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

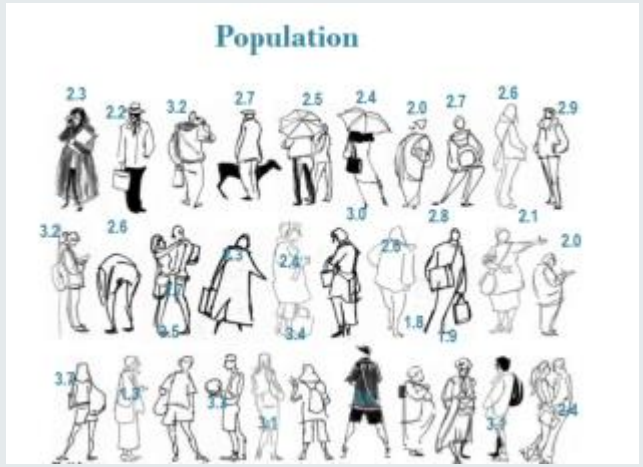
Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

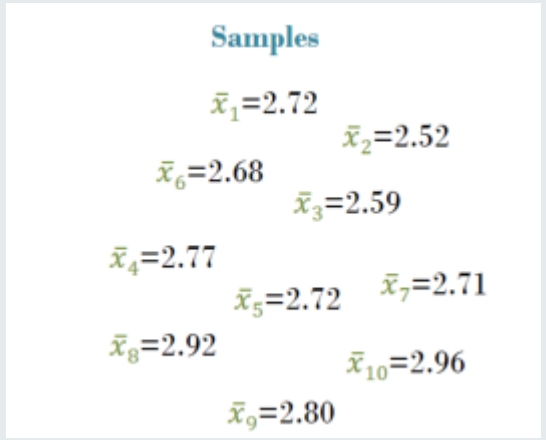
Learning Outcomes

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

The Normal (Gaussian) Distribution

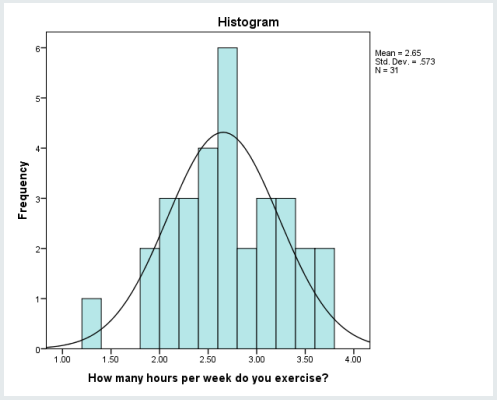
The distribution of these estimated means, from a number of samples, is called the **sampling distribution** of the mean.



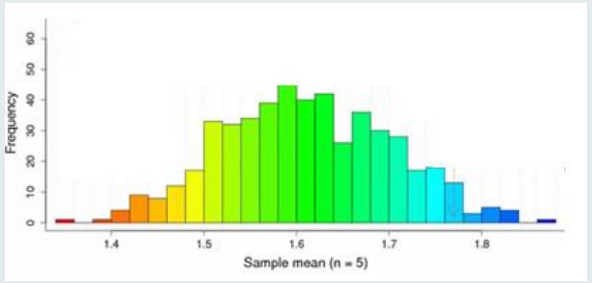


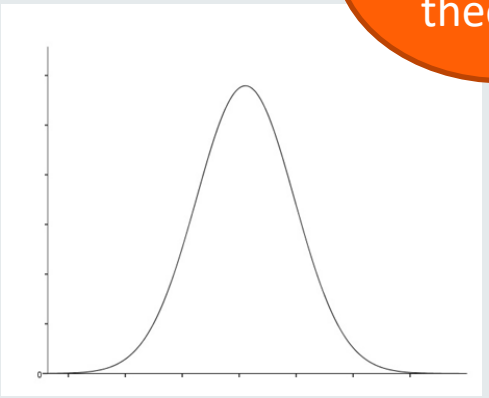
Theory states that if we take more and more and more samples from the same population, then the sampling distribution of the statistic mean will be a **normal distribution**.

