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:

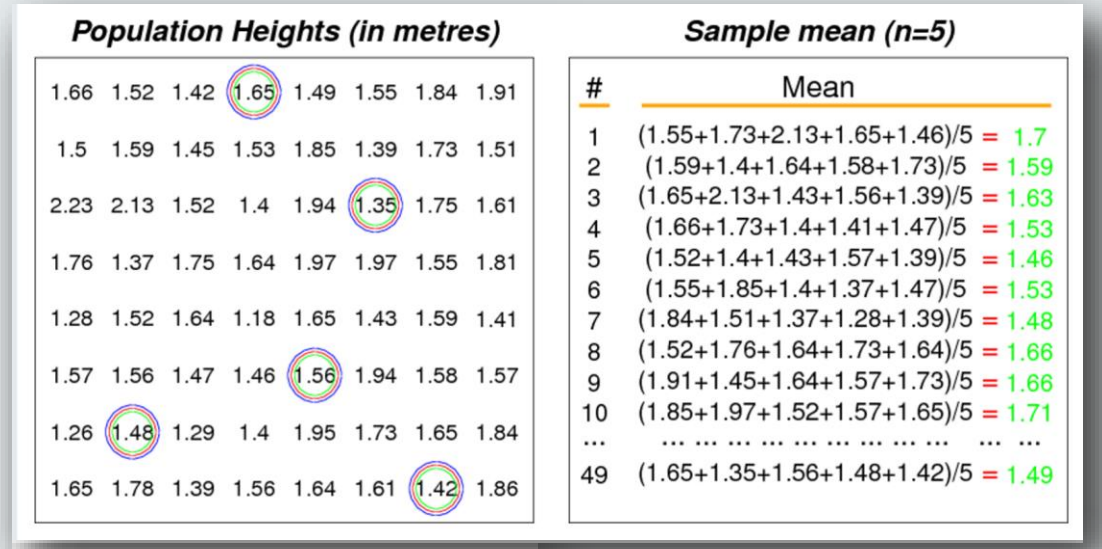
*We wish to **infer** on the value of a **parameter** in the population*

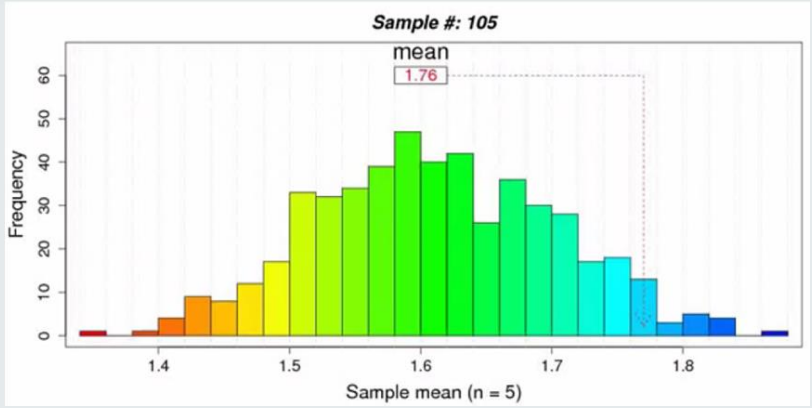
*As we do not have access to the entire population, we use a **sample** to estimate the parameter.*

