

Quantitative Methods for Political Economy

Week 2: Normal distributions and sampling distributions

Wouter Leenders

wouter.leenders@kcl.ac.uk

Recap: random variables and distributions

A **random variable** is a numerical summary of the outcome of a random process.

For discrete variables, the **probability distribution** lists all possible values and their probabilities:

$$p(x) = \Pr(X = x), \quad \sum_x p(x) = 1.$$

The **cumulative probability distribution** gives the probability that a value is at or below a threshold:

$$F(x) = \Pr(X \leq x).$$

Recap: expected value and dispersion

- **Expected value**: the long-run average of a random variable.
- **Variance**: a random variable's expected squared deviation from the mean.
- **Standard deviation**: the square root of variance.

$$\mu_X = E(X) = \sum_x x \Pr(X = x),$$

$$\sigma_X^2 = \text{Var}(X) = E[(X - \mu_X)^2],$$

$$\sigma_X = \sqrt{\text{Var}(X)}.$$

Recap: joint, marginal and conditional probabilities

	Did not vote	Voted	Total
Below bachelor's ($E = 0$)	600	1,054	1,654
Bachelor's or higher ($E = 1$)	46	296	342
Total	646	1,350	1,996

$$\text{Joint: } \Pr(E = 1, V = 1) = \frac{296}{1996} \approx 0.148,$$

$$\text{Marginal: } \Pr(V = 1) = \frac{1350}{1996} \approx 0.676,$$

$$\text{Conditional: } \Pr(V = 1 | E = 1) = \frac{296}{342} \approx 0.865.$$

Recap: independence, covariance and correlation

Two random variables are **independent** if one variable provides no information about the other.

$$\Pr(Y = y \mid X = x) = \Pr(Y = y).$$

Covariance measures how variables move together.

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

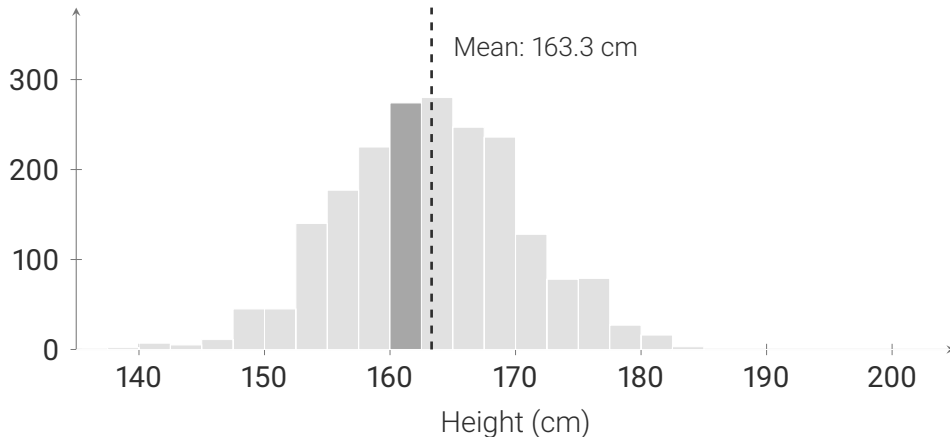
Correlation standardises covariance:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The normal distributions

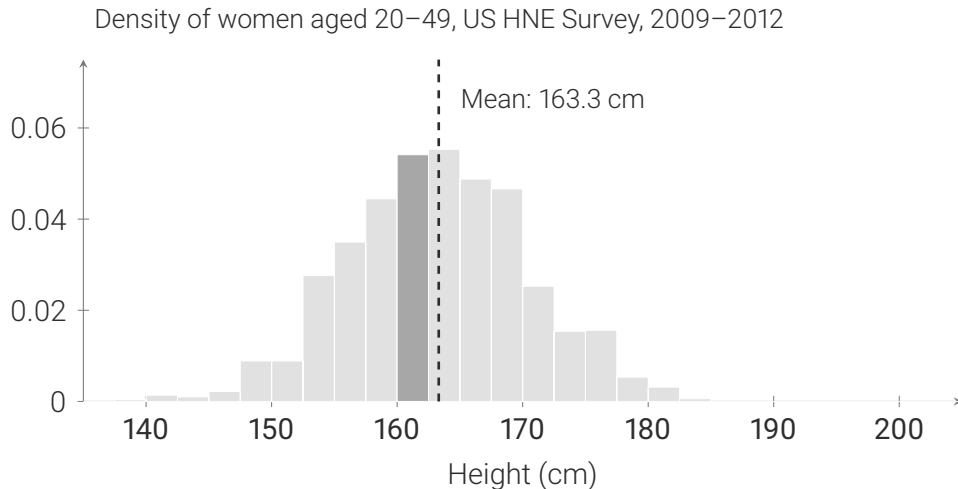
The distribution of women's height

Number of women aged 20–49, US HNE Survey, 2009–2012



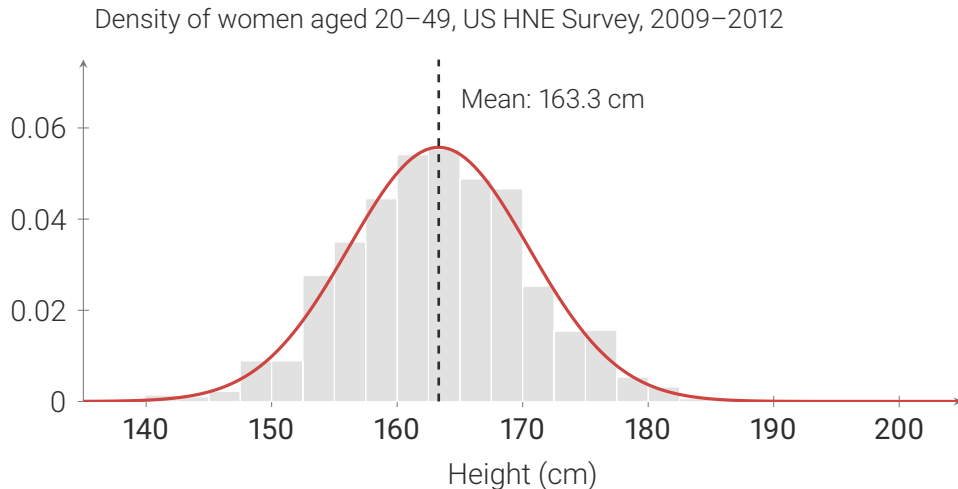
274 women are between 160 and 162.5 cm tall.

The distribution of women's height



0.135 (13.5%) of women are between 160 and 162.5 cm tall.

