Institute of Psychiatry, Psychology and Neuroscience

KING'S College LONDON

Dr Silia Vitoratou

Department: Biostatistics and Health Informatics

Module Title: Introduction to Statistics

Session Title: Sampling and error

Topic title: Confidence and significance (I)

Topic materials:
Silia Vitoratou

Contributions:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

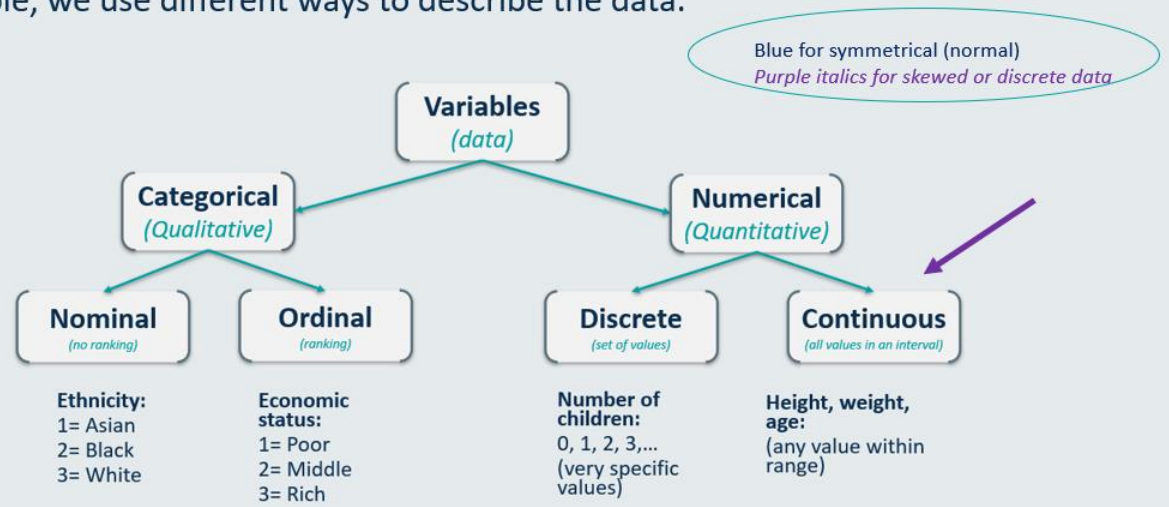
Learning Outcomes

- To understand how sampling works
- To understand the difference in random and systematic error
- To understand the sampling variability
- To introduce the sampling distribution

Previously on 'introduction to statistics'.....

1. To understand a characteristic which varies from person to person (a '**variable**') in a population, we study **a representative sample**.
2. The first step is to familiarise with our data, that is, the variables in our sample data set. We can do this with descriptive statistics, which will also allow us to **clean up the data from any typos**.

Based on the type of each variable, we use different ways to describe the data.



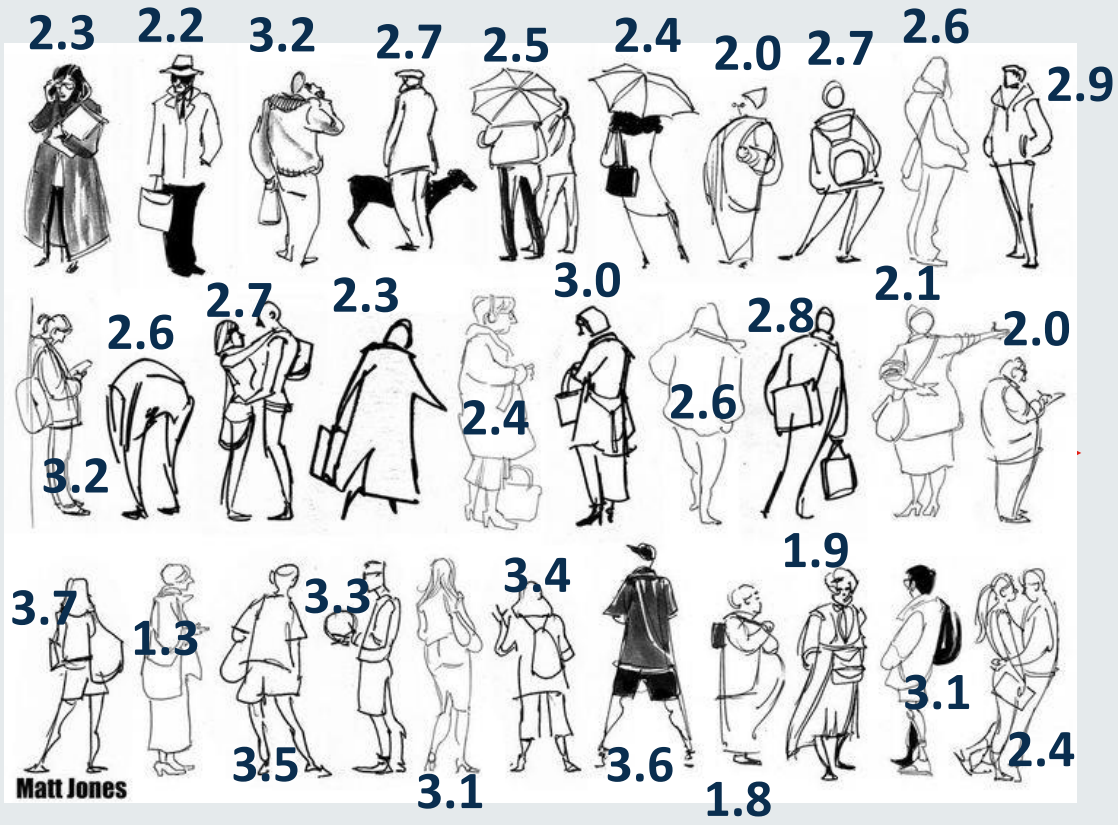
• Descriptive indices	Frequencies (Percentages %)	location: mean, *median*, mode Dispersion: SD, *min-max*, range
• Charts/plots	Bar Chart	Histogram, Box plot

