

# Confidence intervals

## *Lecture 3*

Quantitative Methods for Political Economy  
7SSPN107

# Recap: normal and standard normal distributions

A **normal distribution** is a probability distribution with a symmetric, bell-shaped density:

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

The mean  $\mu$  determines the centre of the distribution and the standard deviation  $\sigma$  determines its spread.

# Recap: normal and standard normal distributions

A **normal distribution** is a probability distribution with a symmetric, bell-shaped density:

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

The mean  $\mu$  determines the centre of the distribution and the standard deviation  $\sigma$  determines its spread.

**Standardisation** replaces values with z-scores, which measure distance from the mean in standard deviations:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

The **standard normal distribution** has mean 0 and variance 1.

# Recap: independent and identically distributed

Let  $H_1, \dots, H_n$  be the heights of women in a random sample.

$H_1, \dots, H_n$  are **independent** if knowing some sampled heights provides no information about the others.

$H_1, \dots, H_n$  are **identically distributed** if, before sampling, each  $H_i$  has the same distribution of possible heights, with mean  $\mu_H$  and variance  $\sigma_H^2$ .

Together, these properties are called **i.i.d.**

# Recap: the distribution of the sample mean

The sample mean is the arithmetic average of the observations:

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

If  $H_1, \dots, H_n$  are i.i.d. with mean  $\mu_H$  and finite variance  $\sigma_H^2$ , then

$$E(\bar{H}) = \mu_H, \quad \text{Var}(\bar{H}) = \frac{\sigma_H^2}{n}, \quad \text{SD}(\bar{H}) = \frac{\sigma_H}{\sqrt{n}}.$$

## Recap: the distribution of the sample mean

The sample mean is the arithmetic average of the observations:

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

If  $H_1, \dots, H_n$  are i.i.d. with mean  $\mu_H$  and finite variance  $\sigma_H^2$ , then

$$E(\bar{H}) = \mu_H, \quad \text{Var}(\bar{H}) = \frac{\sigma_H^2}{n}, \quad \text{SD}(\bar{H}) = \frac{\sigma_H}{\sqrt{n}}.$$

If individual heights are normally distributed, then so is the sample mean:

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

# Recap: the LLN and the CLT

Suppose  $H_1, \dots, H_n$  are i.i.d. with finite, positive variance.

**Law of large numbers:** the sample mean,  $\bar{H}$ , will be close to the population mean,  $\mu_H$ , with very high probability if  $n$  is large.

# Recap: the LLN and the CLT

Suppose  $H_1, \dots, H_n$  are i.i.d. with finite, positive variance.

**Law of large numbers:** the sample mean,  $\bar{H}$ , will be close to the population mean,  $\mu_H$ , with very high probability if  $n$  is large.

**Central limit theorem:** the distribution of the sample mean,  $\bar{H}$ , is well approximated by a normal distribution when  $n$  is large.

# Uncertainty in estimating the population mean

We can draw samples from a population and use the sample mean to estimate the population mean.

Since we rely on samples rather than on the full population, there is uncertainty as to how accurate our estimate is.

Today, we will quantify this uncertainty using **confidence intervals**.

## 3.1

Estimation of the population mean

# Estimators and estimates

An **estimator** is a function of sample data to estimate a population parameter, e.g., the sample mean or the first observation:

$$\bar{H} = \frac{H_1 + \cdots + H_n}{n}, \quad H_1$$

The value of the function will differ across samples, which is to say that estimators have sampling distributions.

# Estimators and estimates

An **estimator** is a function of sample data to estimate a population parameter, e.g., the sample mean or the first observation:

$$\bar{H} = \frac{H_1 + \cdots + H_n}{n}, \quad H_1$$

The value of the function will differ across samples, which is to say that estimators have sampling distributions.

An **estimate** is the numerical value of that function for a particular sample.

Hypothetical sample:  $\bar{H} = 164.0$  cm.

