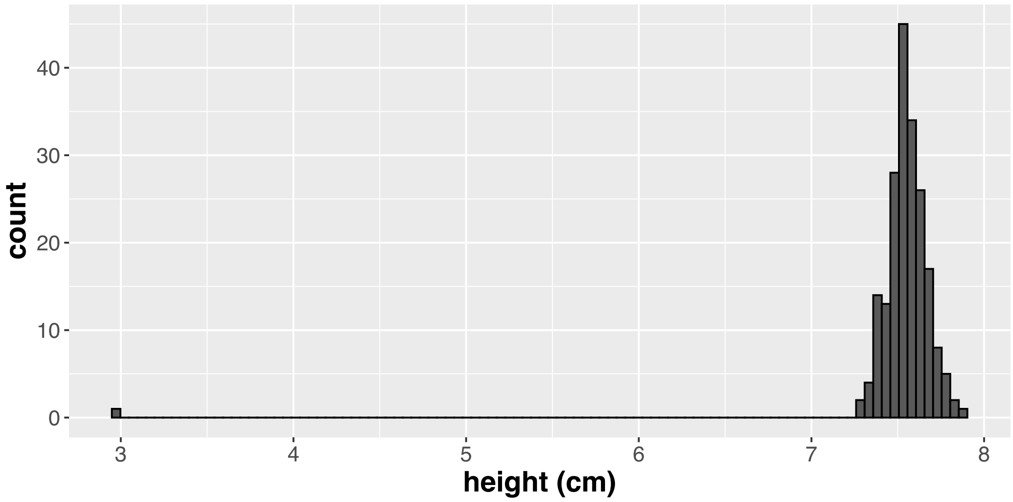
binwidth = .01

Histograms

- The Good
  - Way to examine the *distribution* of the data
  - Easy to interpret (*y* axis = counts)
  - Sometimes helpful in spotting weird data (**outliers**)

Histograms

- The Good
  - Way to examine the *distribution* of the data
  - Easy to interpret (y axis = counts)



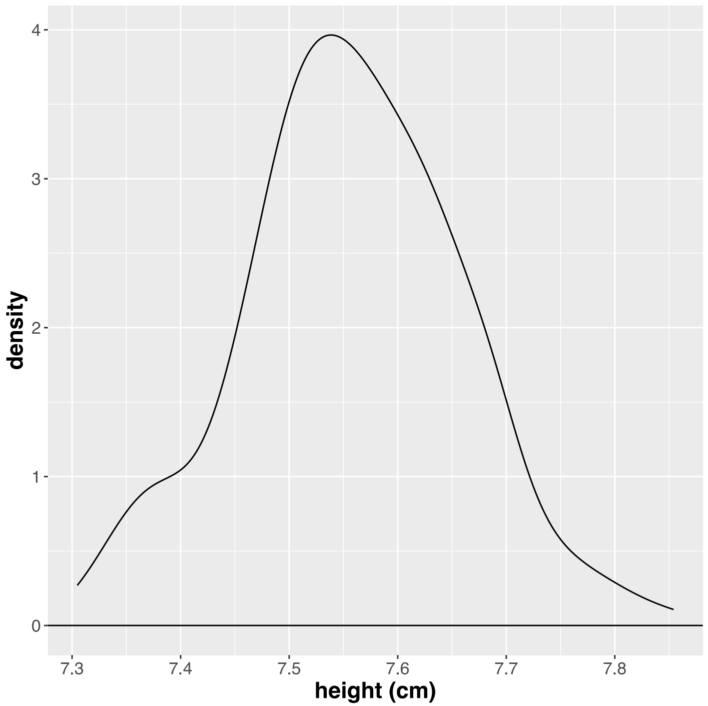
  - Sometimes helpful in spotting weird data (**outliers**)

Histograms

- The Good
    - Way to examine the *distribution* of the data
    - Easy to interpret (y axis = counts)
    - Sometimes helpful in spotting weird data (outliers)
- The Bad
    - Only gives us information about distribution and mode; doesn't give us other information
        - E.g., the mean or median
    - Changing bin width can completely change the graph