Sampling and error

To study a variable in a population we use the values in a representative sample from that population

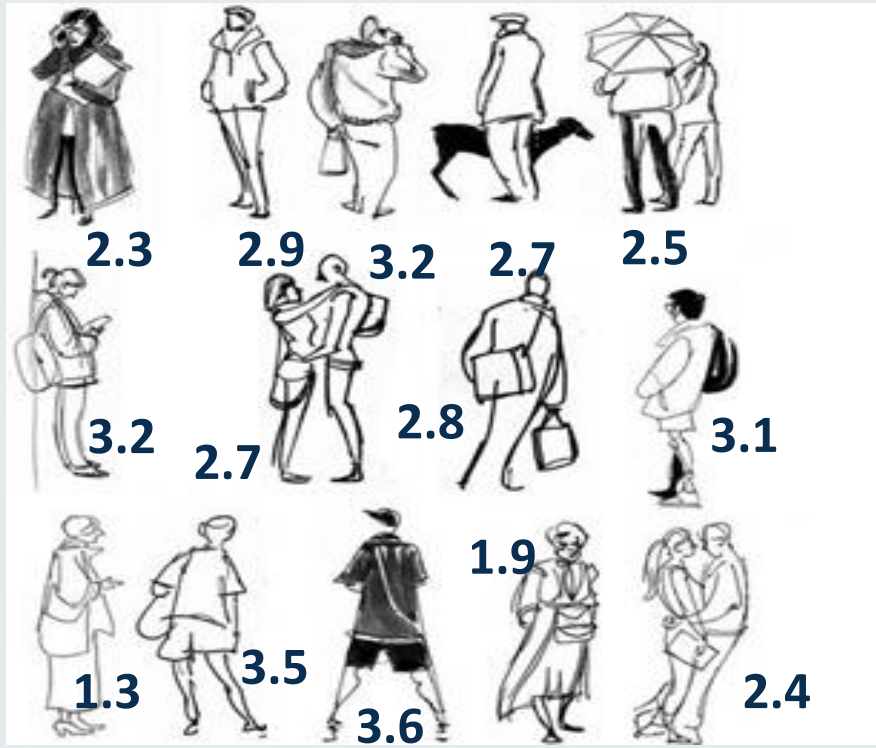
‘How many hours per week do you exercise?’

Population mean μ=2.66



Population

Sample mean $\bar{x}$*=2.72*



Sample

Sampling and error

In order to draw conclusions about an underlying population we study a sample or subset of the data. This process is called **statistical inference**.

We compute the **statistic** in the sample to estimate the **parameter** in the population, some examples

Parameter	
Population mean	$\mu = 2.66$
Population SD	$\sigma = 0.572$
Population variance	$\sigma^2 = 0.333$
Population proportion	$\pi = 0.20$

Statistic	
Sample mean	$\bar{x} = 2.72$
Sample SD	$s = 0.624$
Sample variance	$s^2 = 0.382$
Sample proportion	$p = 0.18$

The statistic is the **estimator** of the model **parameter**. *The estimator, for instance, of the population mean μ, is the statistic 'sample mean' $\bar{x}$.*

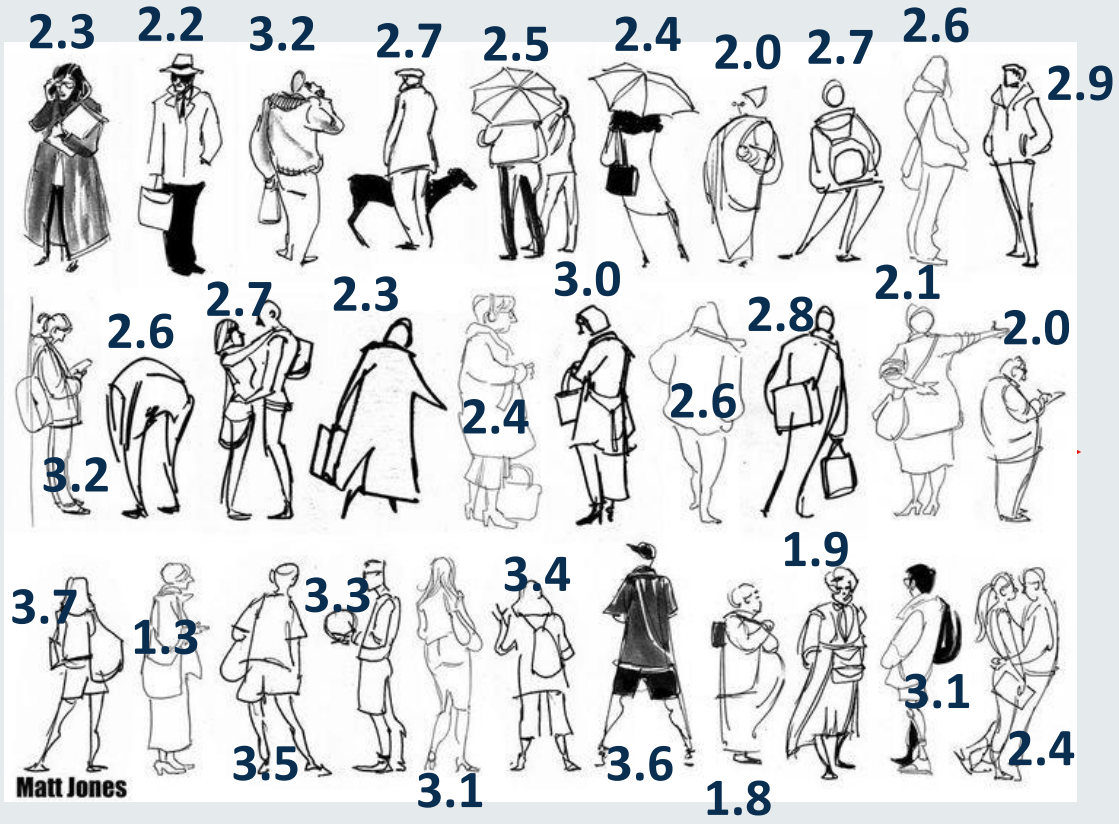
The **estimated** value of the population mean μ, is 2.72.

