SOC6302 Statistics for Sociologists

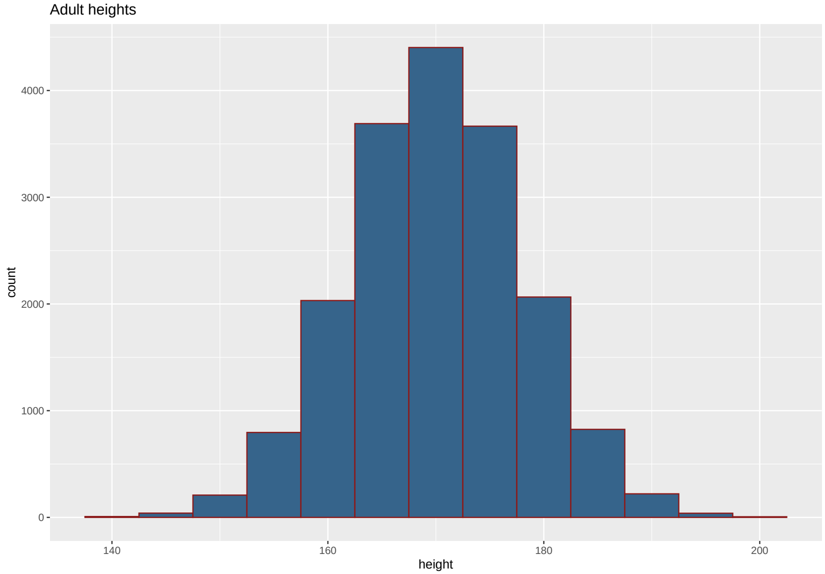
Monica Alexander

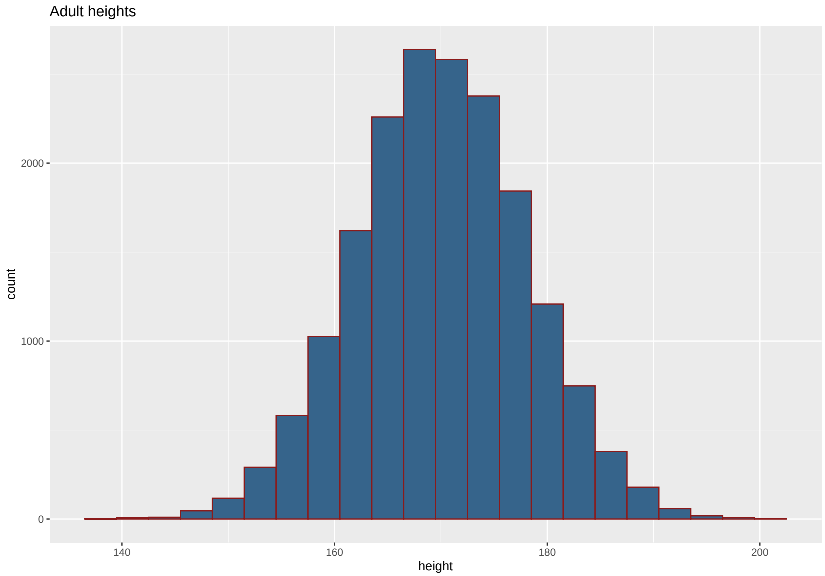
Week 5: Probability and sampling distributions

Where are we at

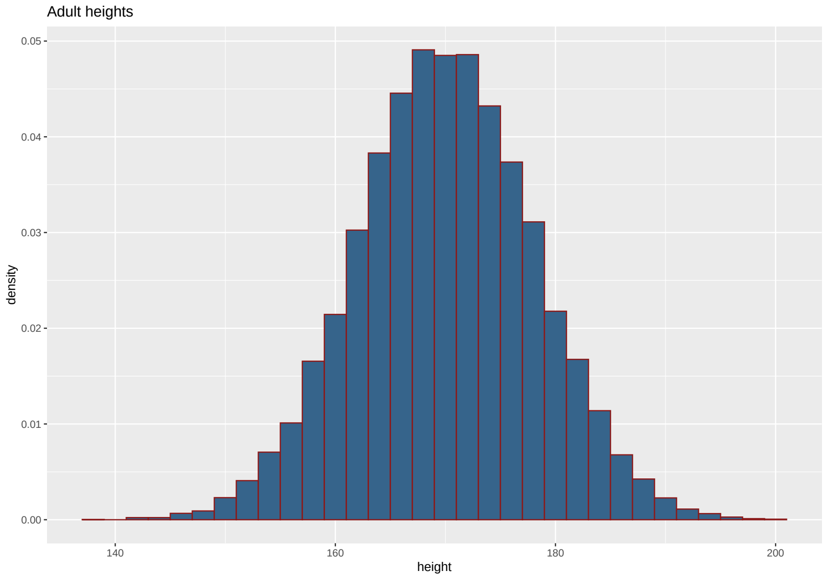
- Probability concepts
  - Additive rule, mutually exclusive events, multiplicative rule, independence, complements
- Probability distributions
  - Discrete RV = probability mass function
  - Continuous RV = probability density function
- Probabilities as areas