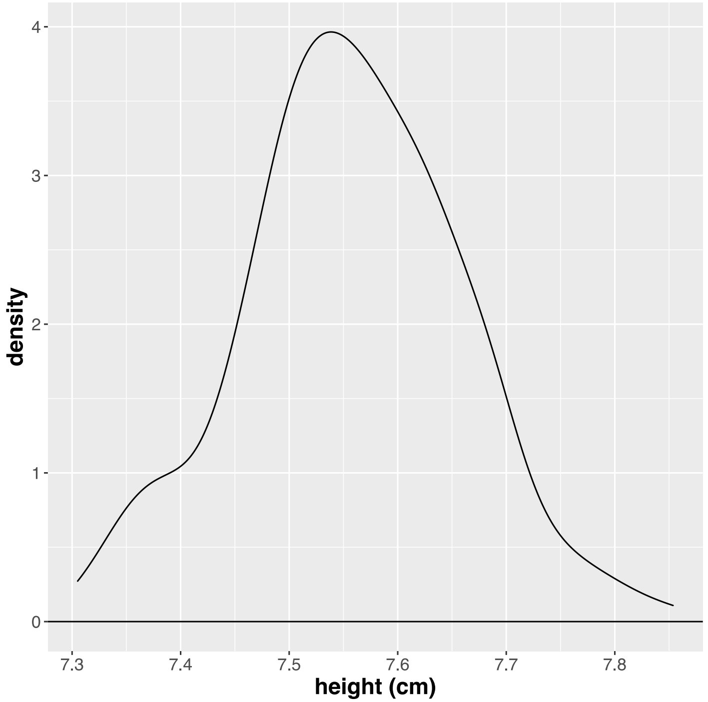
Density Plots

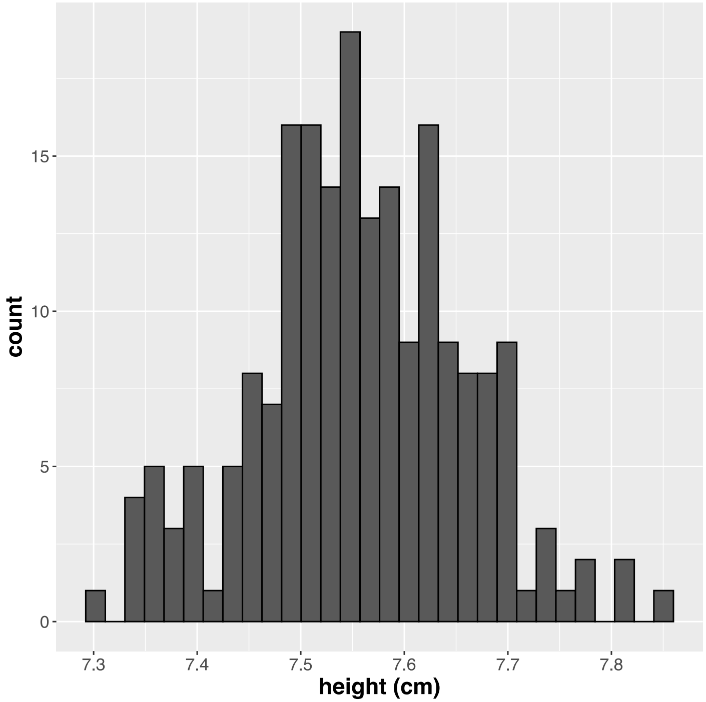
- Similar to histogram in that it shows the distribution of the data
- However, the *y* axis is no longer a count, but represents a **proportion** of cases.
- The area under the curve is equal to 1 (or 100%, reflecting all cases)



Density Plots

- Similar to histogram in that it shows the distribution of the data
- However, the y axis is no longer a count, but represents a **proportion** of cases.
- The area under the curve is equal to 1 (or 100%, reflecting all cases)