Sampling and error

Let us go back to our example

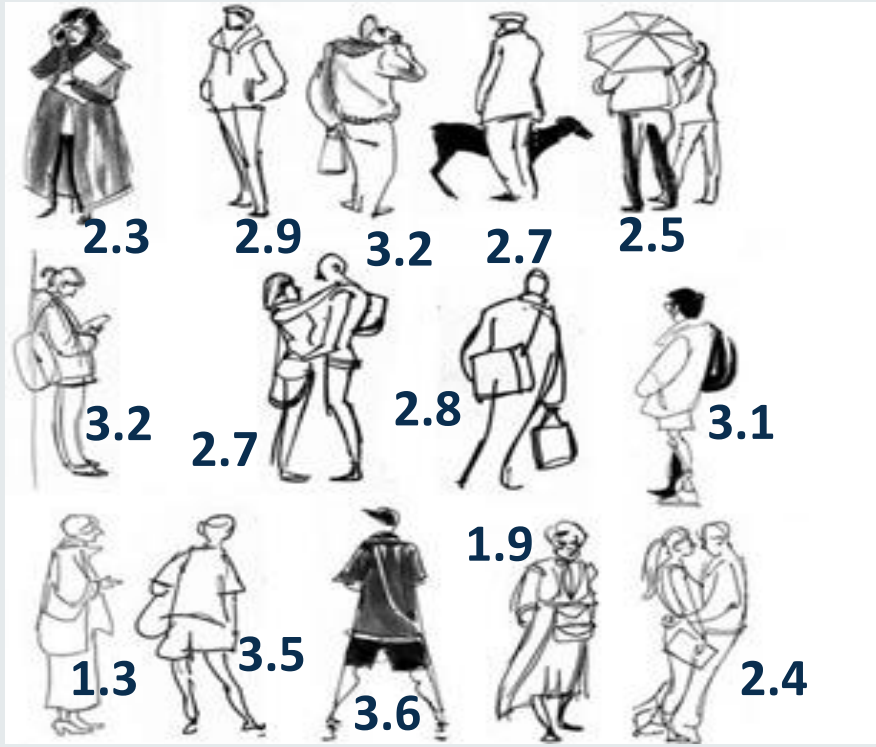
‘How many hours per week do you exercise?’

Population mean μ=2.66



Population

Sample mean $\bar{x}$=2.72



Sample

Sampling and error

What if I collect more than one sample?