A continuous probability distribution is just a histogram with infinitely small bins

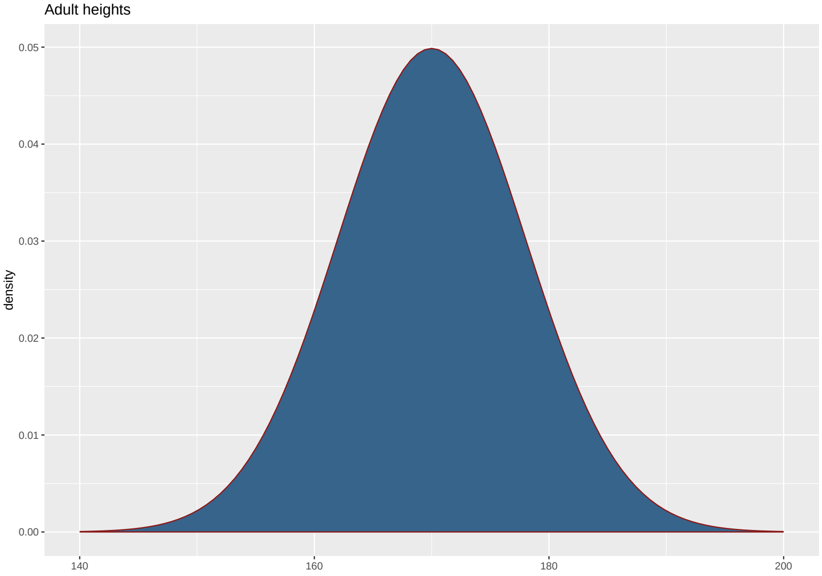




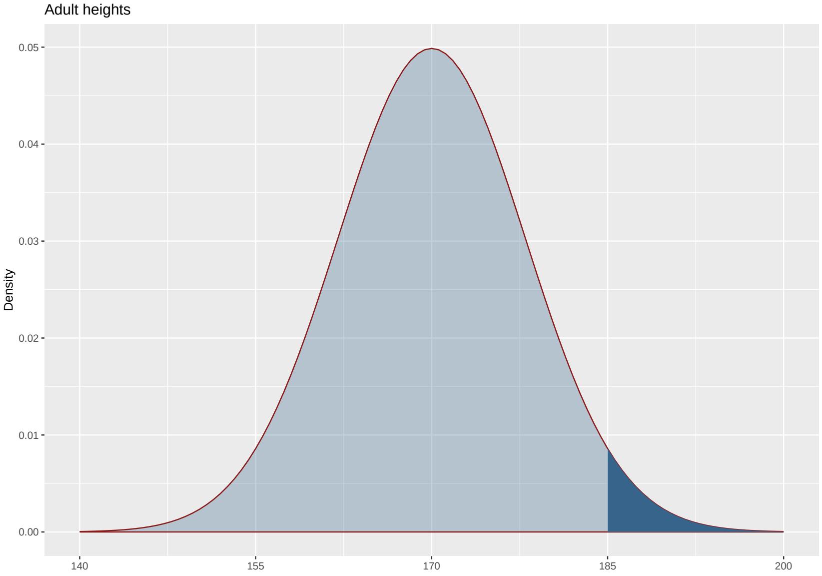
Adult heights
count
2000
1000
0
140
160
180
200
height