Central limit theorem



Sampling distribution





The Normal (Gaussian) Distribution

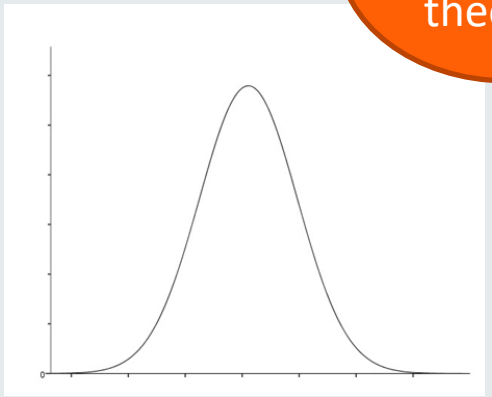
The normal distribution is called as such because a lot of events in real life follow this bell-shaped pattern. It is the norm (the rule).

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.

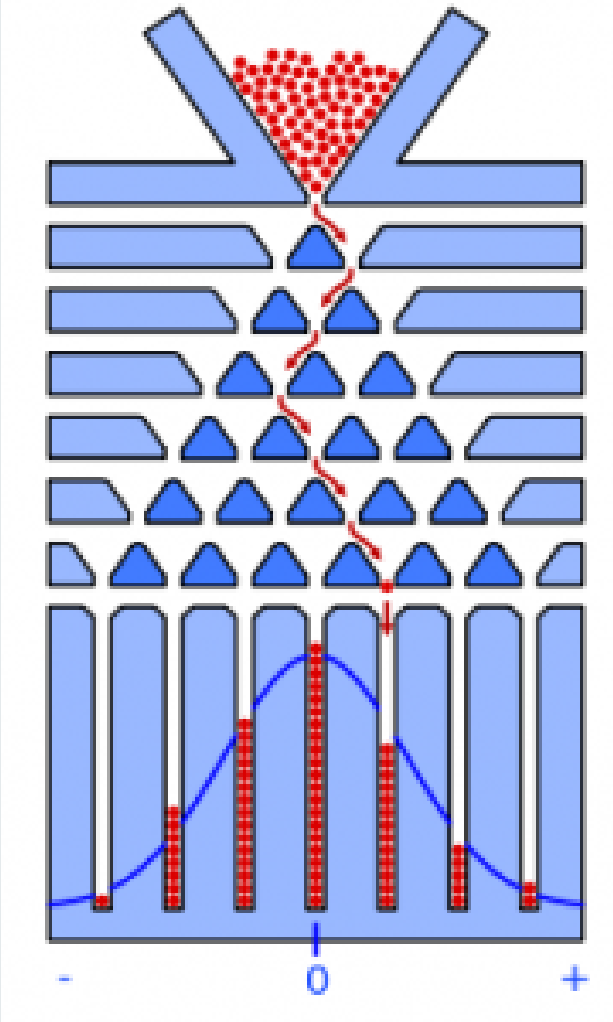
Video demonstrating the Galton Board from the Large Maths Outreach and Careers Kit developed by the Institute of Mathematics and its Applications as part of the National HE STEM Programme.
https://www.youtube.com/watch?v=6YDHBFVIvIs