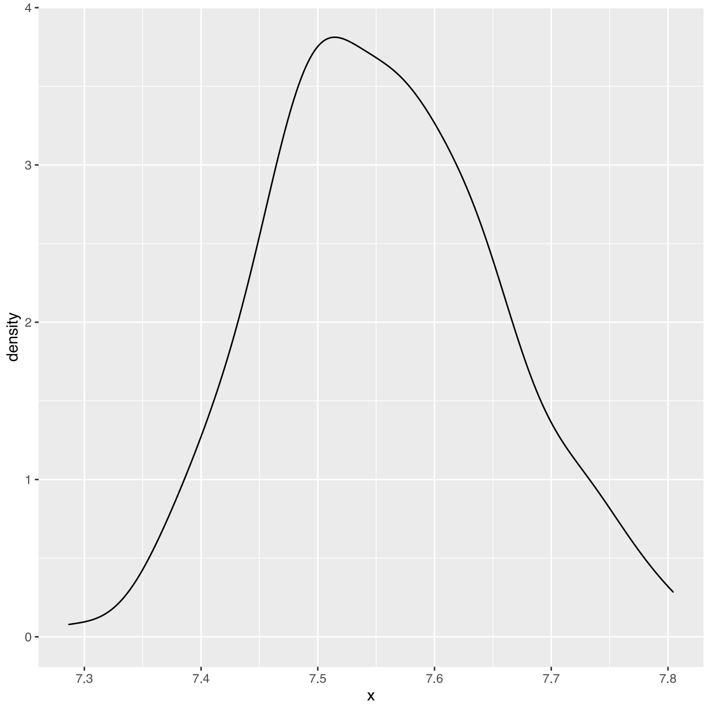




Density Plots in R

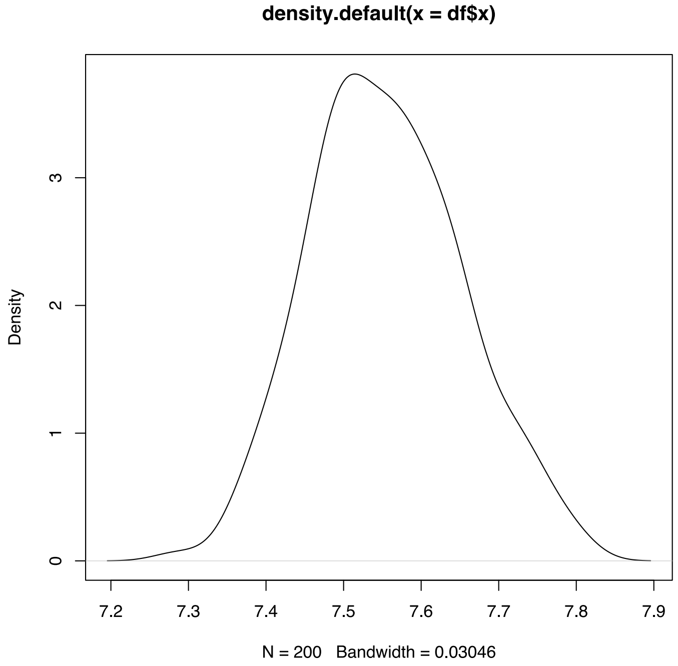
ggplot

```
ggplot(df, aes(x)) + geom_density()
```



