Lecture 01b - Introduction

ENVX2001 Applied Statistical Methods

Dr. Januar Harianto
The University of Sydney

Feb 2024

Outline

- Download PDF
- Samples, populations and study design
- Designs: why do we care?
- Mean and standard deviation
- The sampling distribution
- Central limit theorem

Download PDF

Click above to access a .pdf version of this presentation.

Otherwise, continue navigating through the slides.

Samples, populations and study design

> “To call in a statistician after the experiment has been done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.”

– Sir Ronald Fisher

Revision

- **Population**: the entire group of individuals or instances about whom we want to draw conclusions.
- **Sample**: a *subset* of the population.

Parameter

- A numerical **measure** that describes an aspect of a population.
- *Not known* (unless we sample the entire population), therefore we **estimate** them using a **sample statistic**.
- **What information does the *sample statistic* give about the *population parameter*, and how reliable is that information?**

Confused?

Visit the ENVX Resources organisation on GitHub.

- Probability distributions (ENVX1002) - 2024 version
- Sampling distributions (ENVX1002) - 2024 version

Tip

You will explore more experimental design principles next week, and in **Module 2**.

Designs: why do we care?

> (On a failed experiment)
>
> *That is not an experiment you have there, that is an experience*.

– Sir Ronald Fisher

Measure everything?

Why not measure every individual in a population, instead of designing a sampling strategy?

- **Impractical** to measure every individual in a population, and some populations are *infinite* – practically impossible to measure all.
- **Costly** to measure every individual in a population – time, money, resources.
- **Destructive** in many biological cases e.g. to measure the age of a plant, you may have to cut it down, so you want to respect the loss of life.

Important

Importantly, sampling from a population – when done correctly – can give a **good estimate of the population parameter**, give or take some *uncertainty*. Apart from a census study, there should be no reason to measure every individual in a population.

Sampling designs

Can be done in two general ways:

1. **Observational study**
2. **Controlled experiment**

When designed correctly, both can give us a good estimate of the population parameter while saving time and resources.

Considerations

- Samples should be **representative** of the population and **randomly** selected.
- **Bias** can be introduced if the sampling design is not carefully considered.
- **Confounding** variables can also affect the results.

We will explore these concepts in more detail over the next few weeks.

Observational study vs. controlled experiment

Aspect	Observational study	Controlled experiment
Control	No control over the variables of interest - **Mensurative** and **Absolute**	Control over the variables of interest - **Comparative** and **Manipulative**
Causation	Cannot establish causation, but perhaps **association**	Can establish **causation**
Feasibility	Can be done in many cases	May be destructive and cannot always be done

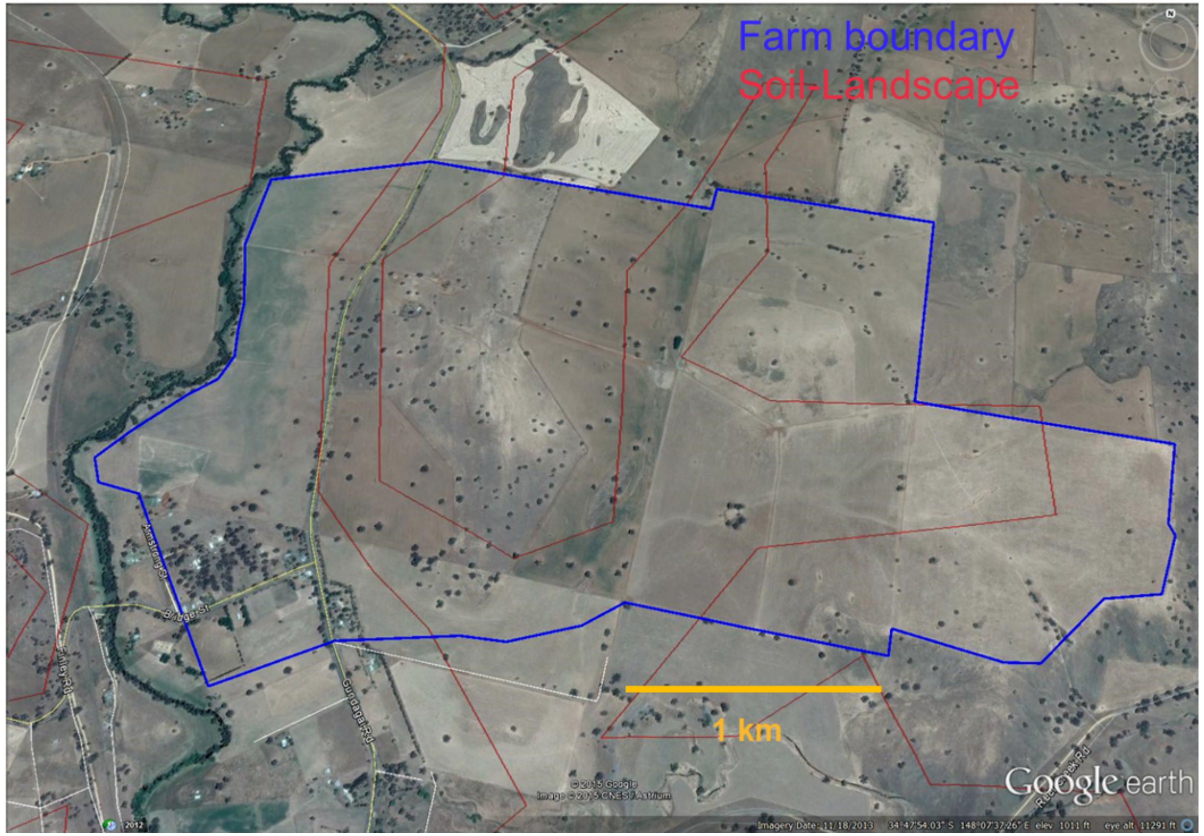
Other designs exist

- **Theoretical models** (e.g. mathematical models): useful for understanding the system, often used in ecology and epidemiology. No data collection.
- **Simulation studies**: useful for figuring out experimental design and understanding the system. Some data collection may be involved to inform the model.
- **Case studies**: Similar to observational studies, but often with a single case! Useful for understanding a unique situation, often used in medicine and psychology. No control over the variables of interest and sometimes no statistical inference is made.

