Statistical Sampling

What is Probability?

In layman terms - degree of belief that an event will happen.

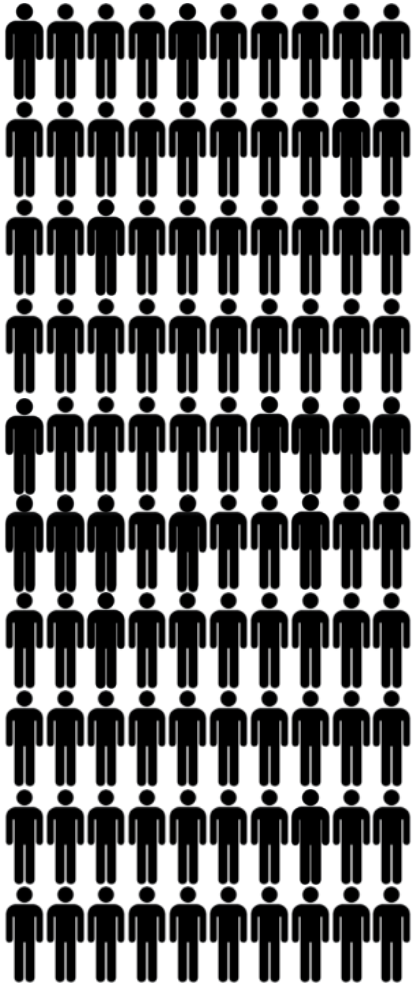
In statistical terms - probability (denoted as **P** or **p**) is the long term frequency of an event occurring.

Probability can take any value between **0** and **1**

0 (the event will definitely not occur) - **1** (the event will definitely occur).

Sampling	▪ Describe the concept of a source population ▪ Explain random sampling ▪ Explain estimation of population statistics ▪ Describe standard error of a sample mean and of a proportion, and their differences ▪ Define and use confidence intervals ▪ Explain reference ranges

Population



Random Sample

Select random sample

*Use sample to make an **inference** about the population*

Distinction between population and sample

In any experiment or survey:-

the ***target population*** is the group of individuals (or objects) ***to which the results will be generalised***

the ***sample*** is a subset of the population ***on which the data are collected.***

A sample of 6 week old Burmese kittens

What is the target population?

Why is sampling used?

Target population is often ***large***

impractical to study the whole population without substantial resources (e.g. time, money, research staff, computing equipment).

often ***not necessary***. All the information required ***in practice*** can be obtained from a carefully chosen sample.

the smaller volume of data collected from a sample allows more attention to be given to the ***validity***, ***reliability*** and ***completeness*** of the data.

Simple random sample:

every possible sample from the target population has an ***equal probability of being chosen*** *(this implies that every individual has an equal probability of being chosen)*

requires a ***sampling frame*** *(think of a census or phone book)*

sample of the required size can be chosen using ***random number tables***.

Various modifications exist:

stratified sampling multi-stage sampling
cluster sampling systematic sampling

At the core of sampling is ***randomness.***

Bias

Bias is a type of error that skews results in a certain direction.

Random selection helps reduce **selection bias** by ensuring each individual has an equal chance of being selected.

It is associated with research where the selection of participants isn't random.