Adult heights
density
0.05
0.04
0.03
0.02
0.01
0.00
140
160
180
200
height

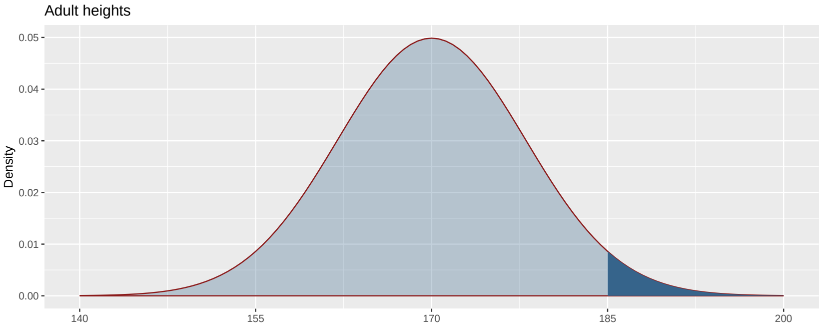
Probability density function



Probabilities as areas



Probabilities as areas

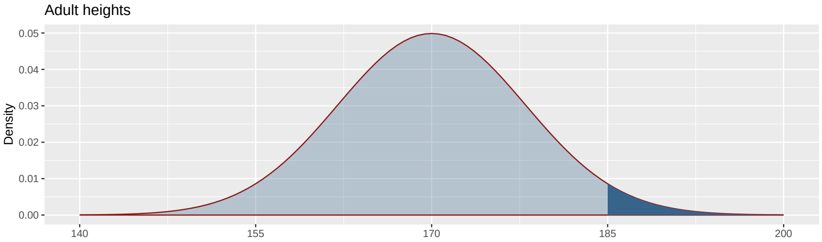


- The probability that height is greater than 185cm i.e. $P(X > 185)$
- Like summing up very tiny histogram bins above a certain point

Probability as areas

Important notes

- The sum of the whole area under the curve is equal to 1 (because we know all probabilities have to sum to one)
- A value is either greater than or less than/equal to a number
- So can express probabilities as the complement e.g. $P(X > 185) = 1 - P(X \leq 185)$



Where are we going

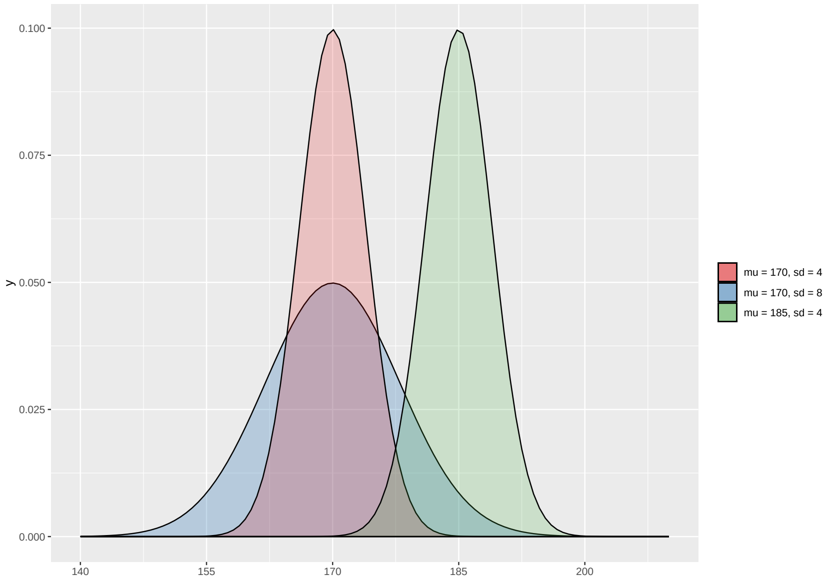
The normal distribution

The normal distribution

- One of the most important continuous probability distributions
- Is described by the formula

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- The shape is determined by two **parameters**, μ and σ
- If we were to plot $f(x)$ as a function of x, we would obtain a normal distribution that would be centered at whatever value of μ we specified, and it would have a standard deviation equal to σ.