Soil carbon

Farm boundary
Soil-Landscape
1 km
Armstrong St
Bridge St
Gundagai Rd
Finley Rd
Google earth
© 2015 Google
Image © 2015 CNES / Astrium
Imagery Date: 11/18/2013
34°47'54.03" S 148°07'37.26" E elev 1011 ft
eye alt 11291 ft
2012

Soil carbon



What is the best way to sample?

- Sequestering carbon in soil is a potential way to mitigate climate change, and provides nutrients and resilience to crops. Worth $50/tonne if measured.
- Collecting soil samples is costly and time-consuming, about $100/sample.
- We want a way to estimate the soil carbon content in a large area - some kind of **summary statistic**.

Summary statistics

Central tendency

- **Mean**: the average of the data.
- **Median**: the middle value of the data.
- **Mode**: the most frequent value in the data.

Variability

- **Range**: the difference between the largest and smallest value.
- **Interquartile range**: the difference between the 75th and 25th percentile.
- **Variance**: the average of the squared differences from the mean.
- **Standard deviation**: the square root of the variance.

Mean and standard deviation

> Statistics always remind me of the fellow who drowned in a river whose average depth was only three feet (~0.9 m).

– *Woody Hayes, American football coach*

Mean and standard deviation

- The most *common* measures of central tendency and variability.
- Works well for **symmetric** and **unimodal** distributions, therefore the assumption is that the data is normally distributed.

▶ Code

Arithmetic mean

> Sum of all the values, divided by the number of values.

Population mean

If we measure the entire population, the population mean μ is:

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

where y_i is the ith observation and N is the number of individuals in the population.

Sample mean

Sample mean is based on the same principle, but we use n instead of N and $\bar{y}$ instead of μ.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where y_i is the ith observation and n is the number of sample observations.

Variance

> The average of the squared differences from the mean.

Population variance:

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

Sample variance

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

Standard deviation

> The square root of the variance.

Population standard deviation

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

Sample standard deviation

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

Why $n - 1$?

- The sample variance and standard deviation calculations use $n - 1$ in the denominator, not n.
- This is called **Bessel's correction**.
- It is used to correct the bias in the estimation of the population variance from a sample, **as n number of observations have $n - 1$ independent residuals**.
  - *You will learn more about this – and degrees of freedom – in the next module*.

Soil carbon

Sampling design: Soil carbon content was measured at 7 locations across the area. The amount at each location was 48, 56, 90, 78, 86, 71, 42 tonnes per hectare (t/ha).

```r
soil <- c(48, 56, 90, 78, 86, 71, 42)
soil
```

```
[1] 48 56 90 78 86 71 42
```

Calculating mean and standard deviation

```r
mean(soil)
```

```
[1] 67.28571
```

```r
sd(soil)
```

```
[1] 18.8566
```

What do these numbers tell us? How confident are we that they represent the entire area?

The sampling distribution

Distributions

- The **population distribution** is the distribution of all the individuals in the population.
- From the population distribution, we can sample it to get a **sample distribution**.
- If we summarise the sample distribution, we get a single value - the **sample statistic**.
- The sample statistic is part of a **sampling distribution**, based on the idea that given unlimited resources, we could sample the population many times and calculate the sample statistic each time.

Example

- We want to measure the mean height of trees in a forest, which contains 1000 trees. **1000 possible height values make up the population distribution**.
- We can't measure all the trees, so we take a sample of 100 trees and calculate the average height.**100 height values make up the sample distribution**.
- The mean height of the 100 trees is calculated. This is the **sample statistic** - a single value for the sample.
- To make up the **sampling distribution**, we could repeat the process of taking a sample of 100 trees and calculating the mean height many times…

