Institute of Psychiatry, Psychology and Neuroscience

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

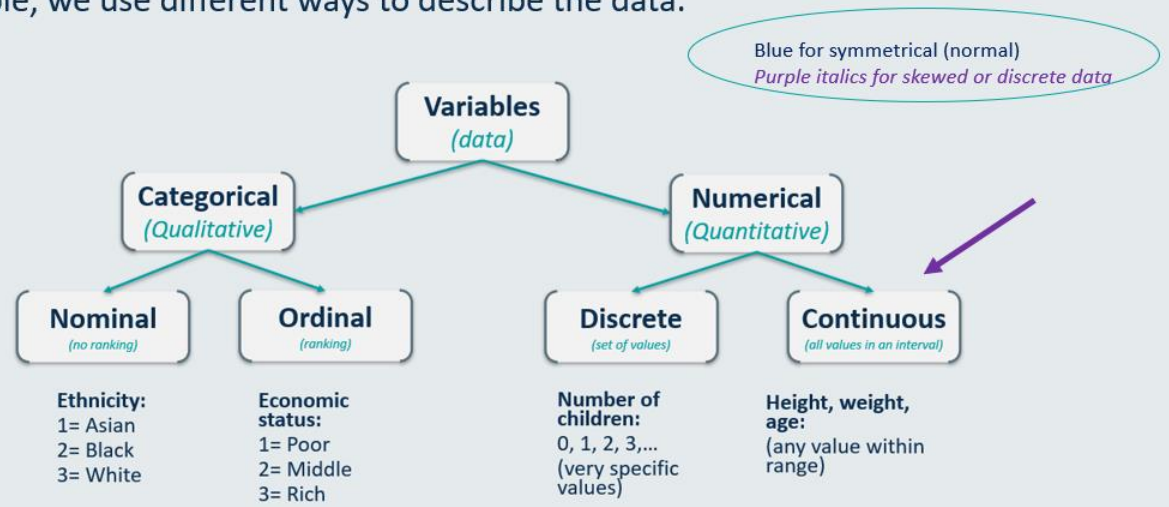
Learning Outcomes

- 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.



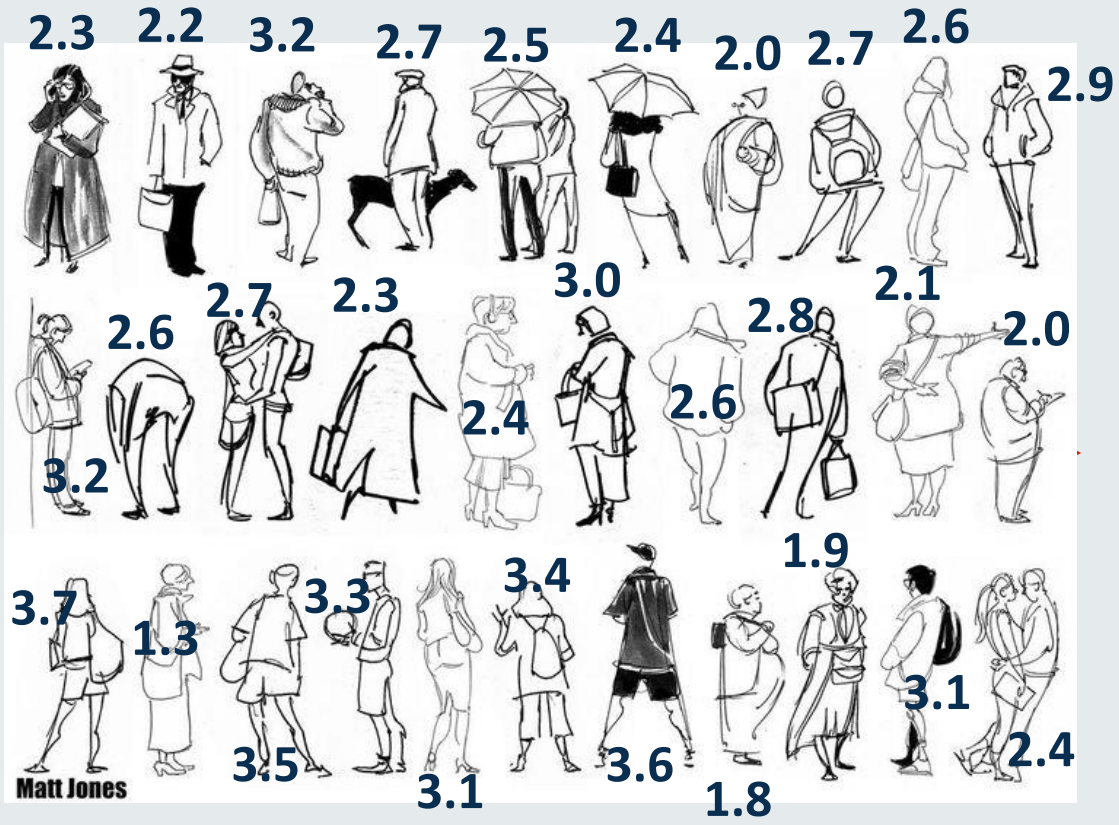
• Descriptive indices	Frequencies (Percentages %)	location: mean, *median*, mode Dispersion: SD, *min-max*, range
• Charts/plots	Bar Chart	Histogram, Box plot

Sampling and error

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

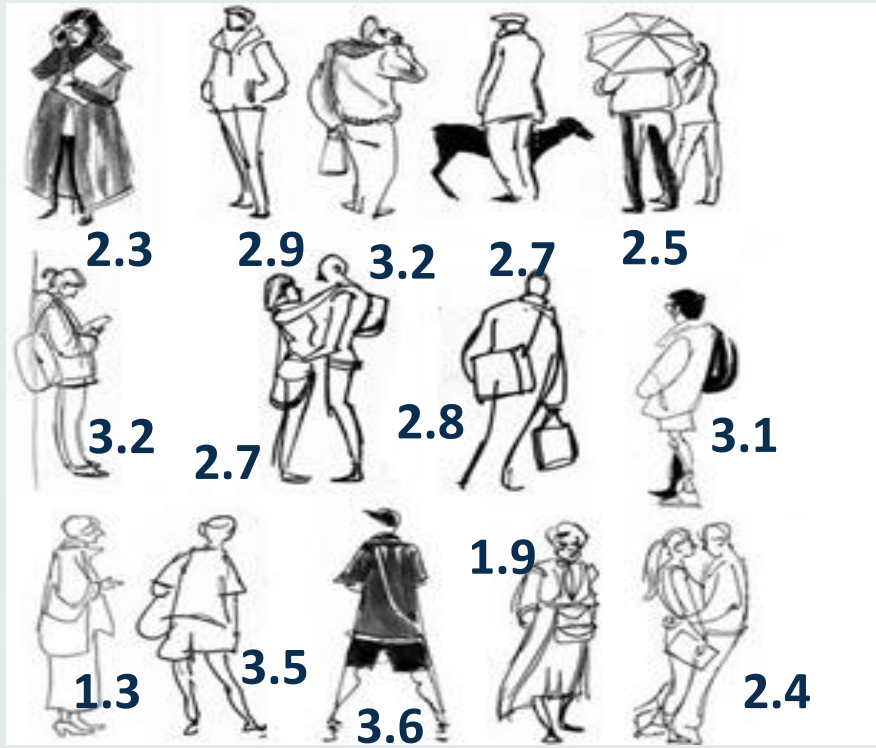
‘How many hours per week do you exercise?’

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' $\bar{x}$.*

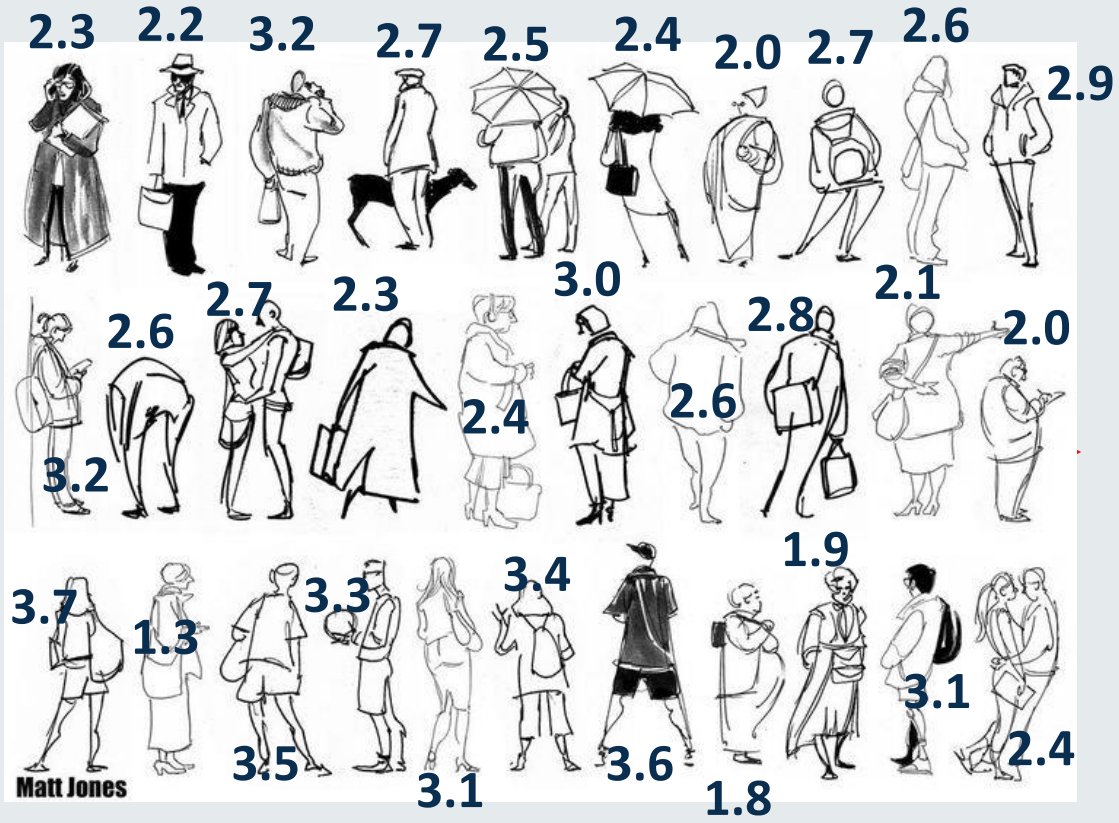
The **estimated** value of the population mean μ, is 2.72.

Sampling and error

Let us go back to our example

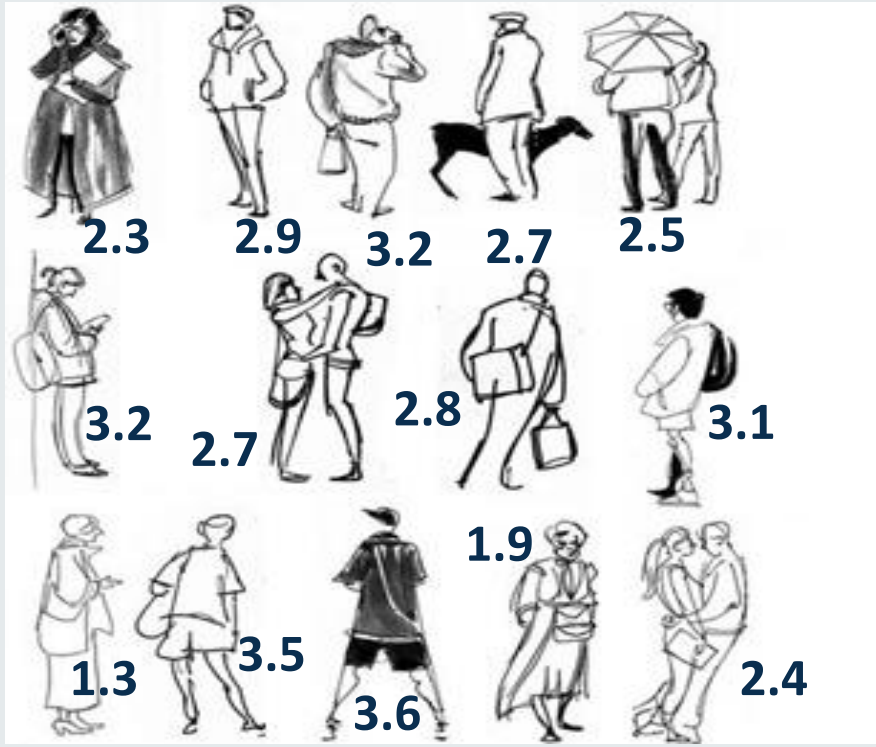
‘How many hours per week do you exercise?’