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**

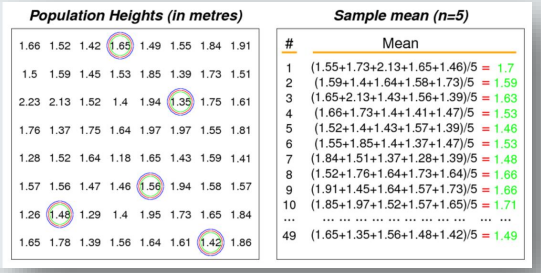
$$\text{mean}(\bar{x})=\mu$$

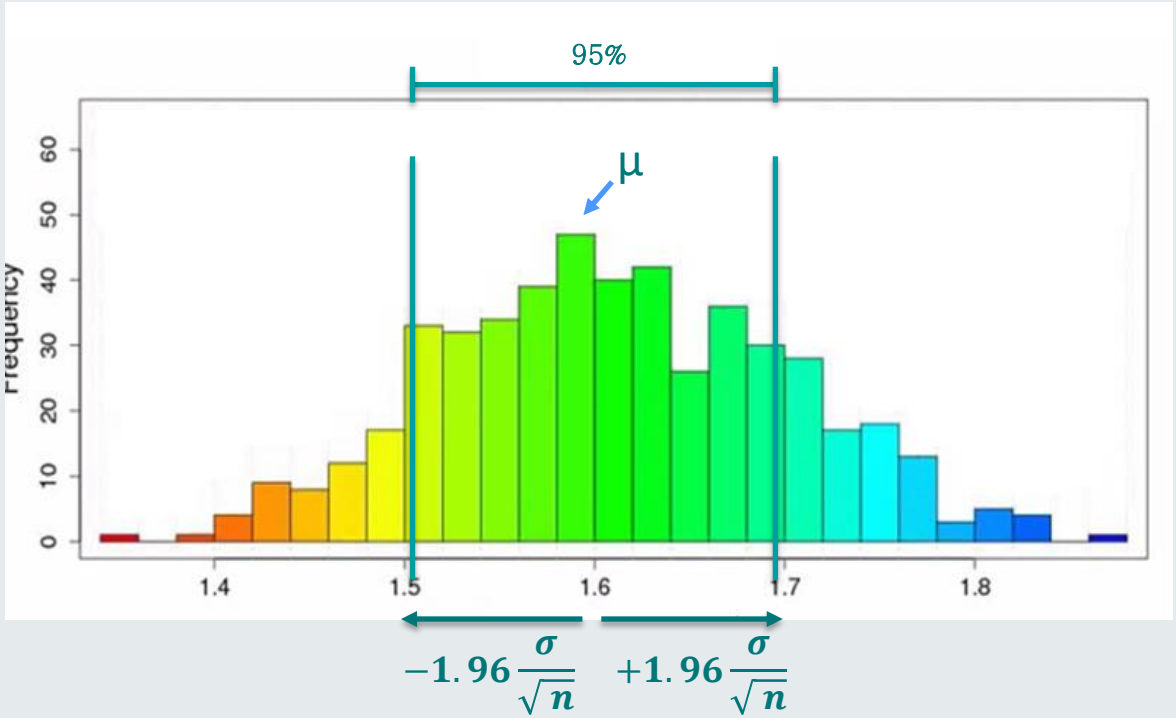
and its standard deviation is the standard deviation of the population σ divided by the square root of our sample size (we call this the standard error).

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

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.

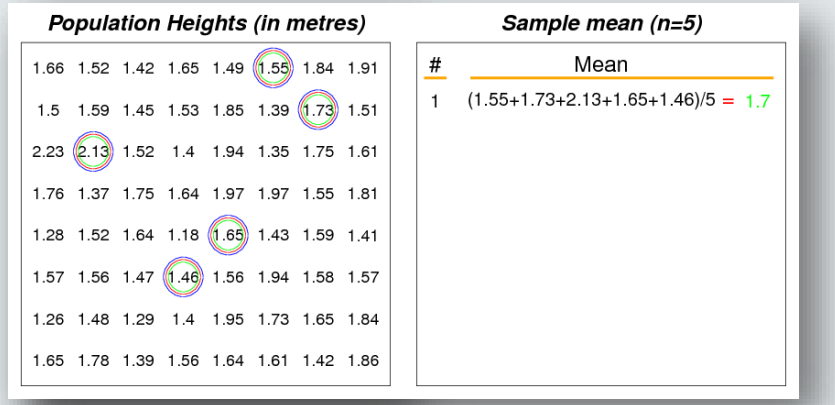
That is, if we sample 100 samples from the same population of size n, then in 95 of them the estimated mean will be within this range.