Distributions - visualised

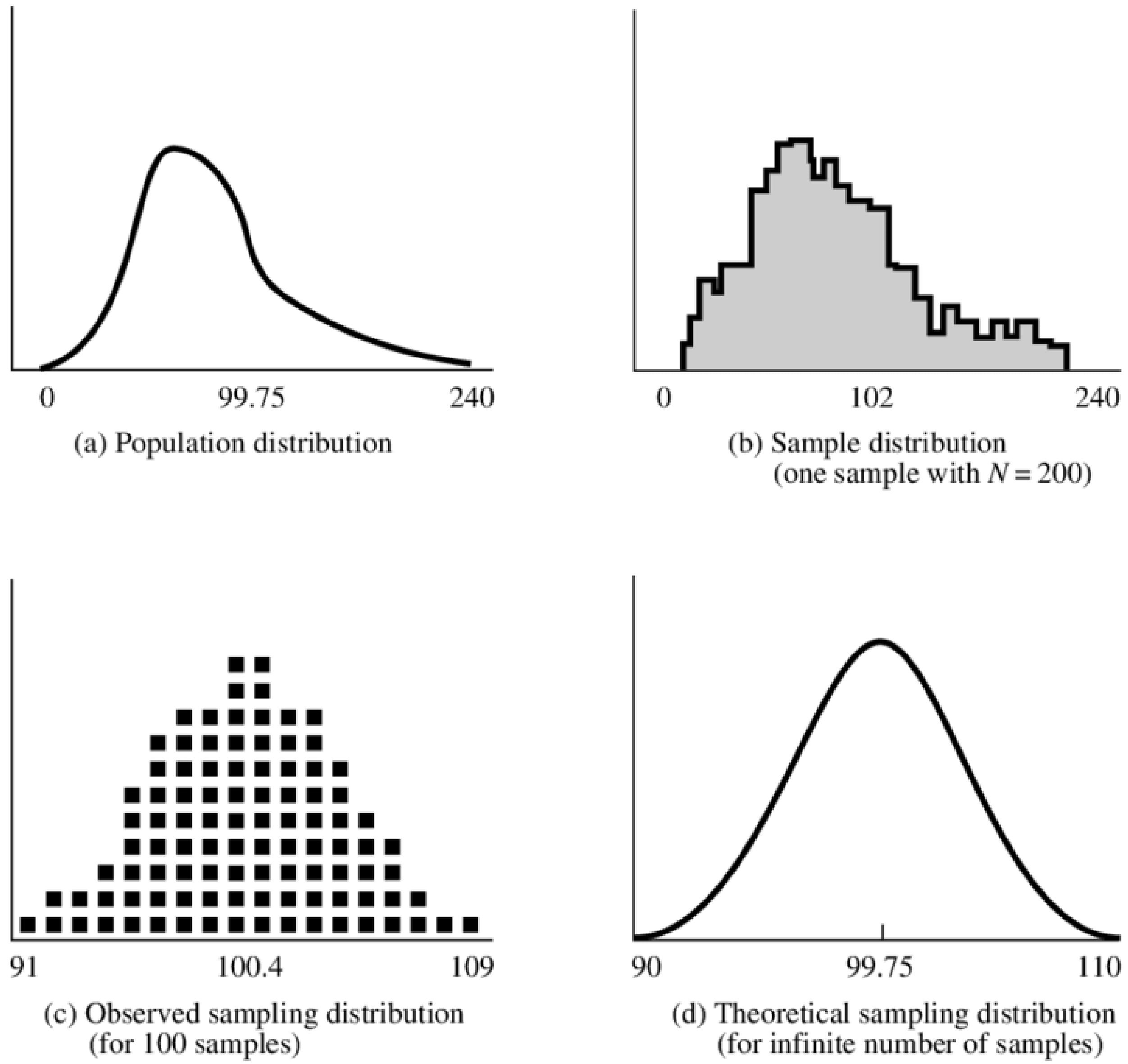


Figure 1: Population, sample and sampling distributions. Source.

How can distributions help us answer the question?

> What information does the *sample statistic* give about the *population parameter*, and how reliable is that information?

We need to standardise the sample statistic to the *number of observations* in the sample.

Standard error

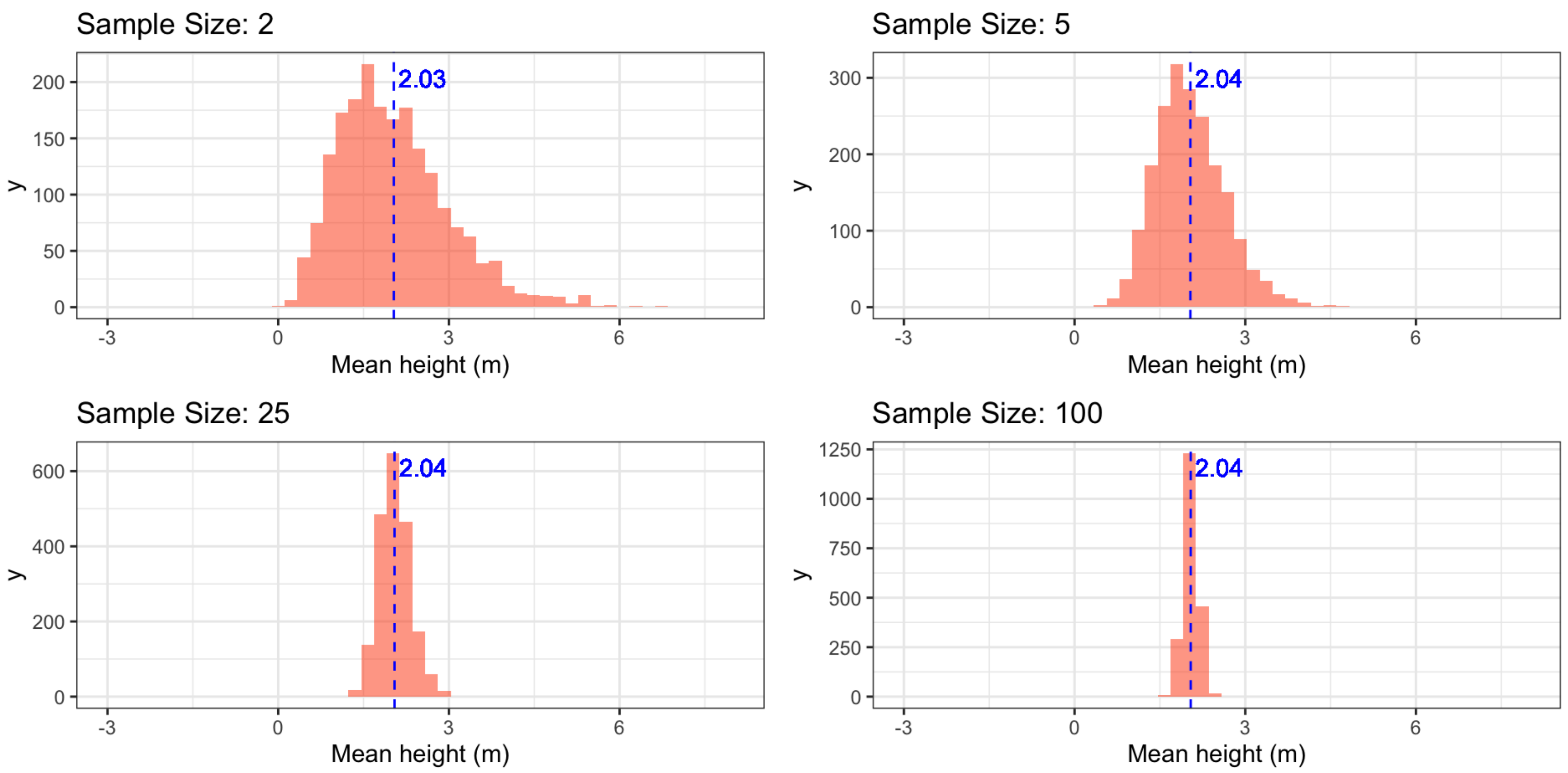
$$SE = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation and n is the number of observations in the sample.