base R

```
densDat <- density(df$x)
plot(densDat)
```

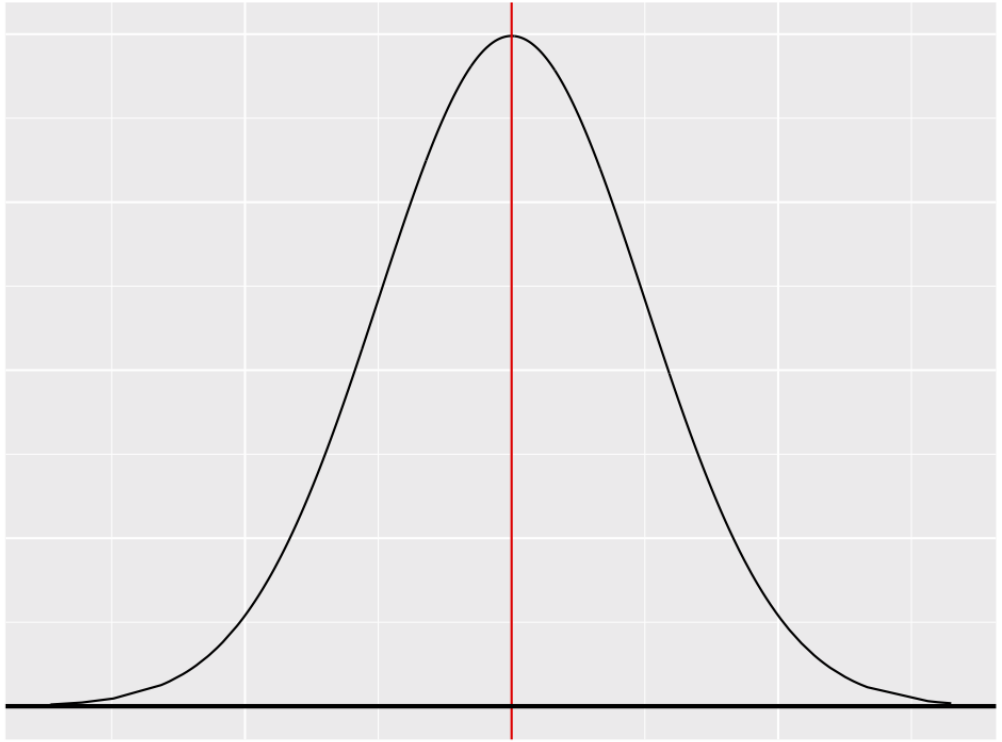


Part 2

The Normal Distribution

The Normal Distribution

- A hypothetical density plot
  - Probability distribution of a random variable
- Normal curves are unimodal, with values symmetrically distributed around the peak
  - Centered around the mean
  - A higher proportion of cases near the mean and a lower proportion of cases with more extreme values



The Normal Distribution

Normal curves can be defined in terms of *two parameters*:

- The **mean** of the distribution ($\bar{x}$, or sometimes μ)

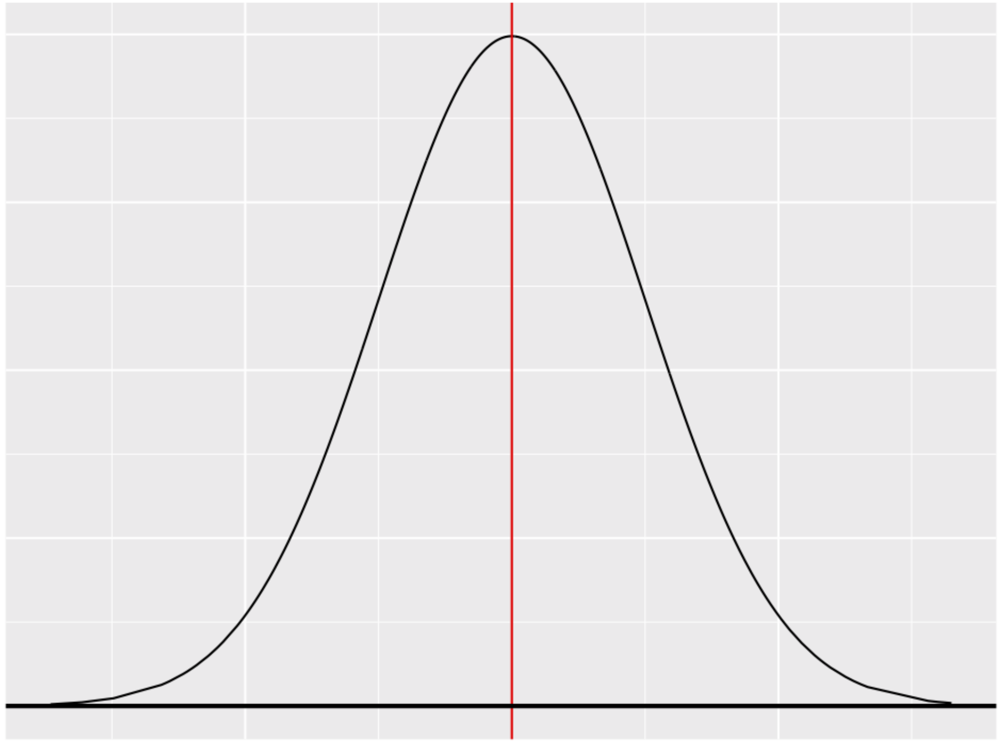
```
x <- c(22, 24, 21, 19, 22, 20)
mean(x)
```

```
## [1] 21.33
```

- The **standard deviation** of the distribution (sd, or sometimes σ)

```
sd(x)
```

```
## [1] 1.751
```