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



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

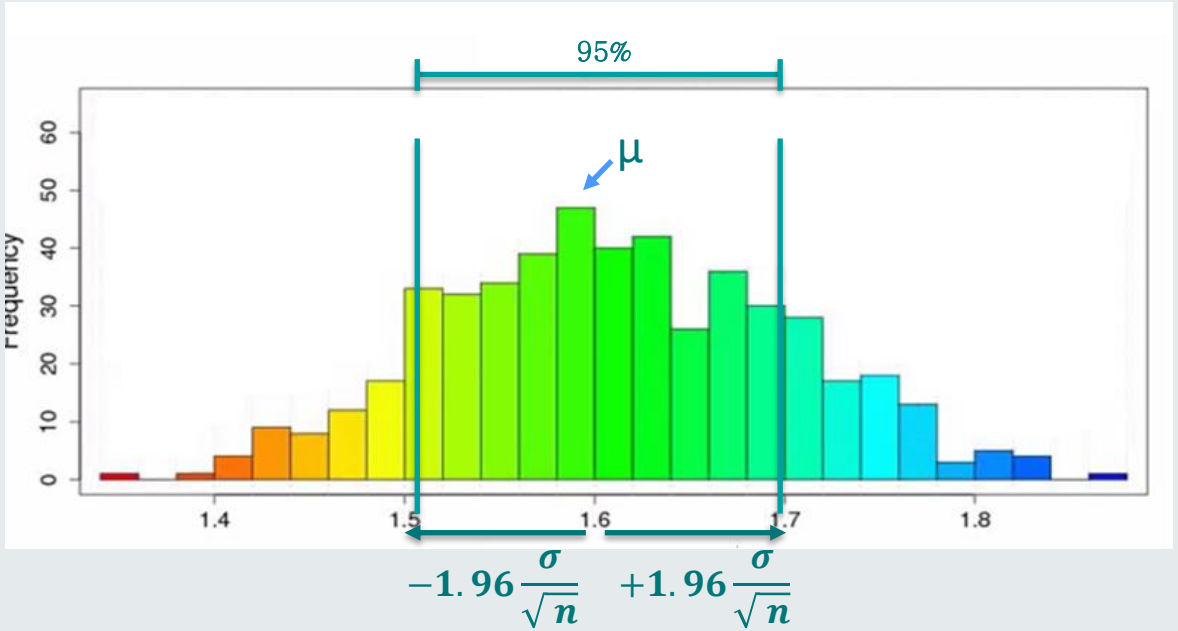
I don't know those!

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

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

I can compute those!

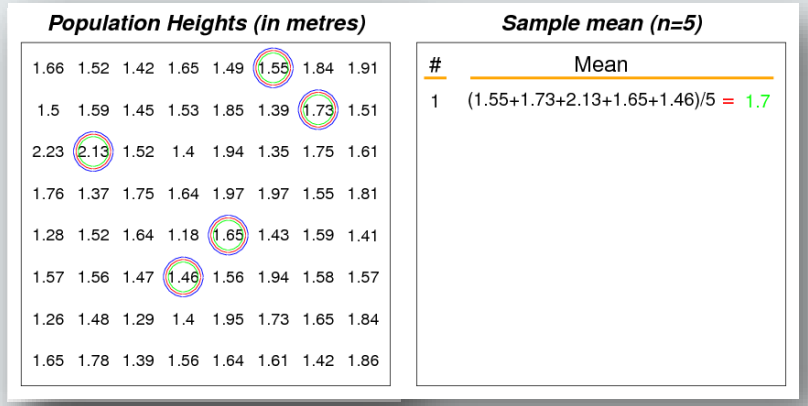
Sampling distribution mean($\bar{x}$)=**μ** and standard deviation **SE**=$\frac{\sigma}{\sqrt{n}}$