- The standard deviation value is *standardised* to the number of observations in the sample.
- Tells us how much the sample statistic varies from sample to sample, i.e. **how well we know the mean**.
- If standard error is "small", we are more confident in the sample statistic – **more on this next week**.

Effect of sample size

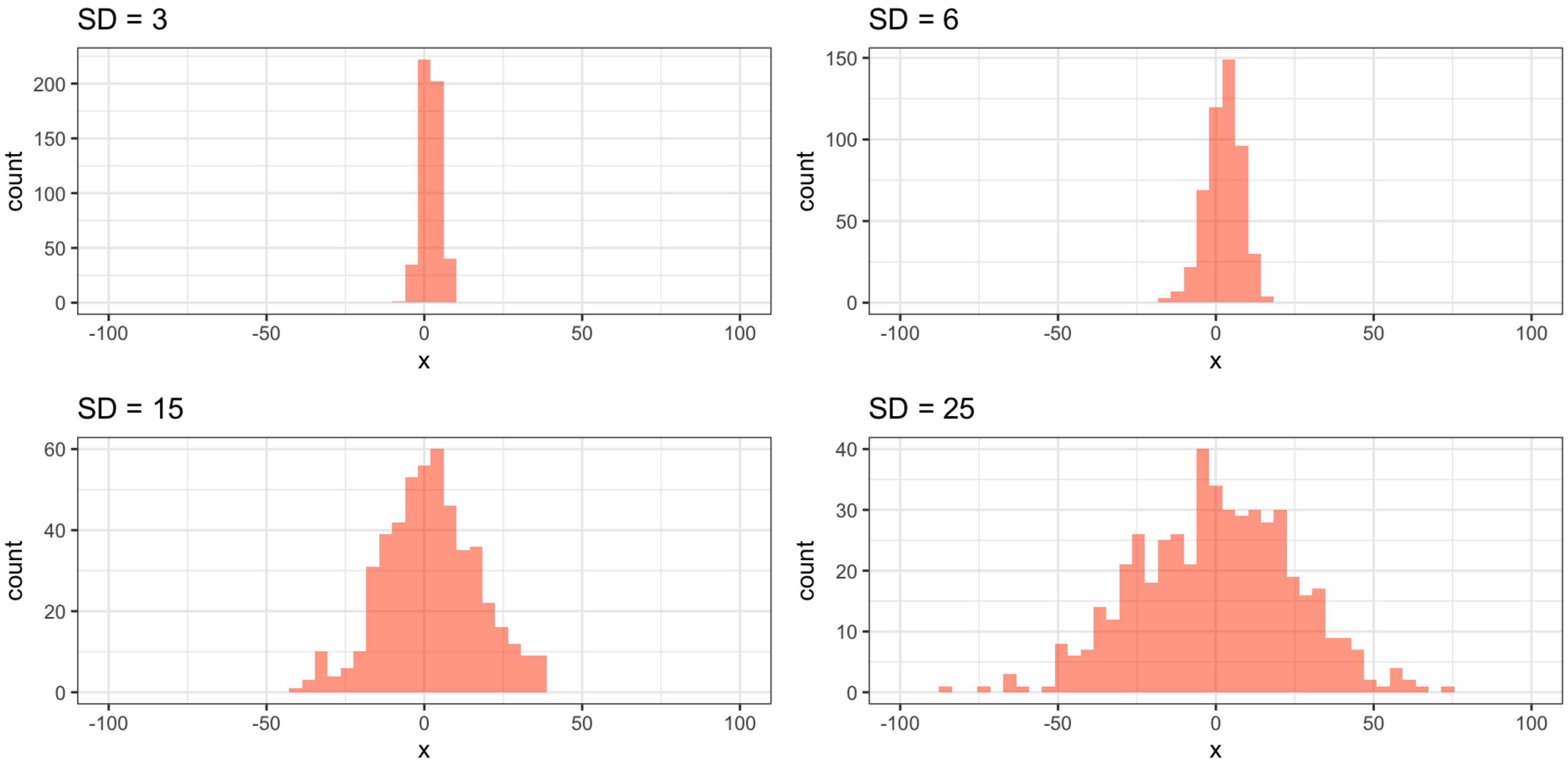
▶ Code



- Increased sample size leads to a more accurate estimate of the population mean, reflected by the **narrower distribution** of the sample mean, which is captured by the **standard error**.

Effect of variability

▶ Code



- Increased variability leads to a wider distribution of the sample mean (i.e. less precision), which is *also* reflected by the **standard error**.

Central limit theorem

> I know of scarcely anything so apt to impress the imagination as the **wonderful form of cosmic order** expressed by the Central Limit Theorem. The law would have been personified by the Greeks and deified, if they had known of it."