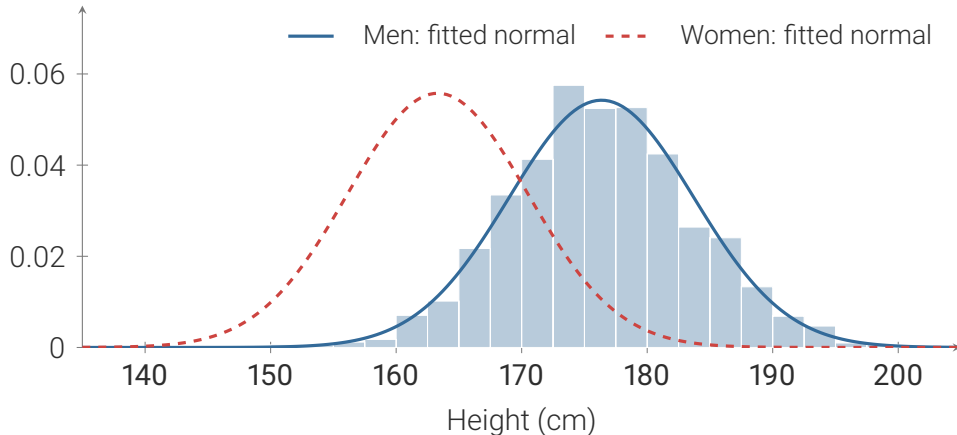
The distribution of women's height



Normal approximation fitted using $\bar{h} = 163.3$ cm and $s = 7.2$ cm.

The distribution of men's and women's height

Density of men and women aged 20–49, US HNE Survey, 2009–2012



Men: $\bar{h} = 176.4$ cm, $s = 7.4$ cm; women: $\bar{h} = 163.3$ cm, $s = 7.2$ cm.

The normal distribution

A **normal distribution** is a probability distribution with a symmetric, bell-shaped density. Its curve is completely determined by its mean μ and standard deviation σ .

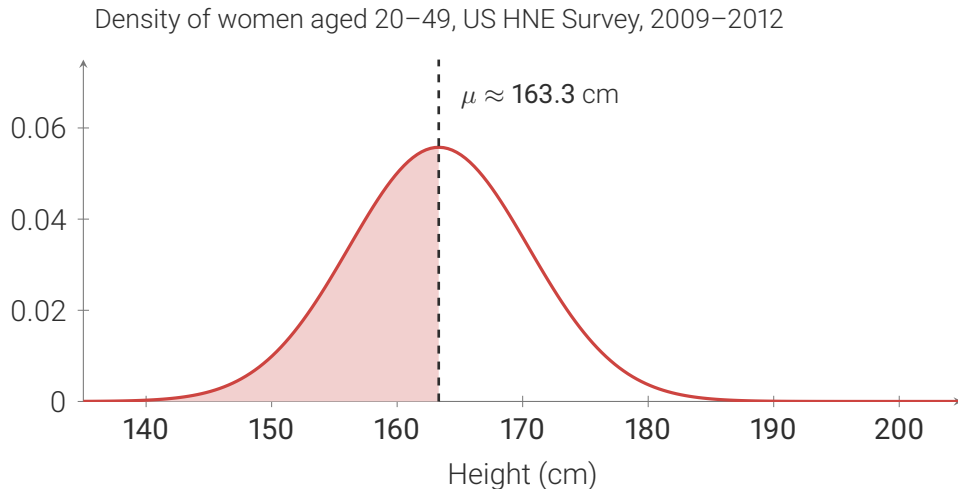
When a random variable X is normally distributed, we write:

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

A few properties of normal distributions:

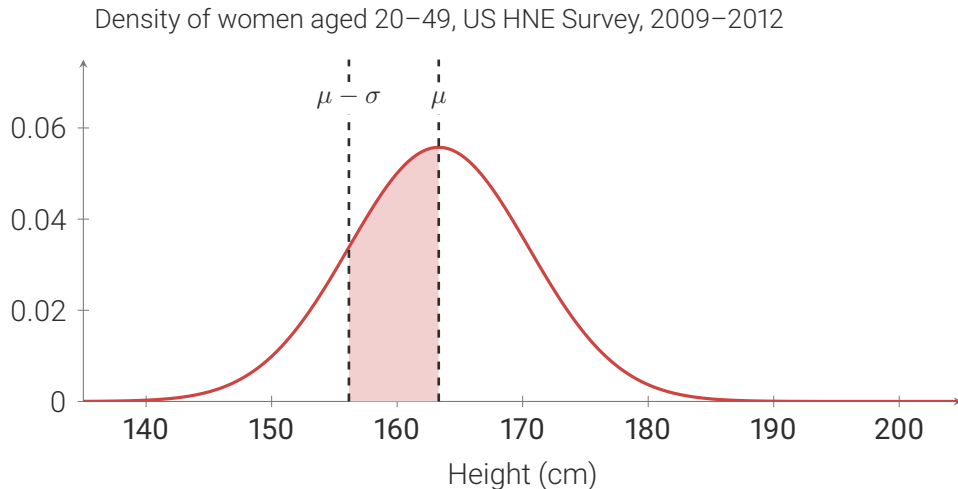
- The mean, median and mode coincide, and are equal to μ .
- μ determines the centre of the distribution; σ determines its spread.
- The probability that a value lies between two values is equal to the area between those values: 68% of observations lie within one standard deviation of the mean; 95% within two standard deviations.

Areas under the normal curve



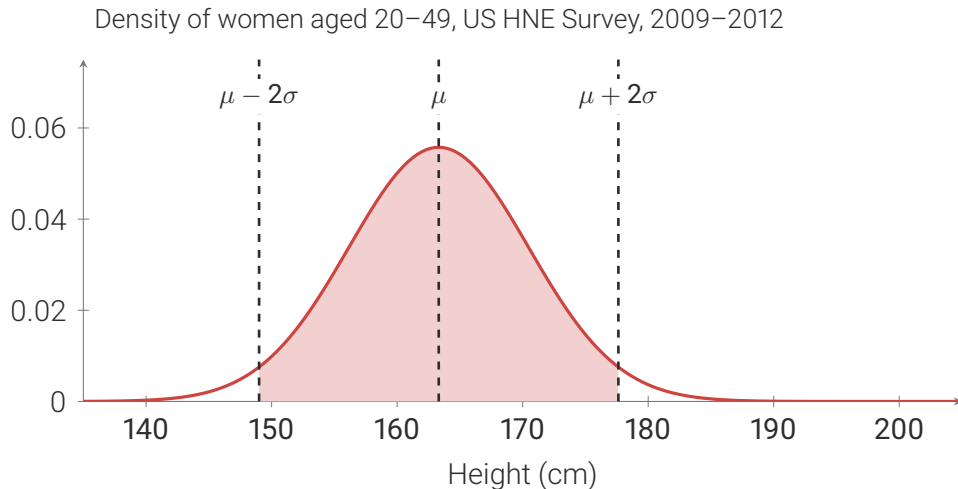
In the fitted normal model, 50% of women are shorter than the average.

Areas under the normal curve



About 34% of observations lie between $\mu - \sigma$ and μ .