# Estimators and estimates

An **estimator** is a function of sample data to estimate a population parameter, e.g., the sample mean or the first observation:

$$\bar{H} = \frac{H_1 + \cdots + H_n}{n}, \quad H_1$$

The value of the function will differ across samples, which is to say that estimators have sampling distributions.

An **estimate** is the numerical value of that function for a particular sample.

Hypothetical sample:  $\bar{H} = 164.0$  cm.

How do we choose an estimator?

# Estimators and estimates

An **estimator** is a function of sample data to estimate a population parameter, e.g., the sample mean or the first observation:

$$\bar{H} = \frac{H_1 + \cdots + H_n}{n}, \quad H_1$$

The value of the function will differ across samples, which is to say that estimators have sampling distributions.

An **estimate** is the numerical value of that function for a particular sample.

Hypothetical sample:  $\bar{H} = 164.0$  cm.

How do we choose an estimator? **Unbiasedness**, **consistency**, and **efficiency**.

# Unbiasedness

An estimator of a population parameter,  $\hat{\mu}_H$ , is **unbiased** if its expected value is equal to that population parameter,  $\mu_H$ :

$$E(\hat{\mu}_H) = \mu_H.$$

# Unbiasedness

An estimator of a population parameter,  $\hat{\mu}_H$ , is **unbiased** if its expected value is equal to that population parameter,  $\mu_H$ :

$$E(\hat{\mu}_H) = \mu_H.$$

Both  $\bar{H}$  and  $H_1$  are unbiased under i.i.d. sampling.

$$E(\bar{H}) = \mu_H \quad \text{and} \quad E(H_1) = \mu_H.$$

# Unbiasedness

An estimator of a population parameter,  $\hat{\mu}_H$ , is **unbiased** if its expected value is equal to that population parameter,  $\mu_H$ :

$$E(\hat{\mu}_H) = \mu_H.$$

Both  $\bar{H}$  and  $H_1$  are unbiased under i.i.d. sampling.

$$E(\bar{H}) = \mu_H \quad \text{and} \quad E(H_1) = \mu_H.$$

Can you think of other unbiased estimators of the population mean?

# Consistency

An estimator of a population parameter is **consistent** if the probability that it is close to the population parameter goes to **1** when the sample size,  **$n$** , grows.

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

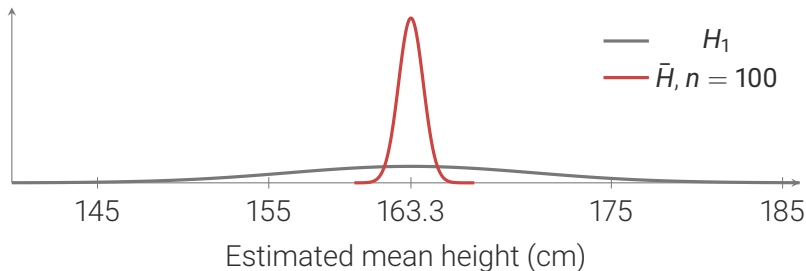
The sample mean is a consistent estimator of the population mean by the Law of Large Numbers.

Because the first observation of a sample,  **$H_1$** , does not change when the sample is made bigger, it does not get closer to the population mean as  **$n$**  grows: it is not consistent.

# Efficiency

Among unbiased estimators, the estimator with the lowest variance is **efficient**.

$$\text{Var}(\bar{H}) = \frac{\sigma_H^2}{n} \quad \text{Var}(H_1) = \sigma_H^2.$$



# The sample mean is BLUE

Under i.i.d. sampling with finite variance, the sample mean,  $\bar{H}$ , has the lowest variance of any linear, unbiased estimator: it is the **Best Linear Unbiased Estimator** of  $\mu_H$ .

# The sample mean as a least-squares estimate

Suppose we want to find an estimator,  $m$ , that minimises the squared difference with individual observations,  $H_i$ , the **least squares estimator**:

$$\sum_{i=1}^n (H_i - m)^2.$$

The estimator that minimises the sum of squared differences is the sample mean,  $m = \bar{H}$ .

