MM104/MM106/BM110
Statistics and Data Presentation

Lecture 6-1:

Estimators

Confidence Intervals

Confidence Intervals for the mean

Chris Robertson

DEPARTMENT OF MATHEMATICS & STATISTICS

University of Strathclyde Science

Estimators

Estimators

- Statistics from samples are *estimators* for population *parameters*.

Population mean → Sample mean
Population proportion → Sample proportion

- The sample mean is the estimator for the population mean

- The sample proportion is the estimator for the population proportion

Estimators

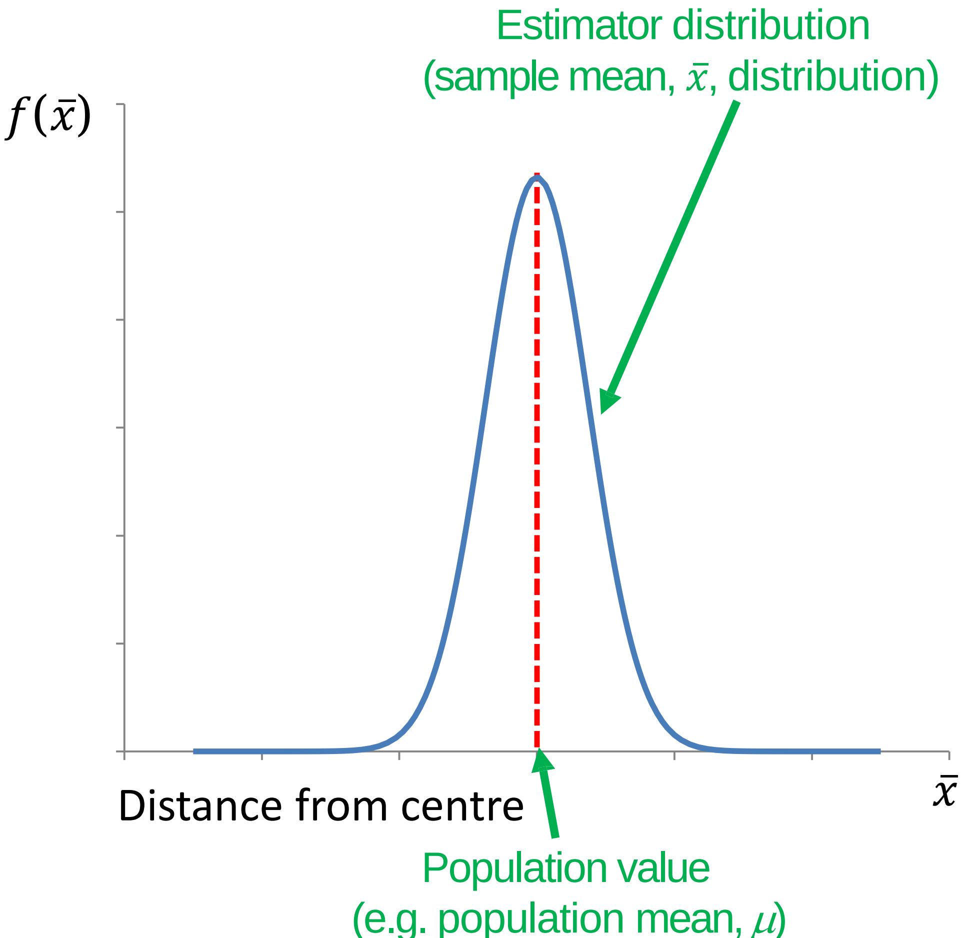
- Sampling distributions provide information to calculate two types of estimators:
  - Point estimator:
    - single number to estimate the parameter.
  - Interval estimator:
    - two numbers (interval) to quantify precision in the estimate of the population parameter.

Point estimation

- Ideally, if we provide just a single point, we want
  The estimate to be very close to the population value

Bias and Precision

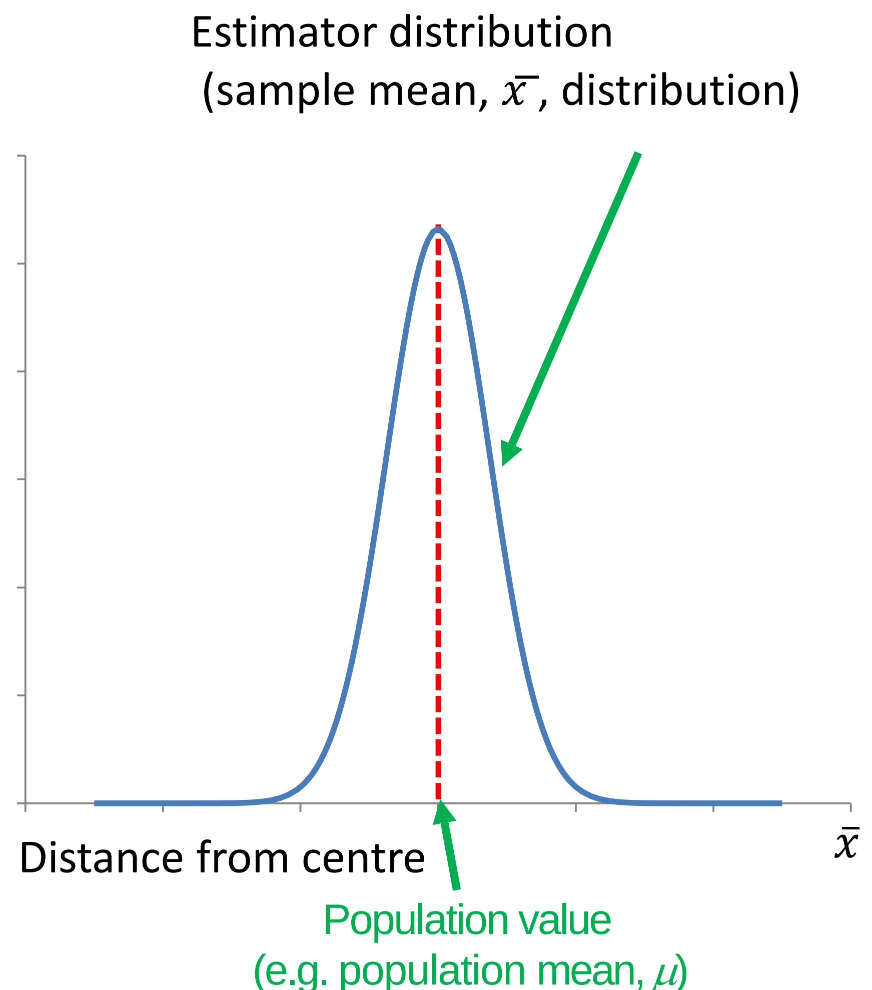
- When sampling, we hope for low *bias* and low spread - :