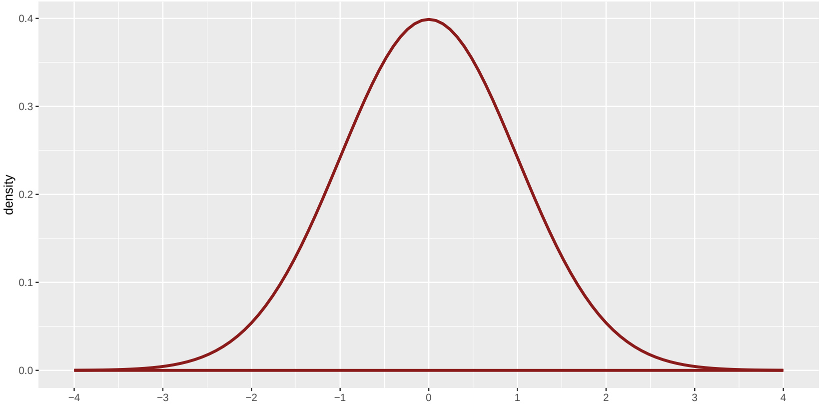

0.100
0.075
0.050
0.025
0.000
y
140
155
170
185
200
mu = 170, sd = 4
mu = 170, sd = 8
mu = 185, sd = 4

The normal distribution

- Many variables naturally resemble the normal distribution (or can be transformed to be so)
- Height, weight, intelligence...
- Strong relationships with other distributions
- Many sample statistics are normally distributed (more later)

The standard normal distribution

A special case of the normal distribution with $\mu = 0$ and $\sigma = 1$.



The standard normal distribution

- ~68% of area within 1 standard deviation

The standard normal distribution

- ~95% of the area within 2 standard deviations

Any normal distribution can be transformed into the standard normal

Say X is normally distributed with mean μ and variance σ^2. We can write this as

$$X \sim N(\mu, \sigma^2)$$

We can transform X using the *z*-**transformation**

$$\frac{X - \mu}{\sigma}$$

Call this transformed version Z i.e. $Z = \frac{X-\mu}{\sigma}$. Then

$$Z \sim N(0, 1)$$

we can refer to the transformed version as **Z-scores**.

Z-scores

- Z-scores tell you the number of standard deviations by which the value of a raw score is above or below the mean value.
- In the heights example, the mean $\mu = 170$ and standard deviation $\sigma = 8$.

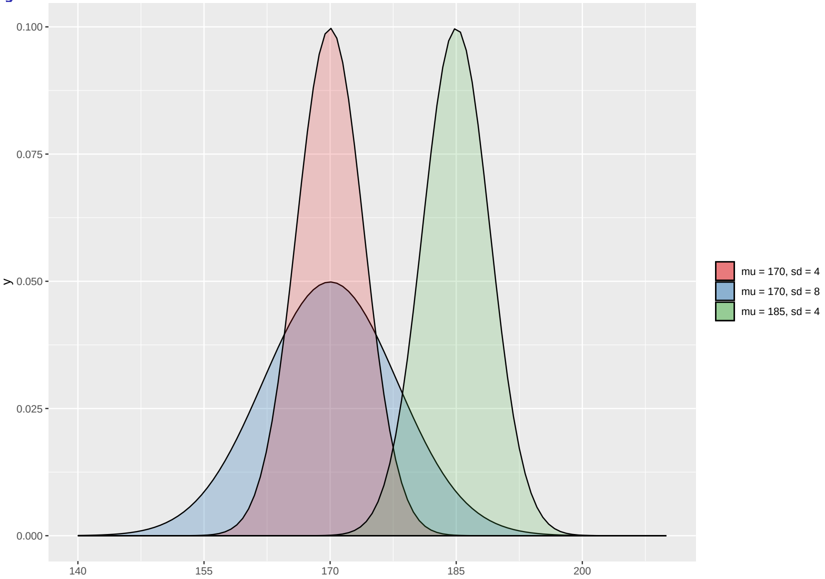
Rohan is 180cm. What is his Z-score?

$$Z = \frac{180 - 170}{8} = 1.25$$

So Rohan is 1.25 standard deviations above the mean height.

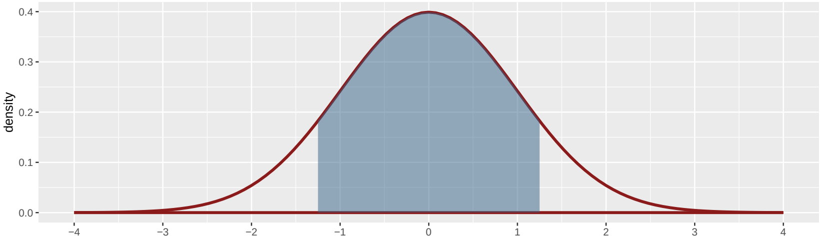
- Monica is 168cm, so her Z-score is -0.25. So she is 0.25 standard deviations below the mean height.