Population mean μ=2.66



Population

Sample mean $\bar{x}$=2.72



Sample

Sampling and error

What if I collect more than one sample?

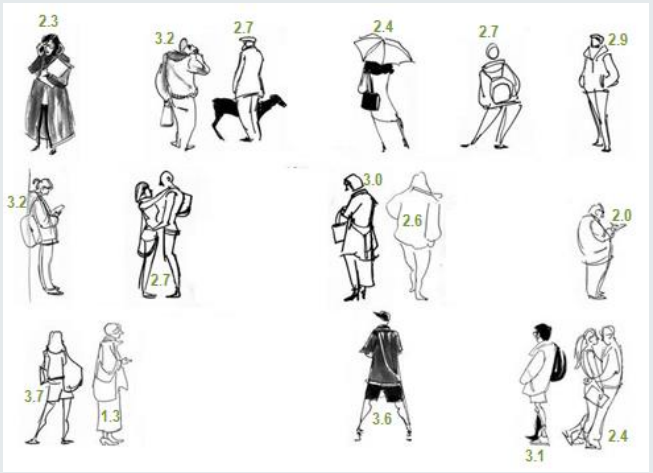
Population μ=2.66
σ=0.572

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

Random sample 1

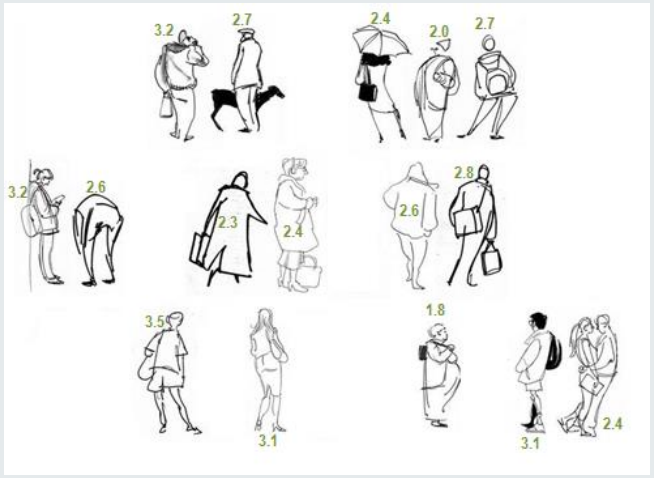
$\bar{x}_1$=2.72
s_1=0.622

Random sample 2



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

Random sample 3



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

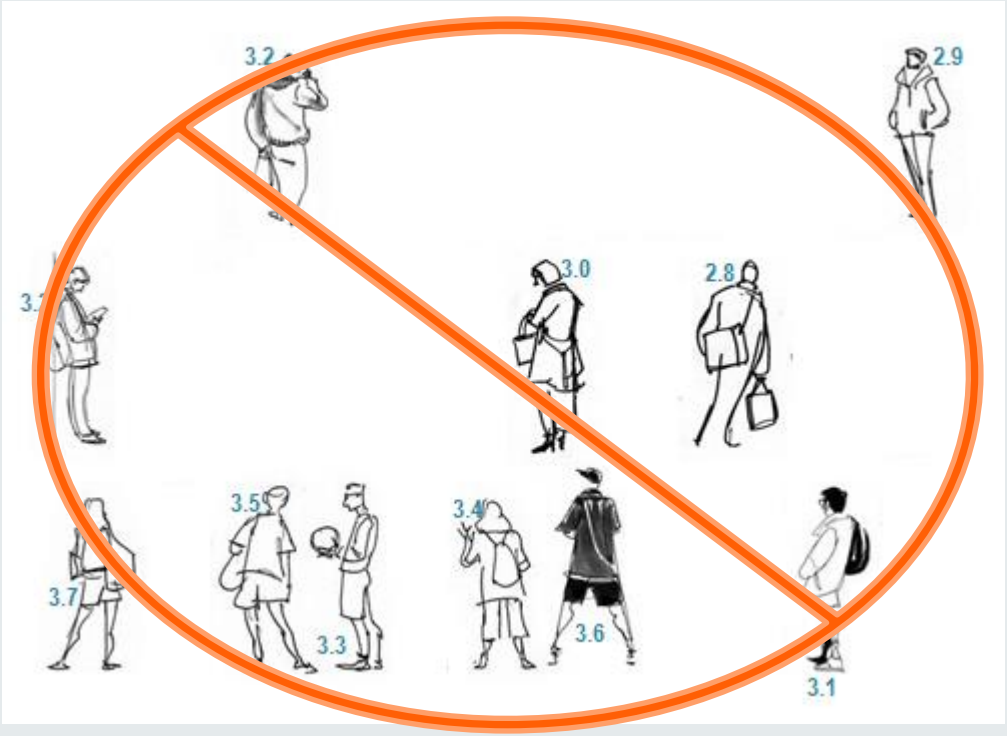
Sampling and error

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)

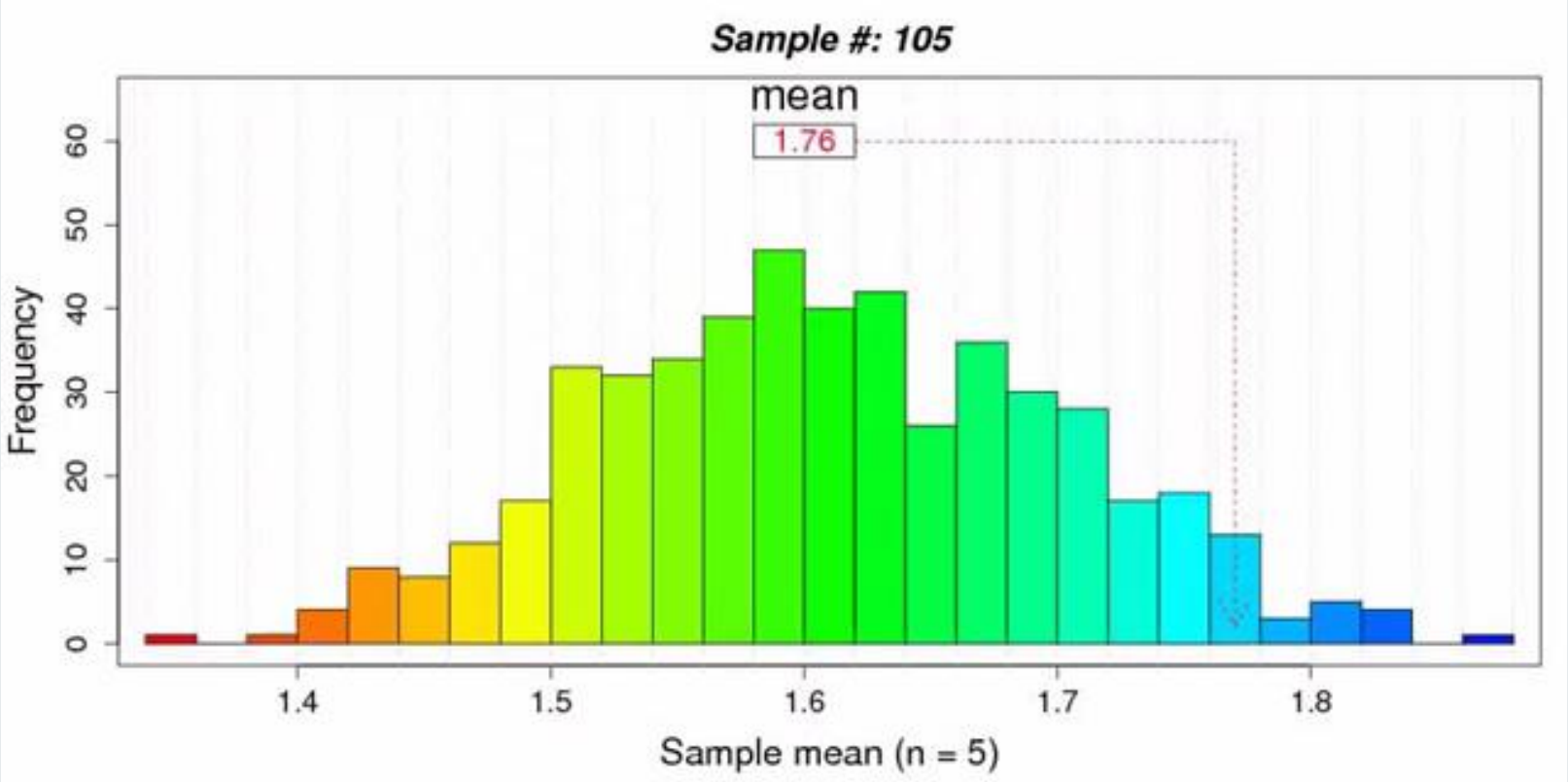
1.66	1.52	1.42	1.65	1.49	1.55	1.84	1.91
1.5	1.59	1.45	1.53	1.85	1.39	1.73	1.51
2.23	2.13	1.52	1.4	1.94	1.35	1.75	1.61
1.76	1.37	1.75	1.64	1.97	1.97	1.55	1.81
1.28	1.52	1.64	1.18	1.65	1.43	1.59	1.41
1.57	1.56	1.47	1.46	1.56	1.94	1.58	1.57
1.26	1.48	1.29	1.4	1.95	1.73	1.65	1.84
1.65	1.78	1.39	1.56	1.64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

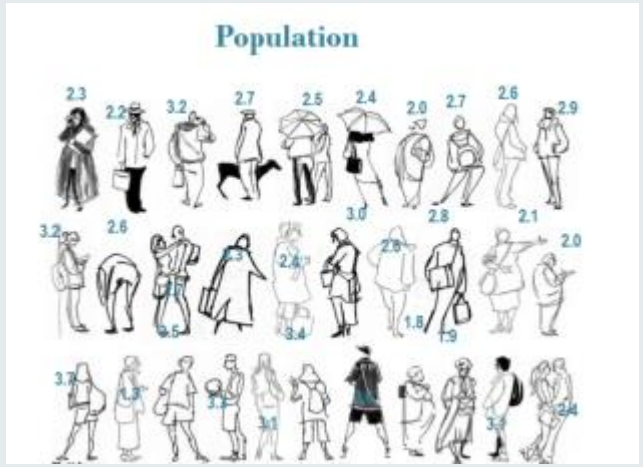
Sampling distribution

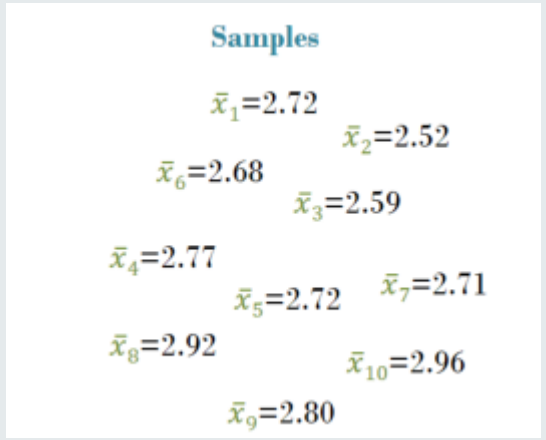
Let us see those values graphically, say sample 500 times from that population



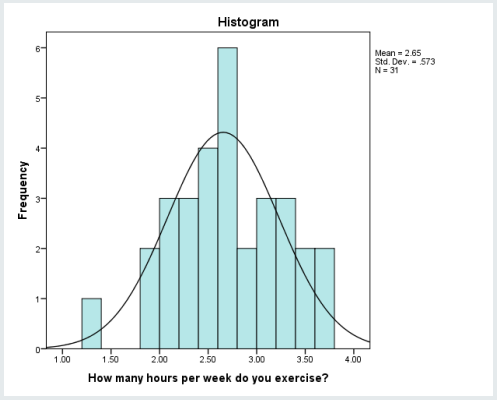
Sampling distribution

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.