Bias and Precision

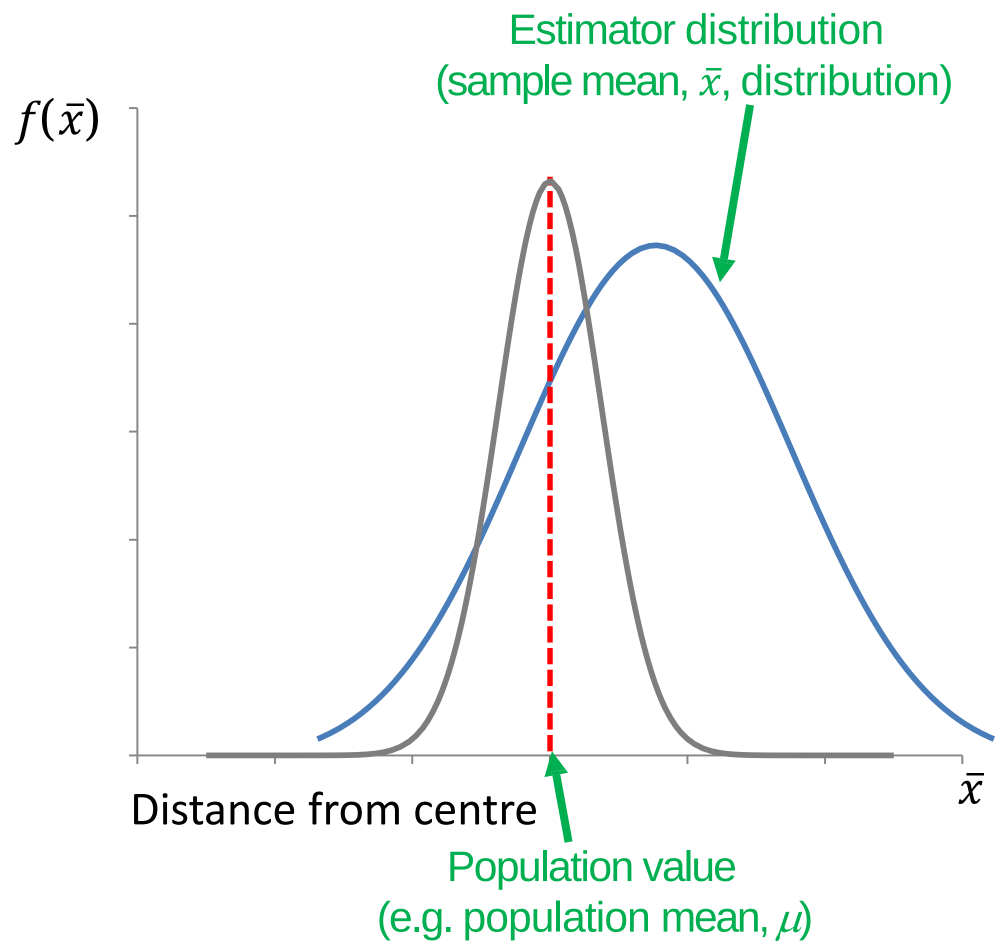
Low bias and high precision means that the sample values are clustered close to the population mean (high precision) and the average of the sample values is the same as the population value (no bias)