Example

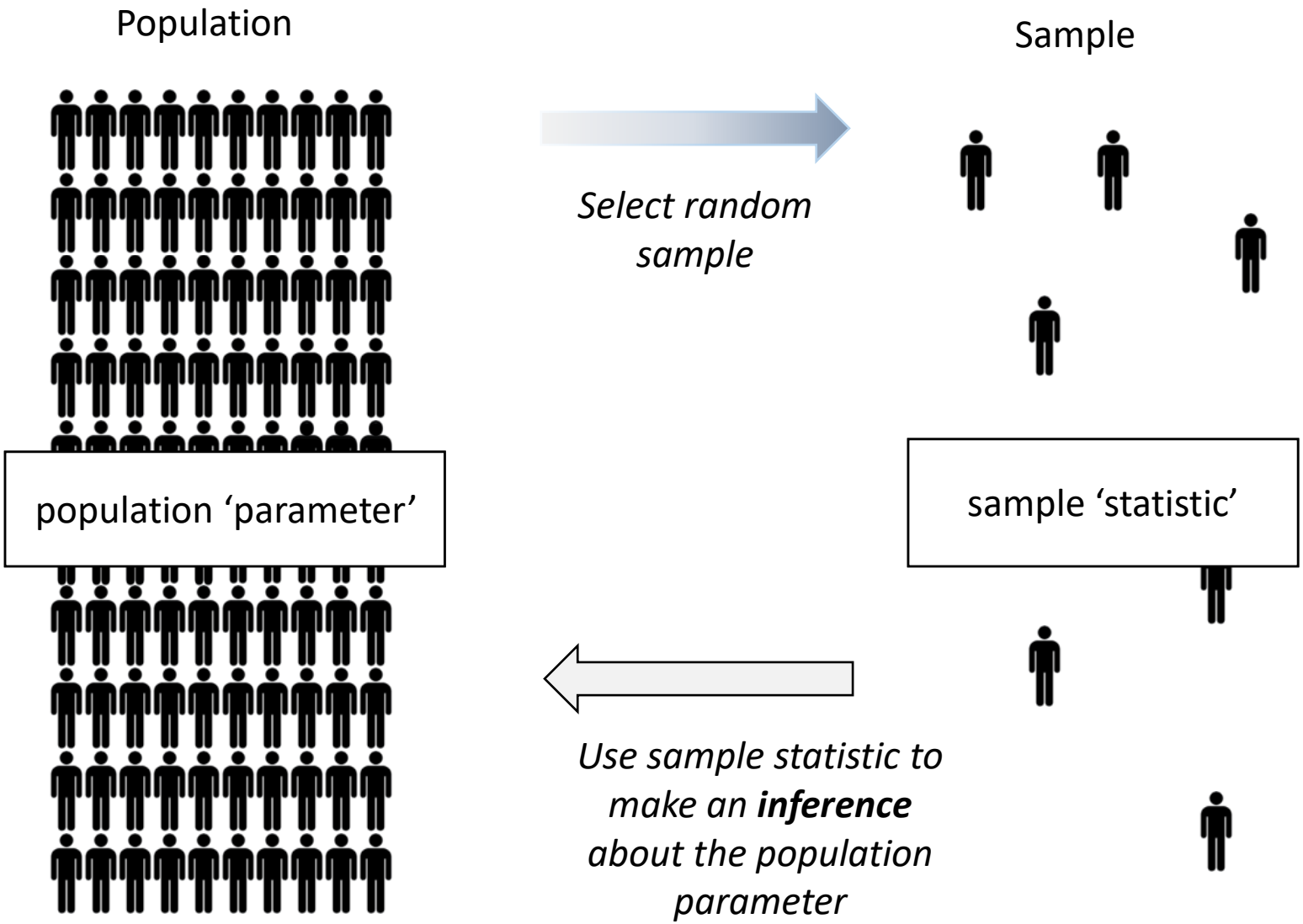
Aim: to *estimate the average value* for triceps skinfold thickness of 10 year old boys in Glasgow (for eventual comparison with other areas).

Target population: *all* 10 year old boys in Glasgow

Sample: a ***random sample*** of 40 ten year old boys are selected for study

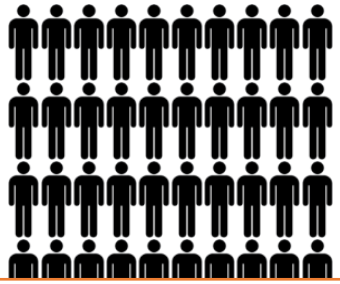
The ***sample mean*** is *8.71 mm* (s.d. *2.77* mm)

What can we infer about the ***population mean*** based on these results?

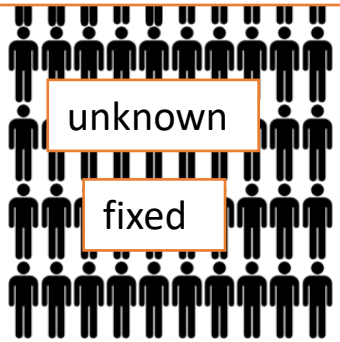

Population
Sample
Select random sample
population 'parameter'
sample 'statistic'
Use sample statistic to make an **inference** about the population parameter

Population Parameters and Sample Statistics

Population Parameter



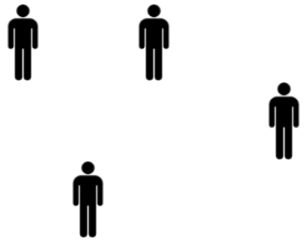
μ = mean triceps skinfold of *population* of 10-year old boys in Glasgow