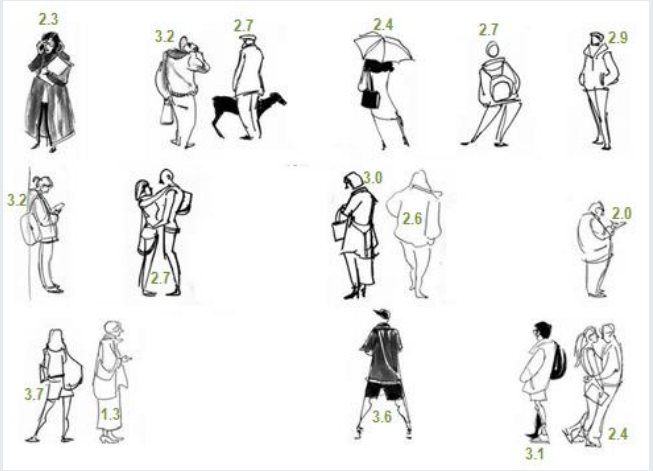
Population μ=2.66
σ=0.572

Difference due to <u>random variation (or sampling error)</u>

Random sample 1

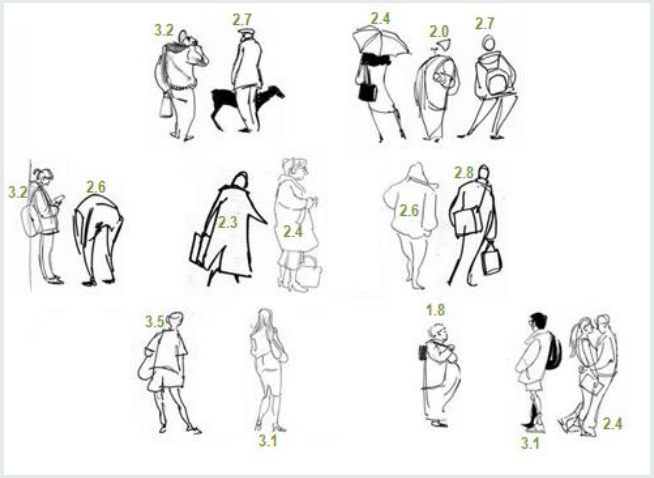
$\bar{x}_1$=2.72
s_1=0.622

Random sample 2



$\bar{x}_2$=2.48
s_2=0.593

Random sample 3



$\bar{x}_3$=2.68
s_3=0.461

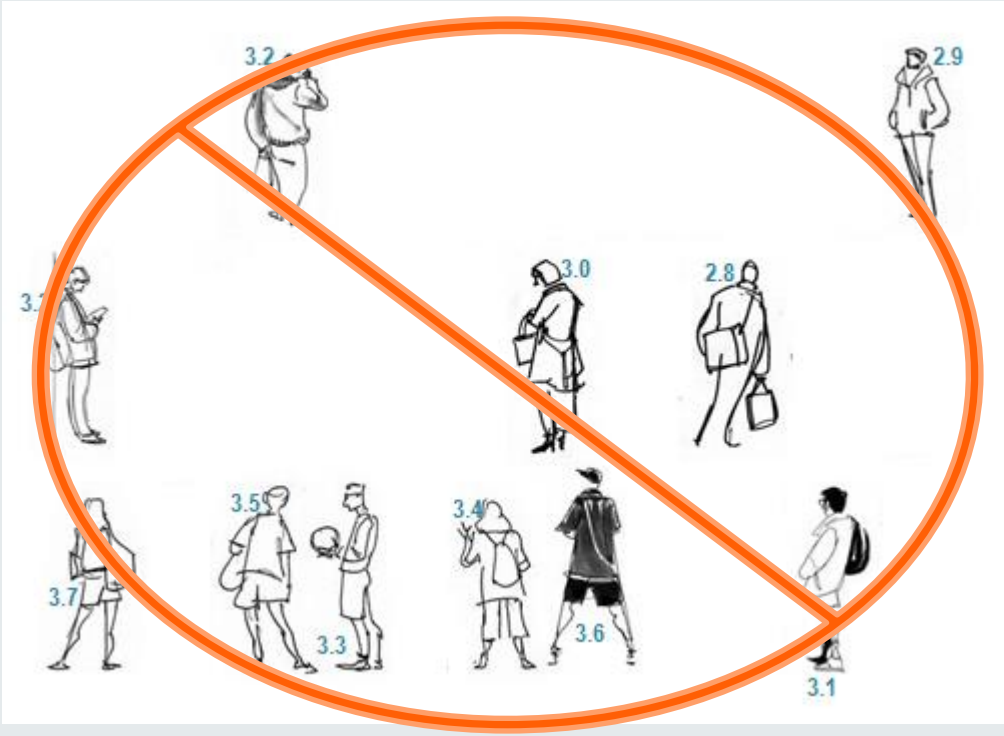
Sampling and error

What if I selected my sample standing outside a gym?

Population

Population mean $\mu=2.66$

Population stand. dev. $\sigma=0.572$



Difference due to Systematic error (BIAS)

Random Sample 1

Sample 1 mean $\bar{x}_1=2.72$

Sample 1 stand. dev. $s_1=0.622$

Random Sample 2

Sample 2 mean $\bar{x}_2=2.48$

Sample 2 stand. dev. $s_2=0.593$

Random Sample 3

Sample 3 mean $\bar{x}_3=2.68$

Sample 3 stand. dev. $s_3=0.461$

Random Sample 4

Sample 4 mean $\bar{x}_4=3.25$

Sample 4 stand. dev. $s_4=0.294$

Sampling and error

Random error (noise, random)

Unpredictable:

- Goes on either direction (your measurement one time emerges randomly to be higher and another time lower than the actual value).
- It is due to unknown factors

Systematic error (bias)

Consistent:

- You consistently, repeatedly underestimate or overestimate the true value (always same direction: one or the other).
- It is due to factors that can be traced (wrong sample, malfunctioning scale etc) in the experimental design.

Sampling and error

From this point onwards we will assume that the researcher has planned the study appropriately and that there is no bias in the experiments. We are going to focus on the uncertainty due to random variation.

Population $\mu=2.66$ $\sigma=0.572$

Random sample 1

Random sample 2

Random sample 3

$\bar{x}_1=2.72$ $s_1=0.622$

$\bar{x}_2=2.48$ $s_2=0.593$

$\bar{x}_3=2.68$ $s_3=0.461$

How are these values distributed?

Sampling distribution

- Suppose we have a population of heights as shown on the right.
- This could, for example, be the heights of all participants in this classroom.
- In that case, all participants in this class constitute the population.