High Accuracy is the combination of no bias and high precision



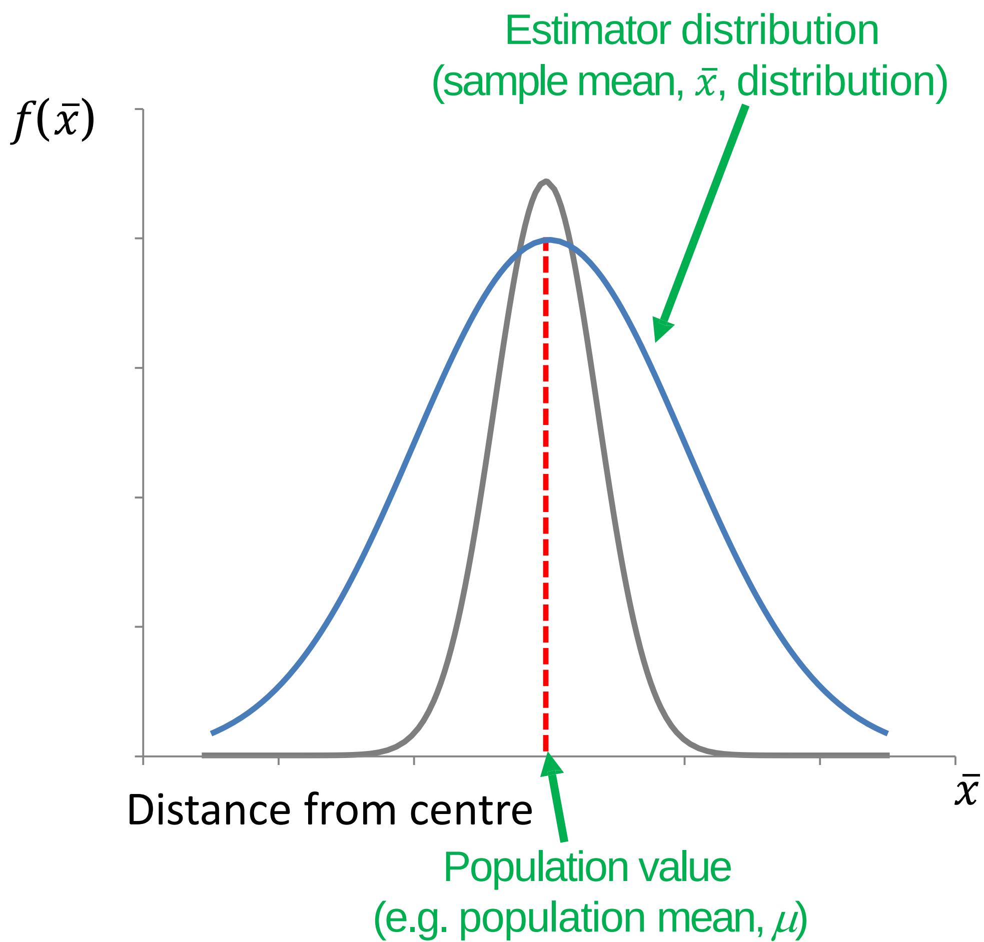
High Bias

The location of the sample mean is far away from the population mean



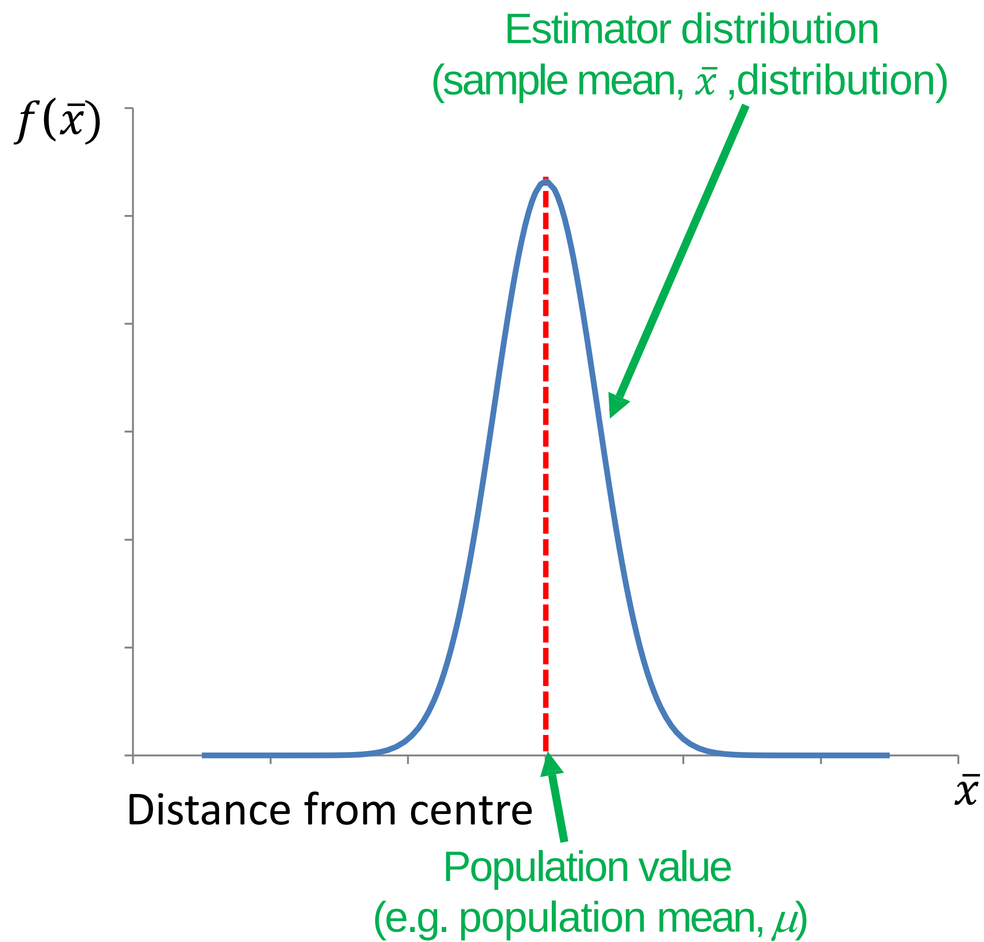
High Spread – Low Precision

Here we have a sample which is unbiased – the average in the sample is the same as the population mean but the variability in the sample is high and the data are very spread out – low precision – high variance



Accurate Estimator has low bias and high precision

- $\bar{x}$ and $\hat{p}$ are unbiased estimators, and spread decreases with sample size!!



Interval estimation

- But remember…we normally do not know the real value of the population parameter.

How do we know whether our estimation is close to the population parameter?

Confidence Intervals!!

Confidence Intervals

- Interval estimation provides a *confidence interval* (CI).
- A range of values to indicate the precision of the estimator
- Two numbers: *upper* confidence limit (UCL) and *lower* confidence limit (LCL).

CI $\longrightarrow$ [LCL, UCL]

Confidence Intervals

- Two numbers: *upper* confidence limit (UCL) and *lower* confidence limit (LCL).