Why calculate Z scores?



- Z-score for person 165cm tall from red v blue v green distribution?

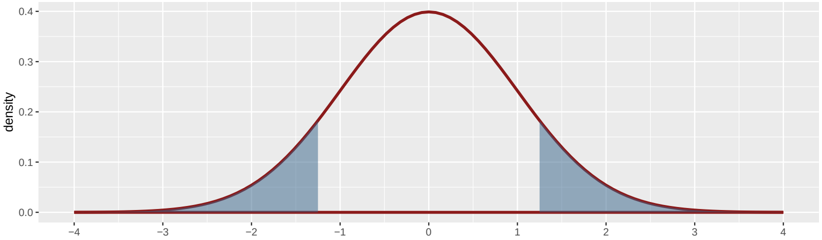
Z-scores



- Recall that ∼95% of the area was within 2 standard deviations.
- We can flip this and ask what proportion of the area of a standard normal is within 1.25 standard deviations?
- The answer is ∼79%

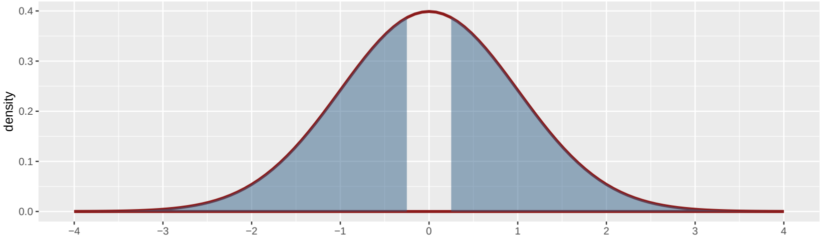
Z-scores

- ~79% of the area is within 1.25 standard deviations
- Alternative interpretation: If we randomly draw a value from a standard normal distribution, the value will be between [-1.25,1.25] 79% of the time.
- Conversely, 21% of values will fall outside that range



Z-scores

- If the Z-score is small (i.e. the absolute value is close to zero), then the observed value is likely to come from that distribution
- E.g. Monica's Z-score was -0.25
  - 40% of the total area falls outside the [-0.25, 0.25] range
  - So it would be quite likely to observe outside this range, just by chance



- In contrast it would be less likely to observe something outside the range of Rohan's Z-score [-1.25, 1.25]
- If we observed an even bigger Z-score, it would be even less likely

Finding areas under the curve using R

You can find the area under a normal curve in R using the `pnorm` function, which returns the cumulative probability from $-\infty$ to the value supplied to the `pnorm` function.

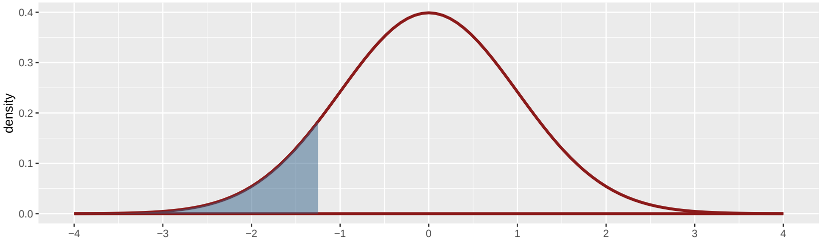
For example, find the area below z = -1.25:

```
pnorm(-1.25)
```

```
## [1] 0.1056498
```

```
# note that this is short for pnorm(-1.25, mu = 0, sd = 1)
```

This corresponds to



Finding areas under the curve using R

To find the area above 1.25:

```
1-pnorm(1.25)
```

```
## [1] 0.1056498
```

```
# or
# pnorm(1.25, lower.tail = FALSE)
```

This corresponds to

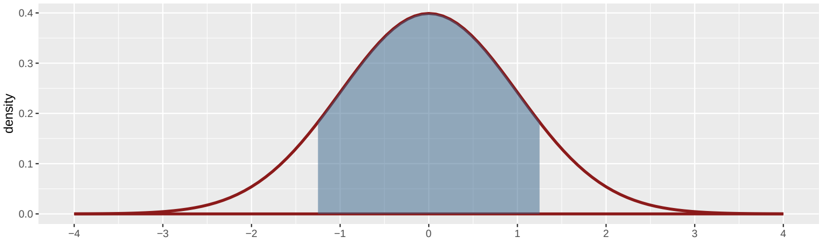
Finding areas under the curve using R

To find the area between -1.25 and 1.25

```
pnorm(1.25) - pnorm(-1.25)
```

```
## [1] 0.7887005
```

This corresponds to



Remember that probability is just counting!