Population Heights (in metres)

1.66	1.52	1.42	1.65	1.49	1.55	1.84	1.91
1.5	1.59	1.45	1.53	1.85	1.39	1.73	1.51
2.23	2.13	1.52	1.4	1.94	1.35	1.75	1.61
1.76	1.37	1.75	1.64	1.97	1.97	1.55	1.81
1.28	1.52	1.64	1.18	1.65	1.43	1.59	1.41
1.57	1.56	1.47	1.46	1.56	1.94	1.58	1.57
1.26	1.48	1.29	1.4	1.95	1.73	1.65	1.84
1.65	1.78	1.39	1.56	1.64	1.61	1.42	1.86

Sampling distribution

Let's take a **single sample** of size **5** and calculate the **mean height** (1.7m)

Population Heights (in metres)

1.66	1.52	1.42	1.65	1.49	1.55	1.84	1.91
1.5	1.59	1.45	1.53	1.85	1.39	1.73	1.51
2.23	2.13	1.52	1.4	1.94	1.35	1.75	1.61
1.76	1.37	1.75	1.64	1.97	1.97	1.55	1.81
1.28	1.52	1.64	1.18	1.65	1.43	1.59	1.41
1.57	1.56	1.47	1.46	1.56	1.94	1.58	1.57
1.26	1.48	1.29	1.4	1.95	1.73	1.65	1.84
1.65	1.78	1.39	1.56	1.64	1.61	1.42	1.86

Sample mean (n=5)

#	Mean
1	(1.55+1.73+2.13+1.65+1.46)/5 = 1.7

A **point** estimate of the average population height based on this sample is 1.7 metres

Sampling distribution

Let's take **another** sample and calculate the mean height (1.59 m)

Population Heights (in metres)

1.66	1.52	1.42	1.65	1.49	1.55	1.84	1.91
1.5	1.59	1.45	1.53	1.85	1.39	1.73	1.51
2.23	2.13	1.52	1.4	1.94	1.35	1.75	1.61
1.76	1.37	1.75	1.64	1.97	1.97	1.55	1.81
1.28	1.52	1.64	1.18	1.65	1.43	1.59	1.41
1.57	1.56	1.47	1.46	1.56	1.94	1.58	1.57
1.26	1.48	1.29	1.4	1.95	1.73	1.65	1.84
1.65	1.78	1.39	1.56	1.64	1.61	1.42	1.86

Sample mean (n=5)

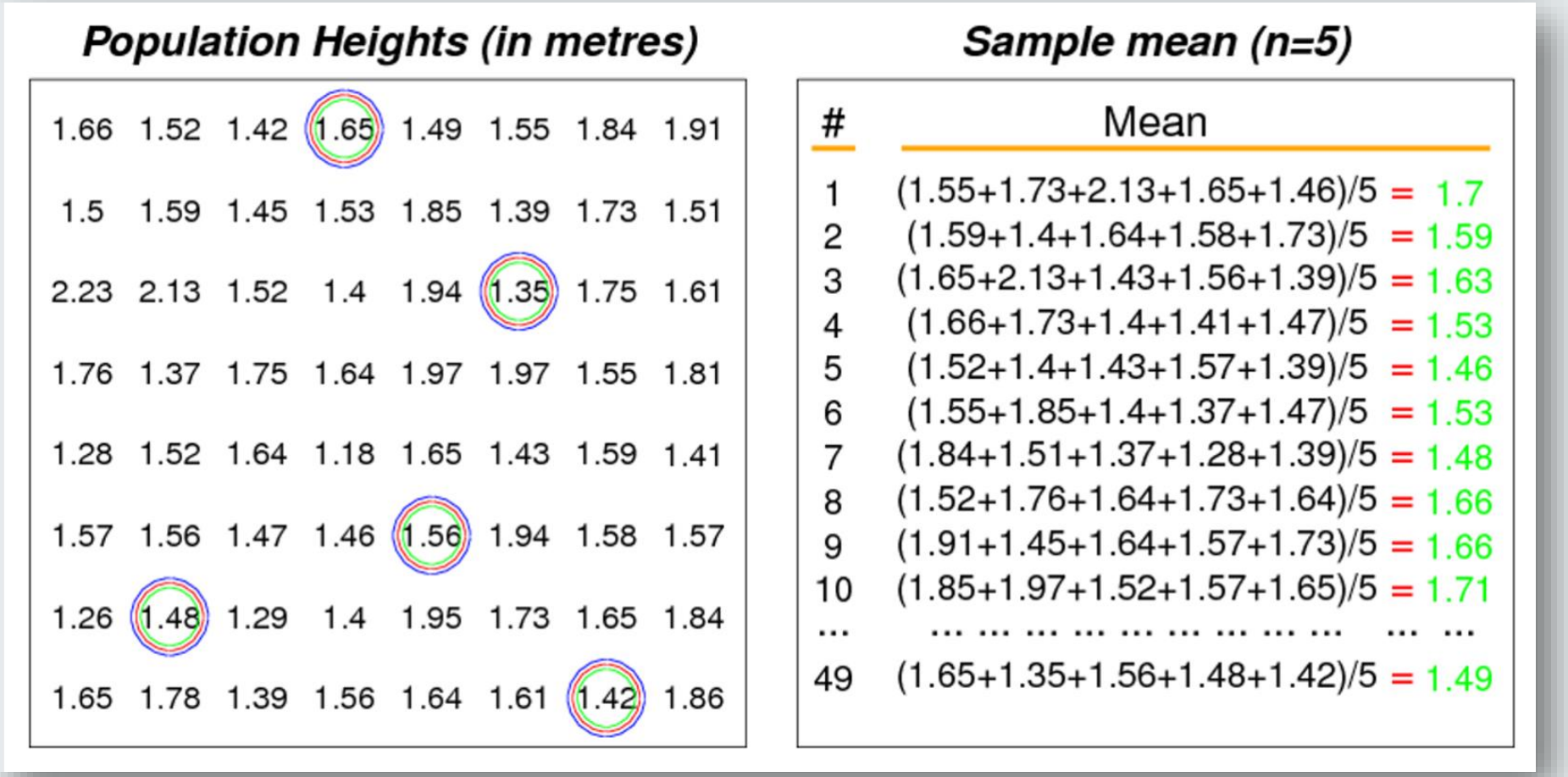
#	Mean
1	(1.55+1.73+2.13+1.65+1.46)/5 = 1.7
2	(1.59+1.4+1.64+1.58+1.73)/5 = 1.59

Why do the two means differ?