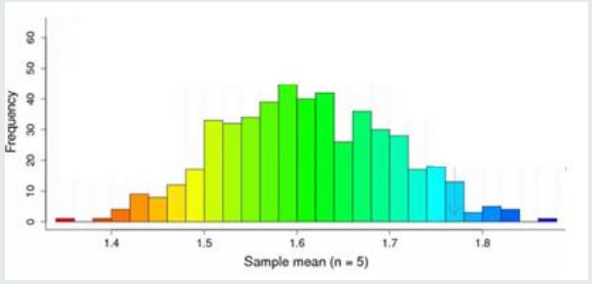


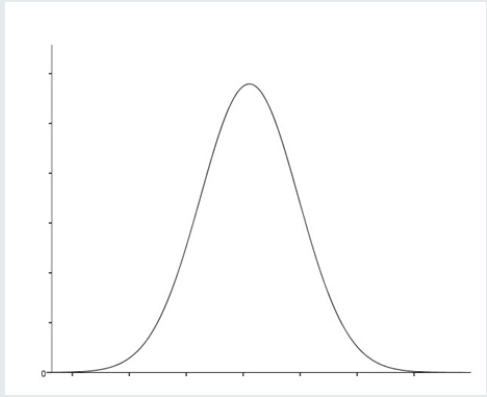


Theory states that if we take more and more and more samples from the same population, then the sampling distribution of the statistic mean will be a **normal distribution**.



Sampling distribution





Central limit theorem

Summary

Let us summarise what we learned in this session:

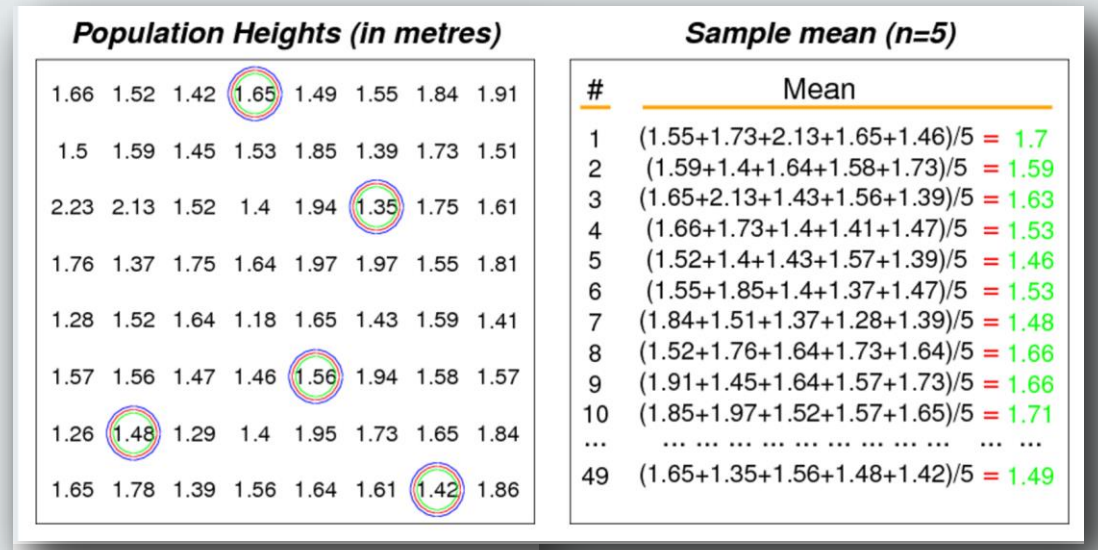
*We wish to **infer** on the value of a **parameter** in the population*

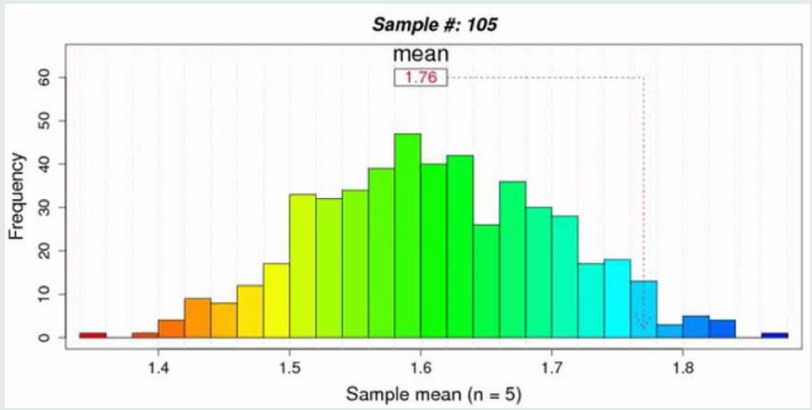
*As we do not have access to the entire population, we use a **sample** to estimate the parameter.*

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**

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 Distribution

Topic title: Confidence and significance (I)

Topic materials:
Silia Vitoratou

Contributions:
Zahra Abdula

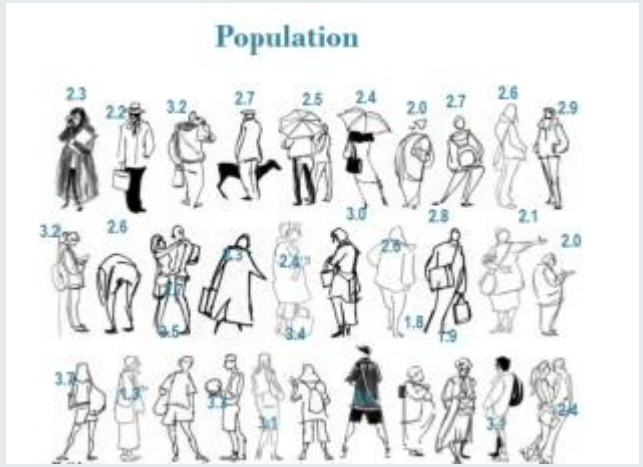
Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

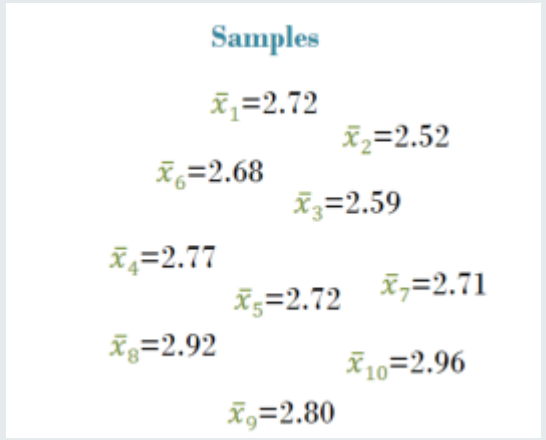
Learning Outcomes

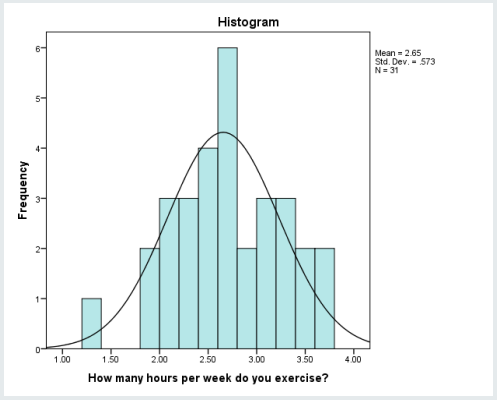
- To understand the sampling distribution
- To understand the central limit theorem
- To understand the normal distribution

The Normal (Gaussian) Distribution

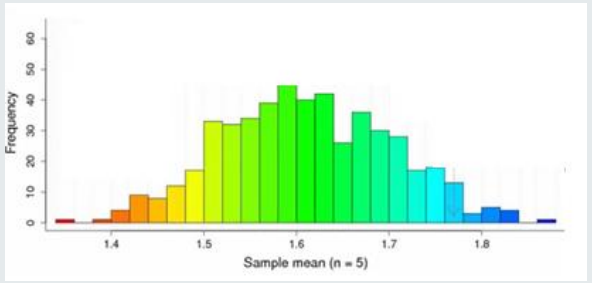
The distribution of these estimated means, from a number of samples, is called the **sampling distribution** of the mean.





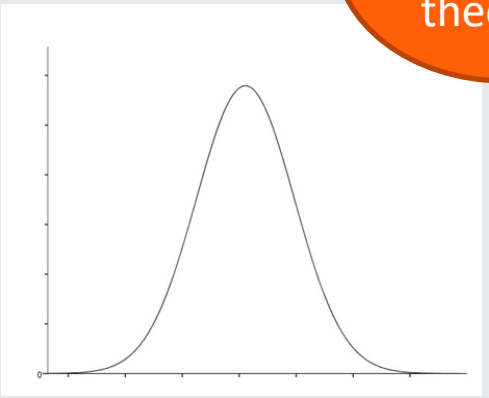


Sampling distribution



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**.

Central limit theorem



The Normal (Gaussian) Distribution

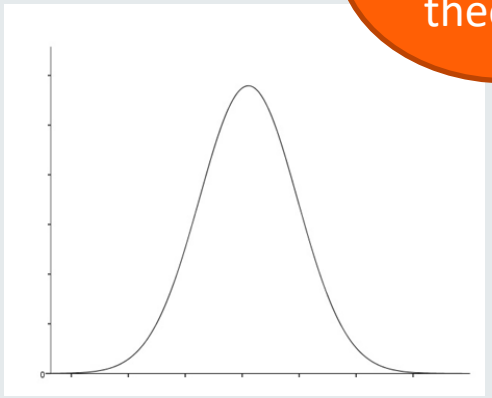
The normal distribution is called as such because a lot of events in real life follow this bell-shaped pattern. It is the norm (the rule).

The **Central Limit Theorem** in probability theory states that, given a sufficiently large amount of repetitions, the sampling distribution will approximate the normal distribution — no matter if the events follow different distributions.

Let us watch a magnificent demonstration of the central limit theorem, using the Galton board.

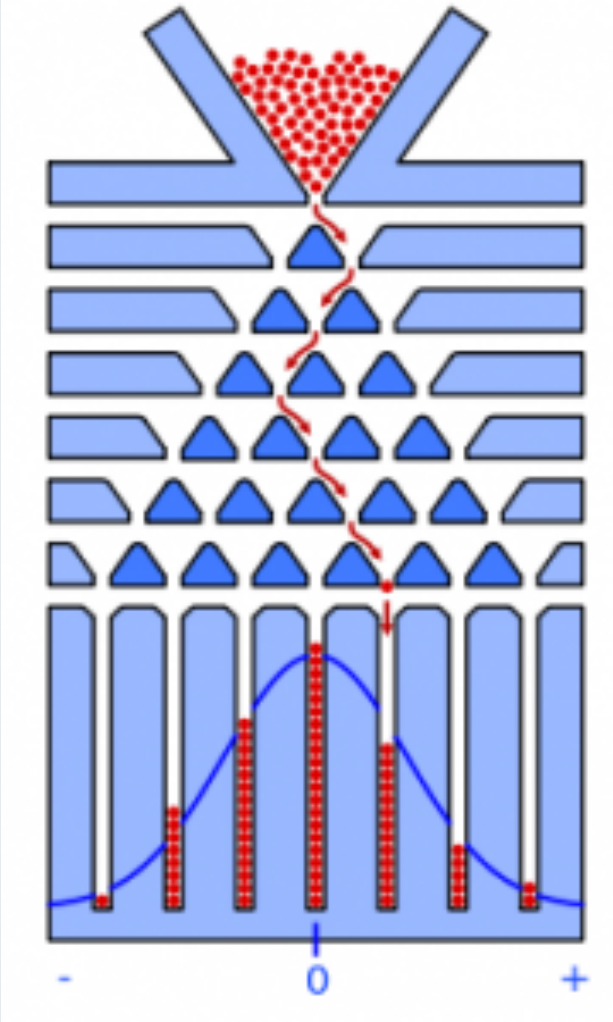
Video demonstrating the Galton Board from the Large Maths Outreach and Careers Kit developed by the Institute of Mathematics and its Applications as part of the National HE STEM Programme.
https://www.youtube.com/watch?v=6YDHBFVIvIs