Probability is just counting

- Last week when we first started looking at probability, we were just counting things
- We can approximate Normal probabilities by counting the number of observations above/below a value
- Imagine we had a dataset of people's heights
- Calculate Z scores, then calculate the proportion of observations below a Z score of -1.25
- This would approximate $P(Z<-1.25)$

Back to simulation

- Last week we were simulating coin flips using the sample function
- We can also simulate 'draws' (i.e. observations) from a Normal distribution

Example simulation in R

- We can generate random draws from a normal distribution in R using the `rnorm` function.
- For example, the following code generates 1000 observations from a standard normal

```
set.seed(1889) # makes sure the random numbers generated are the same each time
# rnorm allows you to simulate values from a normal distribution
z_scores <- rnorm(n = 1000, mean = 0, sd = 1)
z_scores[1:10] # show the first ten draws
```

```
##  [1]  2.4552434 -0.2762338 -0.9449433 -0.8251957  0.4853165  0.3160066
##  [7] -1.1079699 -0.7399001 -1.0937371  0.1744686
```

Calculating probabilities by summing

We can then calculate what proportion of the simulated values are above a Z-score of 1.25 or below -1.25:

```
(sum(z_scores< -1.25) + sum(z_scores> 1.25))/1000
```

```
## [1] 0.211
```

Note that this is very similar to

```
pnorm(-1.25) + 1- pnorm(1.25)
```

```
## [1] 0.2112995
```

Sampling distributions and the central limit theorem

Why do we care about the Normal distribution so much

- It's just one distribution
- It tends to over-simplify the real distribution of outcomes/variables

BUT it turns out that the distribution of summary statistics that we're interested in (e.g. means) tend towards being Normal

- this is important for regression and inference

Sampling distributions

A **sampling distribution** is a probability distribution for a statistic based on repeated samples.

Say we are interested in taking a random sample of people's heights, X and calculating the mean height for that sample. So our statistic of interest is the mean height, $\bar{X}$.

Randomly sampling heights

We first take a random sample of 12 people and get the following heights

```
##  [1] 180.9 177.2 179.4 181.8 159.2 182.2 184.7 171.3 178.1 170.4 157.5 167.8
```

The observed mean height of this sample is 174.2.

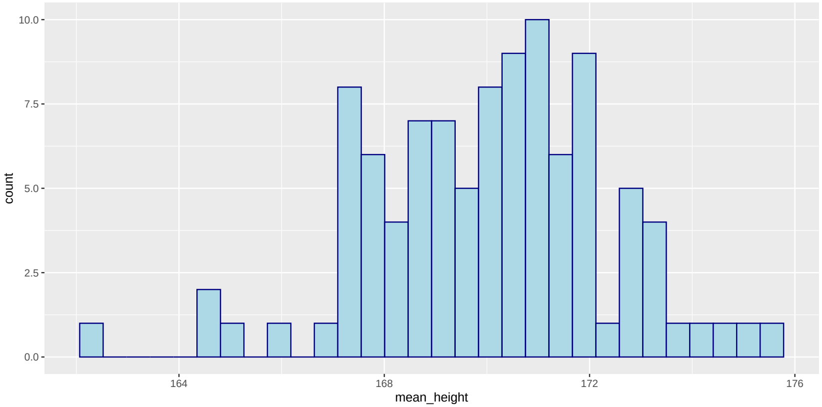
We take a random sample of another 12 people and get the following heights:

```
##  [1] 163.8 158.5 175.5 162.5 166.5 183.0 168.0 174.8 160.2 181.4 173.8 167.6
```

The observed mean height of this sample is 169.6.