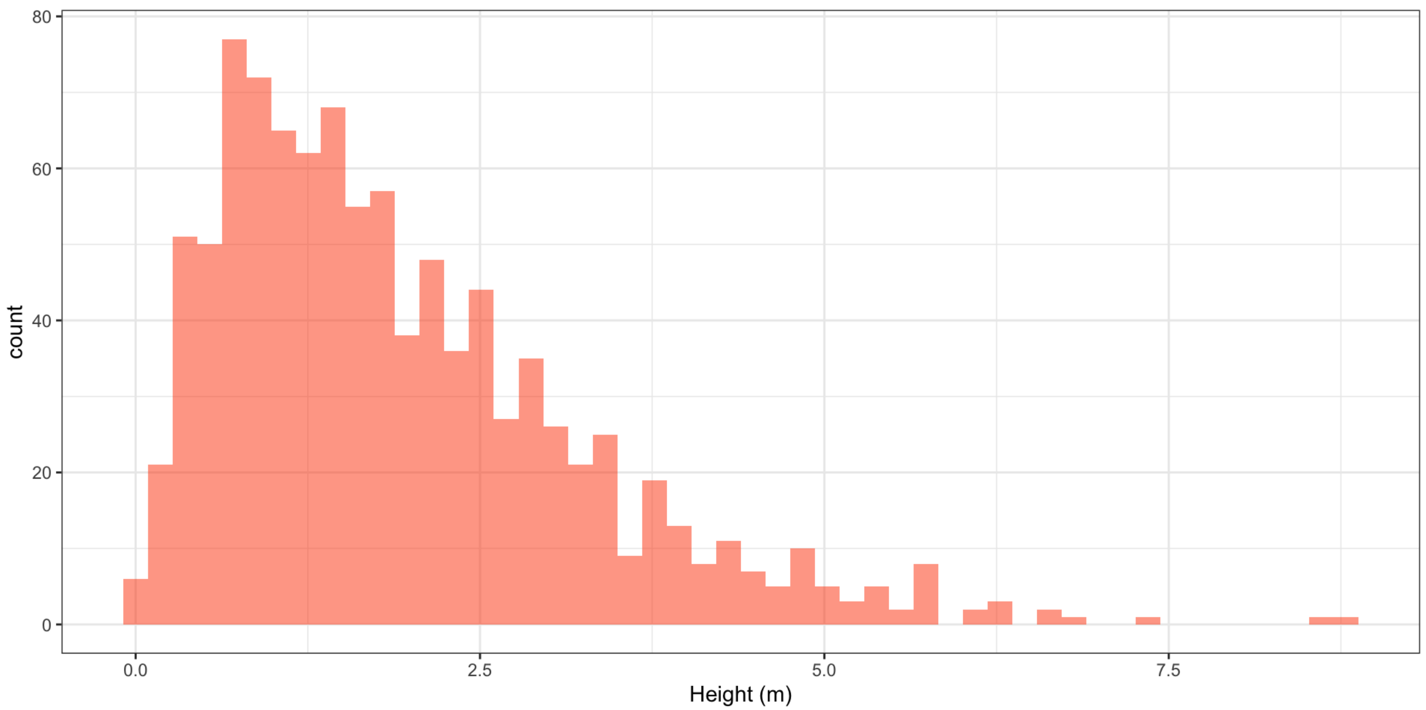
– Sir Francis Galton, 1889, Natural Inheritance* (emphasis added)

CLT

- A fundamental theorem in statistics.
- Regardless of the shape of the population distribution, the sampling distribution of the **sample mean** will be approximately normally distributed **if the sample size is large enough**.
- Because of this, we can make predictions about the population by assuming that the sampling distribution is normally distributed – **a core assumption in many statistical tests**.

Example

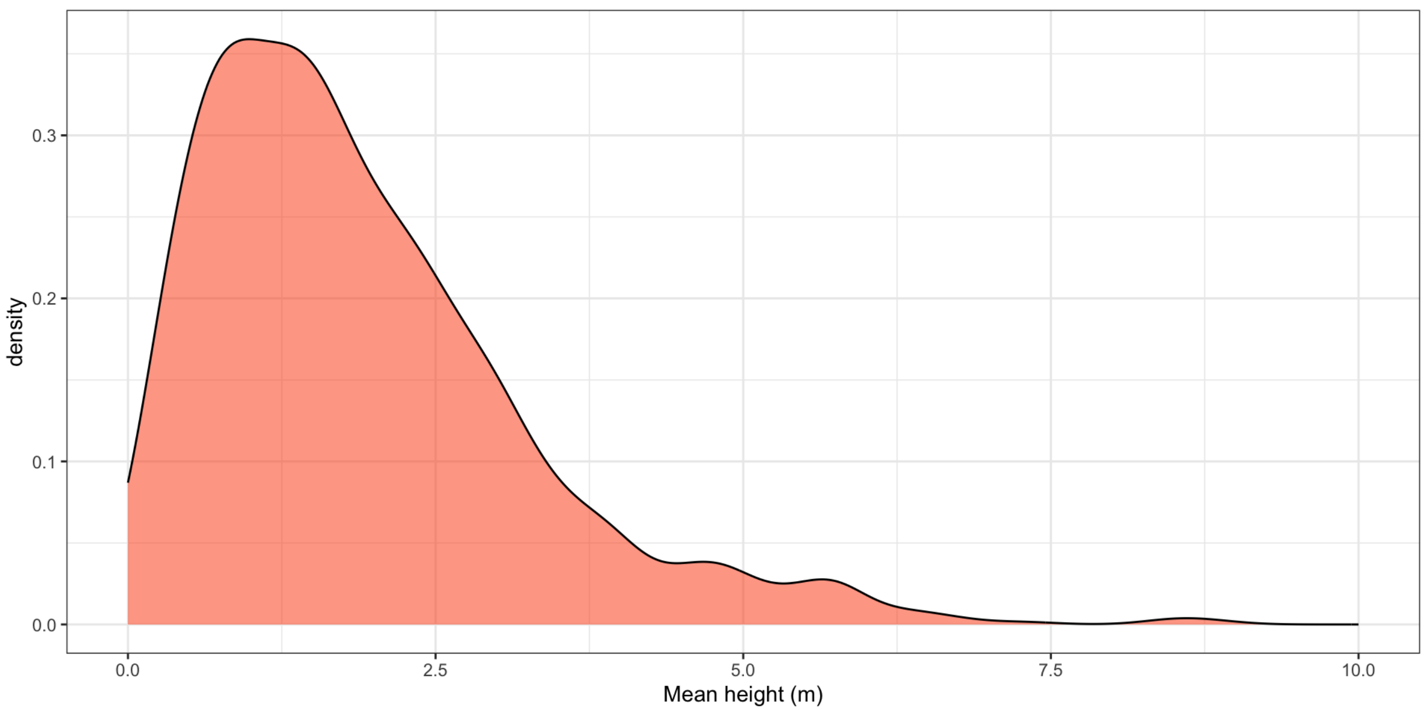
```r
set.seed(239)
library(ggplot2)
library(dplyr)
# Generate a skewed distribution
skewed <- tibble(
    x = rgamma(1000, shape = 2, scal
)

# plot in ggplot2
ggplot(data = skewed, aes(x = x)) +
    geom_histogram(
        fill = "orangered",
        alpha = 0.5, bins = 50
    ) +
    xlab("Height (m)") +
    theme_bw()
```