Central limit theorem

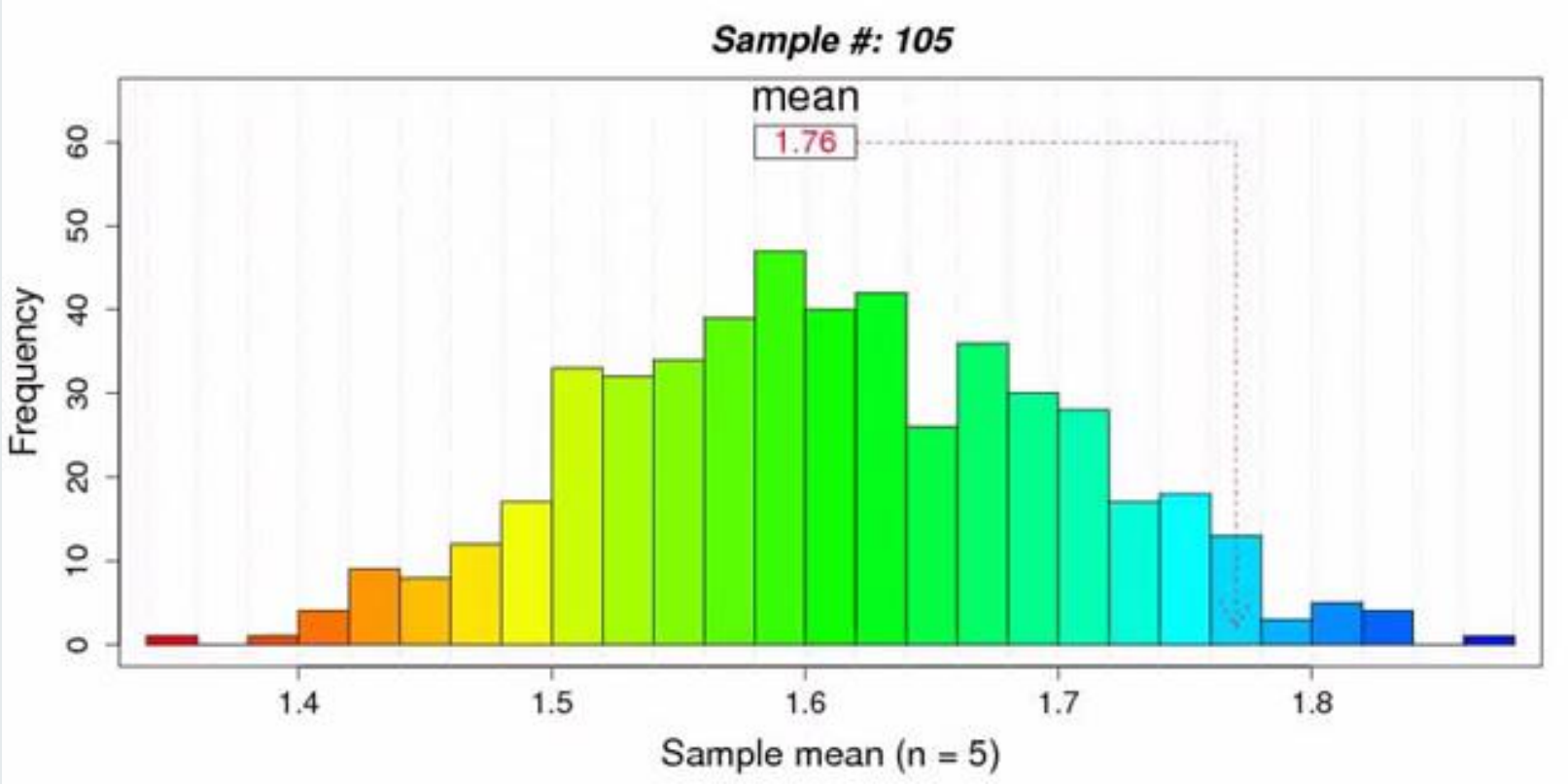


The Normal (Gaussian) Distribution

Galton Board



Sampling distribution of the means

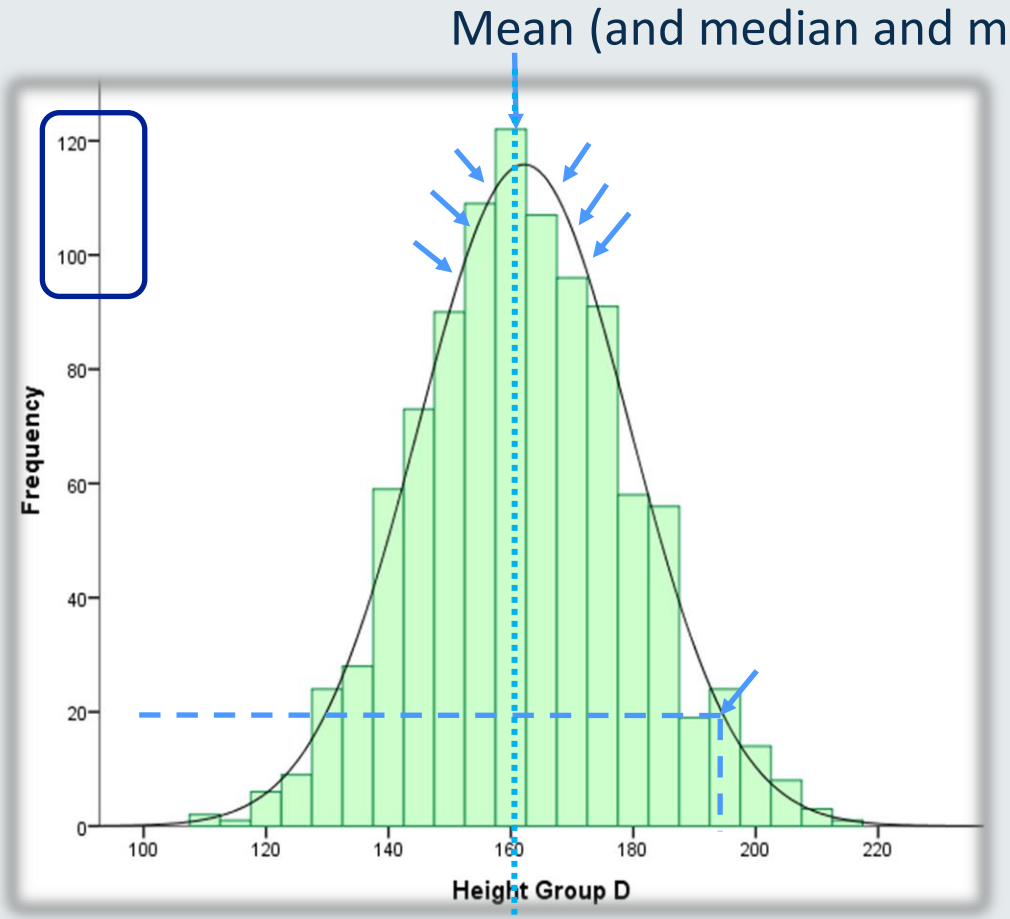


The Normal (Gaussian) Distribution

Because of its tremendous importance, we will focus on this distribution.

The normal distribution looks like a bell. In a normal distribution the mean, median, and mode values coincide.

Mean = 162cm
SD = 17cm

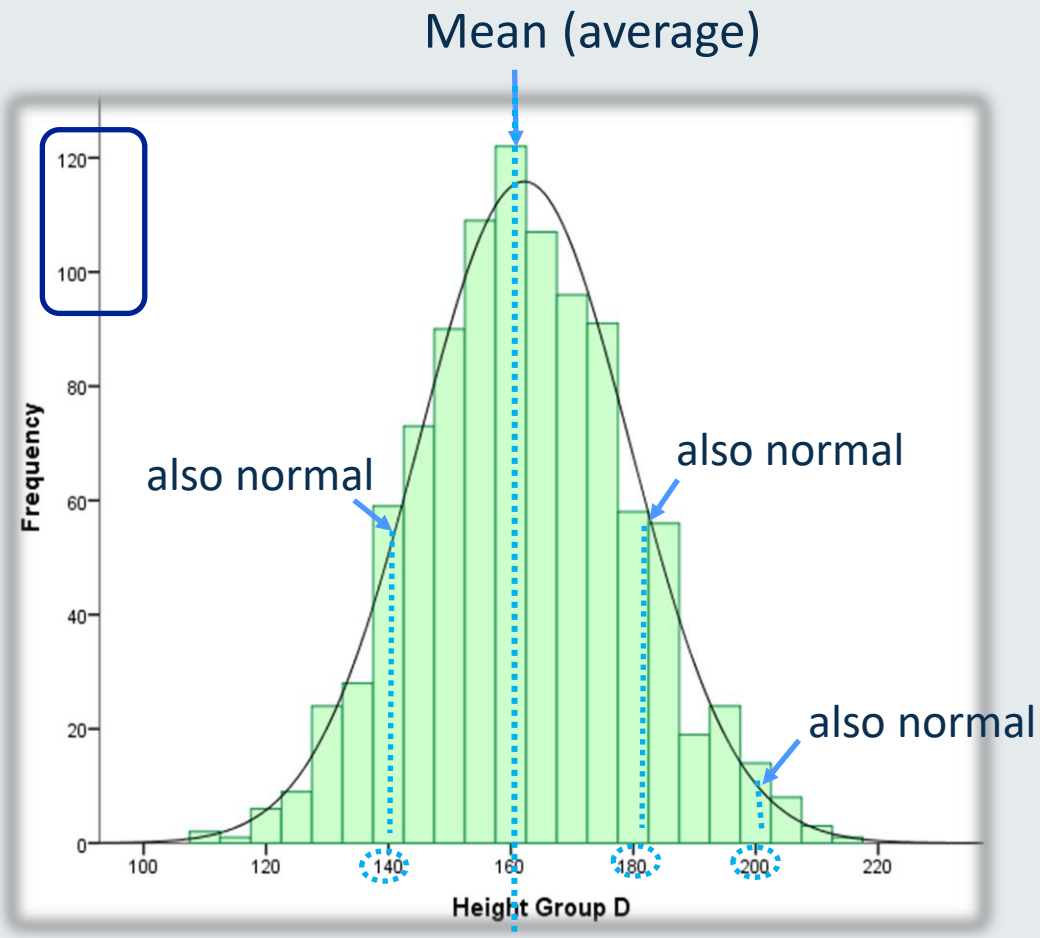


- half of the people (median) have values lower than the average (and half higher than the **average**)
- the most common value (mode) is the average
- The **majority of the people** are close to the average
- As we move away from the average, we have **less** observations.

The Normal (Gaussian) Distribution

So the dominant value in a normal distribution is the average. But beware:

Mean = 162cm
SD = 17cm



The average is a normal value

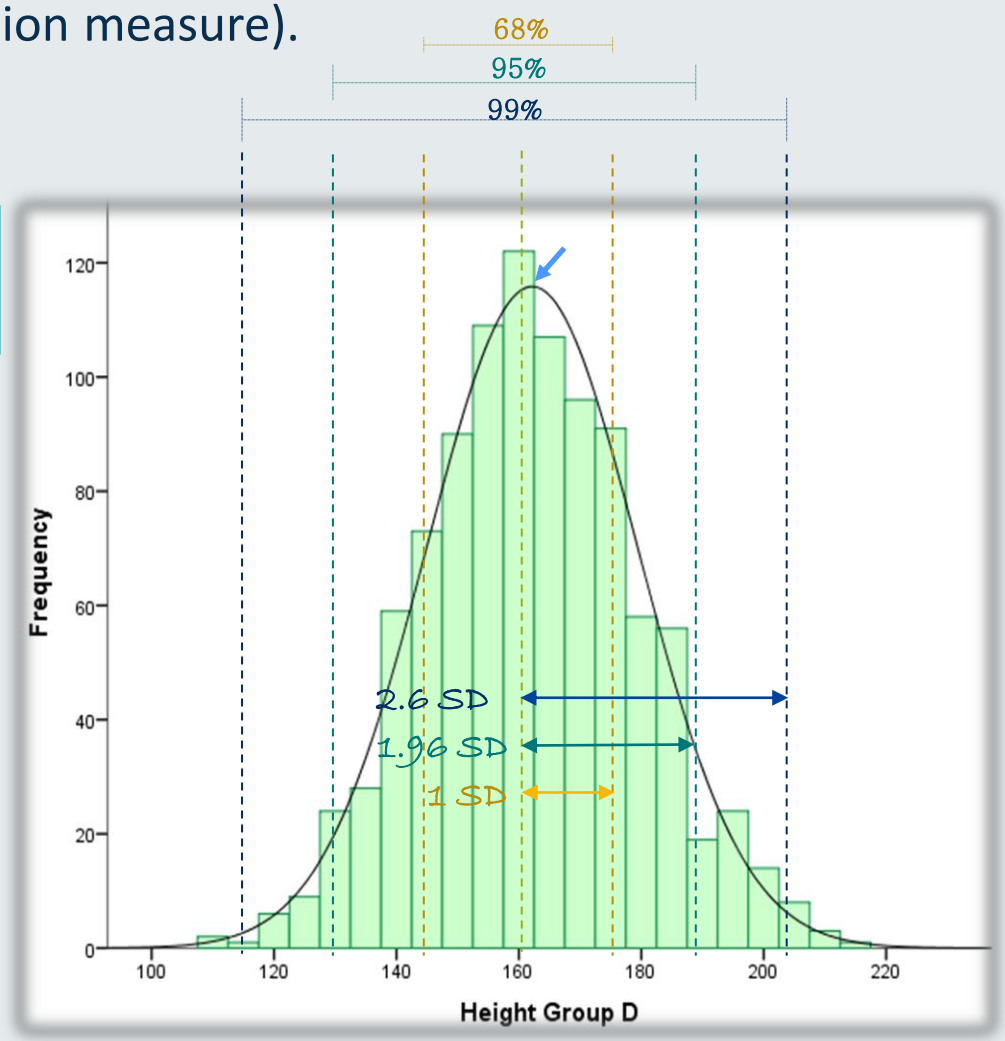
but normal value is not only the average!