Select random sample

Use sample statistic to estimate the population parameter

Sample statistic



$\bar{x}$ = mean triceps skinfold of *random sample* of 40 10-year old boys in Glasgow

known - 8.71mm calculated from data collected

not fixed – will differ between samples

Illustration of the Relationship between Population Parameters and Sample Statistics

Routine physical examination of *all* 5 year old school entrants in Glasgow (complete census)

mean height is 110 cm (SD 6 cm)

These are ***population parameters.***

Data were entered into a computer and a simple random sample of 30 children chosen:

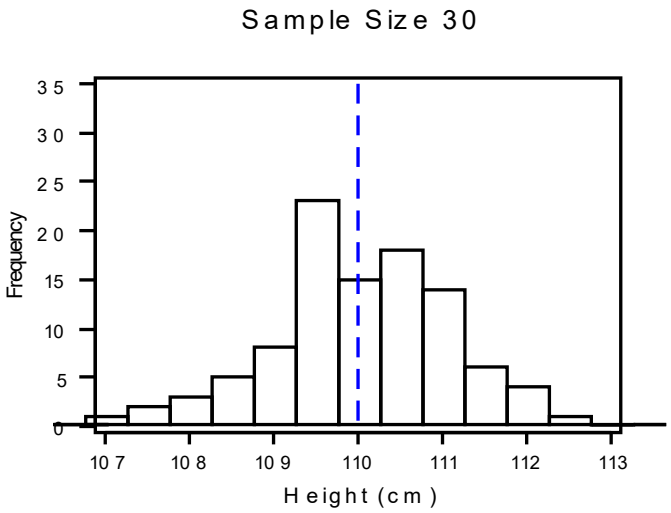
108 114 101 114 108 99 118 111 115 106 108
101 110 118 108 98 103 110 107 122 112 96
116 105 115 108 105 114 106