- Skewed population distribution for tree heights.
- We want to estimate the mean height of the trees in the forest.

1 sample (no summary statistic)

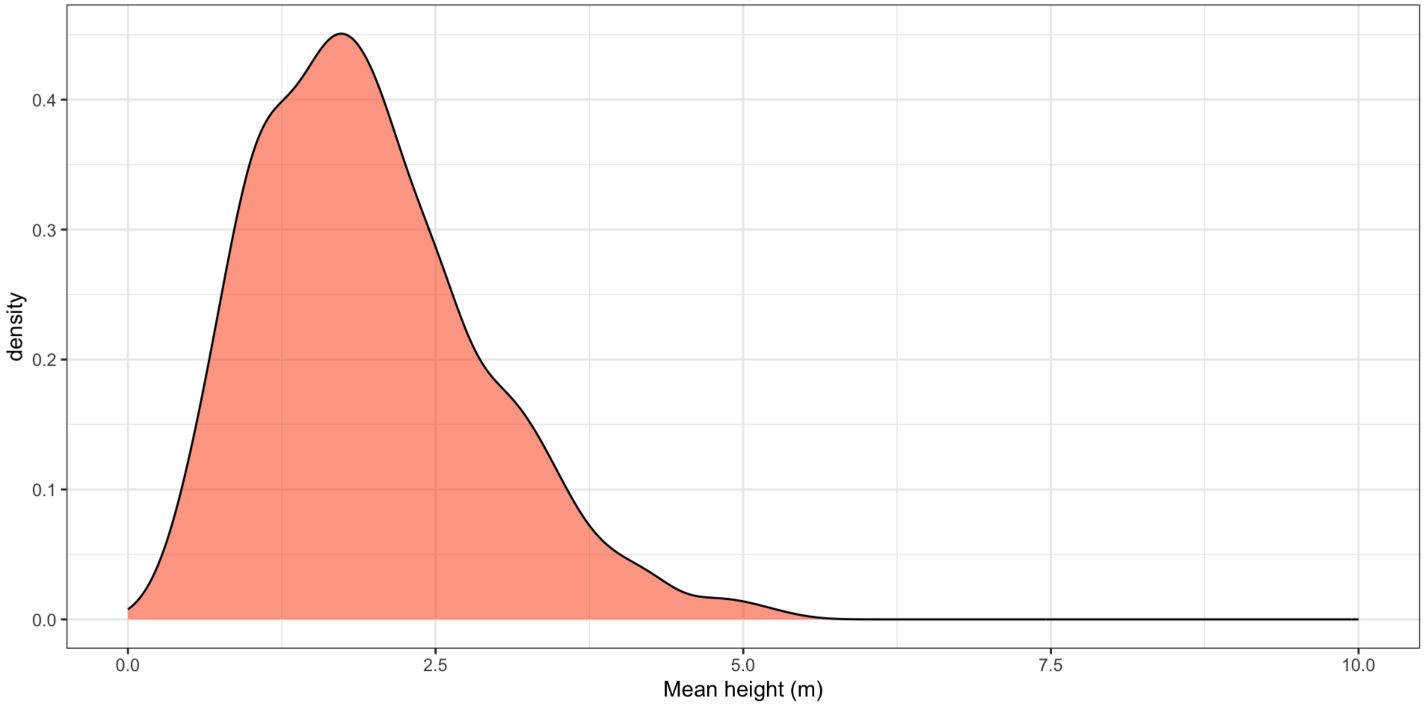
```r
skewed |>
    infer::rep_sample_n(
        size = 1,
        reps = 1000
    ) |>
    group_by(replicate) |>
    summarise(xbar = mean(x)) |>
    ggplot(aes(x = xbar)) +
    geom_density(
        fill = "orangered",
        alpha = 0.5, bins = 50
    ) +
    xlim(0, 10) +
    xlab("Mean height (m)") +
    theme_bw()
```



- A single random sample per calculated mean, repeated 1000 times, gives us a distribution of sample means that will likely mirrors the population distribution.