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**

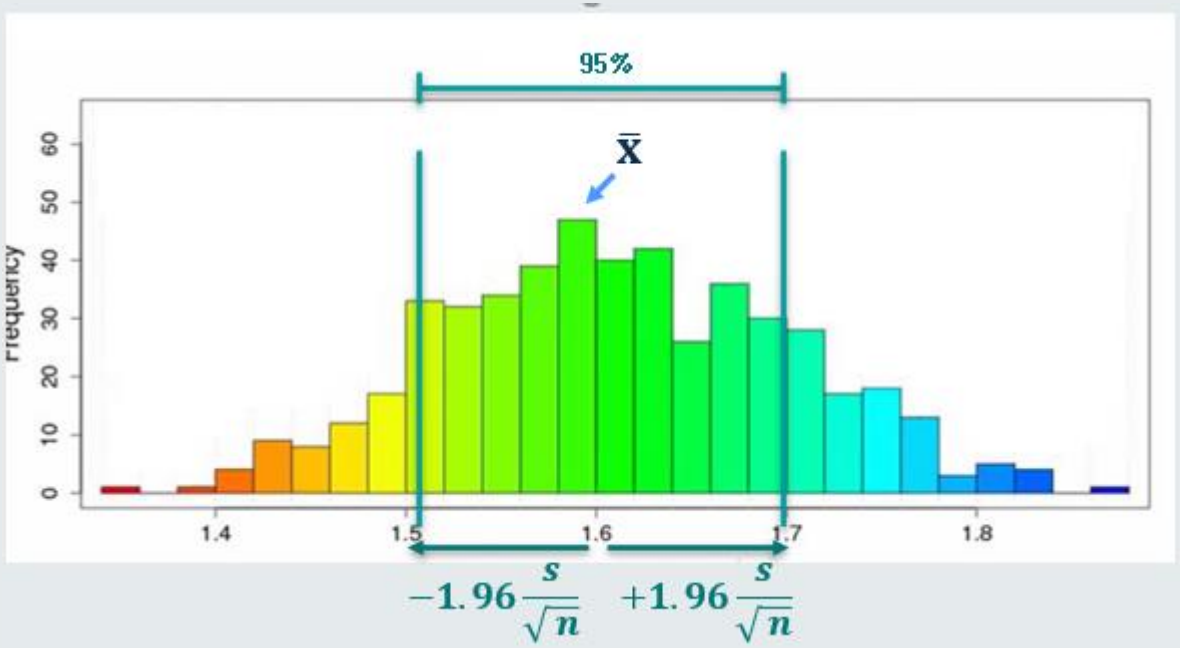
Sample mean $\bar{x}$ and standard deviation **s**

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.