A quick note on standard deviation

The **standard deviation** is the average distance of observations from the mean

$$\text{sd} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

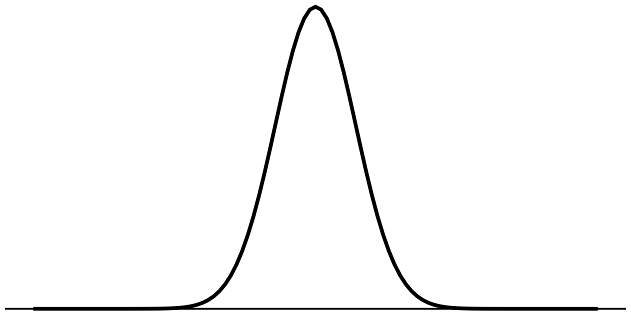
x = individual observation

$\bar{x}$ = mean

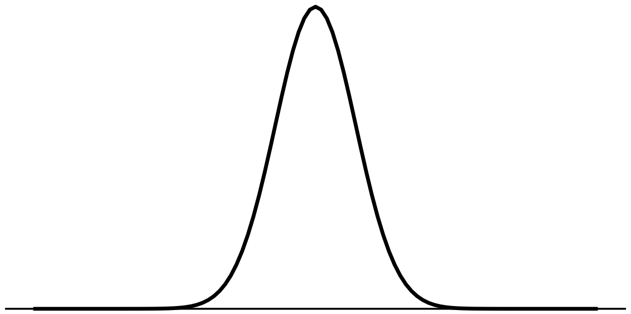
n = sample size

$\sum$ = add it up

How do these features affect the normal curve?

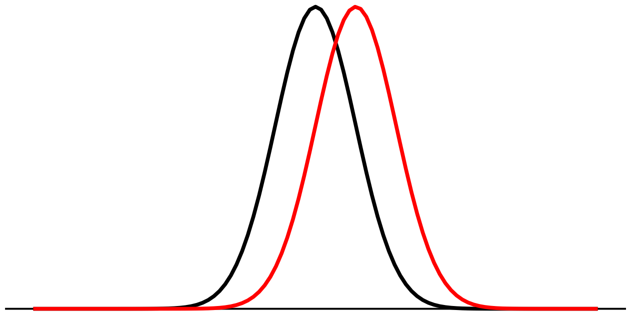


How do these features affect the normal curve?



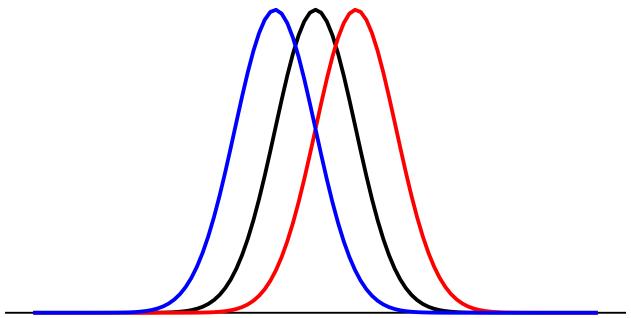
The *mean* determines where the curve is centered.

How do these features affect the normal curve?



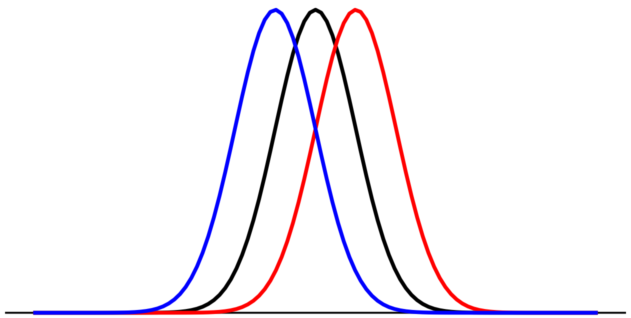
The *mean* determines where the curve is centered.

How do these features affect the normal curve?



The *mean* determines where the curve is centered.

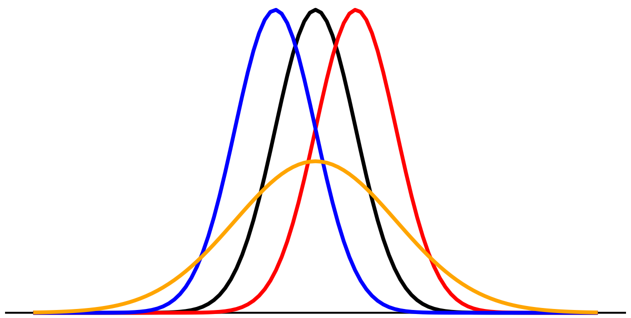
How do these features affect the normal curve?



The *mean* determines where the curve is centered.