CI → [LCL, UCL]

- Calculated using three elements: point estimator, z-statistics, and standard error.
- Every sample can produce an interval estimate (i.e. a CI, that is, one LCL and one UCL).
- Z-value comes from a normal distribution

Confidence Level

- Specifically, LCL and UCL are the limits for a $100(1-\alpha)\%$ *confidence interval*.
- $100(1-\alpha)\%$ is the *level of confidence*:

$100(1-\alpha)\%=95\%$ confidence level (i.e. $\alpha=0.05$)
(if we have 100 samples, and calculate their 100 CIs, *95* of the CIs will contain the population value)

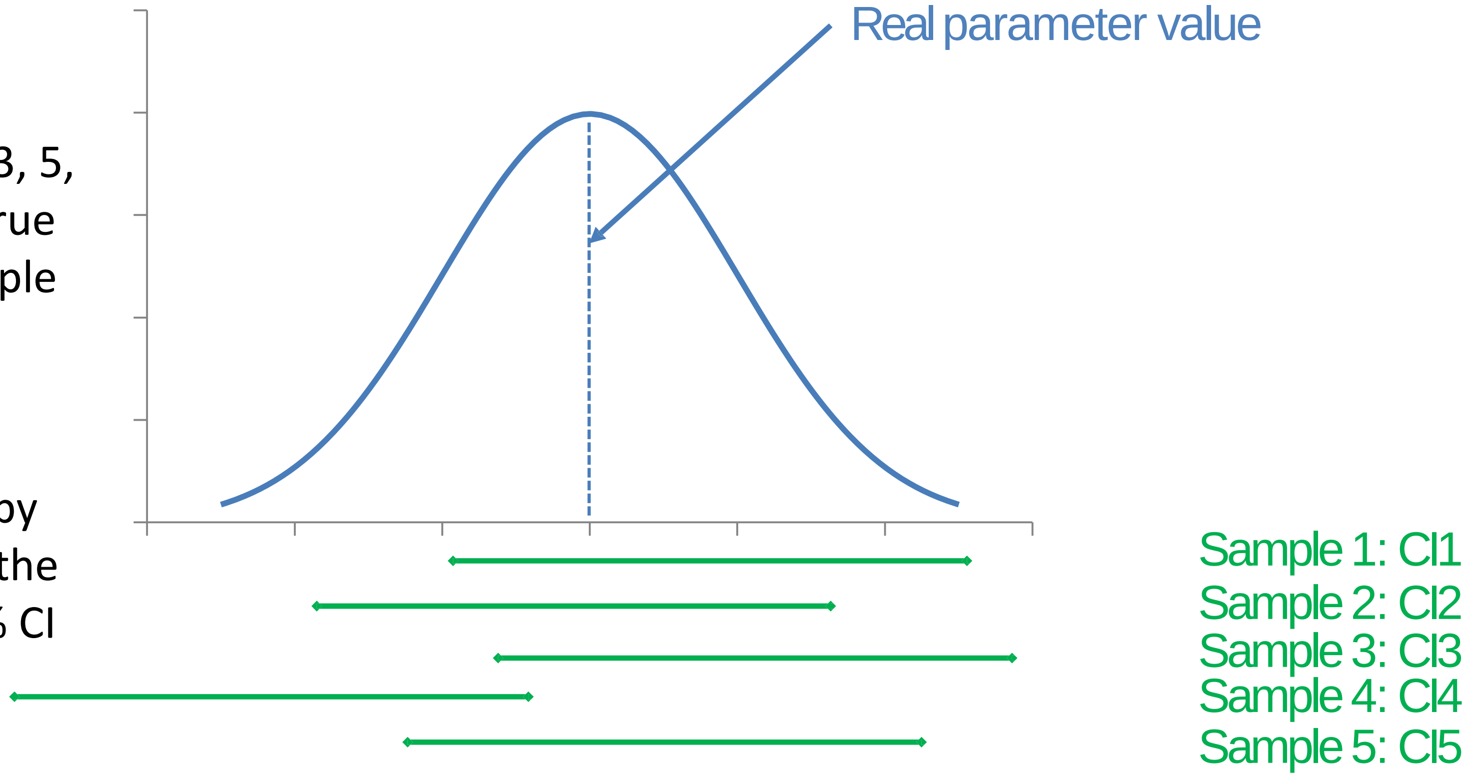
$100(1-\alpha)\%=99\%$ confidence level (i.e. $\alpha=0.01$)
(if we have 100 samples, and calculate their 100 CIs, *99* of the CIs will contain the real value)

Confidence Intervals

- 95% confidence interval: if sample many times, 95 out of 100 times the population (unknown) parameter will fall within our interval.

Samples 1, 2, 3, 5, all cover the true value but sample 4 misses it completely

This happens by chance 5% of the time for a 95% CI



Confidence Intervals

Confidence limits

- Calculated using three elements:
- estimator, z- statistics, and standard error:

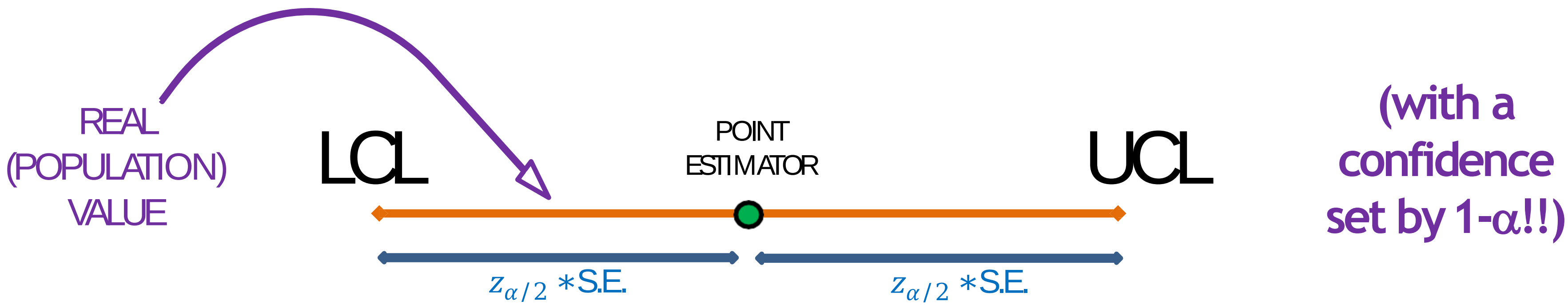
$$(\text{Point estimator}) \pm z_{\alpha/2} * (\text{Standard error estimator})$$

$$\text{UCL} = \text{Point estimator} + z_{\alpha/2} * \text{S.E.}$$
$$\text{LCL} = \text{Point estimator} - z_{\alpha/2} * \text{S.E.}$$