→ **Due to sampling variation (Random Error)**

Sampling distribution

Take **more** samples - almost each one has a different sample mean



Population Heights (in metres)

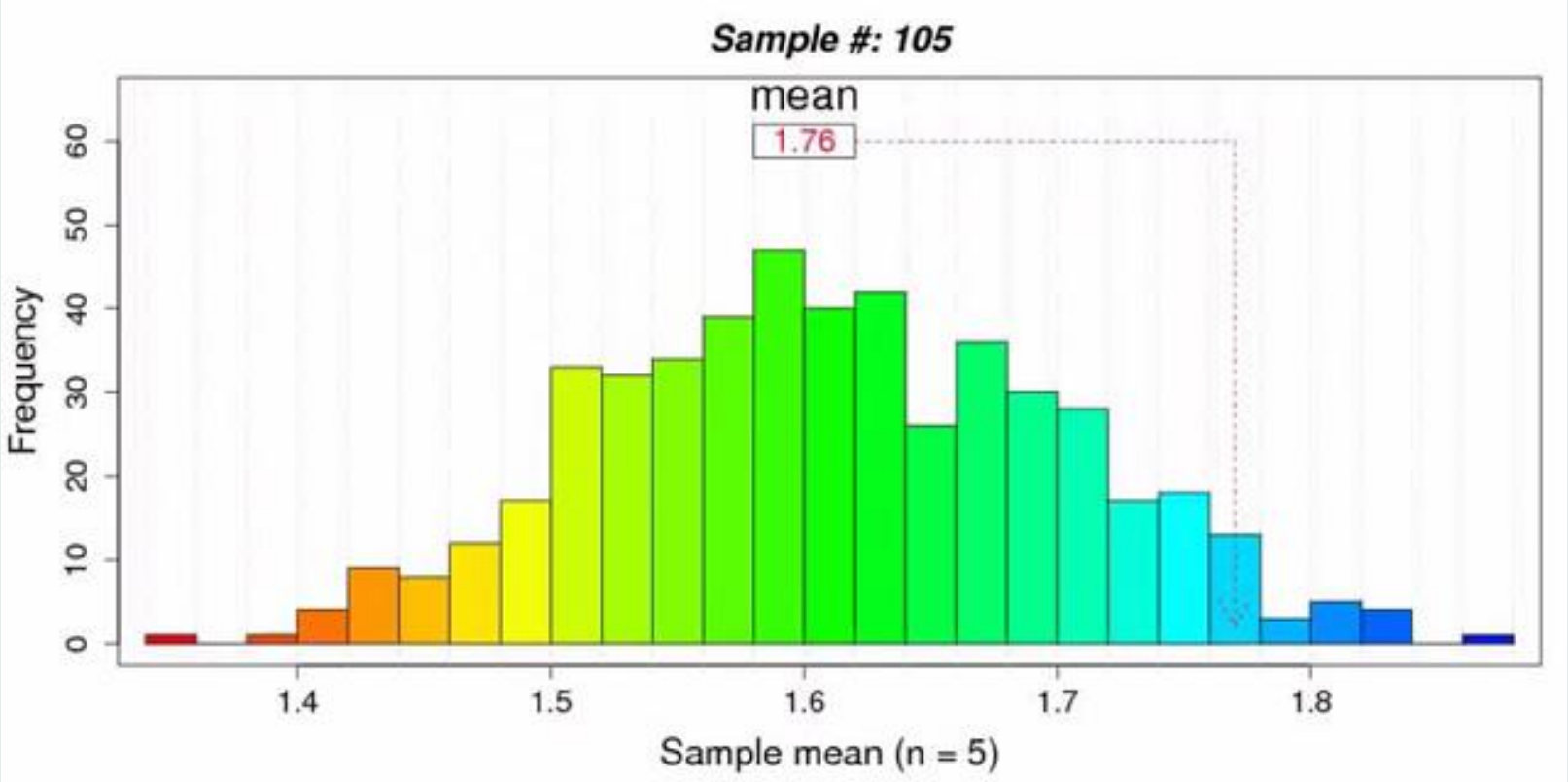
1.66	1.52	1.42	1.65	1.49	1.55	1.84	1.91
1.5	1.59	1.45	1.53	1.85	1.39	1.73	1.51
2.23	2.13	1.52	1.4	1.94	1.35	1.75	1.61
1.76	1.37	1.75	1.64	1.97	1.97	1.55	1.81
1.28	1.52	1.64	1.18	1.65	1.43	1.59	1.41
1.57	1.56	1.47	1.46	1.56	1.94	1.58	1.57
1.26	1.48	1.29	1.4	1.95	1.73	1.65	1.84
1.65	1.78	1.39	1.56	1.64	1.61	1.42	1.86

Sample mean (n=5)