Areas under the normal curve



About 95% of observations lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

The standard normal distribution

The **standard normal distribution** is a normal distribution with mean 0 and standard deviation 1:

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

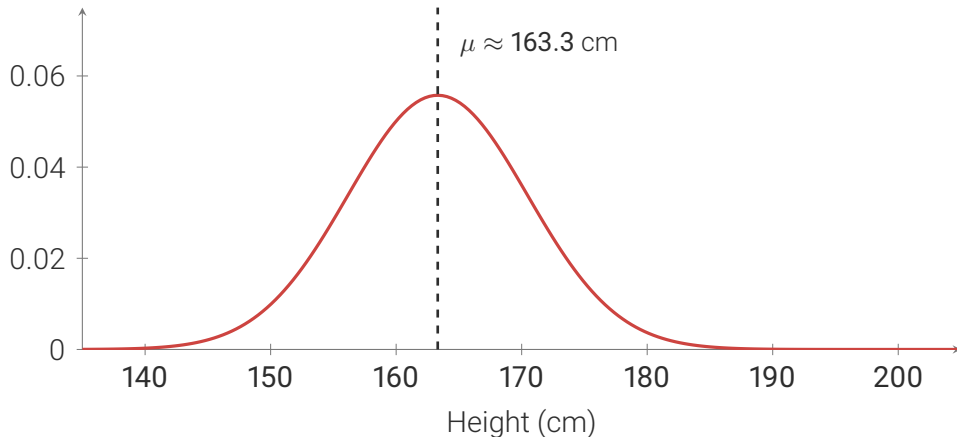
Standardisation replaces values with z-scores, which measure distance from the mean in terms of standard deviation units.

$$z = \frac{x - \mu}{\sigma}.$$

The **standard normal cumulative distribution function**, $\Phi(z) = \Pr(Z \leq z)$, gives the probability that a specific value of Z is less than or equal to z .

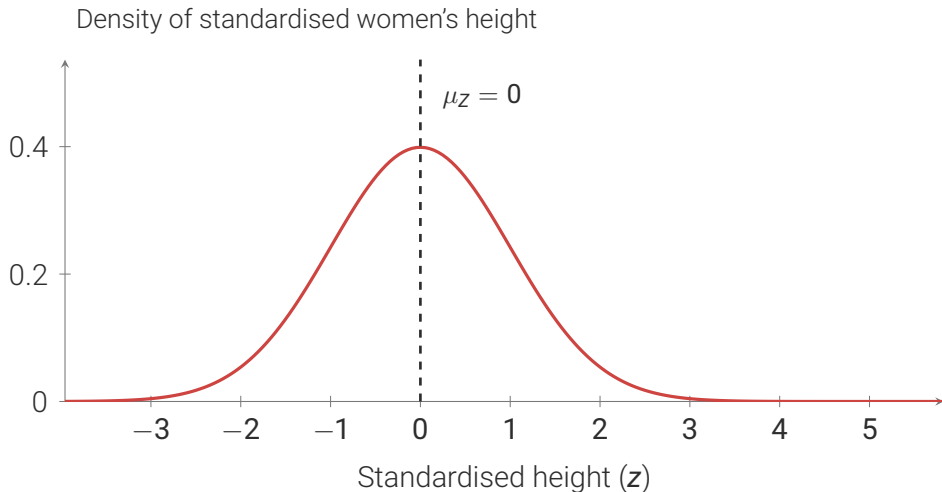
Standardising women's height

Density of women aged 20–49, US HNE Survey, 2009–2012



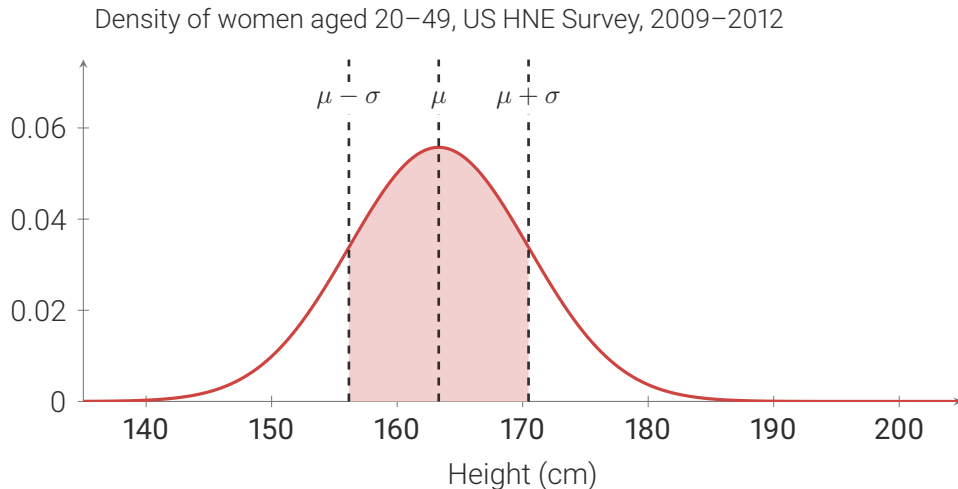
Fitted normal curve: $\mu \approx 163.3$ cm and $\sigma \approx 7.2$ cm.

Standardising women's height



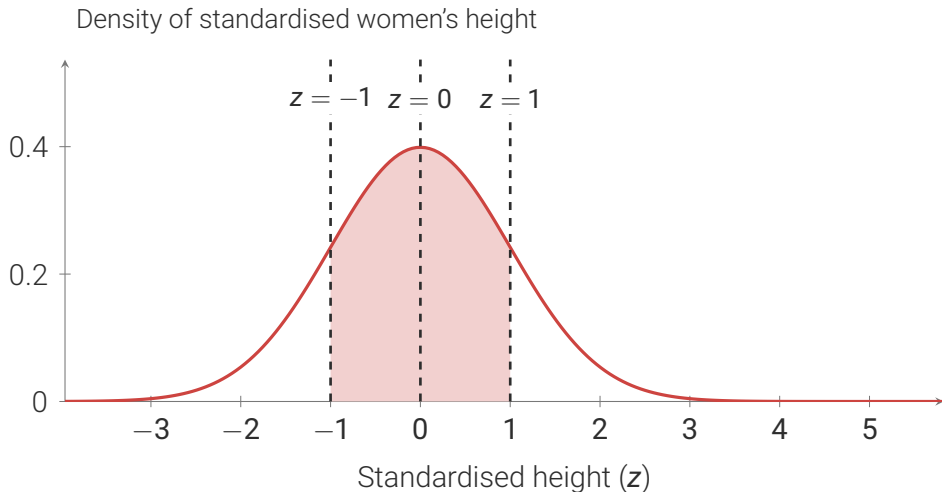
After $z = (h - \mu)/\sigma$, the fitted curve has mean 0 and standard deviation 1.