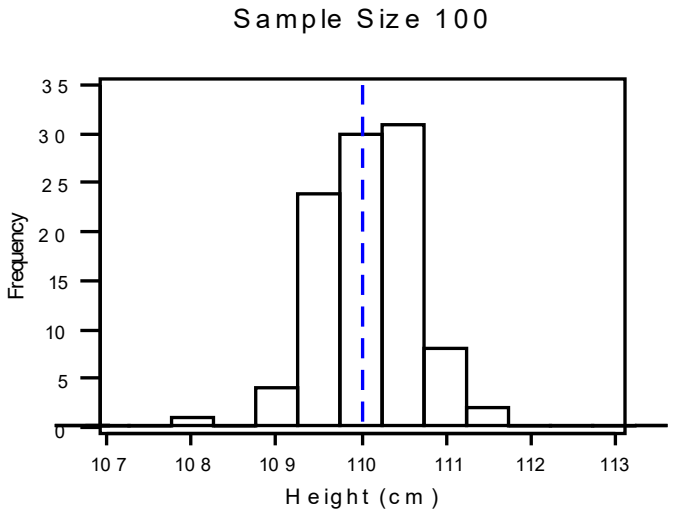
mean = 108.6cm sd = 6.4cm

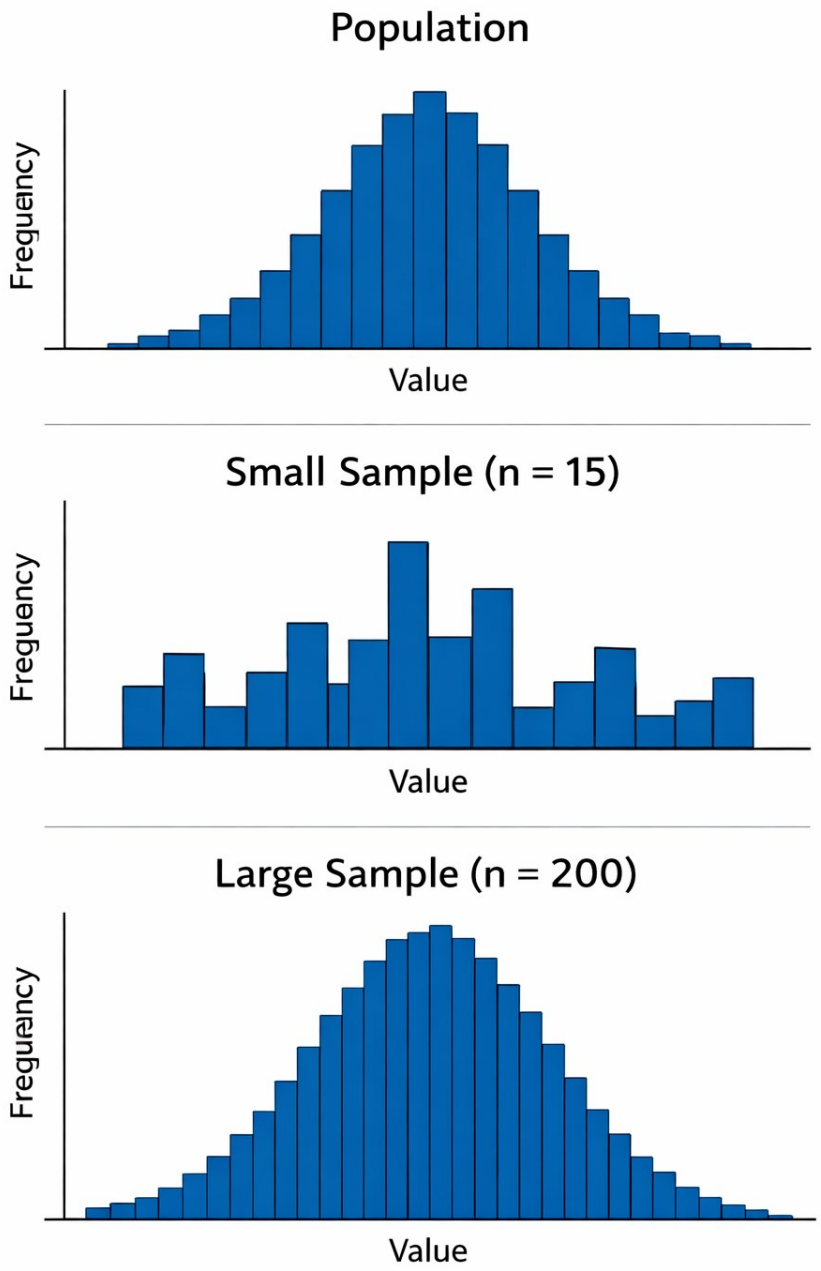
Sample statistics for the first 10 samples were

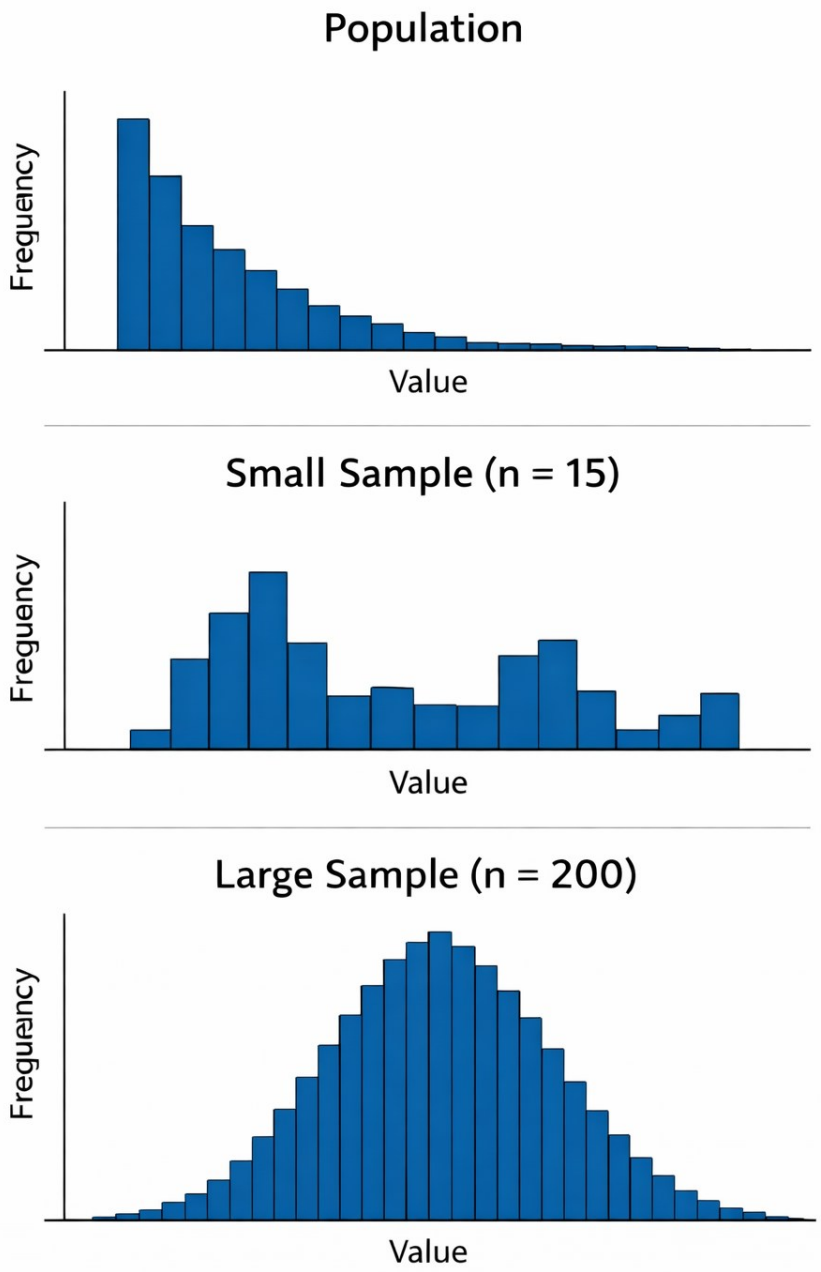
Sample number	Sample size	mean height	standard deviation
1	30	108.6	6.4
2	30	109.1	4.6
3	30	108.5	5.9
4	30	109.4	6.1
5	30	110.5	4.8
6	30	110.0	4.6
7	30	109.7	7.6
8	30	110.1	6.9
9	30	111.4	7.0
10	30	108.2	6.6