Central limit theorem

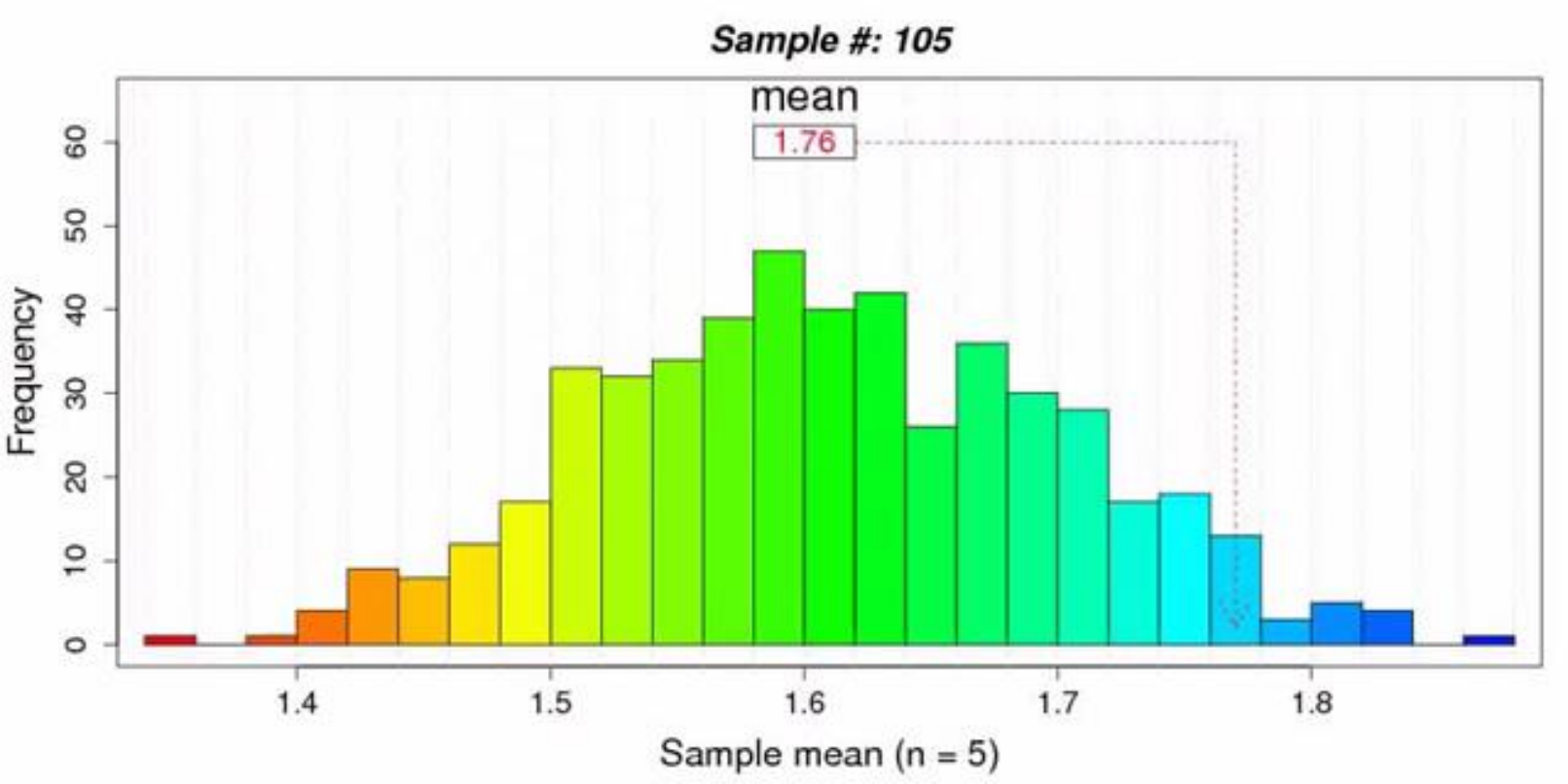


The Normal (Gaussian) Distribution

Galton Board



Sampling distribution of the means

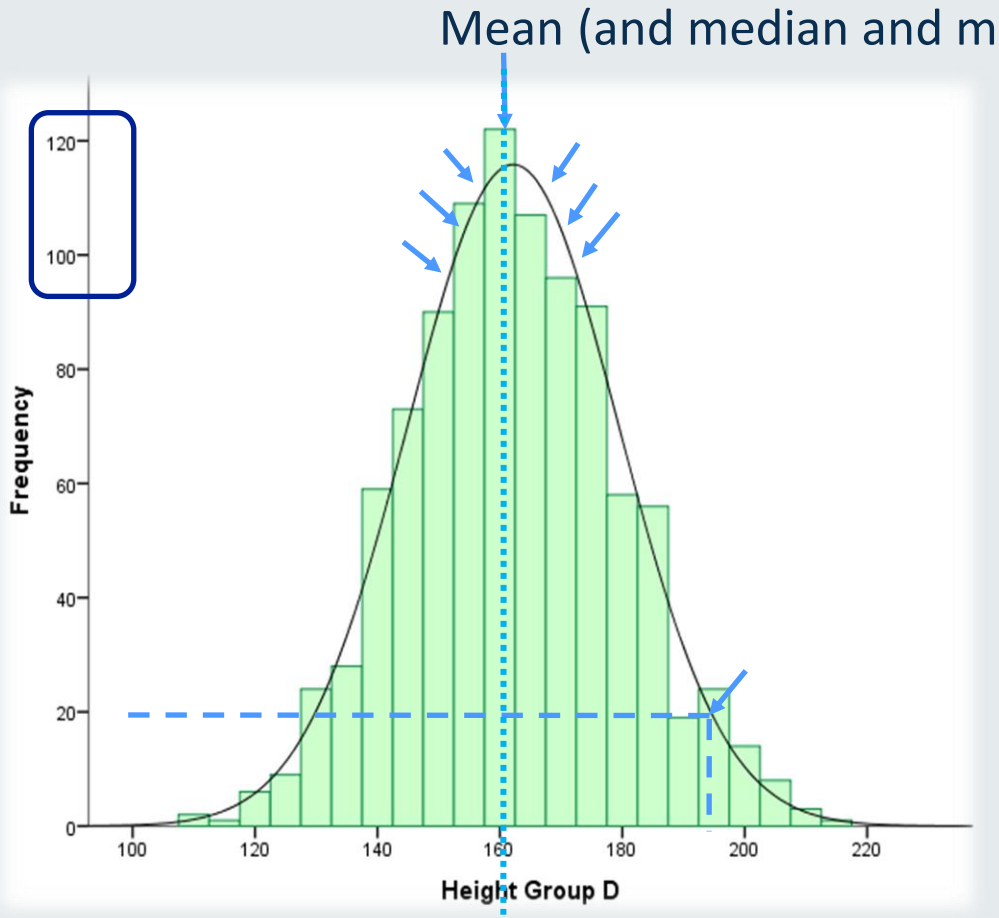


The Normal (Gaussian) Distribution

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