Standardising women's height



In the fitted normal model, about 68% of women are between 156.2 and 170.5 cm

Standardising women's height



$$\Pr(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) \approx 0.68.$$

From z-scores to probabilities

$\Phi(z) = \Pr(Z \leq z)$ gives the area to the **left** of z .

z	$\Phi(z)$
-2.00	0.0228
-1.96	0.0250
-1.00	0.1587
0.00	0.5000
1.00	0.8413
1.96	0.9750
2.00	0.9772

The central 68% area is

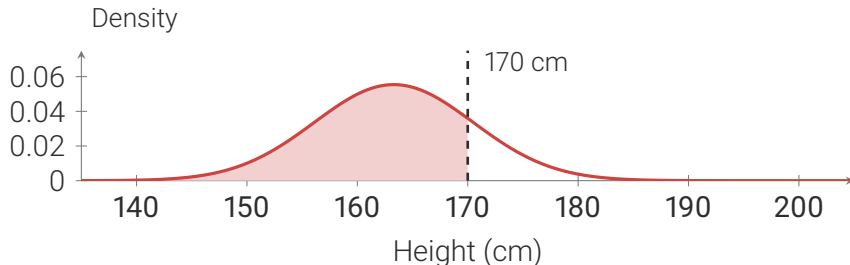
$$\Phi(1) - \Phi(-1) = 0.8413 - 0.1587 = 0.6826.$$

Exercise 1: height and probability

Assume height $H \sim N(163.3, 7.2^2)$.

[Open normal table](#)

What proportion of women are shorter than 170 cm?



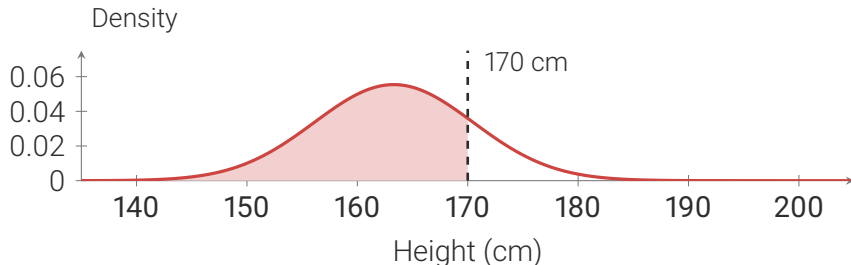
Calculate the z-score and use the normal table to find the shaded probability.

Exercise 1: height and probability

Assume height $H \sim N(163.3, 7.2^2)$.

[Open normal table](#)

What proportion of women are shorter than 170 cm?



$$z = \frac{170 - 163.3}{7.2} \approx 0.93,$$

$$\Pr(H < 170) \approx \Phi(0.93) \approx 0.8238 \quad (\text{row 0.9, column 0.03}).$$

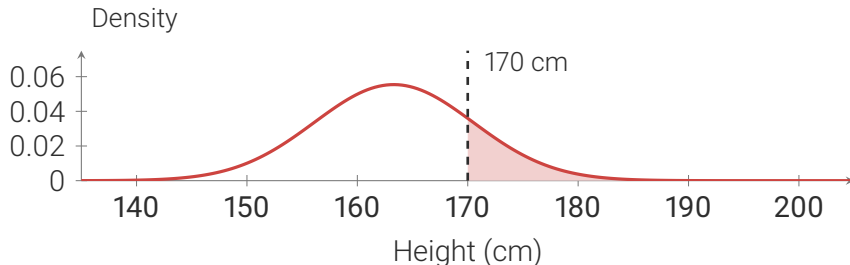
About **82.4%** of women are shorter than 170 cm.

Exercise 1: height and probability

Assume height $H \sim N(163.3, 7.2^2)$.

[Open normal table](#)

What proportion of women are taller than 170 cm?



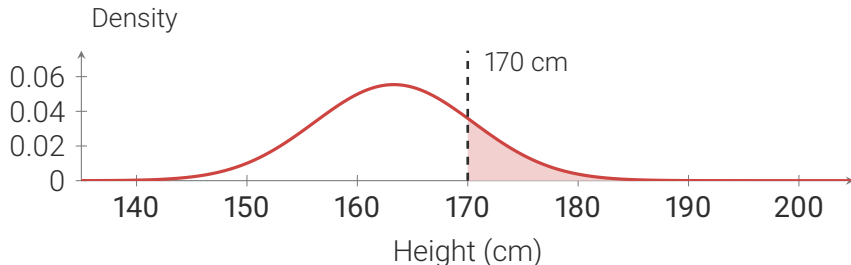
Calculate the z-score and use the normal table to find the shaded probability.

Exercise 1: height and probability

Assume height $H \sim N(163.3, 7.2^2)$.

[Open normal table](#)

What proportion of women are taller than 170 cm?



$$\begin{aligned}\Pr(H > 170) &= 1 - \Pr(H \leq 170), \\ &\approx 1 - \Phi(0.93) \approx 1 - 0.8238 = 0.1762.\end{aligned}$$

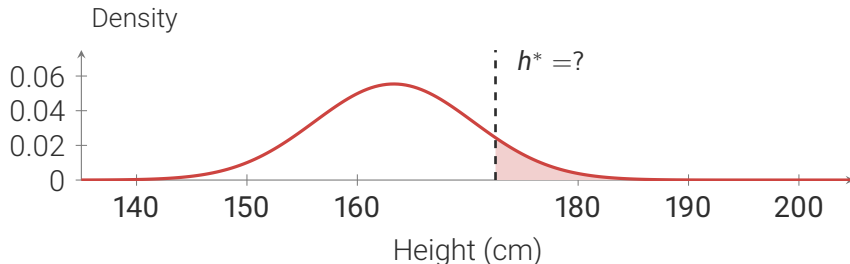
About 17.6% of women are taller than 170 cm.

Exercise 2: probability and height

Assume height $H \sim N(163.3, 7.2^2)$.

[Open normal table](#)

What height marks the **tallest 10%** of women?



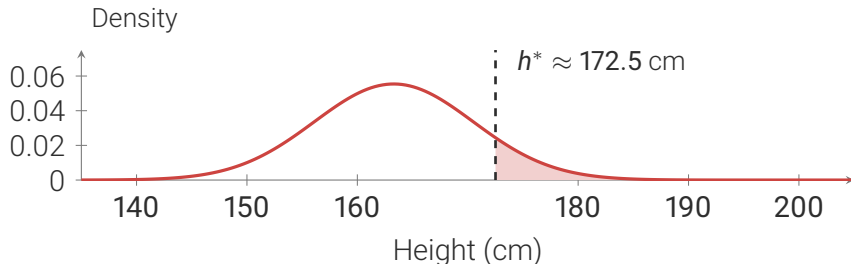
Find the z-score with 90% of the distribution to its left, then convert it to centimetres.

Exercise 2: probability and height

Assume height $H \sim N(163.3, 7.2^2)$.