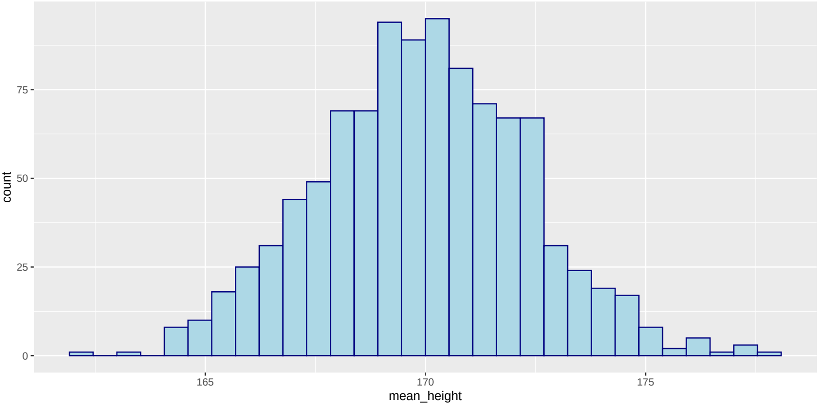
Randomly sampling heights

Say that I keep doing this process again and again and again, and end up with 100 observations of mean height. I can plot a histogram of these means:



Randomly sampling heights

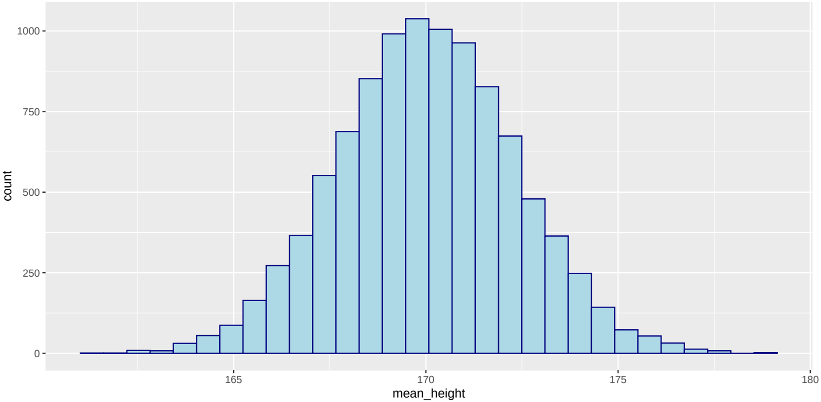
Okay, what if I take 1000 samples. I plot a histogram of the means again



What do you notice?

Randomly sampling heights

What about 10,000 samples?



The central limit theorem

The distribution of the sum (or mean) of a set of independent random variables will **tend towards** a normal distribution.

- "tend towards" means as the number of observations of the sum or mean gets larger, the distribution will become more normal
- The central limit theorem holds even if the original variables themselves are not normally distributed.

The central limit theorem

For a random variable X with $E(X) = \mu$ and $Var(X) = \sigma^2$, the central limit theorem results in the following distribution for the mean $\bar{X}$:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

What does this mean?

- The mean $\bar{X}$ will be centered at the same value as X
- The variance of $\bar{X}$ depends on the variance of the original random variable X and also the number of samples of the mean we have, n.

The quantity $\frac{\sigma}{\sqrt{n}}$ is also called the **standard error of the mean**.

Sampling heights

Before, we were repeatedly taking a random sample of 12 heights.

When we did this 100 times, we had 100 observations of $\bar{X}$. The mean was 170 and the standard error of the mean was 0.078.

When we did this 10000 times, we had 100 observations of $\bar{X}$. The mean was 170 and the standard error of the mean was 8×10^{-4}.

So the more samples we have, the more sure we are of our estimate of the population mean μ

The random variables need not be Normal distributed!

- In the above example, original heights X were Normal distributed, but that isn't required for CLT to hold
- For any sequence of random variables $X_1, \ldots, X_n$ that are drawn randomly from a distribution of expected value μ and variance σ^2
- We can calculate the sample average

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

- For large enough n the distribution of $\bar{X}_n$ gets arbitrarily close to the Normal distribution with mean μ and variance σ^2/n.

Summary

- Normal distribution
- Standard Normal distribution
- Z-scores
- Probabilities of observation a normal variable in a certain range
- Sampling distribution for sample mean
- Central Limit Theorem