The *standard deviation* determines the shape of the curve

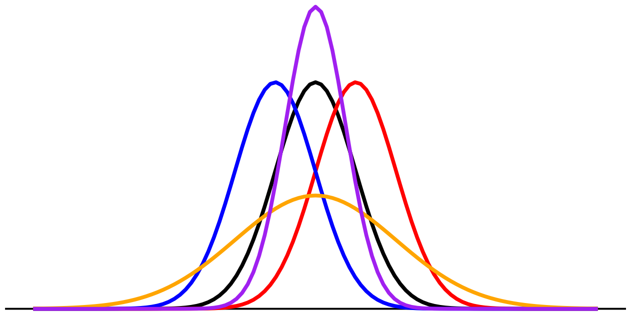
How do these features affect the normal curve?



The *mean* determines where the curve is centered.

The *standard deviation* determines the shape of the curve

How do these features affect the normal curve?



The *mean* determines where the curve is centered.

The *standard deviation* determines the shape of the curve

Part 3

Sampling from a Population

Samples vs Populations

- Population – all members of the group that you are hypothesizing about
- Sample – the subset of the population that you're testing to find the answer
- If we repeatedly sample from a population and measure the mean of each sample, we'll get a normal distribution
  - the mean will be (close to) the population mean
  - the standard deviation ("width") of the distribution of sample means is referred to as the standard error of the distribution

Statistical Estimates

- so far, we've talked about sampling repeatedly from a population
- this might not be possible
- if we only have one sample we can make *estimates* of the mean and standard error
    - the estimated *mean* is the sample mean (we have no other info)
    - the estimated *standard error* of the mean is defined in terms of the sample standard deviation

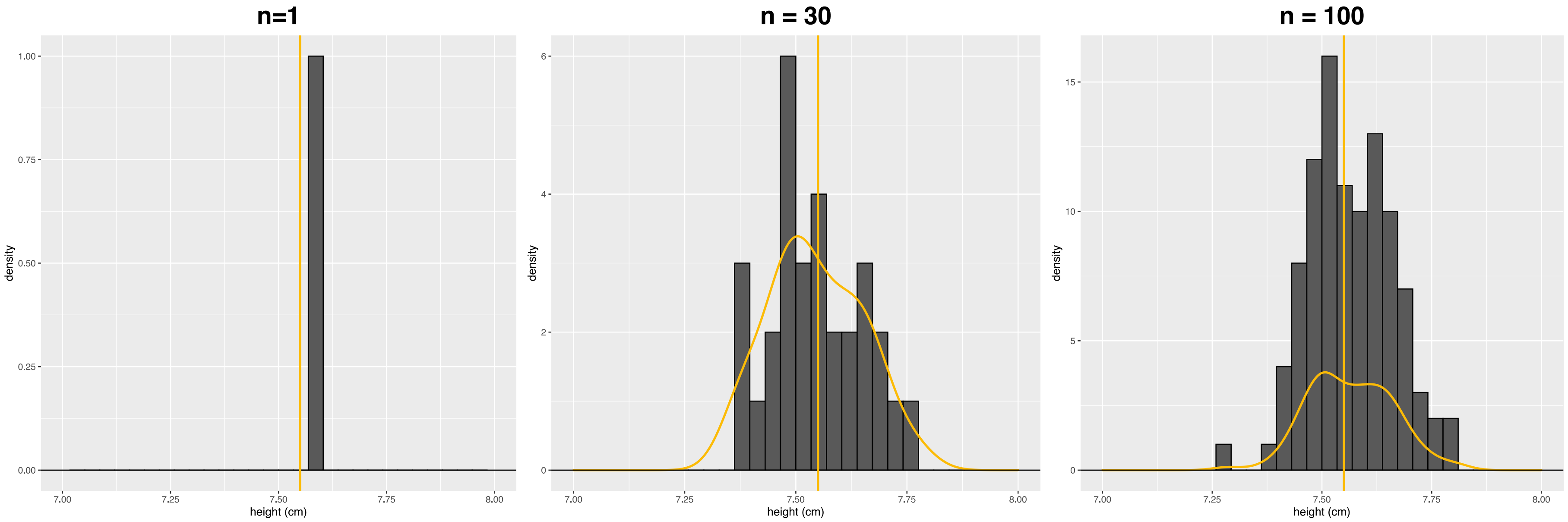
$$\mathrm{se} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}}{\sqrt{n}}$$