2 samples

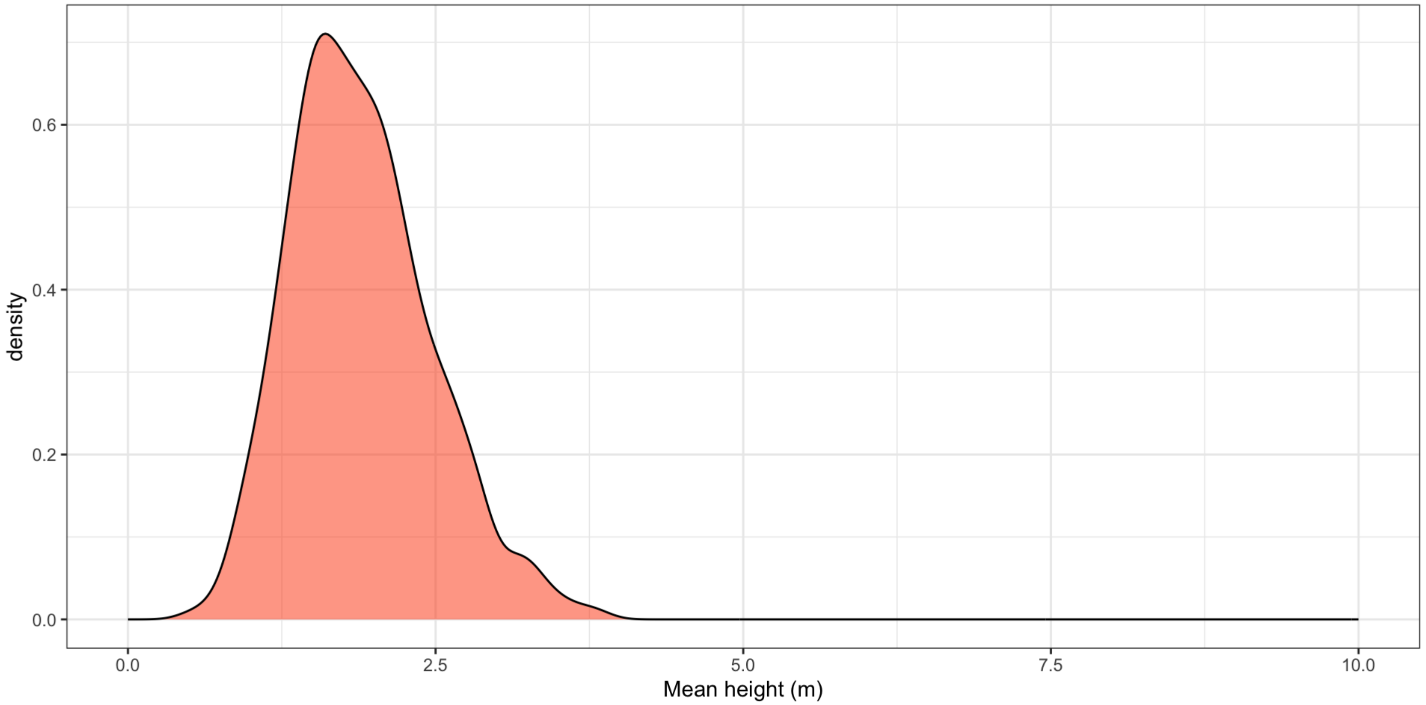
```r
skewed |>
    infer::rep_sample_n(
        size = 2,
        reps = 1000
    ) |>
    group_by(replicate) |>
    summarise(xbar = mean(x)) |>
    ggplot(aes(x = xbar)) +
    geom_density(
        fill = "orangered",
        alpha = 0.5, bins = 50
    ) +
    xlim(0, 10) +
    xlab("Mean height (m)") +
    theme_bw()
```



- Two random samples per calculated mean, repeated 1000 times.
- The distribution of sample means is starting to look more like a normal distribution.

5 samples

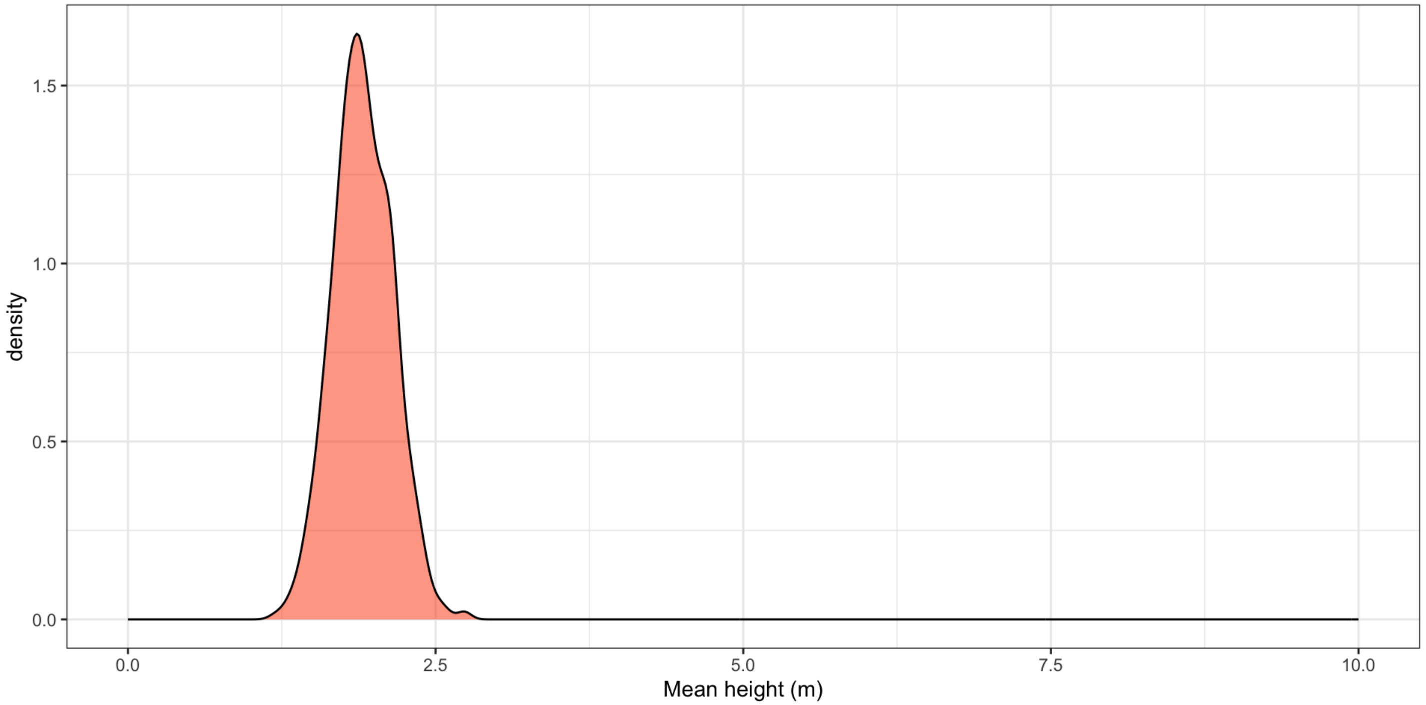
```r
skewed |>
    infer::rep_sample_n(
        size = 5,
        reps = 1000
    ) |>
    group_by(replicate) |>
    summarise(xbar = mean(x)) |>
    ggplot(aes(x = xbar)) +
    geom_density(
        fill = "orangered",
        alpha = 0.5, bins = 50
    ) +
    xlim(0, 10) +
    xlab("Mean height (m)") +
    theme_bw()
```