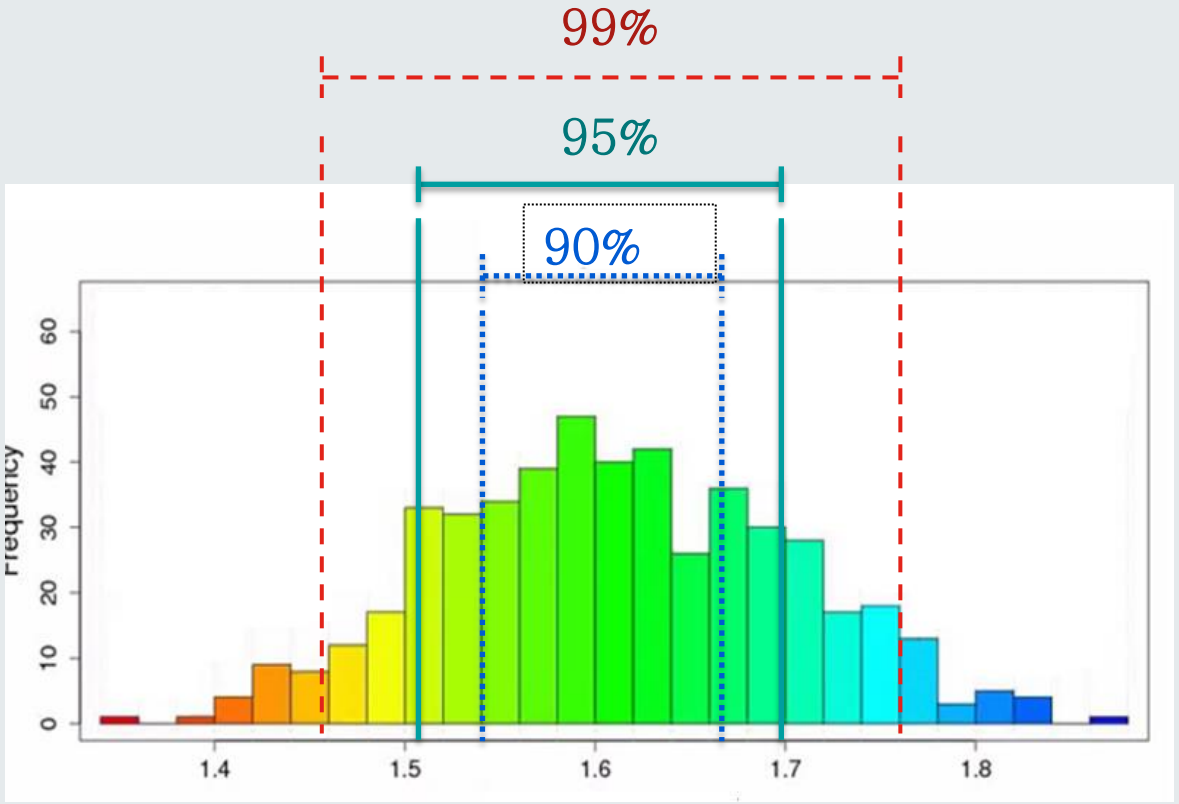
Confidence intervals

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

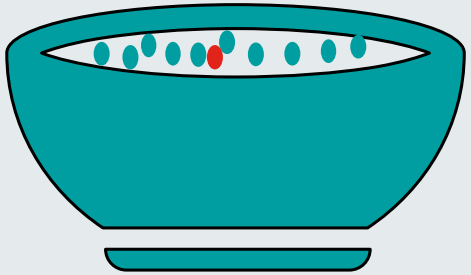
I might say that I am quite confident that my class mean age should range between 20 and 30 years old.

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!