Sampling distribution of Means of 100 Random Samples (population mean 110cm, sd 6)






Population
Frequency
Value
Small Sample (n = 15)
Frequency
Value
Large Sample (n = 200)
Frequency
Value


Population
Frequency
Value
Small Sample (n = 15)
Frequency
Value
Large Sample (n = 200)
Frequency
Value

The central limit theorem

The distribution of sample means will be nearly Normally distributed

It will get closer to a Normal distribution as the sample size increases.

Standard Errors

In practice, we want to make ***inferences*** about the ***parameters of the target population*** using a ***single sample*.**

The precision of the sample mean is measured by the standard deviation of the mean.

This is called the **standard error, SE**.

$$SE = \frac{sample\ standard\ deviation}{\sqrt{sample\ size}} = \frac{SD}{\sqrt{n}}$$

Standard Errors

Based on the first random sample of 30 children we can calculate

$\bar{x}$ = 108.6cm, SD = 6.4cm, n = 30

So that…… SE=6.4/√30

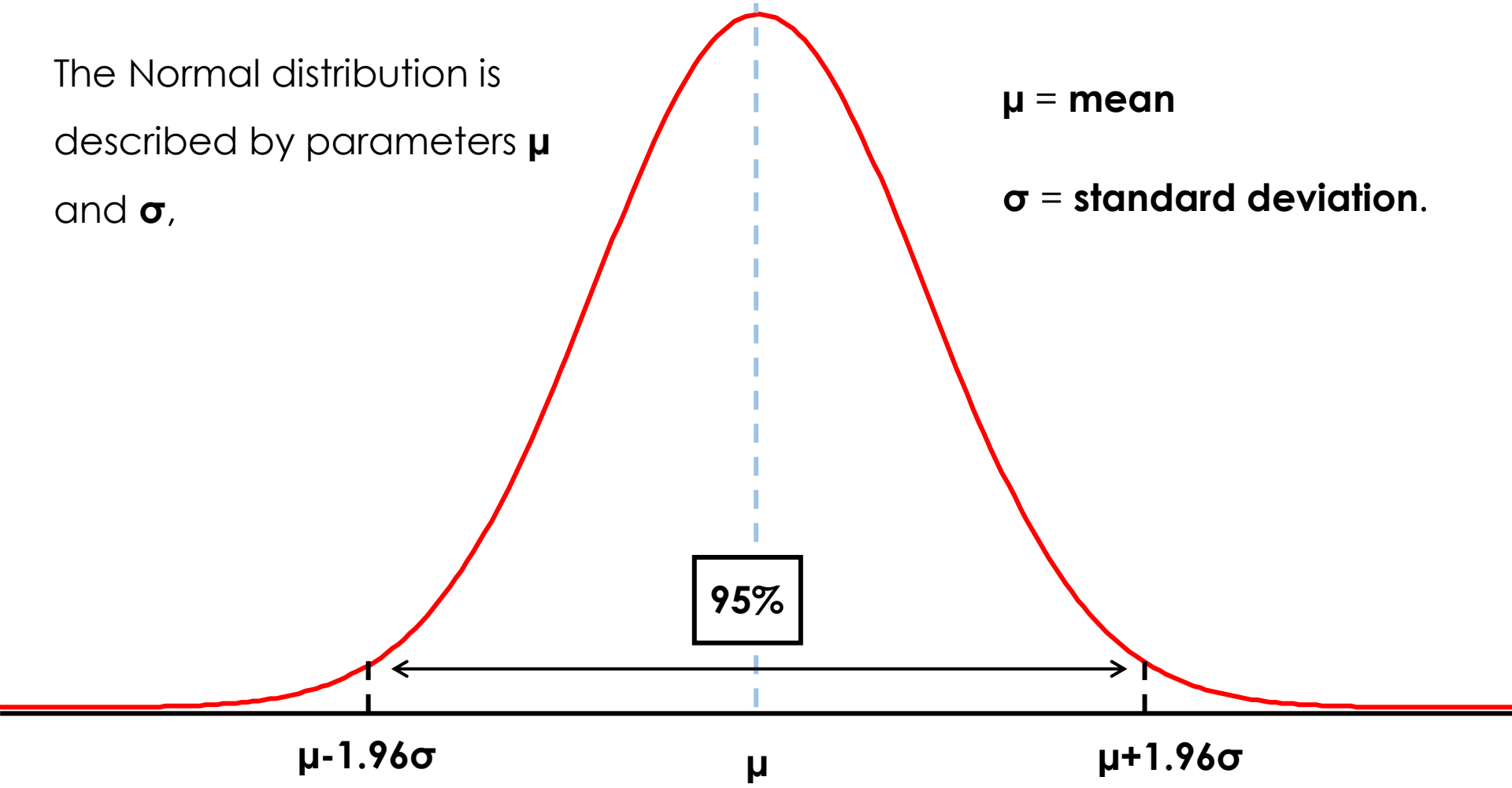
How can this be interpreted?

A standard error is a standard deviation….
the standard deviation of a sampling distribution

The Normal distribution is described by parameters **μ** and **σ**,

μ = **mean**

σ = **standard deviation**.



Confidence intervals

Confidence intervals define the range of values within which the population mean **μ** is likely to lie

A 95% confidence interval for the population mean is defined by

$$\bar{x}\text{-}1.96SE \text{ to } \bar{x} + 1.96SE$$

95% of intervals will contain the true population mean.

Interpreted as 95% chance that the true population mean will be contained in the interval.

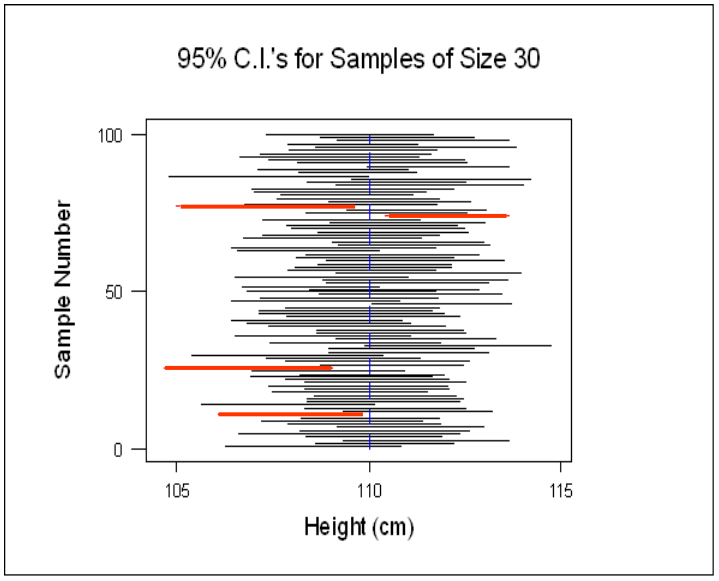
Routine physical examination of *all* 5 year old school entrants in Glasgow (complete census mean height 110cm (SD 6))

95% confident that the true population mean will be contained in the confidence interval

Sample size	mean	SD	SE	95% Confidence Interval
30	108.6	6.4	1.17	106.3 to 110.9

Estimate of population mean is 108.6cm.
95% confident it could be as small as 106.3 or as large as 110.9cm

95% confidence intervals obtained from 100 random samples of 30 children.