# Four measures of spread

---

$\sigma_H$

**Population standard deviation**

Spread of individual heights in the population.

$s_H$

**Sample standard deviation**

Spread of observed heights.

---

$$SD(\bar{H}) = \frac{\sigma_H}{\sqrt{n}}$$

**Standard deviation of the sample mean**

Spread of sample means across repeated samples.

$$SE(\bar{H}) = \frac{s_H}{\sqrt{n}}$$

**Standard error of the sample mean**

Estimates the standard deviation of the sample mean.

---

# Estimating the spread of individual heights

The **sample variance**,  $s_H^2$ , is:

$$s_H^2 = \frac{1}{n-1} \sum_{i=1}^n (H_i - \bar{H})^2.$$

# Estimating the spread of individual heights

Why divide by  $n - 1$ ? [Stock and Watson, Exercise 3.18](#)

$$\begin{aligned}\text{Var}(H_i - \bar{H}) &= E\left[(H_i - \bar{H} - E(H_i - \bar{H}))^2\right] \\ &= E[(H_i - \bar{H})^2].\end{aligned}$$

Then, under i.i.d. sampling with finite variance,

$$\begin{aligned}E[(H_i - \bar{H})^2] &= \text{Var}(H_i - \bar{H}) \\ &= \text{Var}(H_i) + \text{Var}(\bar{H}) - 2 \text{Cov}(H_i, \bar{H}) \\ &= \sigma_H^2 + \frac{\sigma_H^2}{n} - 2 \frac{\sigma_H^2}{n} = \frac{n-1}{n} \sigma_H^2.\end{aligned}$$

Therefore, for  $n > 1$ ,

$$E(s_H^2) = \frac{n}{n-1} E[(H_i - \bar{H})^2] = \sigma_H^2.$$

# Estimating the spread of sample means

The true standard deviation of the sample mean is

$$\text{SD}(\bar{H}) = \frac{\sigma_H}{\sqrt{n}}.$$

Since  $\sigma_H$  is unknown, we replace it by  $s_H$  to obtain its estimate, the standard error of the sample mean:

$$\text{SE}(\bar{H}) = \frac{s_H}{\sqrt{n}}.$$

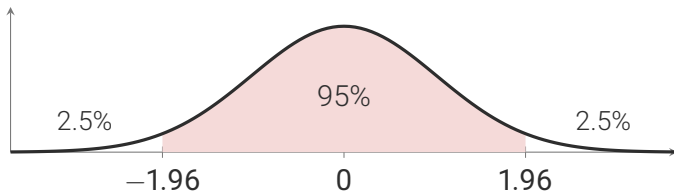
## 3.3

Confidence intervals for the population mean

# The normal approximation

For large  $n$ , the Central Limit Theorem and the standard error  $SE(\bar{H})$  give

$$\frac{\bar{H} - \mu_H}{SE(\bar{H})} \approx N(0, 1).$$



About 95% of sample means lie within 1.96 standard errors of  $\mu_H$ .

# A confidence interval

We can construct an interval such that it contains the true population mean  $\mu_H$   $p\%$  of the time, the coverage probability: the confidence interval

For example, the 95% confidence interval can be found as follows:

$$-1.96 \leq \frac{\bar{H} - \mu_H}{SE(\bar{H})} \leq 1.96.$$

# A confidence interval

We can construct an interval such that it contains the true population mean  $\mu_H$   $p\%$  of the time, **the coverage probability**: the **confidence interval**

For example, the 95% confidence interval can be found as follows:

$$-1.96 \leq \frac{\bar{H} - \mu_H}{SE(\bar{H})} \leq 1.96.$$

Rearranging this statement puts  $\mu_H$  between two limits:

$$\bar{H} - 1.96 SE(\bar{H}) \leq \mu_H \leq \bar{H} + 1.96 SE(\bar{H}).$$

Our **95% confidence interval** is

$$\bar{H} \pm 1.96 SE(\bar{H}).$$

where  $1.96 SE(\bar{H})$  is the 95% margin of error.