The α value refers to the confidence level – for a 95% confidence level $\alpha = 0.05$ and this then tells you how to get the *z* value from the tables of the normal distribution

Confidence limits

$$UCL = \text{Point estimator} + z_{\alpha/2} * S.E.$$
$$LCL = \text{Point estimator} - z_{\alpha/2} * S.E.$$



The estimator is at the centre of the confidence interval and the lower and upper limits are an equal distance below and above it.

The real population value is unknown (as you only have a sample of data from the population).

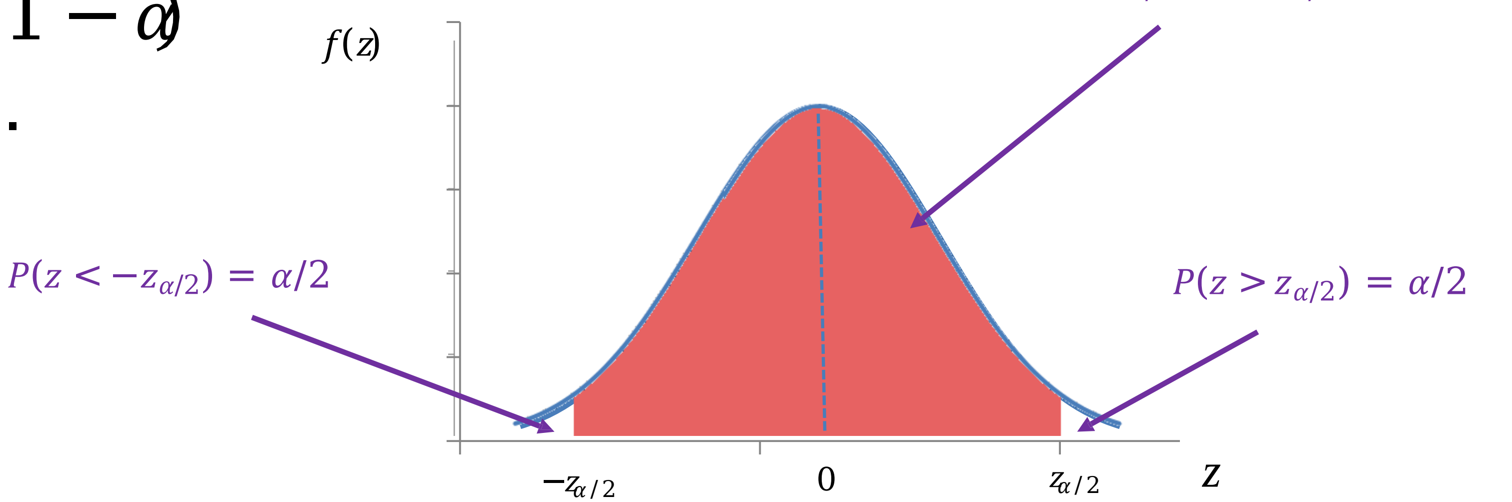
Different samples produce different data, and so different values for the estimator and hence different LCL and UCL

Confidence limits

(Point estimator) $\pm\, z_{\alpha/2}$ $*$(Standard error estimator)

- What $z_{\alpha/2}$ is and is not:
  - $z_{\alpha/2}$ does **not** mean $z * \frac{\alpha}{2}$
  - It is a value for a variable following the **standard normal** distribution, z, such that the area between $-z$ and z is $(1-\alpha)$

.