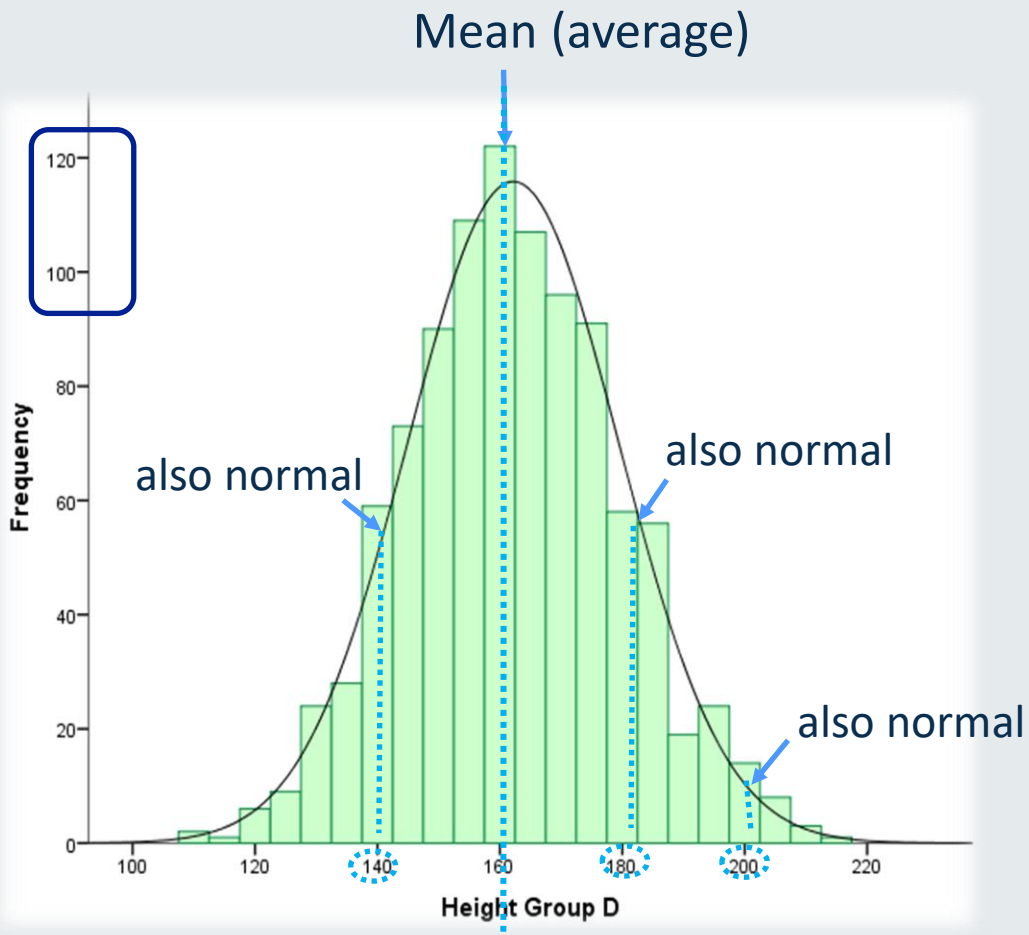


- half of the people (median) have values lower than the average (and half higher than the **average**)
- the most common value (mode) is the average
- The **majority of the people** are close to the average
- As we move away from the average, we have **less** observations.