## Exercise 1: a confidence interval for mean height

A hypothetical random sample of  $n = 100$  women has

$$\bar{H} = 164.0 \text{ cm}, \quad s_H = 7.2 \text{ cm}.$$

Construct the 95% confidence interval for the population mean height,  $\mu_H$ .

## Exercise 1: a confidence interval for mean height

A hypothetical random sample of  $n = 100$  women has

$$\bar{H} = 164.0 \text{ cm}, \quad s_H = 7.2 \text{ cm}.$$

Construct the 95% confidence interval for the population mean height,  $\mu_H$ .

$$\text{SE}(\bar{H}) = \frac{7.2}{\sqrt{100}} = 0.72 \text{ cm},$$

$$\text{margin of error} = 1.96 \times 0.72 = 1.4112 \text{ cm},$$

$$95\% \text{ confidence interval} = [162.59, 165.41] \text{ cm}.$$

## Exercise 2: a confidence interval for mean height

A hypothetical random sample of  $n = 100$  women has

$$\bar{H} = 164.0 \text{ cm}, \quad s_H = 7.2 \text{ cm}.$$

Construct the 90% confidence interval for the population mean height,  $\mu_H$ .

## Exercise 2: a confidence interval for mean height

A hypothetical random sample of  $n = 100$  women has

$$\bar{H} = 164.0 \text{ cm}, \quad s_H = 7.2 \text{ cm}.$$

Construct the 90% confidence interval for the population mean height,  $\mu_H$ .

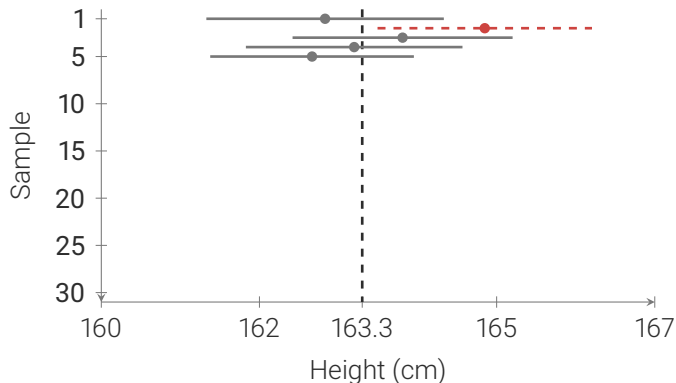
$$SE(\bar{H}) = \frac{7.2}{\sqrt{100}} = 0.72 \text{ cm},$$

$$\text{margin of error} = 1.645 \times 0.72 = 1.1844 \text{ cm},$$

$$90\% \text{ confidence interval} = [162.82, 165.18] \text{ cm}.$$

# Repeated samples, changing intervals

Simulated i.i.d. heights from  $N(163.3, 7.2^2)$ , with  $n = 100$  per sample.

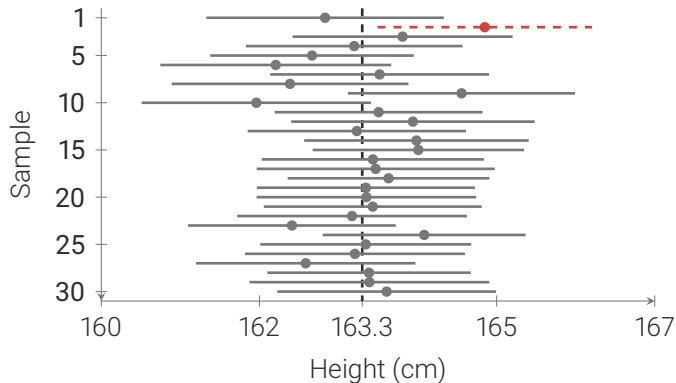


The **population mean** is indicated by the black dashed line.