[Open normal table](#)

What height marks the **tallest 10%** of women?



$$\Phi(z) = 1 - 0.10 = 0.90 \quad \Rightarrow \quad z \approx 1.28,$$

$$h^* = \mu + \sigma z = 163.3 + (7.2 \times 1.28) \approx 172.5 \text{ cm}.$$

The tallest 10% are approximately **172.5 cm or taller**.

The normality assumption

A few important reminders:

- Not all random variables are distributed normally.
- Standardising a variable does not make its distribution normal.
- There are many other distributions: chi-squared, Student t , F , Pareto, etc.

Random sampling and the distribution of the sample average

Populations and samples

We often want to know a **population parameter**, such as the mean height of all US women, but observe only a (not necessarily random) sample.

Parameter of interest	Population	Sample
Proportion intending to vote for Party A	All eligible voters	Voters who respond to a poll
Mean height	All US women	Women in the US HNE survey
Mean floor area	All houses in a city	Houses on Rightmove
Mean temperature	All of the UK	Weather-station locations

Can we use a sample to make inferences about the population?

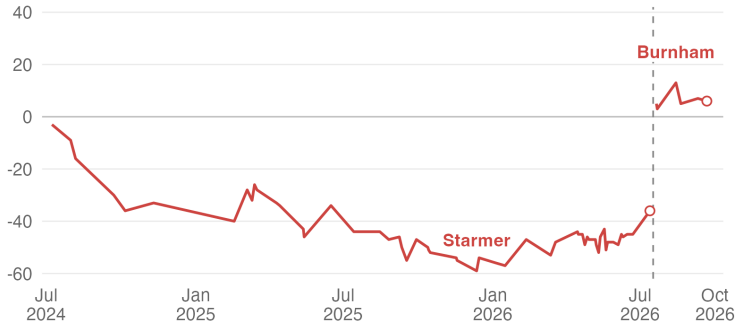
Random sampling

When using a sample to make inferences about the population, the big question is: **how was the sample selected?**

Random sampling

Labour leadership ratings

YouGov net favourability (percentage points)



Net favourability = favourable % minus unfavourable %.

Source: YouGov, collated on [Wikipedia](#).

Random sampling

When using a sample to make inferences about the population, the big question is: **how was the sample selected?**

Example: views of Andy Burnham

Suppose a poll asks voters whether they have a favourable or unfavourable opinion of Andy Burnham.

- Who was invited to take part?
- Who actually responded?
- Might the views of voters in the sample differ from those of other voters?

With **simple random sampling**, every voter has the same chance of being selected.

Modelling observations as i.i.d.

Each randomly selected person's response, A_i (1 if they approve of Andy Burnham, 0 otherwise), is a **random variable**.

If each A_i has the same marginal distribution, then $A_1, \dots, A_n$ are said to be **identically distributed**.

If voters are randomly selected, the A_i are also **independent**: knowing one person's response provides no information about another person's response.

Together, these properties mean that under random sampling $A_1, \dots, A_n$ are **independent and identically distributed**, or **i.i.d.**

The sampling distribution of the sample mean

The **sample mean** is the average of the observations in a sample:

$$\bar{H} = \frac{1}{n} \sum_{i=1}^n H_i.$$

We take three random samples of 10 women from the US HNE Survey:

	Sample 1	Sample 2	Sample 3
Sample mean $\bar{H}$	164.80 cm	162.71 cm	161.56 cm

- Different random samples of women usually give different average heights.
- Repeated samples of the same size produce a distribution of sample means: the **sampling distribution of $\bar{H}$** .

The mean of the sample mean

What is the average sample mean across different samples?

Suppose $H_1, \dots, H_n$ are i.i.d. with mean μ .

$$E(\bar{H}) = \frac{1}{n} \sum_{i=1}^n E(H_i) = \frac{n\mu}{n} = \mu.$$

The expected value of the sampling distribution is the **population mean**.

The variance of the sample mean

What is the variance of the sample mean across different samples?

Suppose $H_1, \dots, H_n$ are i.i.d. with variance σ^2 .

$$\begin{aligned}\text{Var}(\bar{H}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n H_i\right) \\&= \frac{1}{n^2} \left[\sum_{i=1}^n \text{Var}(H_i) + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \text{Cov}(H_i, H_j) \right] \\&= \frac{1}{n^2} [n\sigma^2 + 0] \quad (\text{independence}) \\&= \frac{\sigma^2}{n}.\end{aligned}$$

The variance of the sample mean

What is the variance of the sample mean across different samples?

Suppose $H_1, \dots, H_n$ are i.i.d. with variance σ^2 .

$$\text{Var}(\bar{H}) = \frac{\sigma^2}{n}.$$

The variance of the sample mean

What is the variance of the sample mean across different samples?

Suppose $H_1, \dots, H_n$ are i.i.d. with variance σ^2 .

$$\text{Var}(\bar{H}) = \frac{\sigma^2}{n}.$$

$$\text{SD}(\bar{H}) = \sqrt{\text{Var}(\bar{H})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

The variance of the sample mean

What is the variance of the sample mean across different samples?

Suppose $H_1, \dots, H_n$ are i.i.d. with variance σ^2 .

$$\text{Var}(\bar{H}) = \frac{\sigma^2}{n}.$$

$$\text{SD}(\bar{H}) = \sqrt{\text{Var}(\bar{H})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

The variance of the sample mean is equal to the **population variance divided by the size of the sample**.

The larger the sample, the lower the variance in the sample mean across samples.

Sampling from a normal population

The sum of **independent normal random variables** is itself a normal random variable.

If $H_1, \dots, H_n$ are i.i.d. with $H_i \sim N(\mu, \sigma^2)$, the sample mean is a scaled sum of these normal variables:

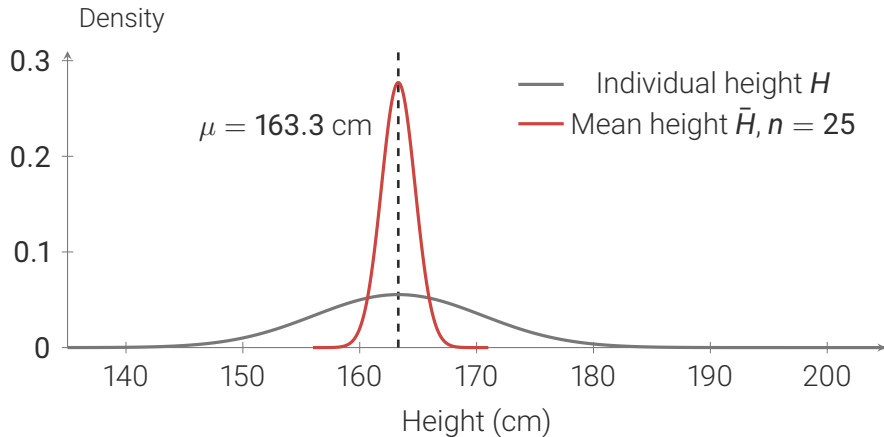
$$\bar{H} = \frac{1}{n} \sum_{i=1}^n H_i.$$

Dividing a normal random variable by a constant also gives a normal random variable.