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

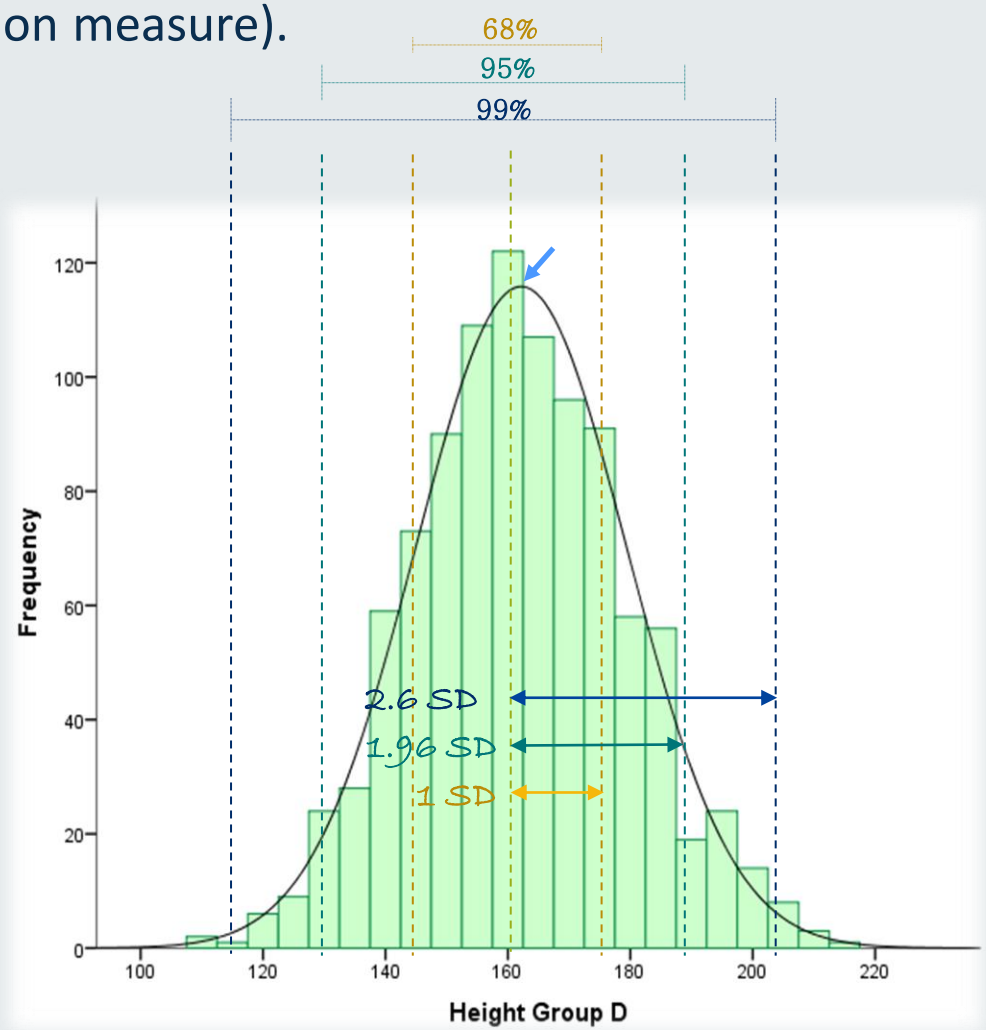
but normal value is not only the average!

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).