The term normal refers to an interval of values, not a point value.

The Normal (Gaussian) Distribution

The normal curve looks like a curve because it is symmetrical around the mean. But what about the standard deviation (dispersion measure).

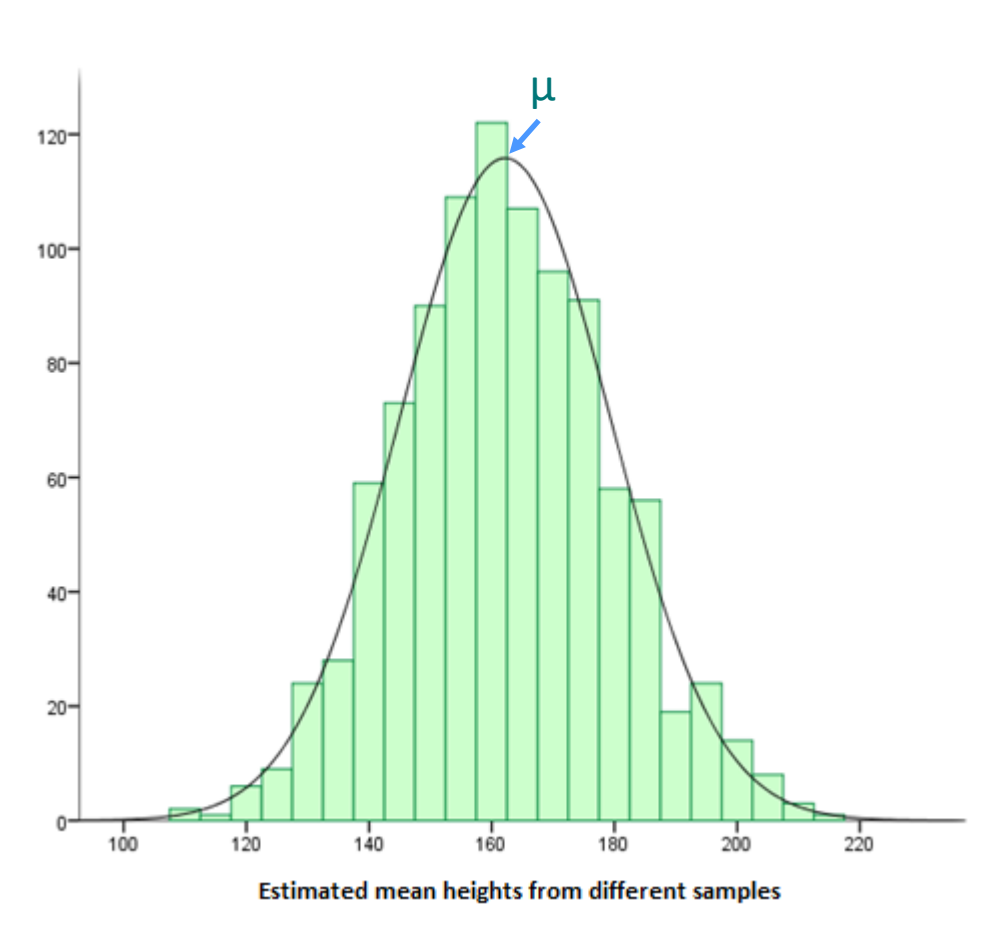
Mean = 162cm
SD = 17cm



- 68% of the observations are in the interval mean plus-minus one SD
- 95% of the observations are in the interval plus-minus 1.96 SD
- 99% of the observations are in the interval plus-minus 2.58 SD

Back to the sampling distribution…

The **sampling distribution** is a normal distribution. Let us now see what are the details of it.



- The mean of sampling distribution will in fact be the mean of the population from which the samples came from.

  That is, the **mean of the samples' means** is actually the **population mean**:

  mean($\bar{x}$)=μ

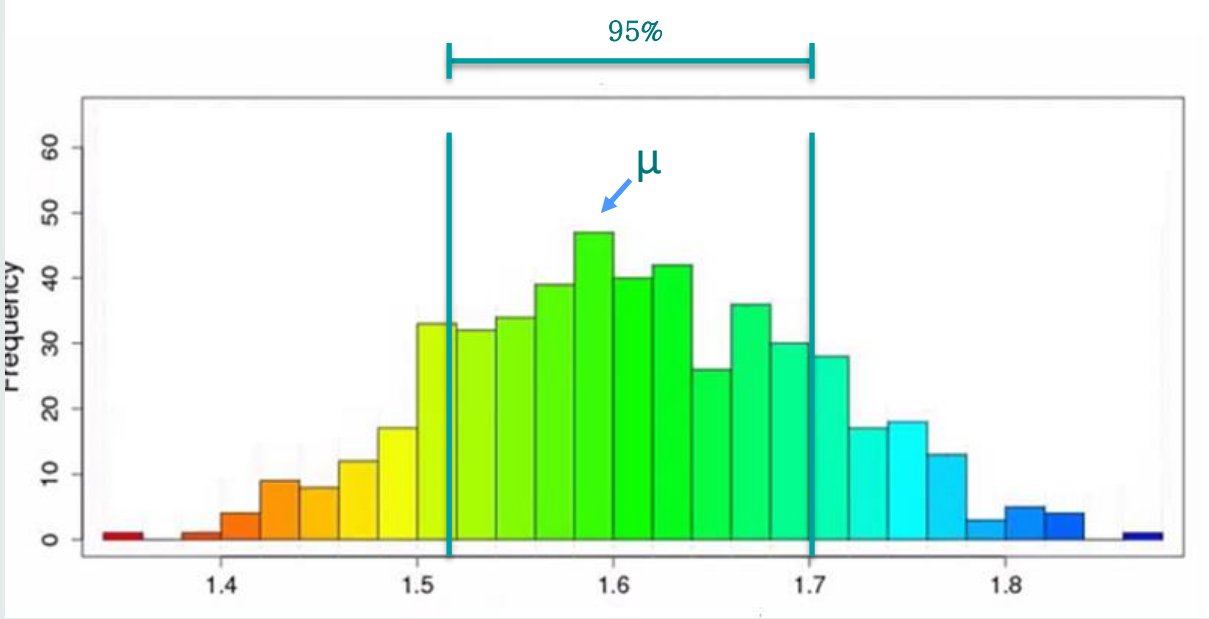
- The **variance** of the **samples' means** is actually the **population variance**, divided by the sample size:

  variance($\bar{x}$)=σ^2 / n

Back to the sampling distribution…

This means, if you and your classmates kept on sampling samples of size n and plot them on a histogram, the distribution of these samples would be a normal distribution with mean and variance:

mean($\bar{x}$)=μ **variance($\bar{x}$)=σ^2 / n**



And because the **sampling** distribution is a **normal** distribution, we also know that 95% of our sampled values will be within the interval plus minus 1.96 SD, that is plus minus 1.96 $\sigma/\sqrt{n}$.

Summarising the sampling distribution…

To summarise, the sampling distribution of the mean is a normal distribution with:

$\text{mean}(\bar{x})=\mu$ $\quad$ $\text{variance}(\bar{x})=\sigma^2 / n$ $\quad$ $SD=\sqrt{\sigma^2 / n}$

The standard deviation of the sampling distribution is called the **standard error**, and it is estimated with the statistic:

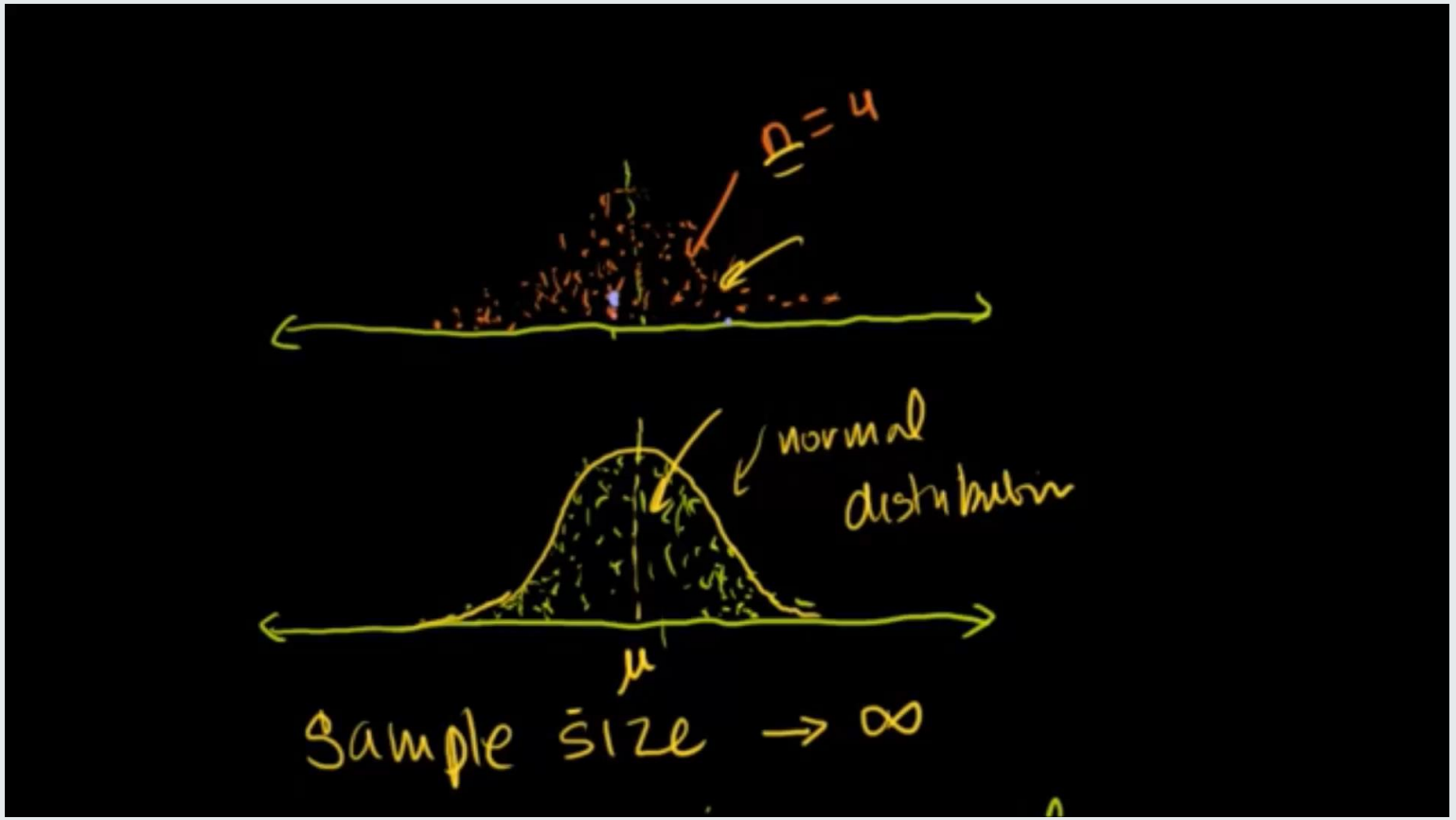
The smaller the variability in our population, the smaller the standard error (random error). Thus, we have greater precision in our estimation.