Each dot represents the mean of a new sample,  $\bar{H}$ . Each line is its 95% confidence interval.

# Repeated samples, changing intervals

Simulated i.i.d. heights from  $N(163.3, 7.2^2)$ , with  $n = 100$  per sample.



The **population mean** is indicated by the black dashed line.

Each dot represents the mean of a new sample,  $\bar{H}$ . Each line is its 95% confidence interval.

In this run, 29 of 30 intervals contain  $\mu_H$ .

## Exercise 3: a confidence interval

An **i.i.d. sample** of 144 women has  $\bar{H} = 162.8$  cm and  $s_H = 7.2$  cm.

1. Calculate the standard error and the 95% margin of error.
2. Find the 95% confidence interval for  $\mu_H$ .
3. If  $s_H$  doubles to **14.4** cm, what happens to the 95% margin of error?

## Exercise 3: a confidence interval

An **i.i.d. sample** of 144 women has  $\bar{H} = 162.8$  cm and  $s_H = 7.2$  cm.

1. Calculate the standard error and the 95% margin of error.
2. Find the 95% confidence interval for  $\mu_H$ .
3. If  $s_H$  doubles to **14.4** cm, what happens to the 95% margin of error?

$$SE(\bar{H}) = \frac{7.2}{12} = 0.60, \quad \text{margin} = 1.96(0.60) = 1.176 \text{ cm.}$$

## Exercise 3: a confidence interval

An **i.i.d. sample** of 144 women has  $\bar{H} = 162.8$  cm and  $s_H = 7.2$  cm.

1. Calculate the standard error and the 95% margin of error.
2. Find the 95% confidence interval for  $\mu_H$ .
3. If  $s_H$  doubles to **14.4** cm, what happens to the 95% margin of error?

$$SE(\bar{H}) = \frac{7.2}{12} = 0.60, \quad \text{margin} = 1.96(0.60) = 1.176 \text{ cm.}$$

$$162.8 \pm 1.176 \approx [161.62, 163.98] \text{ cm.}$$

## Exercise 3: a confidence interval

An **i.i.d. sample** of 144 women has  $\bar{H} = 162.8$  cm and  $s_H = 7.2$  cm.

1. Calculate the standard error and the 95% margin of error.
2. Find the 95% confidence interval for  $\mu_H$ .
3. If  $s_H$  doubles to **14.4** cm, what happens to the 95% margin of error?

$$1.96 \frac{14.4}{\sqrt{144}} = 2.352 \text{ cm.}$$

## Exercise 4: interpretation and width

- A.** A student says: “Our interval contains 95% of women’s heights.” How would you correct the statement?
- B.** A new sample has  $n = 576$  instead of  $n = 144$  and the same  $s_H = 7.2$  cm. For a 95% confidence level, what happens to the margin of error?

## Exercise 4: interpretation and width

**A.** A student says: “Our interval contains 95% of women’s heights.” How would you correct the statement?

**B.** A new sample has  $n = 576$  instead of  $n = 144$  and the same  $s_H = 7.2$  cm. For a 95% confidence level, what happens to the margin of error?

**A.** Across samples, around 95% of intervals made by this procedure contain the population mean.

## Exercise 4: interpretation and width

**A.** A student says: “Our interval contains 95% of women’s heights.” How would you correct the statement?

**B.** A new sample has  $n = 576$  instead of  $n = 144$  and the same  $s_H = 7.2$  cm. For a 95% confidence level, what happens to the margin of error?

**A.** Across samples, around 95% of intervals made by this procedure contain the population mean.

**B.** The sample size quadruples, so the margin halves:

$$1.96 \frac{7.2}{\sqrt{144}} = 1.96 \frac{7.2}{12} = 1.176 \text{ cm.}$$
$$1.96 \frac{7.2}{\sqrt{576}} = 1.96 \frac{7.2}{24} = 0.588 \text{ cm.}$$