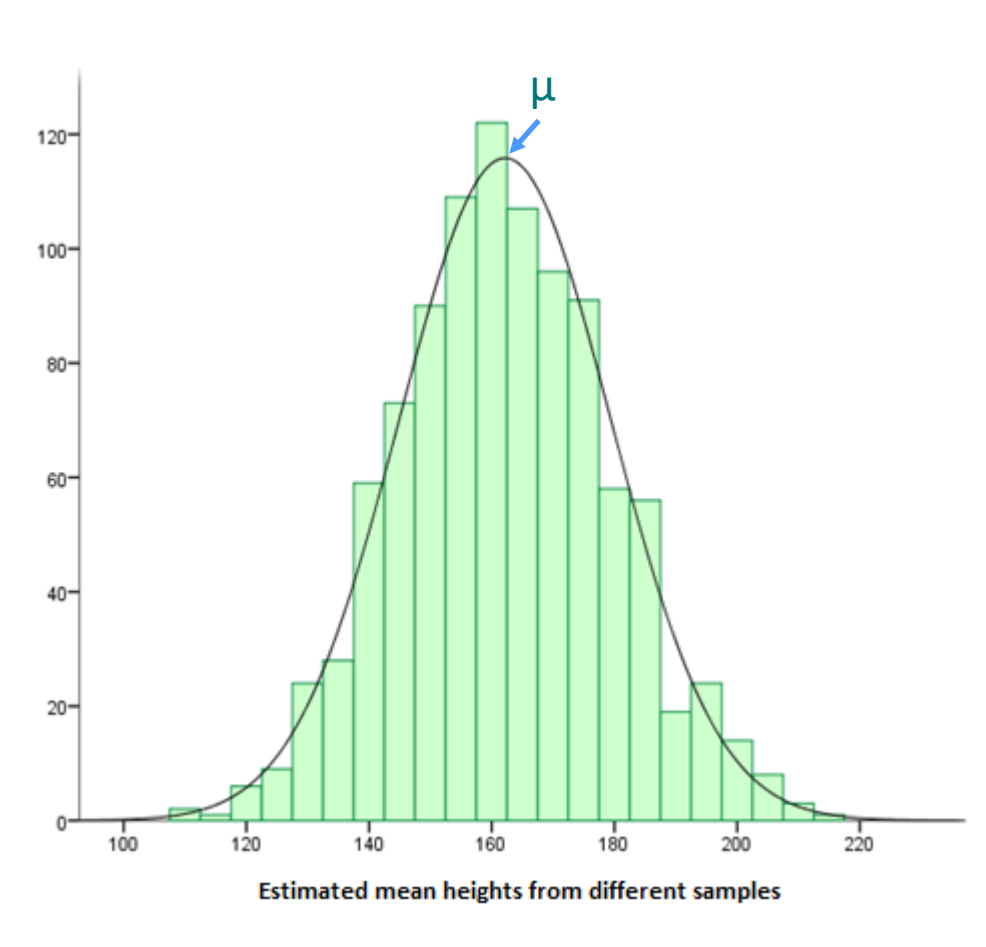
Mean = 162cm
SD = 17cm



- 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

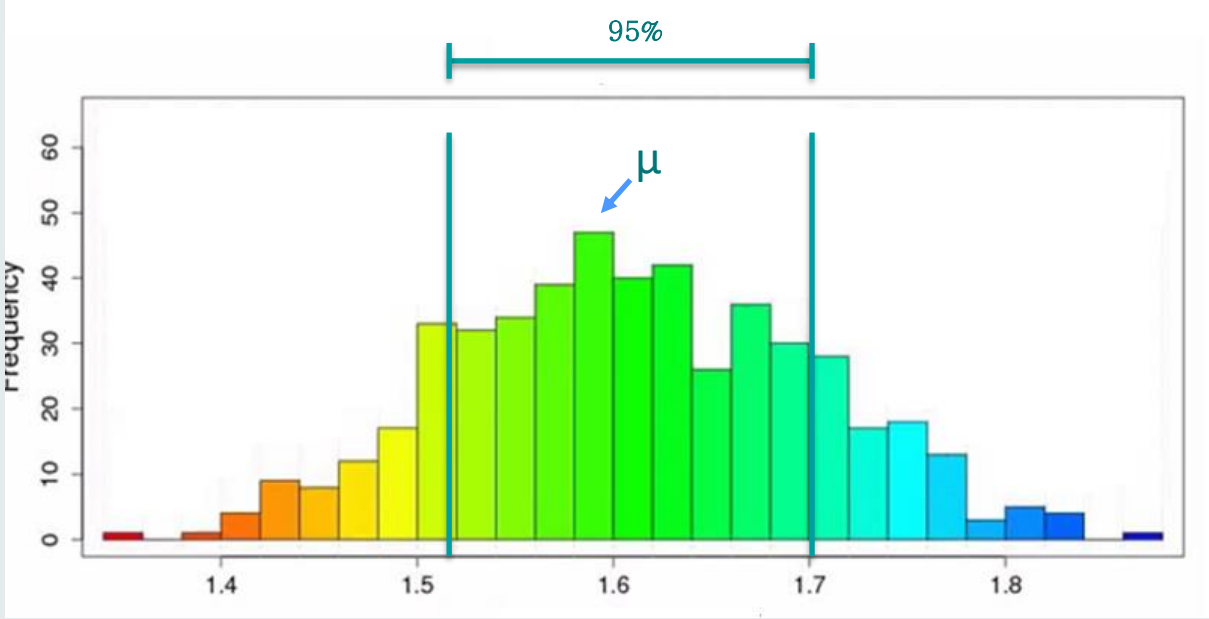
- The **variance** of the **samples' means** is actually the **population variance**, divided by the sample size:

  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:

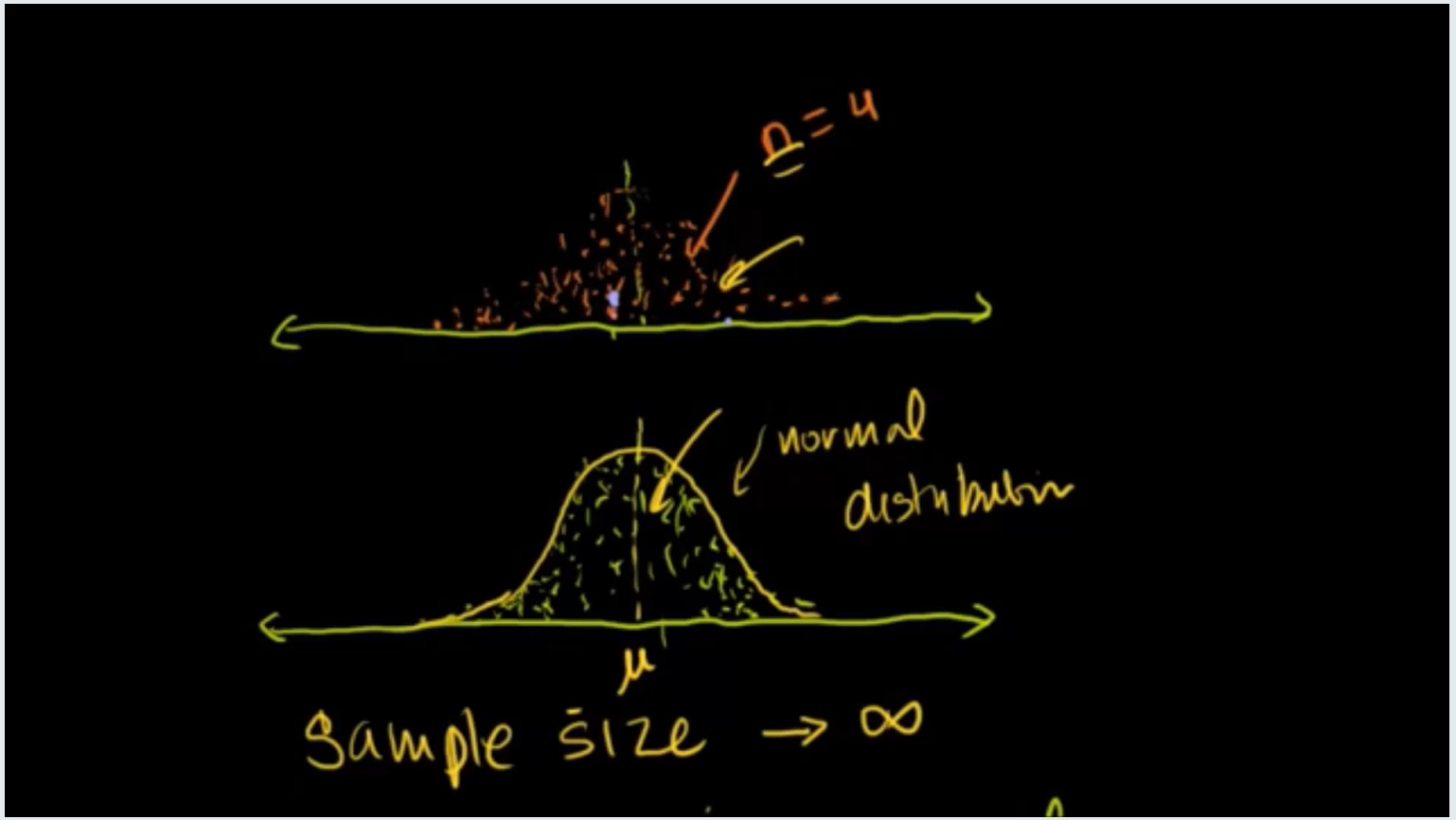
The smaller the variability in our population, the smaller the standard error (random error). Thus, we have greater precision in our estimation.