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

The larger the sample size, the smaller the standard error (random error). Thus, we have greater precision in our estimation.

Summarising the sampling distribution…

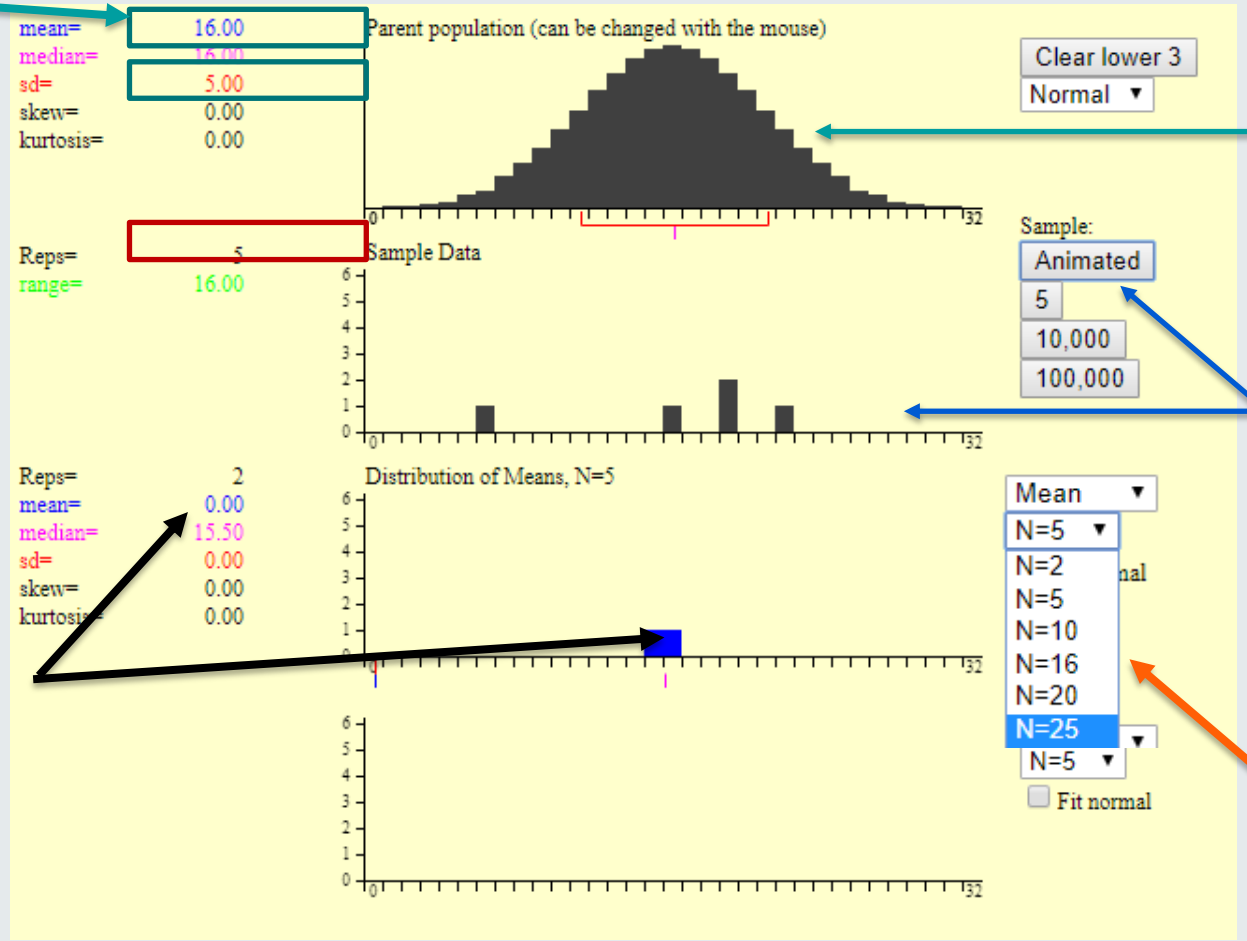
Let us watch a video on the sampling distribution by Khan Academy. It shows that the sampling distribution is a normal distribution regardless of the distribution of the true values.



Summarising the sampling distribution…

Let us explain what you saw (you will have the chance to experiment with the app during the lab)

population
μ and σ



This is the population

By pressing animated, five values are sampled from the population (too small for CLT)

sample mean

You can change how many values are sampled from the population (as this number increases, your sampling distribution will become normally distributed, even if the population is not normally distributed).

Knowledge Check

1. If ten people sample from the same population to estimate the mean weight:
   a) ~~they will all compute the same estimated value~~
   b) the estimated values they will come up with will not be identical

   Due to sample variation (random error) the estimated values will differ. How much they differ depends on the variability in the population and on the sample size.

2. The sampling distribution of the mean is
   a) ~~the distribution of the sampled data~~
   b) the distribution of the means of multiple sampled data

   The sampling distribution is the distribution of the estimated values from different samples

Reflection

Thinking about your own research

- Describe how increasing your sample would affect your results in your study. Why is that?
- Search the literature on your field of research. Does each paper present the same estimated values?

Reference List

For more details of the concepts covered in Session 1, see Chapters 1- 3 of the book:
Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences (4th Edition), Prentice Hall Inc.hapters 1-3

For more details on SPSS implementation see:
Field (2005) Discovering Statistics using SPSS 2nd Edn, Sage, London.
The SPSS Environment, Ch 2.

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

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**

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: Confidence intervals

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 the idea of confidence intervals (CIs)
- To learn how to analytically compute a CI based on one sample
- To learn how to compute a CI on SPSS

Summary (continued)

Let us summarise what we learned so far:

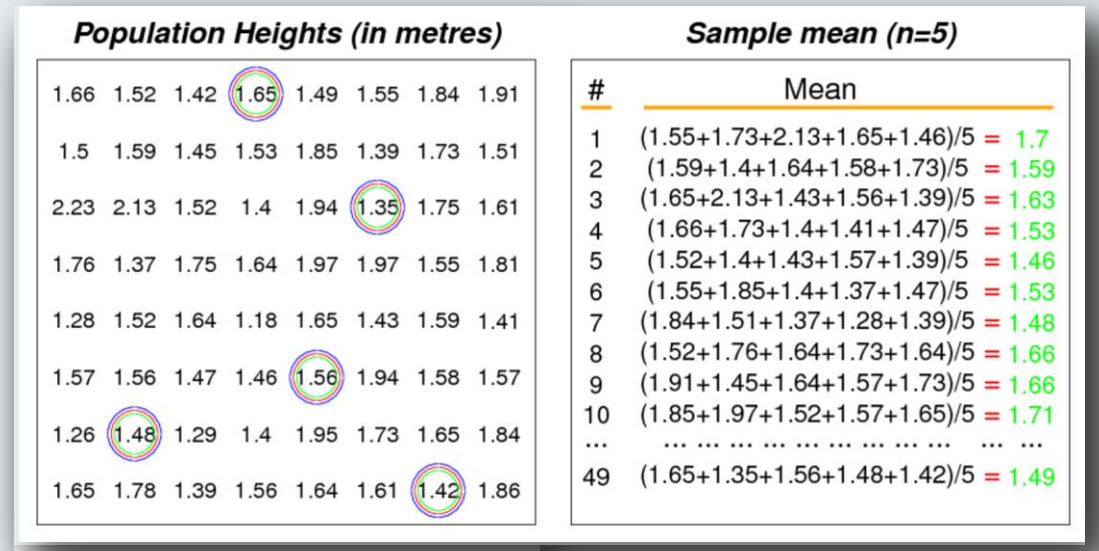
*We wish to **infer** on the value of a **parameter** in the population*

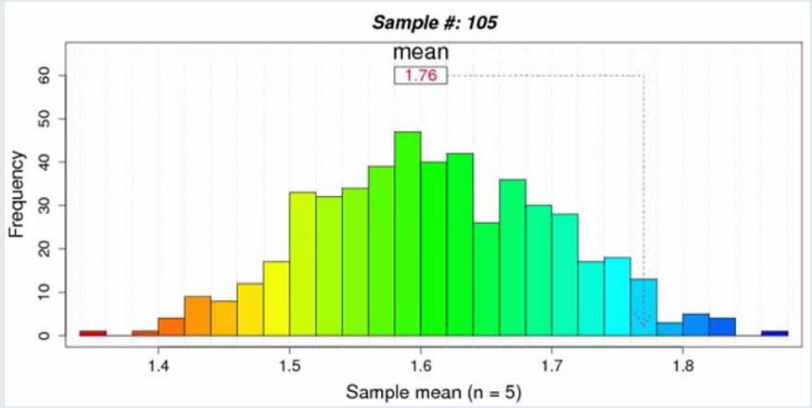
*As we do not have access to the entire population, we use a **sample** to estimate the parameter.*

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).





Summary (continued)

The sampling distribution is a normal distribution

whose mean **mean($\bar{x}$)** *is the population mean* **μ**

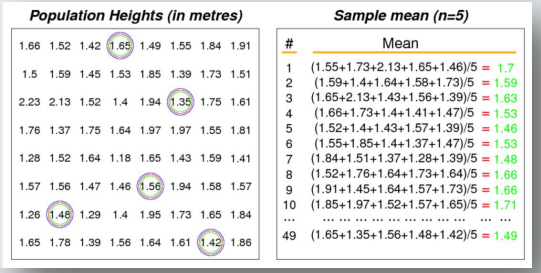
$$\text{mean}(\bar{x})=\mu$$

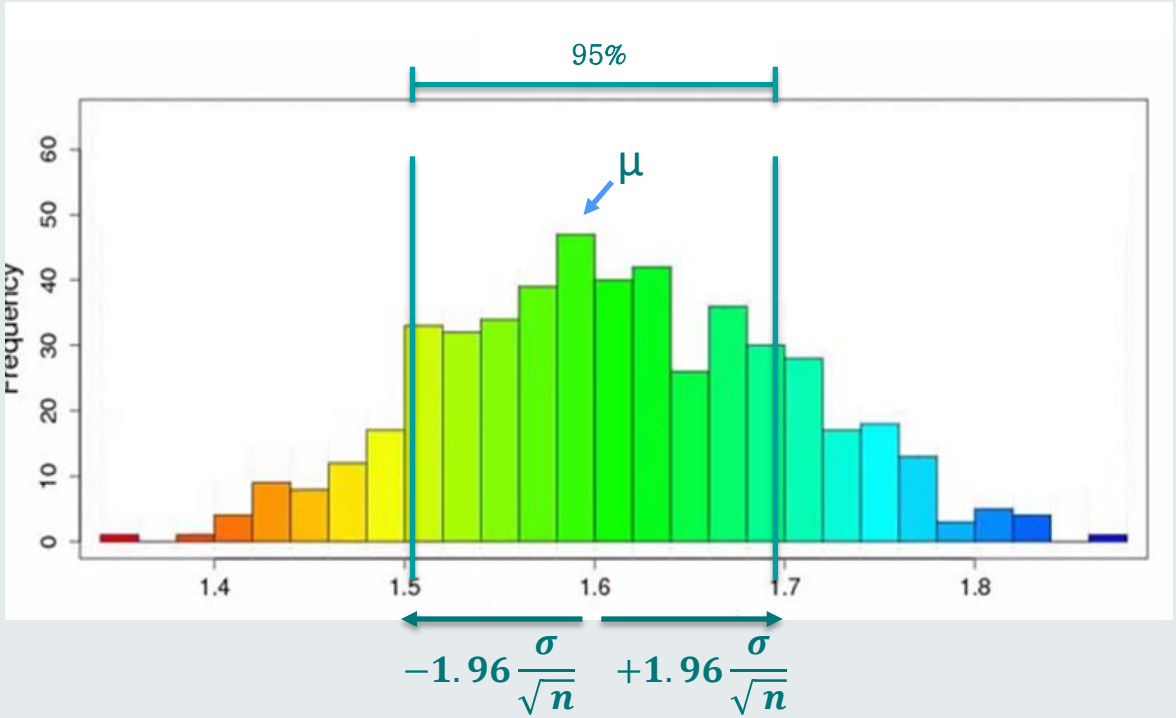
and its standard deviation is the standard deviation of the population σ divided by the square root of our sample size (we call this the standard error).