Therefore, the sample mean is normally distributed with mean μ and variance σ^2/n :

$$\bar{H} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Individual observations and sample means



Both distributions have mean 163.3 cm, but the variance of sample means is a lot lower.

Large-sample approximations to sampling distributions

Law of large numbers

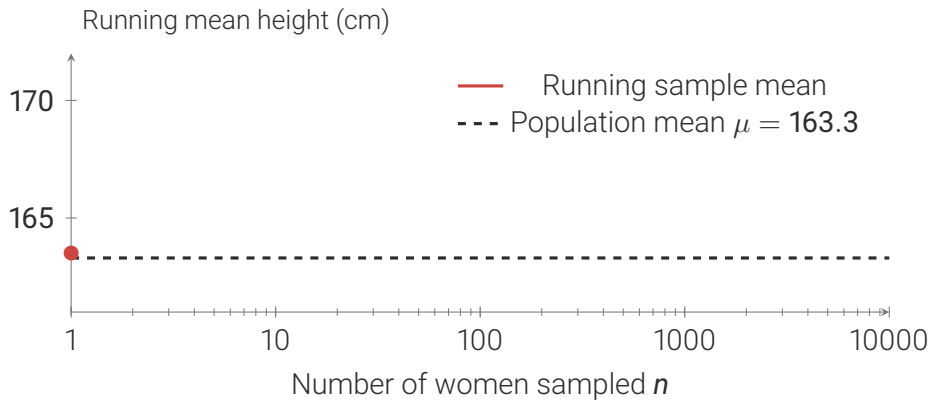
The **Law of Large Numbers** says that the sample mean, $\bar{H}$, will be very close to the population mean, μ_H , with very high probability if the sample size is large.

If $H_1, \dots, H_n$ are i.i.d. with mean μ_H and finite variance, then the sample mean approaches μ_H as the sample size grows:

$$\bar{H} \xrightarrow{p} \mu_H \quad \text{as } n \longrightarrow \infty.$$

Importantly, increasing the sample size does **not** correct bias caused by non-random sample selection.