Part 4

Towards Statistical Testing

Central Limit Theorem

- What we have just seen is a demonstration of **Central Limit Theorem**
- Lay version: *sample means will be normally distributed about the true mean*
- The more samples you take, the more normal the distribution should look, regardless of the variable's distribution in the population



The Standard Normal Curve

We can *standardize* any value on any normal curve by:

- subtracting the mean
    - the effective mean is now *zero*
- dividing by the standard deviation
    - the effective standard deviation is now *one*

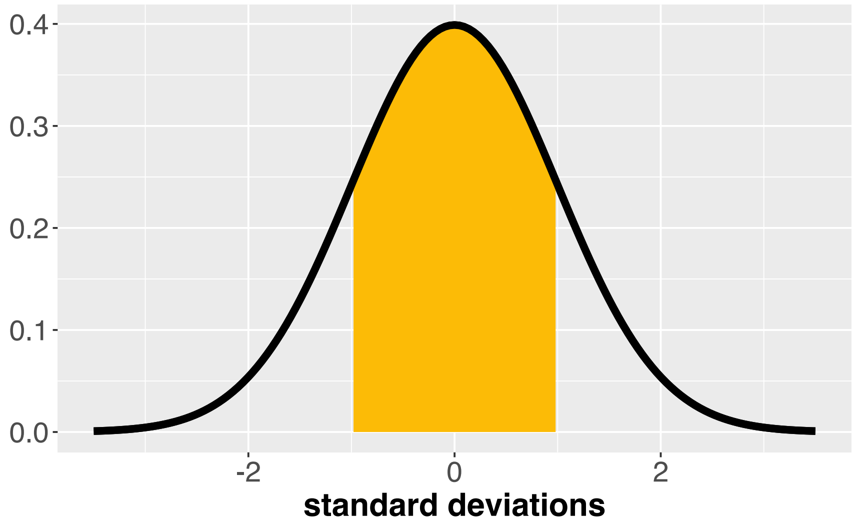
These new standardized values are called **z-scores**.

The standardized normal distribution is also known as **the z-distribution**.

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

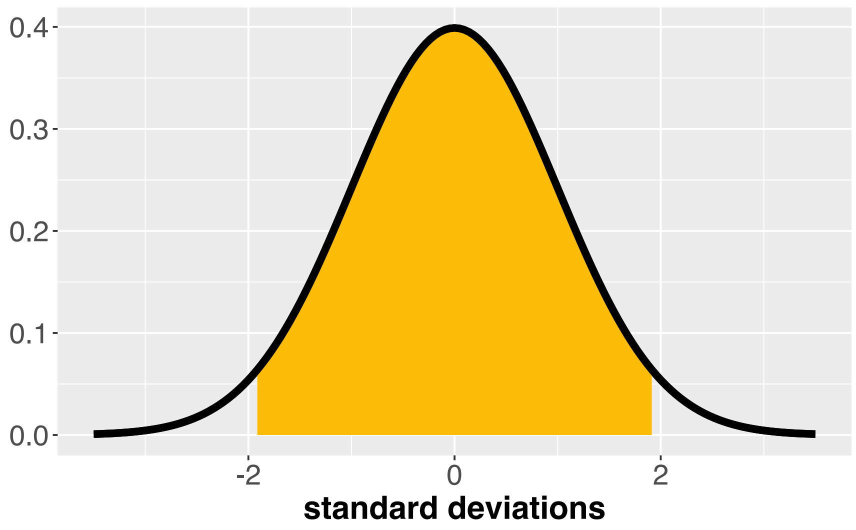
The Standard Normal Curve

- ~68% of observations fall **within one** standard deviation of the mean.
- ~32% of observations fall **greater than one** standard deviation above or below the mean



The Standard Normal Curve

- ~95% of observations fall **within 1.96** standard deviations of the mean
- ~5% of observations fall **greater than 1.96** standard deviations above or below the mean
- We can phrase it another way: *an area of .95* lies between -1.96 and 1.96 standard deviations from the mean
  - "95% of predicted observations" (the 95% confidence interval)



Can We Use This For Real?

- we have some survey data from the USMR class last year, including *height* in cm
- perhaps we're interested in the "average height of a young statistician" (!)

Can We Use This For Real?

- we have some survey data from the USMR class last year, including *height* in cm
- perhaps we're interested in the "average height of a young statistician" (!)
  - "young statisticians" are a **population**

Can We Use This For Real?