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

We also know that as the sampling distribution is a normal distribution, 95% of its values will lie in the interval plus minus 1.96 SE.

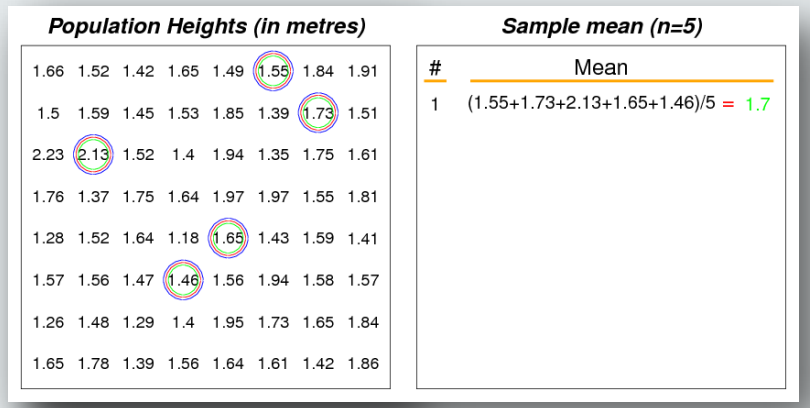
That is, if we sample 100 samples from the same population of size n, then in 95 of them the estimated mean will be within this range.





Confidence intervals

In research what we most often have, is one sample. We compute in this sample the statistic of interest, say in our current example the sample mean



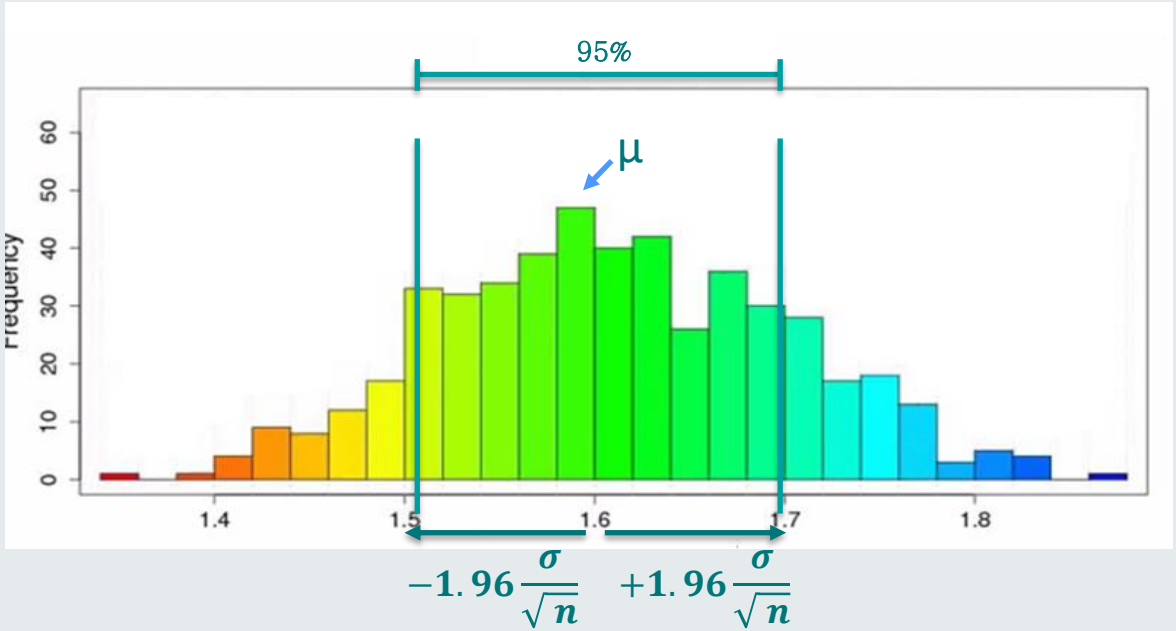
I don't know those!

Population mean μ and standard deviation σ

Sample mean $\bar{x}$ and standard deviation s

I can compute those!

Sampling distribution mean($\bar{x}$)=μ and standard deviation **SE**=$\frac{\sigma}{\sqrt{n}}$



Remember, the sample mean estimates the population mean, and the sample standard deviation estimates the population standard deviation….

Confidence intervals

In research what we most often have, is one sample. We compute in this sample the statistic of interest, say in our current example the 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

I don't know those!

Population mean **μ** and standard deviation **σ**

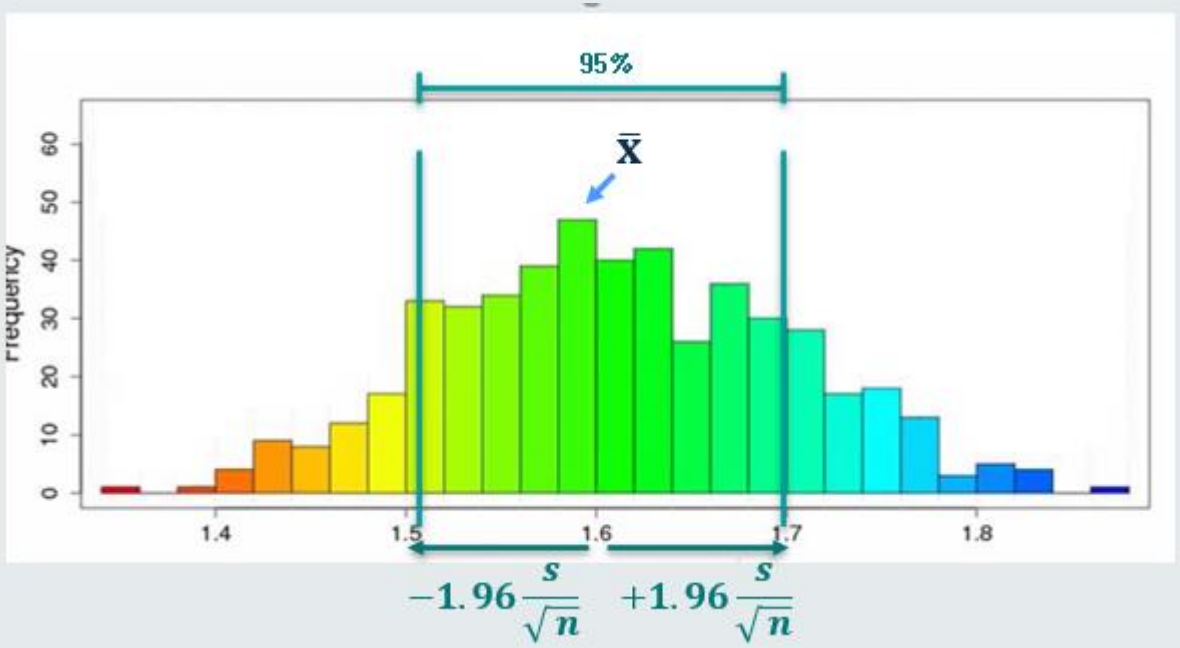
Sample mean $\bar{x}$ and standard deviation **s**

I can compute those!

Estimated sampling distribution mean=$\bar{x}$ and standard deviation **SE=**$\frac{s}{\sqrt{n}}$

Using the estimated values from one sample, we can draw an approximation of the sampling distribution.

Then I can say with 95% confidence, that this interval contains the true, population mean, using one sample.



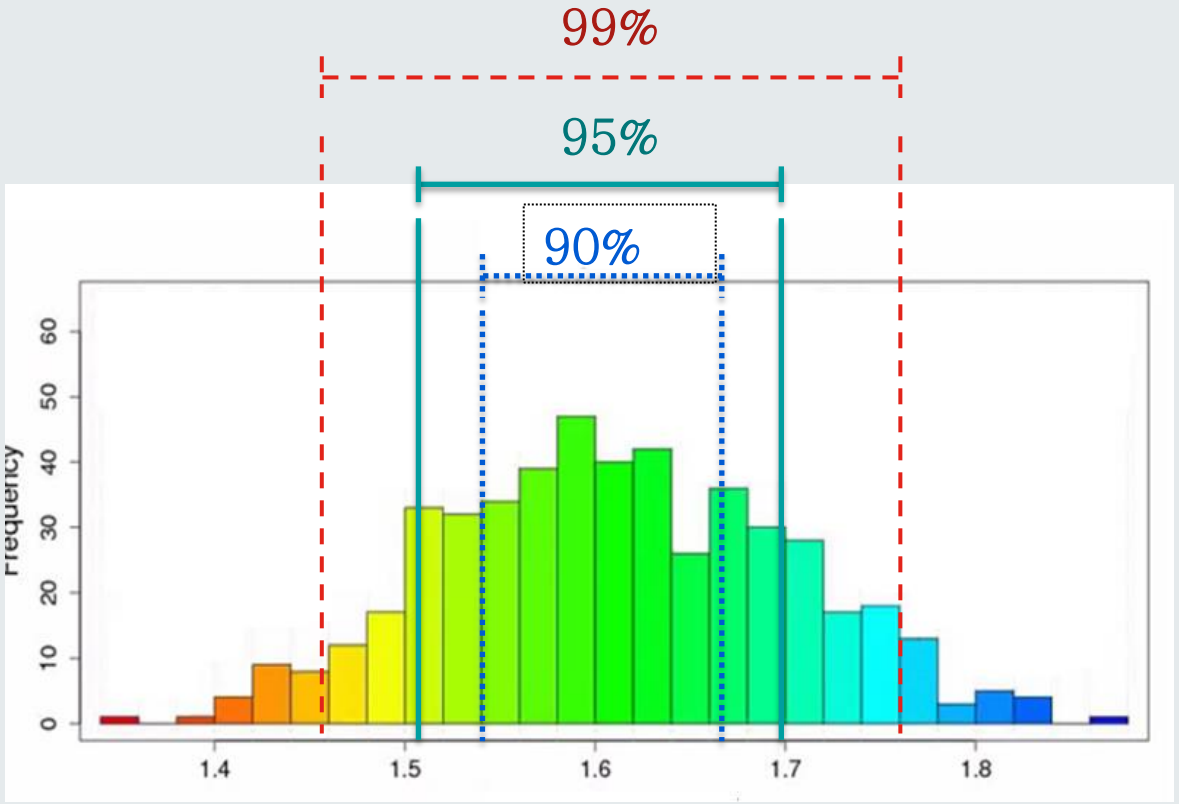
Confidence intervals

So having only one sample, I can estimate the parameter of interest. This estimate is called the point estimate. Using the properties of the sampling distribution (the distribution that I would have had if I had the time and resources to repeat the experiment, say, 100 times) I can also compute a confidence interval for my estimations, that is a range of values that I am confident to a certain value that contains the true, population value.

population	sample		
N	n	$[\bar{x} - 1.65\frac{s}{\sqrt{n}}, \bar{x} + 1.65\frac{s}{\sqrt{n}}]$	*I am 90% confident that this interval contains the population mean*
μ	$\bar{x}$	$[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}]$	*I am 95% confident that this interval contains the population mean*
σ	s	$[\bar{x} - 2.58\frac{s}{\sqrt{n}}, \bar{x} + 2.58\frac{s}{\sqrt{n}}]$	*I am 99% confident that this interval contains the population mean*

Note: To construct these confidence intervals we have approximated the sampling distribution by a normal distribution with mean=$\bar{x}$ and standard deviation **SE=**$\frac{s}{\sqrt{n}}$. This approximation will not work well for small samples (n<30) where instead we preferably use the t-distribution instead. This means, that instead for example to use the so called z-value 1.96 for the 95% CI, we would have used the corresponding t-value for the given sample size. We will not expand further in this module on the t-distribution as this goes beyond the purposes of this course.

Confidence intervals



$$[\bar{x} - 1.65\frac{s}{\sqrt{n}}, \bar{x} + 1.65\frac{s}{\sqrt{n}}]$$

$$[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}]$$

$$[\bar{x} - 2.58\frac{s}{\sqrt{n}}, \bar{x} + 2.58\frac{s}{\sqrt{n}}]$$

Confidence intervals

The wider the interval, the more confident we are the population mean will be included.