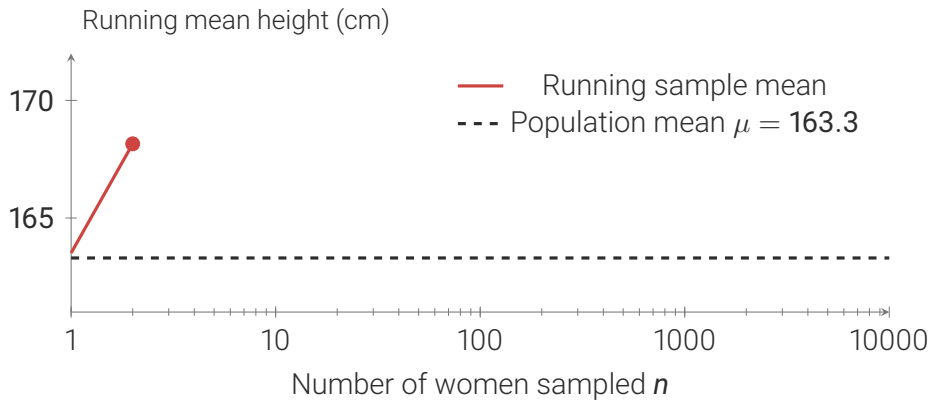
The law of large numbers in action



$$H_1 = 163.51 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

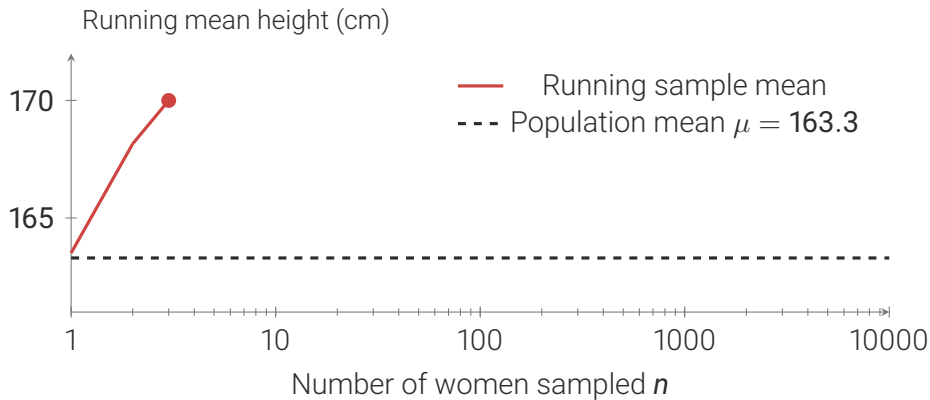
The law of large numbers in action



$$H_2 = 172.80 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

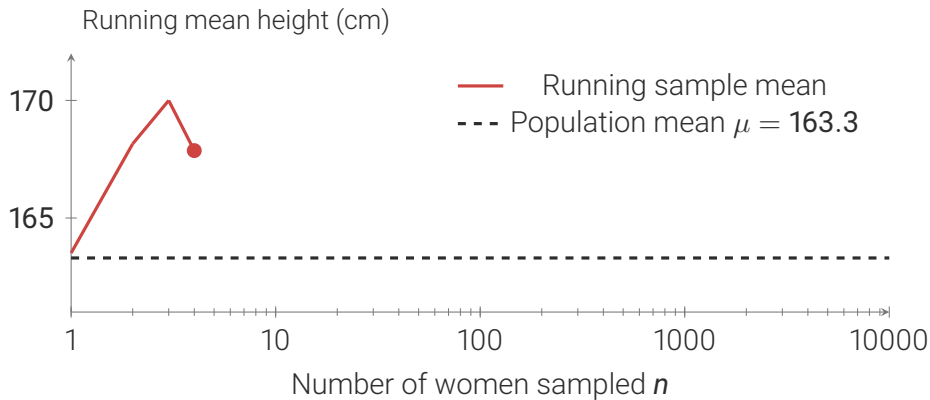
The law of large numbers in action



$$H_3 = 173.68 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

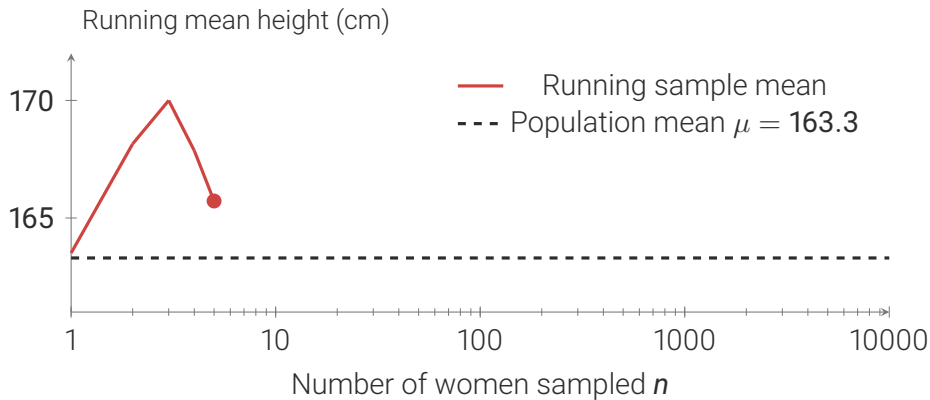
The law of large numbers in action



$$H_4 = 161.47 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

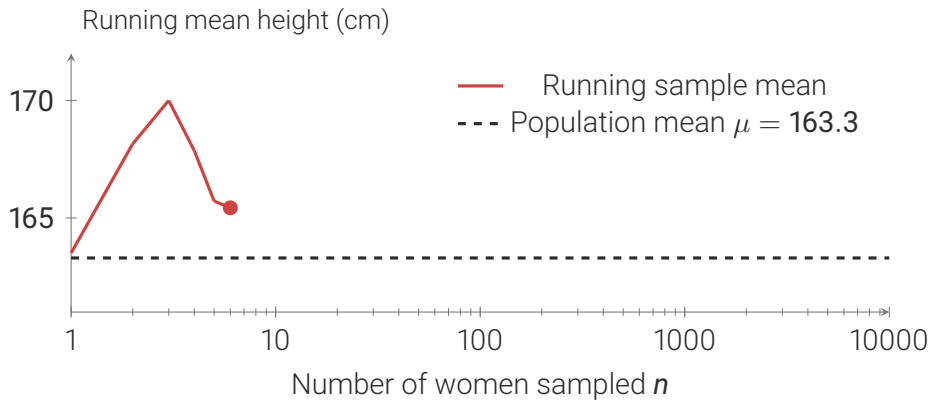
The law of large numbers in action



$$H_5 = 157.15 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

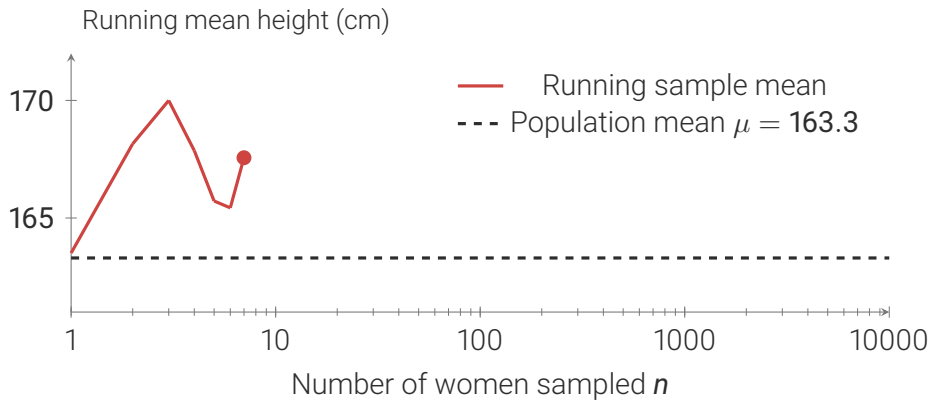
The law of large numbers in action



$$H_6 = 163.98 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

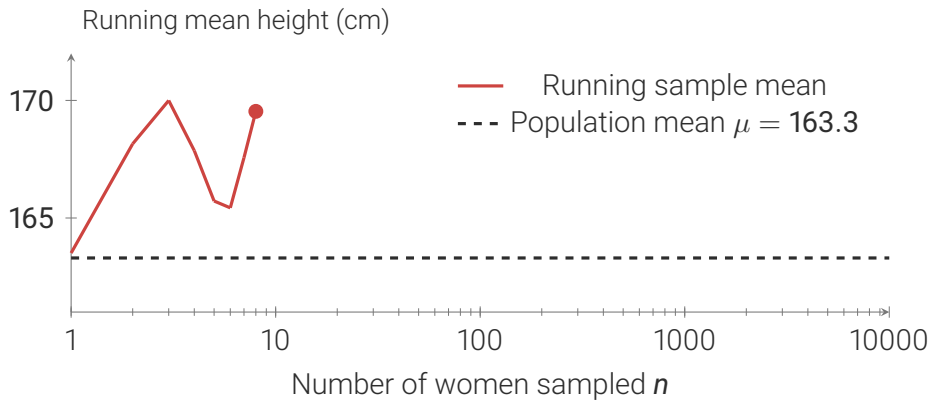
The law of large numbers in action



$$H_7 = 180.35 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

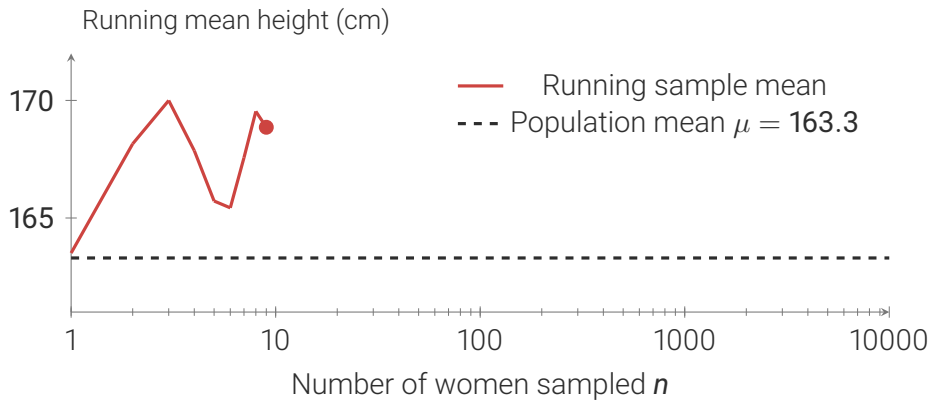
The law of large numbers in action



$$H_8 = 183.34 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

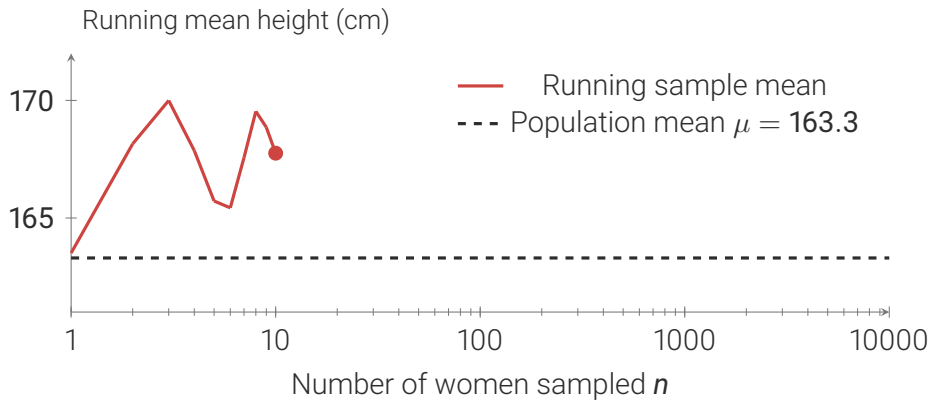
The law of large numbers in action



$$H_9 = 163.45 \text{ cm.}$$

Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

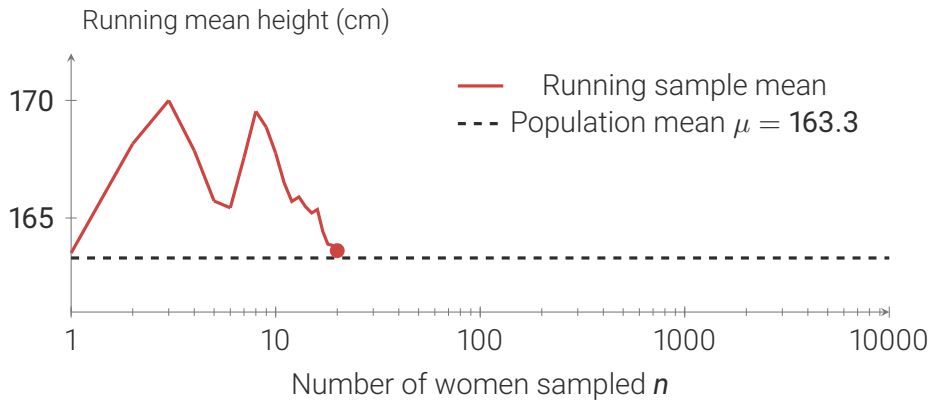
The law of large numbers in action



$$H_{10} = 157.84 \text{ cm.}$$

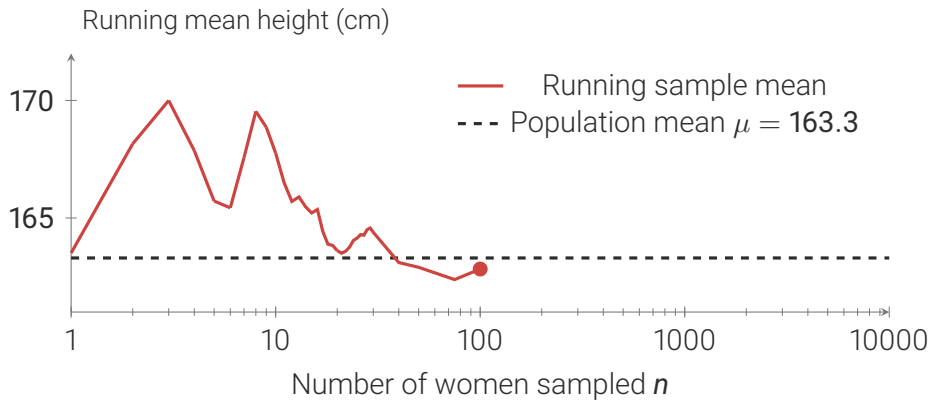
Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

The law of large numbers in action



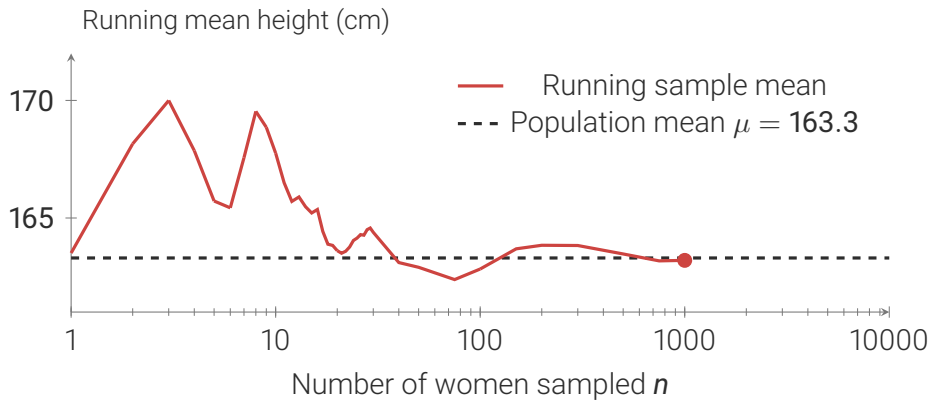
Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

The law of large numbers in action



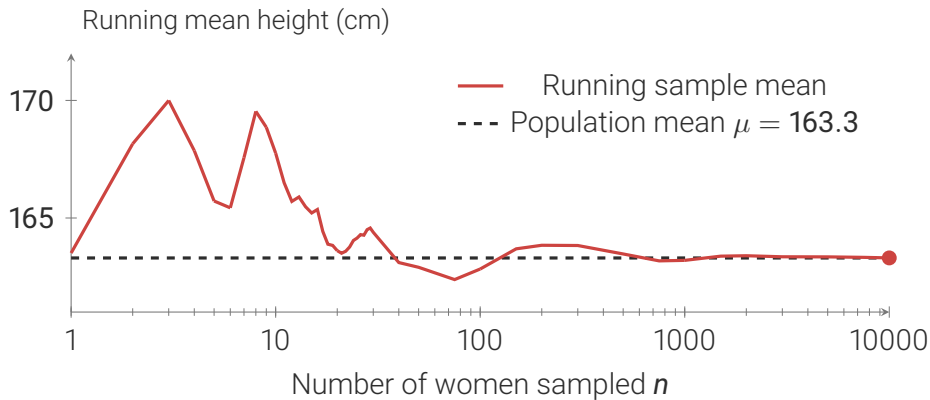
Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

The law of large numbers in action



Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

The law of large numbers in action



Simulated i.i.d. heights from $N(163.3, 7.2^2)$.

The central limit theorem

The **Central Limit Theorem** states that the distribution of the sample mean, $\bar{H}$, is well approximated by a normal distribution when n is large.

This implies that the standardised sample mean has a standard normal distribution:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1).$$

This theorem holds even when $H_1, \dots, H_n$ are **not** normally distributed!

Exercise 3: probability for an average height

We sample $n = 100$ women and assume their heights are i.i.d., with $H_i \sim N(163.3, 7.2^2)$.

What is the probability that the sample average of height exceeds 164 cm?

[Open normal table](#)

Exercise 3: probability for an average height

We sample $n = 100$ women and assume their heights are i.i.d., with $H_i \sim N(163.3, 7.2^2)$.

What is the probability that the sample average of height exceeds 164 cm?

$$SD(\bar{H}) = \frac{7.2}{\sqrt{100}} = 0.72, \quad z = \frac{164 - 163.3}{0.72} \approx 0.97.$$

$$\Pr(\bar{H} > 164) \approx 1 - \Phi(0.97) = 1 - 0.8340 = 0.1660.$$