- we have some survey data from the USMR class last year, including *height* in cm
- perhaps we're interested in the "average height of a young statistician" (!)
    - "young statisticians" are a **population**
    - the USMR class of 2021 is a **sample**

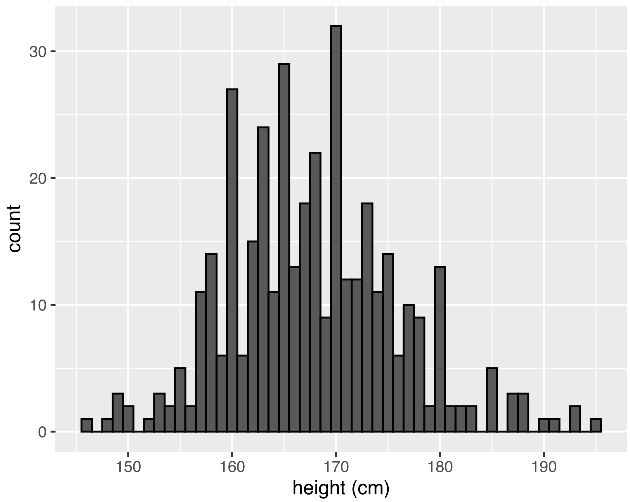
Can We Use This For Real?

- we have some survey data from the USMR class last year, including *height* in cm
- perhaps we're interested in the "average height of a young statistician" (!)
    - "young statisticians" are a **population**
    - the USMR class of 2021 is a **sample**

Can we use the information from the sample of 386 responses we have to say anything about the population?

Looking at the class data

```r
ggplot(hData, aes(height)) +
  geom_histogram(colour = 'black', binwidth = 1) +
  labs(x = 'height (cm)')
```




```r
head(hData$height)
```

```
## [1] 171 149 173 159 157 177
```

```r
mean(hData$height)
```

```
## [1] 167.9
```

```r
sd(hData$height)
```

```
## [1] 8.23
```

Statistically Useful Information

Remember, in normally distributed data, 95% of the data fall between $\bar{x} - 1.96\sigma$ and $\bar{x} + 1.96\sigma$

Statistically Useful Information

Remember, in normally distributed data, 95% of the data fall between $\bar{x} - 1.96\sigma$ and $\bar{x} + 1.96\sigma$

```
mean(hData$height) + 1.96*sd(hData$height)
```

```
## [1] 184
```

```
mean(hData$height) - 1.96*sd(hData$height)
```

```
## [1] 151.7
```

Statistically Useful Information

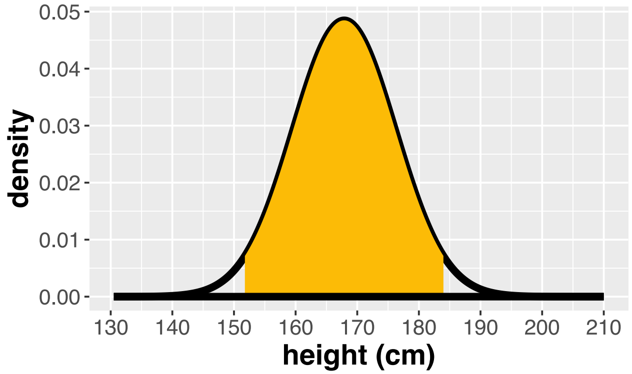
Remember, in normally distributed data, 95% of the data fall between $\bar{x} - 1.96\sigma$ and $\bar{x} + 1.96\sigma$

```
mean(hData$height) + 1.96*sd(hData$height)
```

```
## [1] 184
```

```
mean(hData$height) - 1.96*sd(hData$height)
```

```
## [1] 151.7
```



If we measure the mean height of 386 people from the same population as the USMR class, we estimate that the answer we obtain will lie between **151.7cm** and **184cm** 95% of the time

The Aim of the Game

- As statisticians, a major goal is to infer from **samples** to **populations**
- More about how we do this next time

Today's Key Points

- The distribution of data can be visualized with histograms and density plots
- Normally distributed data are symmetrically distributed, with scores near the mean being measured more often than scores further away
- Populations include every member of a group of interest, while a sample includes only those members being observed or tested
- The Central Limit Theorem states that as sample size increases, a variable's distribution begins to approximate the normal distribution.