Let us consider the scenario that we are asked to estimate the mean age of the participants of this course

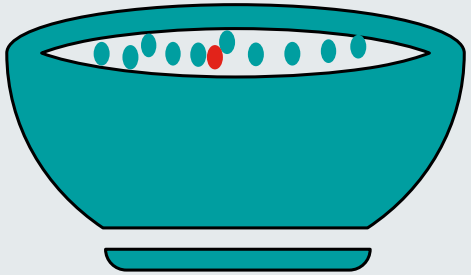
I might say that I am quite confident that my class mean age should range between 20 and 30 years old.

But if my life had dependent on it, I would give an interval between 18 and 100 years old. I would be extremely certain that your mean age is actually in this interval.

remember: it is a confidence, not accuracy or certainty!

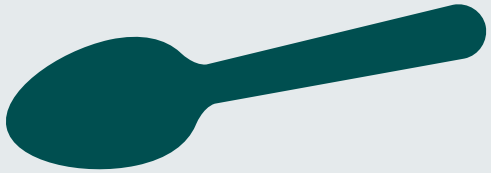
Confidence intervals

Imagine a bowl with candies and you would like to have the red one, but you can not see them.



90%

95%

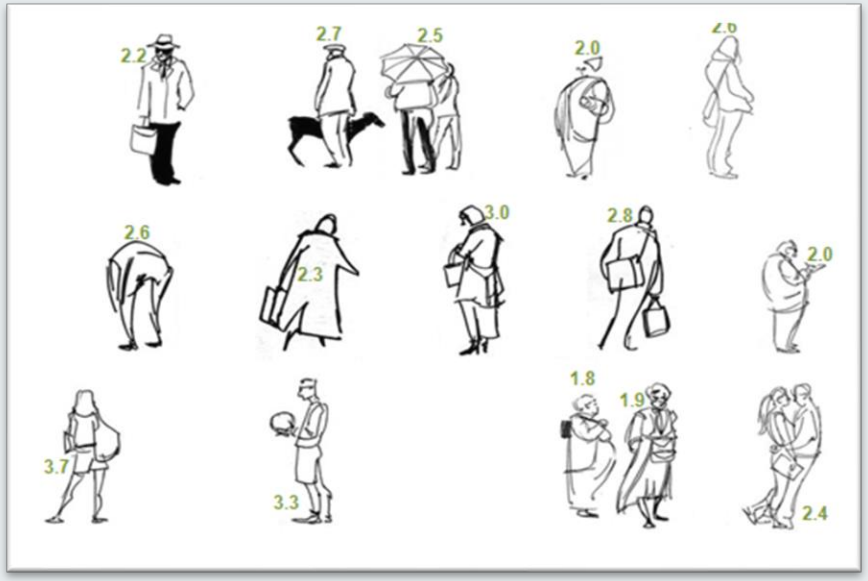


99%

The larger the spoon you use, the more confident you are it might contain the red candy!

Confidence Intervals Example

Say, out of a population in a city, we sampled **140** people. Based on my sample, the estimated mean hours they spend exercising was **2.72** hours per week, with estimated standard deviation **0.62**.



n=140

$\overline{x}$=2.72

s=0.622

s.e.=0.622 / $\sqrt{140}$=0.053

I use these values to estimate the 95% confidence interval

$$[\overline{x} - 1.96\ s/\sqrt{n}, \overline{x} + 1.96\ s/\sqrt{n}]$$

$$[2.72 - 1.96 * 0.053, 2.72 + 1.96 * 0.053]$$

$$= [2.617, 2.823]$$

I can be 95% confident, that the population mean will be in the interval

$$\underset{\textit{lower bound}}{[2.617,} \quad \underset{\textit{upper bound}}{2.823]}$$

90% CI

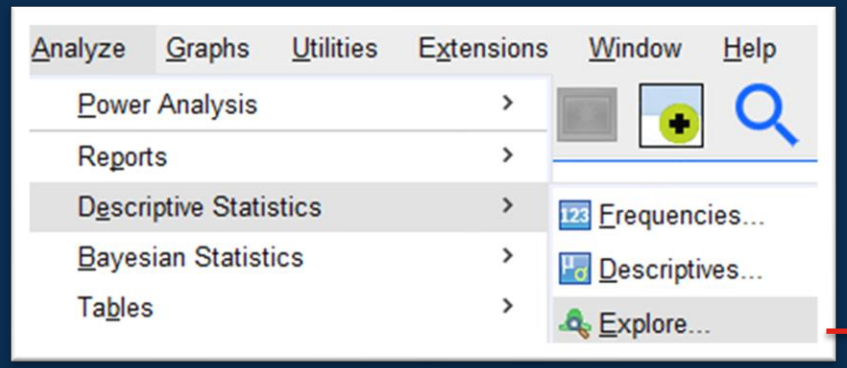
$$[2.633, 2.808]$$

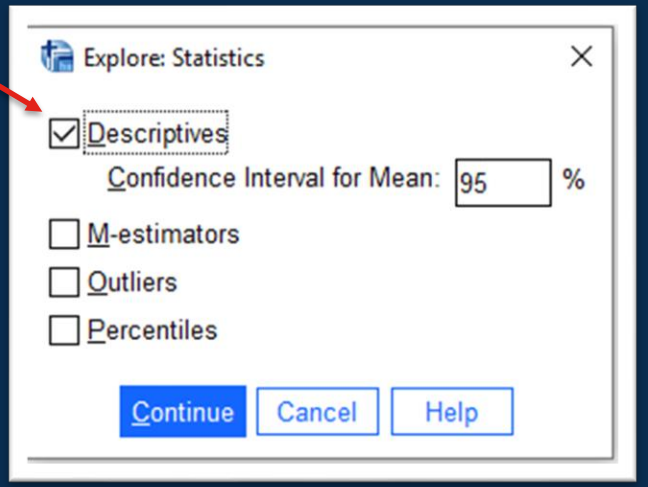
99% CI

$$[2.584, 2.856]$$

SPSS slide: ‘how to’

Analyse-> Descriptive Statistics-> Explore -> put the variable in ‘Dependent list’-> Statistics-> Change the CI if you want to.





Interpretation slide

Descriptives

			Statistic	Std. Error
Height (cm)	Mean		168.5253	1.02324
	95% Confidence Interval for Mean	Lower Bound	166.4886	
		Upper Bound	170.5620	
	5% Trimmed Mean		168.8901	
	Median		168.9280	
	Variance		83.762	
	Std. Deviation		9.15218	
	Minimum		137.03	
	Maximum		191.84	
	Range		54.81	
	Interquartile Range		10.21	
	Skewness		-.712	.269
	Kurtosis		1.316	.532

$$[\bar{x} - 1.96\, s/\sqrt{n}, \bar{x} + 1.96\, s/\sqrt{n}]$$

$$[168.5 - 1.96 * 1.02, 168.5 + 1.960.102) = [166.5, 170.5]$$

Confidence Intervals

We focused on the sampling distribution of the mean. But the same hold for other 'statistics':

Parameter	
Population mean	$\mu = 2.66$
Population SD	$\sigma = 0.57$
Population variance	$\sigma^2 = 0.33$
Population proportion	$\pi = 0.20$

Statistic	
Sample mean	$\bar{x} = 2.72$
Sample SD	$s = 0.62$
Sample variance	$s^2 = 0.38$
Sample proportion	$p = 0.18$

Their sampling distribution is normal, with mean the population parameter.

Confidence Intervals

Let us for example consider a proportion, say the proportion of women in a population.

Let us denote the proportion in the population with π

and the estimated proportion based on our sample with p.

Then the standard error is given by se= $\sqrt{\frac{p(1-p)}{n}}$

The 95% CI for the population proportion π is given by

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{n}},\ p + 1.96\sqrt{\frac{p(1-p)}{n}}\right]$$

The sampling distribution of π will approximate the normal distribution.

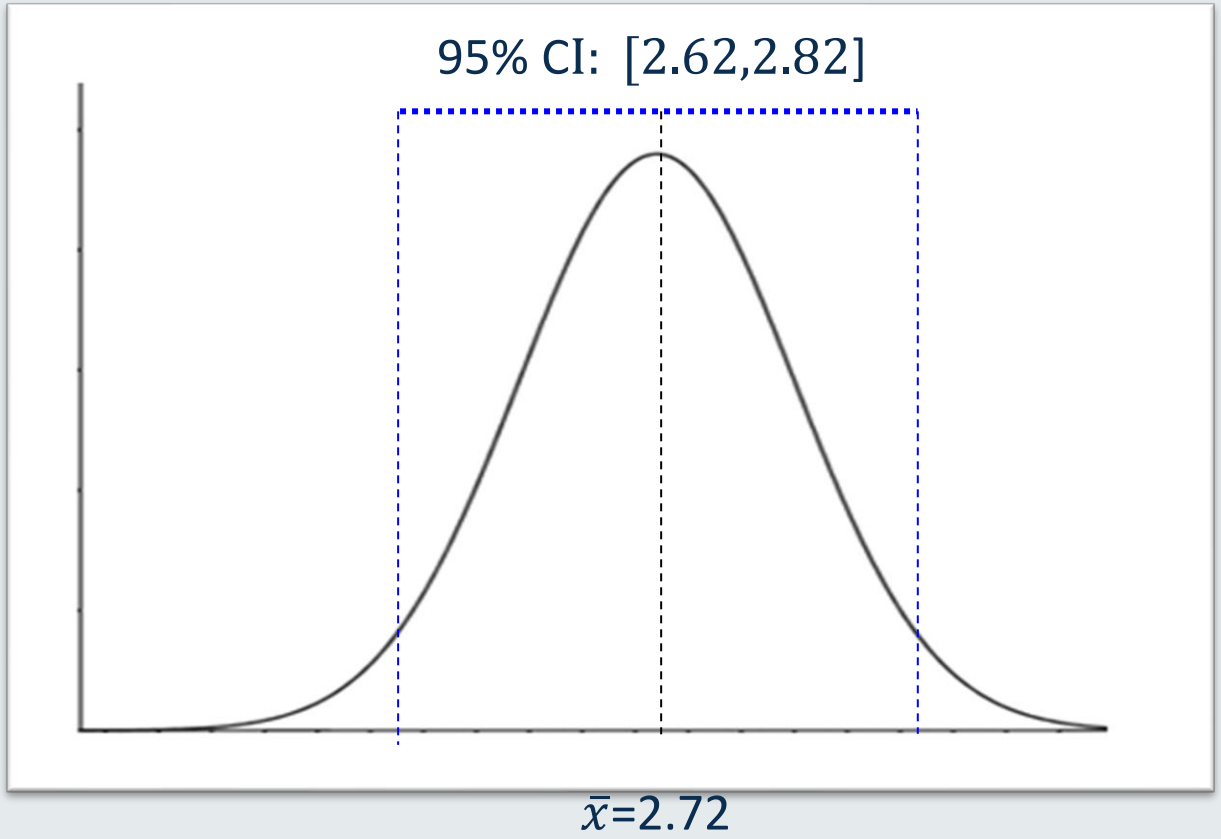
For rare events this may require large sample sizes.

Knowledge Check

1. We sampled 140 people and the mean hours they spend exercising was 2.72 hours per week, with a standard deviation of 0.62. Please compute the 95% confidence interval

$\bar{x}$=2.72
s=0.62
s.e.=0.622/ $\sqrt{140}$ = 0.052
Lower Limit = 2.72 - 1.96 * 0.052 =2.617 h/w
Upper Limit = 2.72 + 1.96 * 0.052 =2.823 h/w

Based on our data, the estimated mean hours per week the people spend to exercise was 2.72 (95% CI: [2.62,2.82]).



2. Between the two intervals below, please select the one that you think it is the 99% confidence interval and which one the 95%.

a) 95% CI: [19, 22]

b) 99% CI: [29,52]

Reflection

A paper provides a 95% confidence interval for the proportion of violent offenders in prisons (in a certain area) as ranging from 0.3 to 0.5. Describe in words what this tells you.

Reference List

For more details of the concepts covered in Session 1, see Chapters 1- 3 of the book:
Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences (4th Edition), Prentice Hall Inc.hapters 1-3

For more details on SPSS implementation see:
Field (2005) Discovering Statistics using SPSS 2nd Edn, Sage, London.
The SPSS Environment, Ch 2.

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

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**