Percentiles of the standard normal distribution

95% confidence interval

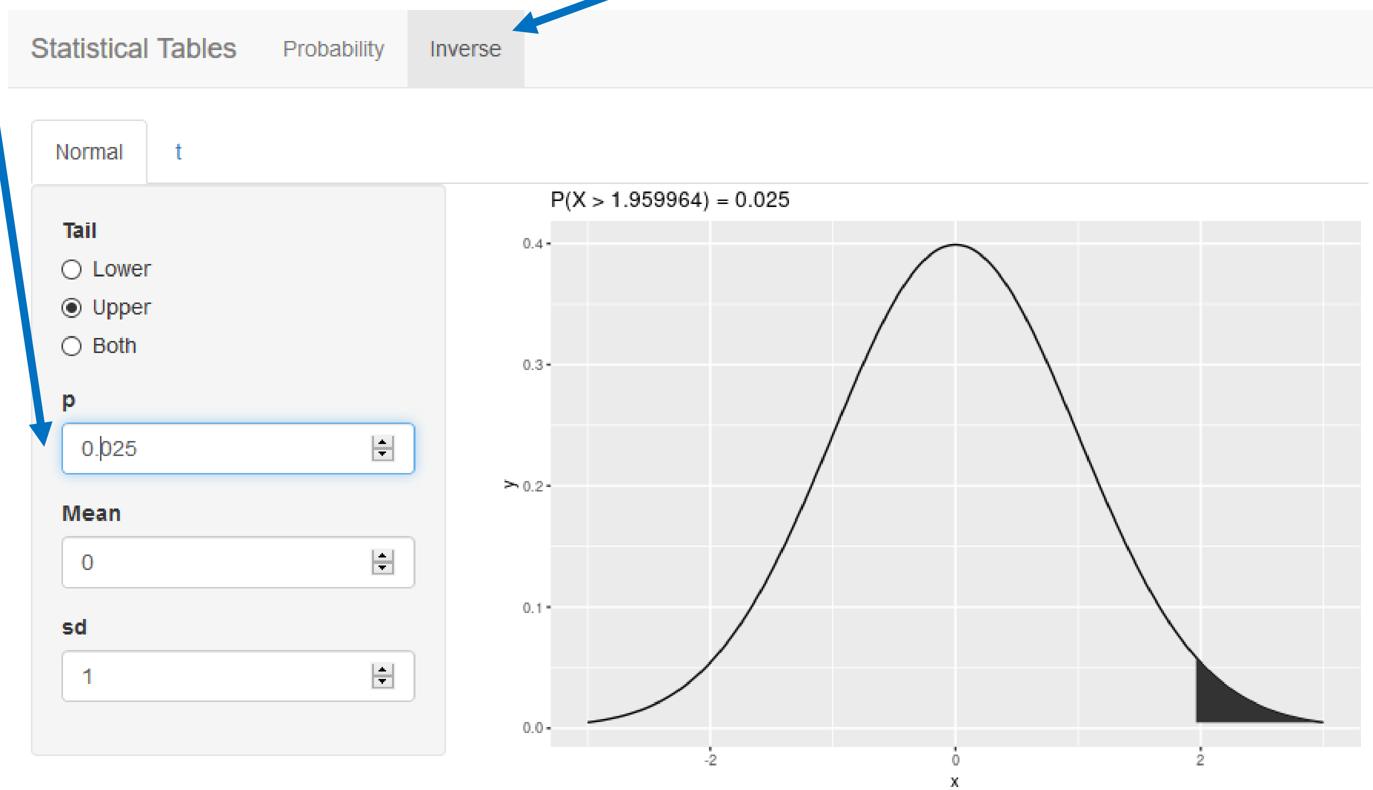
$(1-\alpha) * 100 = 95;$

$\alpha/2 = 0.025$

$z_{\alpha/2}$ is a value for the <u>variable</u>

Inverse standard normal!!

Click the inverse tab



Probability $(\alpha/2)$ goes here; leave mean and sd at 0 and 1, respectively

Confidence Interval for the population mean

Confidence interval for the population mean

Original variable, X, with mean μ and standard deviation σ.

Sample mean $(\bar{X})$ distribution approx. normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, with n = sample size.

1 sample = collection of n measurements (data!!).

Every sample, of size n, generates one value for point estimator $\bar{x}$. All samples share the same standard error, $\sigma/\sqrt{n}$.

Confidence interval for the population mean

The formula follows the same pattern

Point estimator

Standard error estimator

(Sample mean) $\pm z_{\alpha/2}$ $*$(Standard error sample mean)

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

UPPER CONFIDENCE LIMIT

$$\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

LOWER CONFIDENCE LIMIT

$$\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Confidence interval for the population mean

Expressed in probability terms it is clear that the limits are random and the population mean μ is a fixed quantity

$$P(LCL < \mu < UCL) = 1 - \alpha$$

This is the probability definition of a confidence interval – the limits are calculated to ensure that the probability that the lower limit (LCL) is less than the mean (μ) and the upper limit (UCL) is greater than the mean (μ) is equal to the confidence level ($1 - \alpha$)

Confidence interval for the population mean

Expressed in probability terms it is clear that the limits are random and the population mean μ is a fixed quantity

$$P(LCL < \mu < UCL) = 1 - \alpha$$

This is the probability definition of a confidence interval – the limits are calculated to ensure that the probability that the lower limit (LCL) is less than the mean (μ) and the upper limit (UCL) is greater than the mean (μ) is equal to the confidence level ($1 - \alpha$)

$$P\left(\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

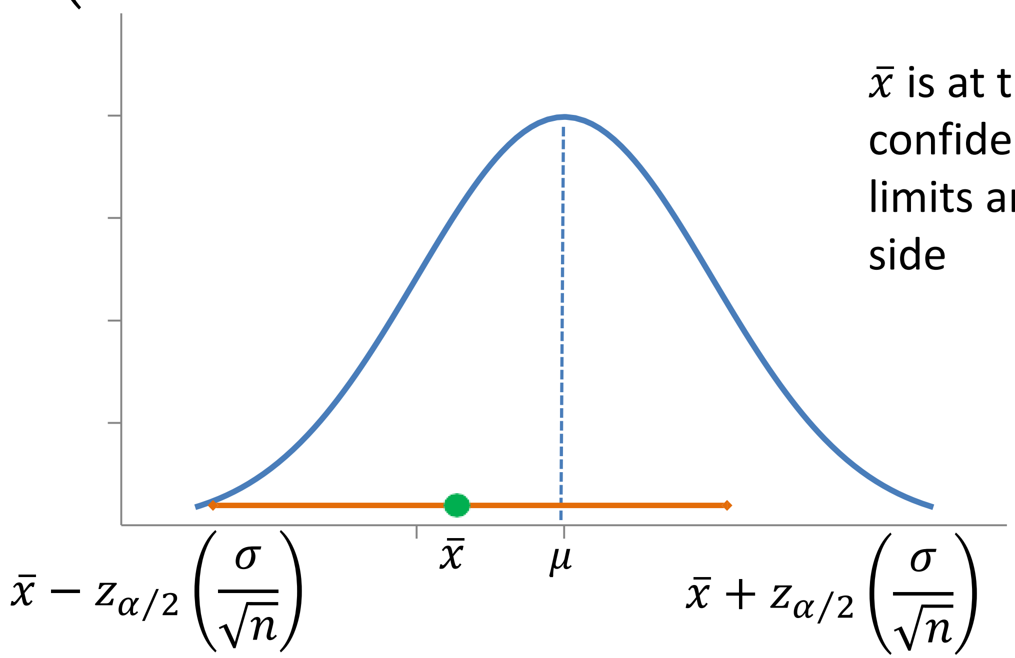
Substituting in the limits reinforces that they are random and vary from one sample to another - σ, μ , n, and $z_{\alpha/2}$ are all fixed quantities – only $\bar{x}$ varies from one sample to the next

Confidence interval for the population mean

Expressed in probability terms it is clear that the limits are random and the population mean μ is a fixed quantity

$$P(LCL < \mu < UCL) = 1 - \alpha$$

$$P\left(\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$



$\bar{x}$ is at the centre of the confidence interval and the limits are equidistant either side

Confidence interval for the population mean

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$P\left(\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha$$

So we want:

$\rightarrow \quad (1-\alpha)$ as close to 1 as possible – 0.95 is usual.