_____ interval contains 110

-------- interval does not contain 110

96 out of 100 contain the population mean

This illustrates the interpretation of a **95%** confidence interval.

Any ***95%CI*** from a ***single sample*** has ***probability 0.95*** of containing the ***population mean.***

Routine physical examination of *all* 5 year old school entrants in Glasgow (complete census mean height 110cm (SD 6))

95% confident that the true population mean will be contained in the confidence interval

Sample size	mean	SD	SE	95% Confidence Interval
60	109.6	6.1	0.78	108.0 to 111.2
70	109.2	5.9	0.70	107.8 to 110.6
100	109.9	6.0	0.60	108.7 to 111.1
200	110.2	6.0	0.43	109.3 to 111.0

Confidence intervals become narrower with increasing sample size

Giving a more ***precise*** estimate of the population mean

Example *estimate the average value* for triceps skinfold thickness of 10 year old boys

Random sample of 40 boys gave:

$\bar{x}$ = 8.71 mm, SD = 2.77 mm

Approximate 95% CI for the population mean

$$\bar{x} \pm 1.96 \times SE \quad = \bar{x} \pm 1.96 \times \frac{SD}{\sqrt{n}} \quad = 8.71 \pm 1.96 \times \frac{2.77}{\sqrt{40}}$$

$$= 8.71 \pm 0.86$$

$$= 7.85 \; to \; 9.57 \; mm$$

"We can be 95% confident that the population mean for 10 year old boys in Glasgow lies between 7.85 and 9.57mm, with a best estimate of 8.71mm."

Example (Two year survival following SABR for lung cancer)

102 patients with lung cancer treated with SABR, 77 were alive after 2 years (75%).

The estimate (75% or 0.75) is subject to ***sampling variability***.

The SE of a sample ***proportion***, p, is given by:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

provided that p is "not too close" to 0 or 1.

An *approximate* 95% CI for the population proportion is then given by:

$$p \pm 1.96 \times SE$$

the same structure as the CI for a population mean.

For survival following SABR, p = 77/102 = 0.75, n=102 and the 95% CI is:

$$p \pm 1.96 \times SE = p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}} = 0.75 \pm 1.96 \times \sqrt{\frac{0.75 \times (1-0.75)}{102}}$$

$$= 0.75 \pm 1.96 \times 0.04$$

$$= 0.67 \; to \; 0.83$$

$$= 67\% \; to \; 83\%$$

Interval is wide, so much **uncertainty** about the value of the population percentage.

A more precise estimate (i.e. a narrower interval) requires a larger sample size.

What degree of precision do you think would be appropriate here?

What sample size, approximately, would be required to achieve this?

Outcome following stroke hospitalisation

The Modified Rankin Scale (mRS):

The scale runs from 0-6, running from perfect health without symptoms to death.

0 - No symptoms

1 - No significant disability.

2 - Slight disability

3 - Moderate disability.

4 - Moderately severe disability

5 - Severe disability.

6 - Dead

Roll the dice...

Let us assume it is known that in the population that the percentage of patients who die within 30 days following their stroke hospitalisation is 16.7%

1. Throw your die. A 6 will indicate that the patient died within 30 days, 1-5 will indicate that the patient did not die within 30 days.
2. Make a note of the outcome for the patient
3. Return to step 1 and continue until you have a sample of size 5
5. Calculate the sample percentage of in-hospital mortality

The standard error & confidence interval for a **proportion** is calculated as follows

$n = sample\ size$

$x = number\ in\ sample\ with\ characteristic$

p is proportion from sample. $p = \frac{x}{n}$ $SE = \sqrt{\frac{p(1-p)}{n}}$

95% confidence interval for p $p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$

For large **n** (sample size).

Proportion of patients who died within 30 days following stroke hospitalisation

Patients (n)	Died (x)	p	se	95%CI		
60	9	15.0%	4.6%	6.0%	*to*	24.0%

2. A 95% confidence interval around a sample mean value

A. shows a range in which 95% of the population will lie
B. shows a range in which 95% of the sample will lie
C. shows a range in which we are 95% confident that the population mean could lie
D. shows a range in which we are 95% confident that the sample mean could lie
E. shows a range in which there is a 95% of population means lie