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

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

Summarising the sampling distribution…

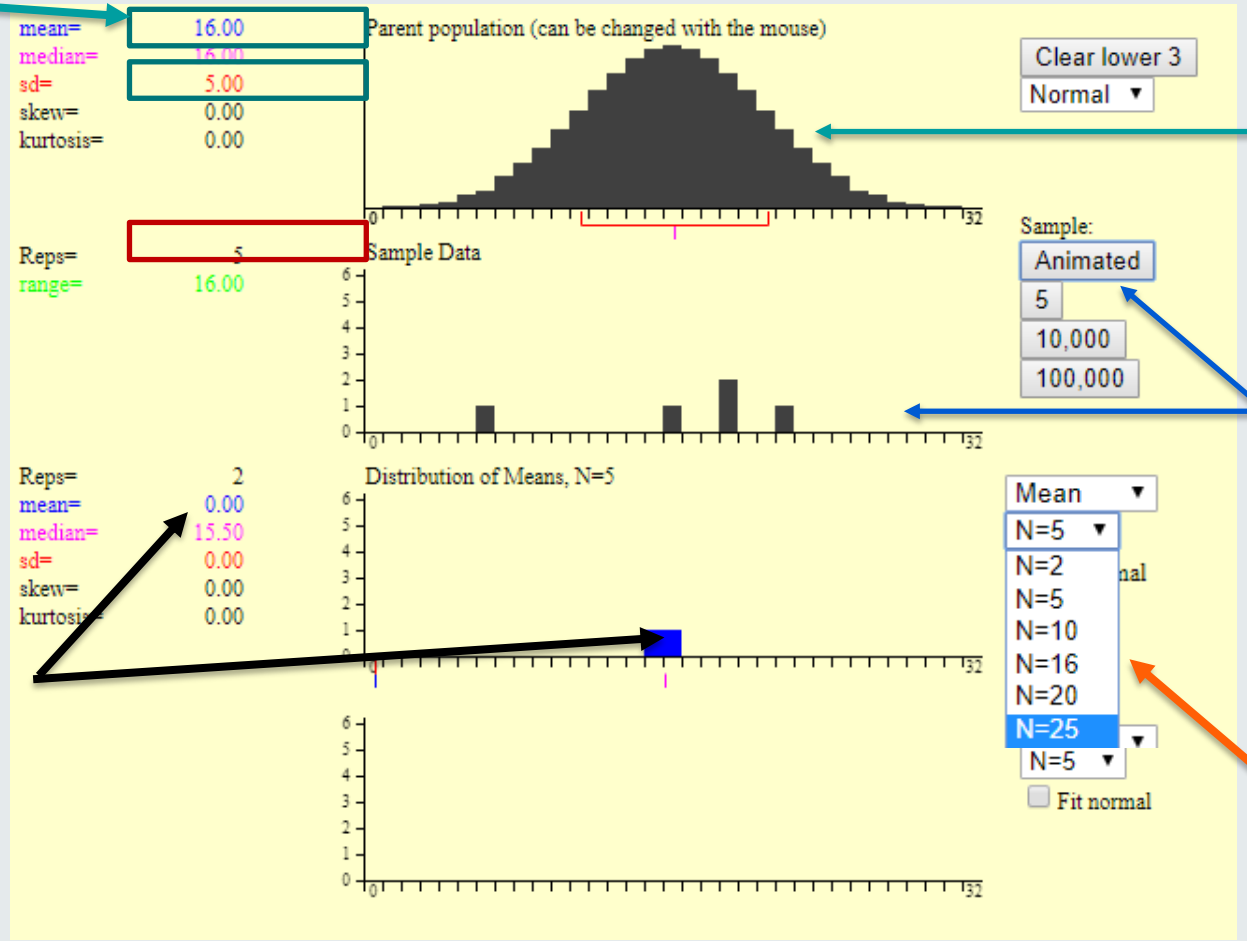
Let us watch a video on the sampling distribution by Khan Academy. It shows that the sampling distribution is a normal distribution regardless of the distribution of the true values.



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**