- Five random samples per calculated mean, repeated 1000 times.
- Not only is the distribution of sample means starting to look more like a normal distribution, but the standard error is also getting smaller.

30 samples

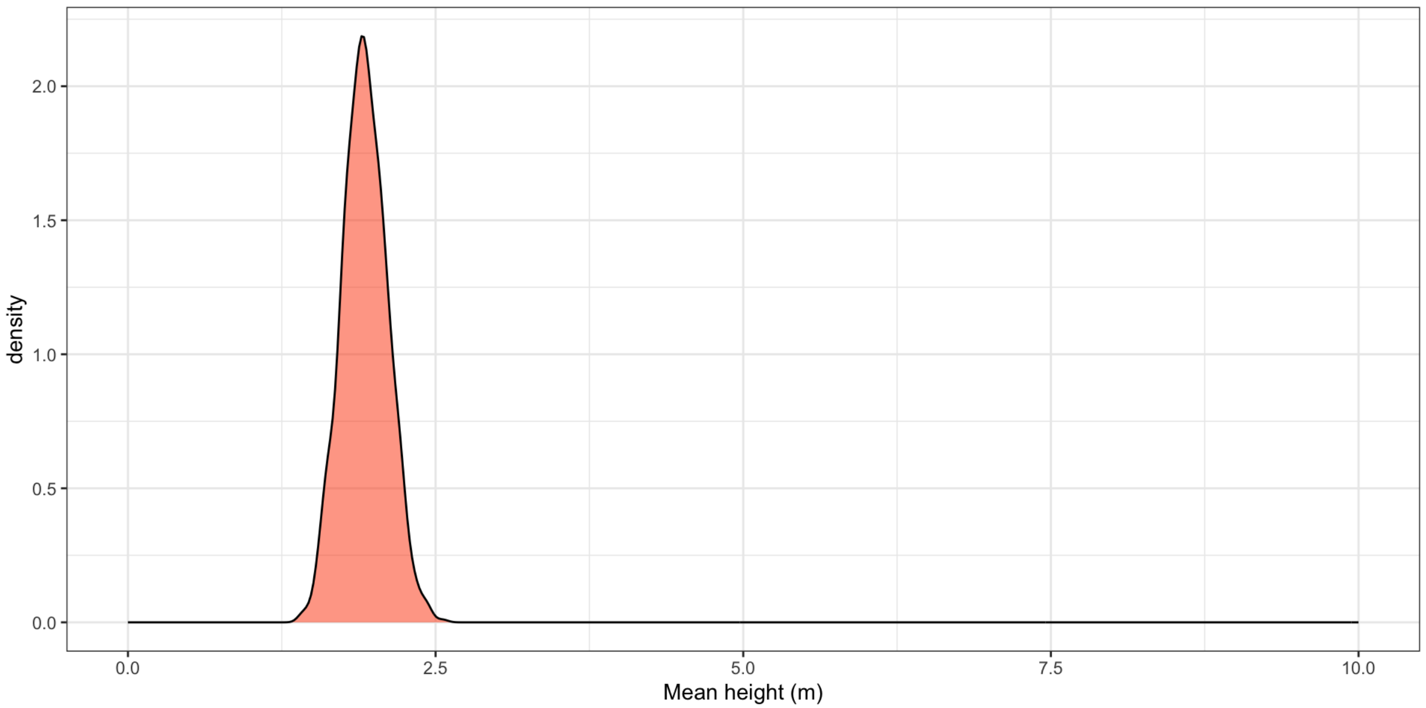
```r
skewed |>
    infer::rep_sample_n(
        size = 30,
        reps = 1000
    ) |>
    group_by(replicate) |>
    summarise(xbar = mean(x)) |>
    ggplot(aes(x = xbar)) +
    geom_density(
        fill = "orangered",
        alpha = 0.5, bins = 50
    ) +
    xlim(0, 10) +
    xlab("Mean height (m)") +
    theme_bw()
```



- Thirty random samples per calculated mean, repeated 1000 times.
- The distribution of sample means is very close to a normal distribution.

50 samples

```
skewed |>
    infer::rep_sample_n(
        size = 50,
        reps = 1000
    ) |>
    group_by(replicate) |>
    summarise(xbar = mean(x)) |>
    ggplot(aes(x = xbar)) +
    geom_density(
        fill = "orangered",
        alpha = 0.5, bins = 50
    ) +
    xlim(0, 10) +
    xlab("Mean height (m)") +
    theme_bw()
```



- Fifty random samples per calculated mean, repeated 1000 times.
- **How many samples is enough?**

How many samples is enough?

- If n is large enough, the sampling distribution of the sample mean will be approximately normally distributed - **allowing us to use the normal distribution to make inferences about the population!**
- **How large is large enough?**
  - **Rule of thumb**: $n \geq 30$ is often used, but this is not a hard and fast rule.
  - **Depends on the population distribution**: if the population distribution is normal, the sampling distribution will be normal for any n.
  - **Depends on the variability**: if the population distribution is highly variable, a larger n is needed to get a normal sampling distribution.

Thanks

Questions?