#	Mean
1	(1.55+1.73+2.13+1.65+1.46)/5 = 1.7
2	(1.59+1.4+1.64+1.58+1.73)/5 = 1.59
3	(1.65+2.13+1.43+1.56+1.39)/5 = 1.63
4	(1.66+1.73+1.4+1.41+1.47)/5 = 1.53
5	(1.52+1.4+1.43+1.57+1.39)/5 = 1.46
6	(1.55+1.85+1.4+1.37+1.47)/5 = 1.53
7	(1.84+1.51+1.37+1.28+1.39)/5 = 1.48
8	(1.52+1.76+1.64+1.73+1.64)/5 = 1.66
9	(1.91+1.45+1.64+1.57+1.73)/5 = 1.66
10	(1.85+1.97+1.52+1.57+1.65)/5 = 1.71
...	... ...
49	(1.65+1.35+1.56+1.48+1.42)/5 = 1.49

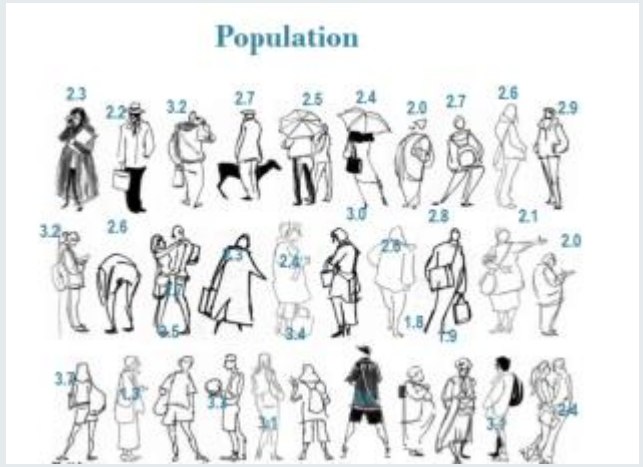
Sampling distribution

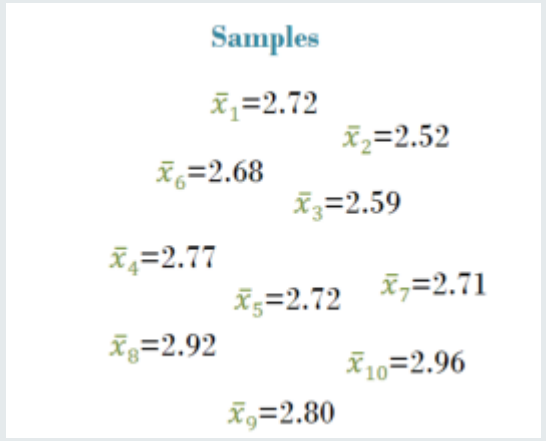
Let us see those values graphically, say sample 500 times from that population



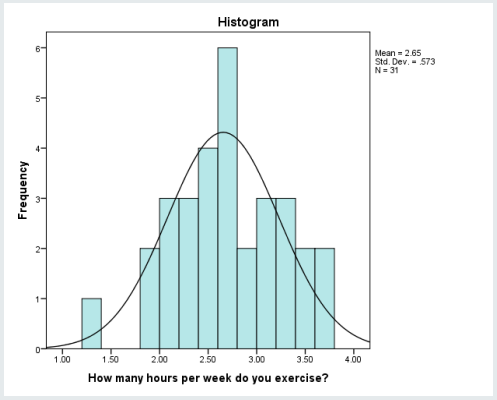
Sampling distribution

The distribution of these estimated means, from a number of samples, is called the **sampling distribution** of the mean. That is, the sampling distribution is the distribution of the estimated means from different samples of the same population.

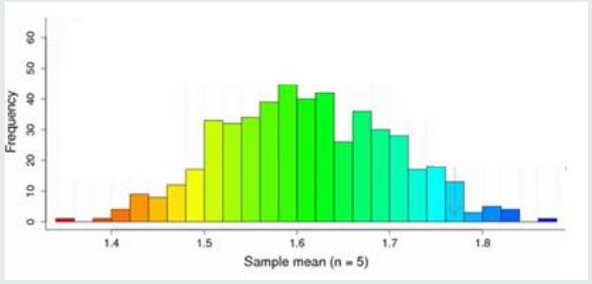


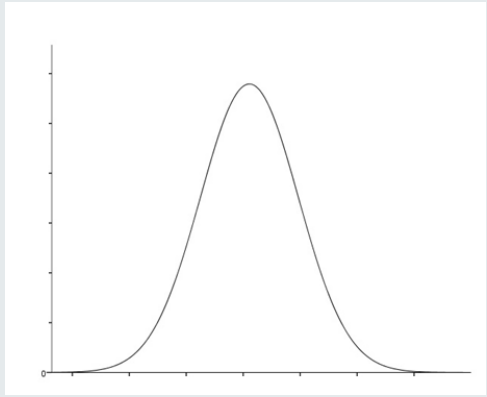


Theory states that if we take more and more and more samples from the same population, then the sampling distribution of the statistic mean will be a **normal distribution**.