This means that we have the greatest probability that the confidence interval contains the true, but unknown, population mean

We can’t specify a 100% interval as that would imply that the interval went from $-\infty$ to $+\infty$

95% intervals as the ones usually used

Confidence interval for the population mean

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

So we want:

→ $(1-\alpha)$ as close to 1 as possible – 0.95 is usual.

→ CI to be as narrow as possible

— Large sample size.

— Small standard deviation.

Narrow intervals imply greater precision – it is better to know the mean time for an air flight to within plus or minus 15 minutes rather than plus or minus 2 hours.
For a fixed confidence level (α) the width is controlled by the term $\frac{\sigma}{\sqrt{n}}$ and this gets smaller as σ decreases (often not possible to adjust as this is a feature of the variable being measured) or as n increases – this is usually within the control of the investigator

Example

— The following data are the number of questions that a sample of 123 students have tried right before the first MM104/106/BM110 test. There are about 350 students in the class. Find a 99% CI for the mean number of questions. You may use the estimate s of the standard deviation as the 'true' standard deviation σ.

77	127	112	116	97	106	91	94	84	121
125	77	77	104	114	98	95	89	93	85
103	64	124	79	64	104	98	118	104	97
106	114	119	92	105	93	99	90	118	117
95	89	80	103	90	96	113	108	97	108
95	112	65	74	93	111	98	104	109	122
99	80	99	121	84	108	120	99	129	99
98	120	116	99	107	109	96	103	130	100
81	97	113	101	121	113	86	96	102	82
84	102	119	104	80	105	83	93	82	101
99	106	86	100	102	102	109	112	76	85
83	94	75	102	125	73	102	103	82	84
87	82	89							

Example

The mean (μ) for the whole class is unknown and we will use the sample to estimate it. As the sample size is large we are going to use the sample standard deviation (s) as an estimate of σ.

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Mean = μ $\longrightarrow$ 99% CI:

Sample size = $n = 123$

Point estimator = $\bar{x}$ = *???*

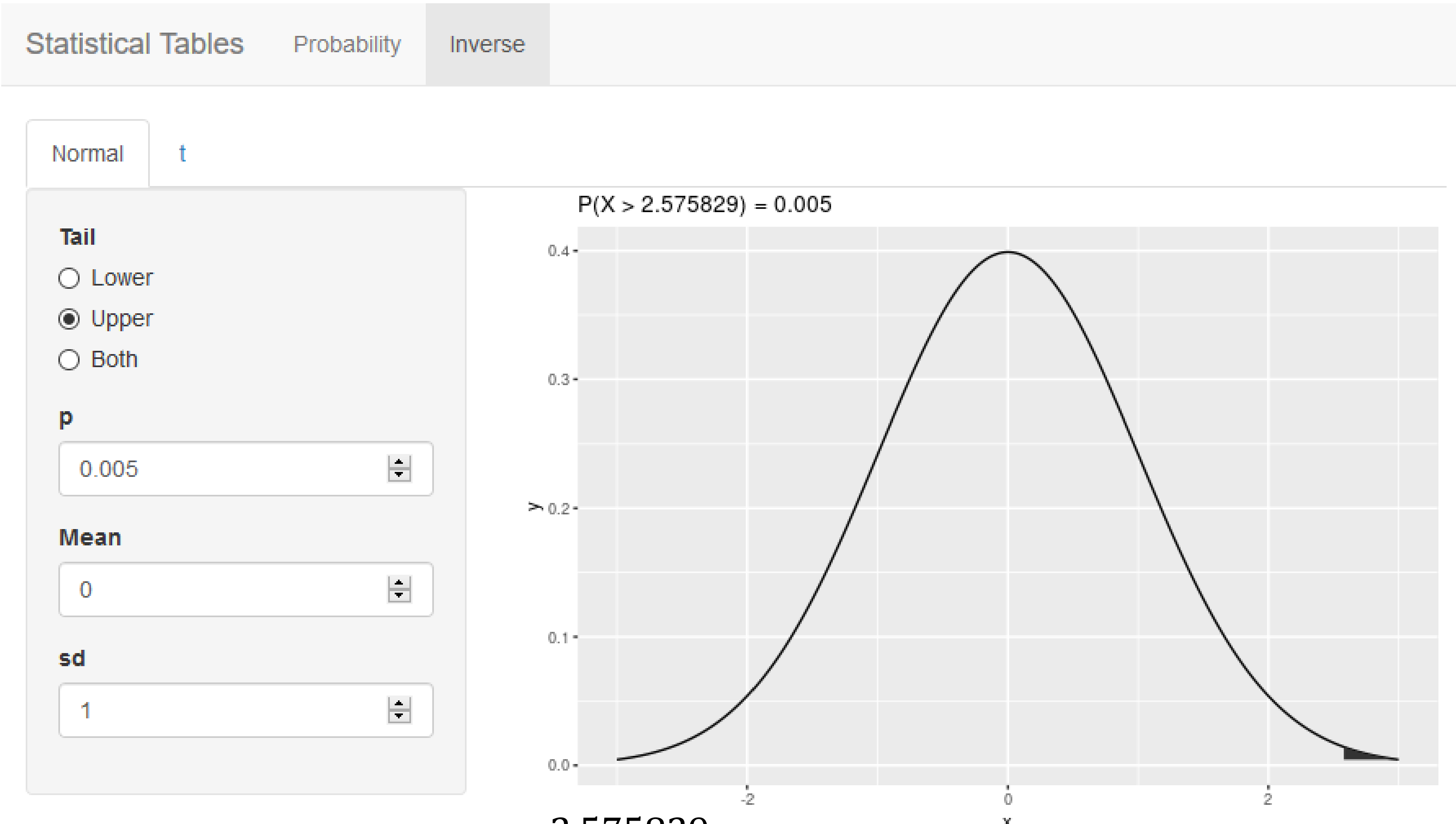
Confidence level = $(1 - \alpha) = 0.99$