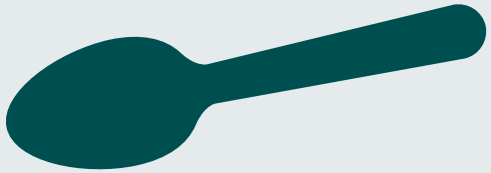
Confidence intervals

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

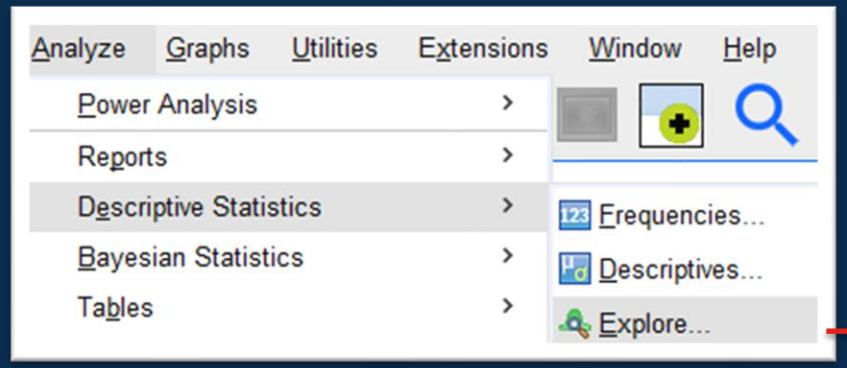
$$[2.633, 2.808]$$

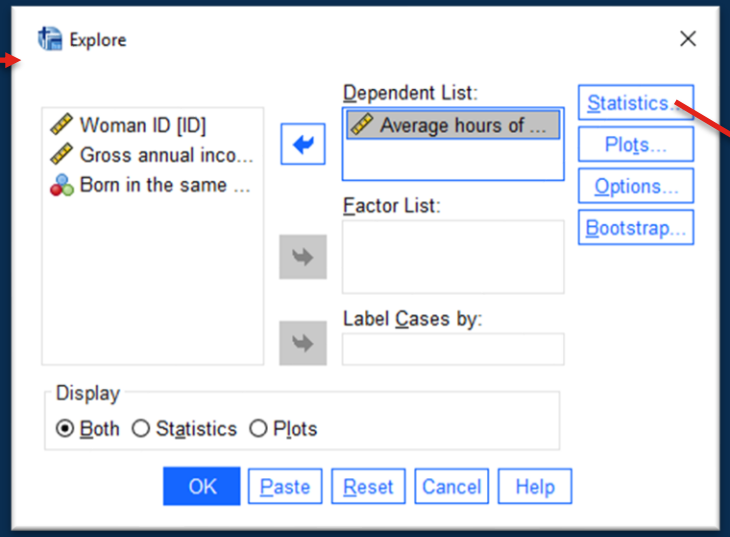
99% CI

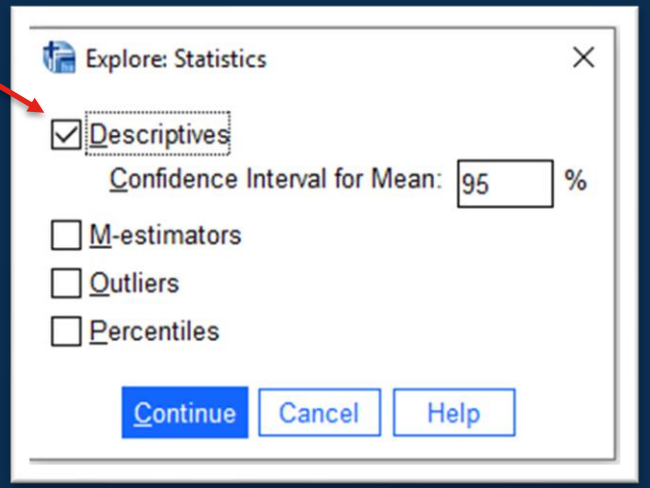
$$[2.584, 2.856]$$

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.

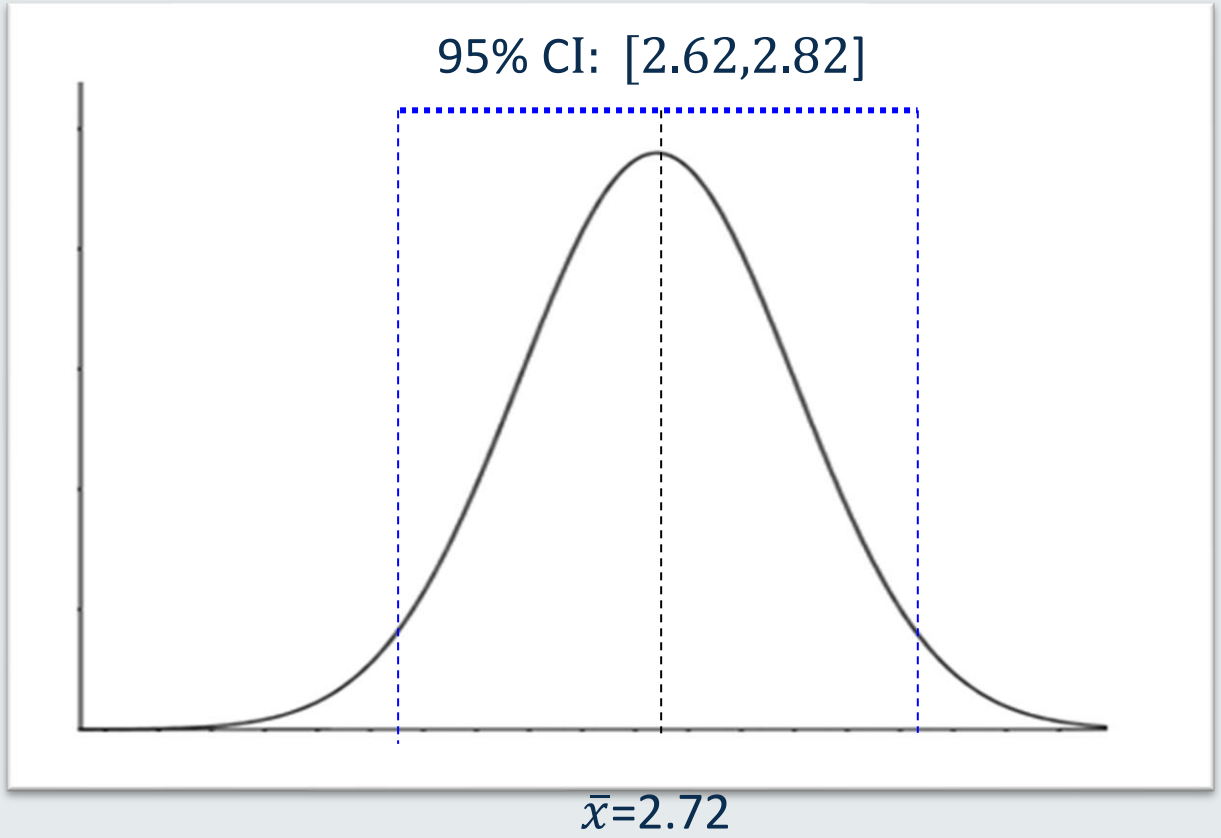
For rare events this may require large sample sizes.

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**