Sampling distribution





Central limit theorem

Summary

Let us summarise what we learned in this session:

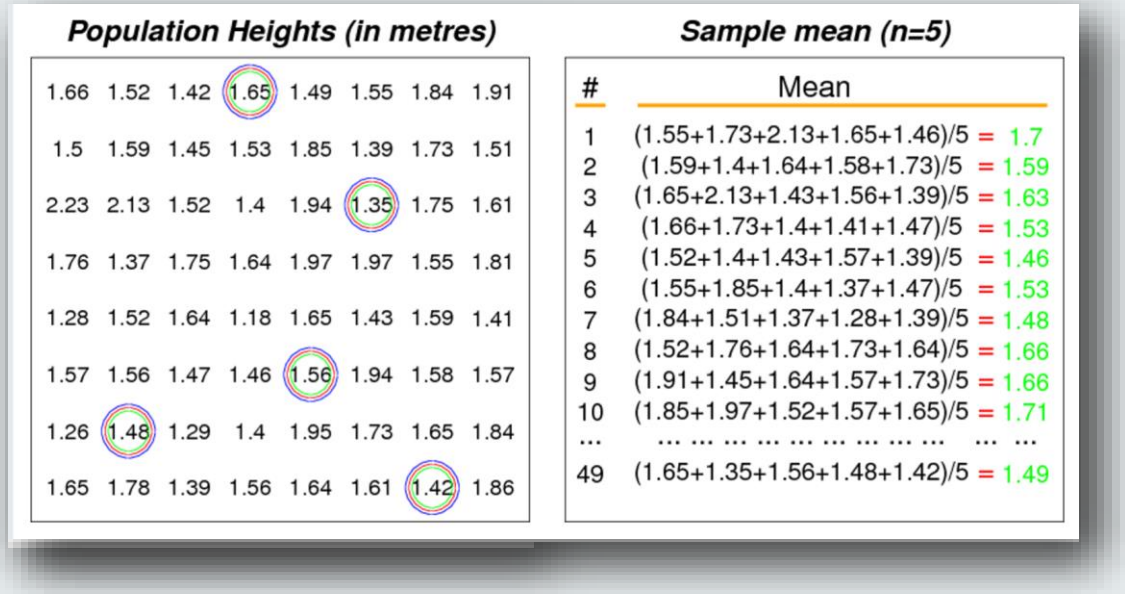
*We wish to **infer** on the value of a **parameter** in the population*

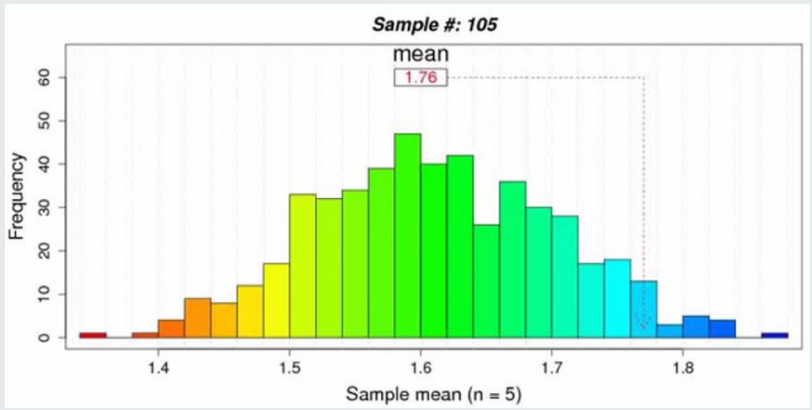
*As we do not have access to the entire population, we use a **sample** to estimate the parameter.*

But different samples lead to different estimated values for the parameter

These different estimated values follow the sampling distribution

For large numbers of samples, this distribution tends to the normal distribution (CLT).





Knowledge Check

1. Tom wants to estimate the mean number of hours people read in his town. Which of the below is correct?
    a) The mean number of hours people read in Tom's town is the **parameter μ**
    b) The mean number of hours people read in Tom's town is the **statistic μ**

2. The people in Tom's sample read 2.4 hours per week. Therefore, the value 2.4h/w is:
    a) the town's **population mean** number of hours reading
    b) the **estimated** town's **mean** number of hours reading
    c) the **estimator** of town's **population mean** number of hours reading

3. Ten of Tom’s classmates also repeat the experiment and they come up with ten more estimate values. The distribution of these values is called
    a) the **sampling** distribution
    b) the **population** distribution

Knowledge Check

1. Tom wants to estimate the mean number of hours people read in his town. Which of the below is correct?

    a) The mean number of hours people read in Tom's town is the **parameter μ**

    b) ~~The mean number of hours people read in Tom's town is the **statistic μ**~~

When we refer to the population we refer to the parameter

1. The people in Tom's sample read 2.4 hours per week. Therefore, the value 2.4h/w is:

    a) ~~the town's **population mean** number of hours reading~~

    b) the **estimated** town's **mean** number of hours reading

    c) ~~the **estimator** of town's **population mean** number of hours reading~~

The sample mean is the test statistic, the estimator, but its value is the estimated value.

3. Ten of Tom's classmates also repeat the experiment and they come up with ten more estimate values. The distribution of these values is called

    a) the **sampling** distribution

    b) ~~the **population** distribution~~

The distribution of the sampled, estimated values is called the sampling distribution

Reflection

Thinking about your own research

- What could be the sources of systematic error in your study?

Reference List

For more details on measurement issues see:
Streiner & Norman (2003) Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford University Press

For more details on SPSS implementation see:
Field (2009) Discovering Statistics using SPSS 3rd Edn, Sage, London. Everything you ever wanted to know about statistics. Ch.2

For more details of the concepts covered in Topic 2, see Chapters 1- 3 of the book:
Agresti and Finlay (2009) Statistical Methods for the Social Sciences (Statistical concepts: Chapters 1-3, Inferences for probabilities: Chapter 5, p110-116, and Chapter 6, p156-159, p169-p173, Inferences for a sample mean: Chapter 6, p147-p156)

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Vitoratou:

Silia Vitoratou, PhD
Psychometrics & Measurement Lab,
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
silia.vitoratou@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**