Standard error = $\frac{\sigma}{\sqrt{N}}$ = $\frac{s}{\sqrt{123}}$ = *???*

A 99% interval is specified so this means that $\alpha/2$ = 0.005 and we use the invers normal tables to find the corresponding value for z

We use the sample data to calculate values for $\bar{x}$ and s

Example



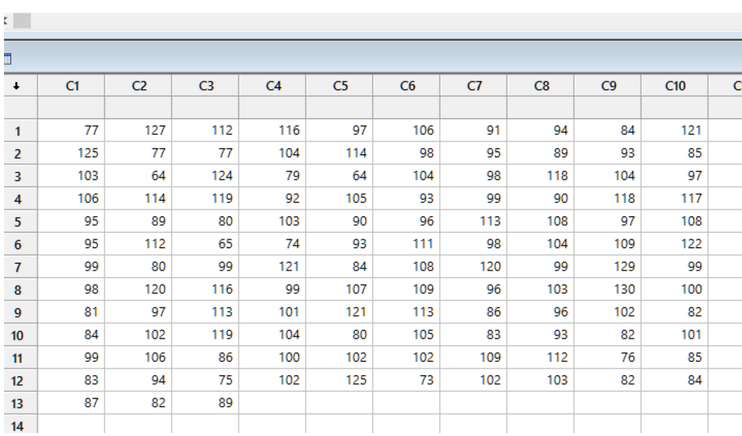
$z_{\alpha/2} = 2.575829$

Example

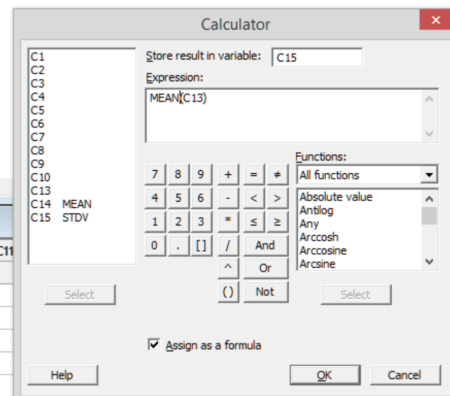
Put the data into Minitab, stack the columns into one column
Use the Calc tab to calculate and store the sample mean and sample standard deviation

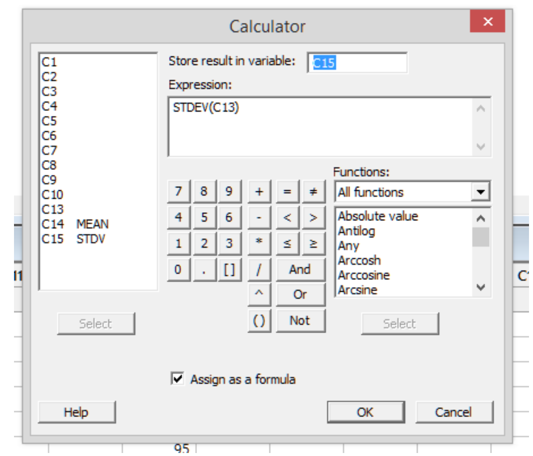
You can also use descriptive statistics and store the result in Minitab – you need to store the result to get sufficient significant digits

(MINITAB "Calc" tab)



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	77	127	112	116	97	106	91	94	84	121
2	125	77	77	104	114	98	95	89	93	85
3	103	64	124	79	64	104	98	118	104	97
4	106	114	119	92	105	93	99	90	118	117
5	95	89	80	103	90	96	113	108	97	108
6	95	112	65	74	93	111	98	104	109	122
7	99	80	99	121	84	108	120	99	129	99
8	98	120	116	99	107	109	96	103	130	100
9	81	97	113	101	121	113	86	96	102	82
10	84	102	119	104	80	105	83	93	82	101
11	99	106	86	100	102	102	109	112	76	85
12	83	94	75	102	125	73	102	103	82	84
13	87	82	89							
14										





C14	C15
MEAN	STDV
98.9512	14.6695

Example

99.9512

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\frac{\sigma}{\sqrt{n}} = \frac{14.6695}{\sqrt{123}} = 1.3227$$

$$(1-\alpha) = 0.99;$$

$$z_{\alpha/2} = 2.578293$$

UPPER C.L.

$$\bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \boxed{102.358}$$

LOWER C.L.

$$\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \boxed{95.544}$$

The mean number of questions attempted is 99.95 with a 99% confidence interval of (95.5, 102.4)

The interval (95.5, 102.4) covers the true, unknown, average with 99% confidence.

Key Points

- Confidence limits for a mean are given by

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- The width of the confidence interval will decrease as the sample size increases.
- As n gets bigger the SE decreases and the width will decrease.
- Hence the precision of the estimate will increase
- From one sample to another the interval will vary as it is a function of $\bar{x}$ which varies in different sample.
- Hence the interpretation of a 95% confidence interval is that the interval contains the true, unknown, population mean with 95 % confidence