Social Science Statistics Module I

Gwilym Pryce

Lecture 3
Introduction to Confidence Intervals

Slides available from *Statistics & SPSS* page of www.gpryce.com

Expectations & Support:

1. **Independent learning:**
   - this is a PG course and a degree of independent learning is assumed.
   - do the reading, attend labs, review the lectures, make use of the computer labs/online help in your own time.
2. **Lab Overview & Feedback:**
   - Please feedback to the tutors & Class Reps how you think that is going, how it could be improved.
   - Tutors and Class Reps will then report back to me how things are going each week.
3. **Talk to tutors if you are struggling:**
   - Let the tutors know if you are struggling (assuming you have done the reading, attended labs etc.)
   - Tutors cannot guarantee extra support, but it might be possible to arrange extra tutorials etc.

4. **Departmental Support:**
   - Struggling students should enquire whether their own dept has support to offer.
   - All the grad school courses are only intended to constitute a generic training component;
   - Individual depts & supervisors should supplement with additional training & support as necessary.
5. **Support from Maths Advisor Shazia Ahmed, University's Maths Adviser:**
   - If you have gone through steps 1 to 4, Shazia has agreed to run one-on-one sessions with students that are struggling with particular mathematical or statistical concepts (though she has made it clear that she cannot advise on SPSS problems, nor will she do the assignment for you).
   - Students who have particular problems in this regard can contact her directly: **Shazia Ahmed**, Maths Adviser, Student Learning Service, McMillan Reading Room, **Tel: 330 5631** Fax: 330 8063
6. **Tutor of Last Resort:**
   - Students who have gone through steps 1 to 5 above, and who still feel they are not receiving enough support, can email me directly
   - I will try to arrange individual or small group meetings for people who have tried all other avenues.
     - You will need to demonstrate that you have gone through steps 1 to 5.

Aims & Objectives

- Aim
  - To introduce the concept of confidence intervals.
- Objectives
  - By the end of this session, students should be able to:
    - Understand the intuition behind confidence intervals;
    - calculate large and small sample confidence intervals for one mean.

Plan

- 1. Intuition Behind CIs
  - a) All normal curves related ⇒ *z* distribution
  - b) Converting x to z values
  - c) Applying z to sampling distributions
  - d) 5 steps of logic behind CI
- 2. Three steps of Confidence Interval Estimation
- 3. Large Sample Confidence Interval for the mean
- 4. Small Sample Confidence intervals for the Population mean

1. Intuition behind CIs

a) All normal curves related ⇒ *z* distribution

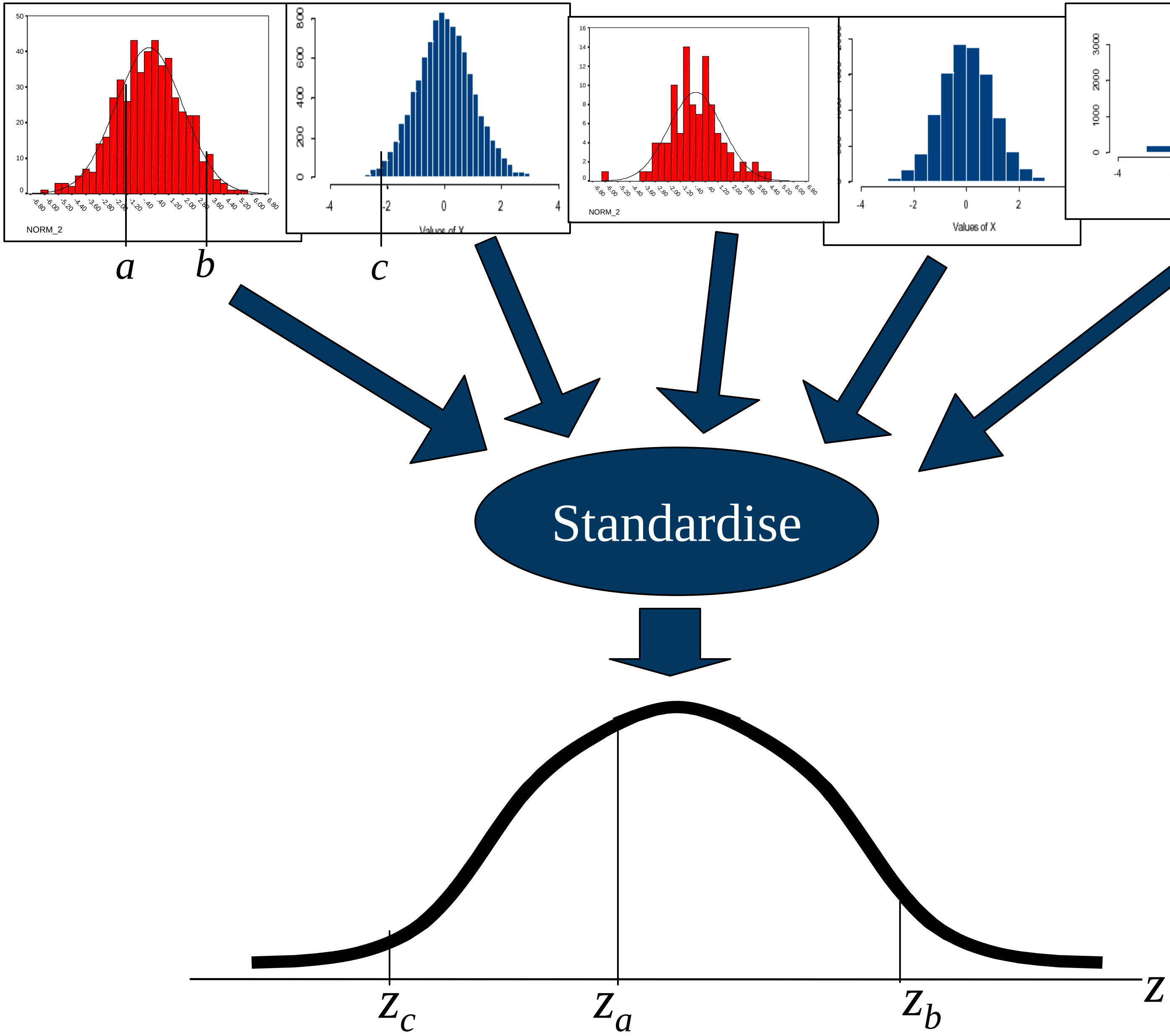
b) Converting x to z values

c) Applying z to sampling distributions

d) 5 steps of logic behind CI

a) All normal curves related ⇒ *z* distribution

- We have said that there are an infinite number of poss. normal distributions
  - but they vary only by *mean* and *S.D.*
    - so they are all related -- just scaled versions of each other
- a baseline normal distribution has been invented:
  - called the **standard normal distribution**
  - has zero mean and one standard deviation



Standardise

Standard Normal Curve

- we can standardise any observation from a normal distribution
  - I.e. show where it fits on the standard normal distribution by:
    - subtracting the mean from each value and dividing the result by the standard deviation.
    - This is called the **z-score** = standardised value of any normally distributed observation.

$$\boxed{z_i = \frac{x_i - \mu}{\sigma}}$$

Where μ = population mean

σ = population S.D.

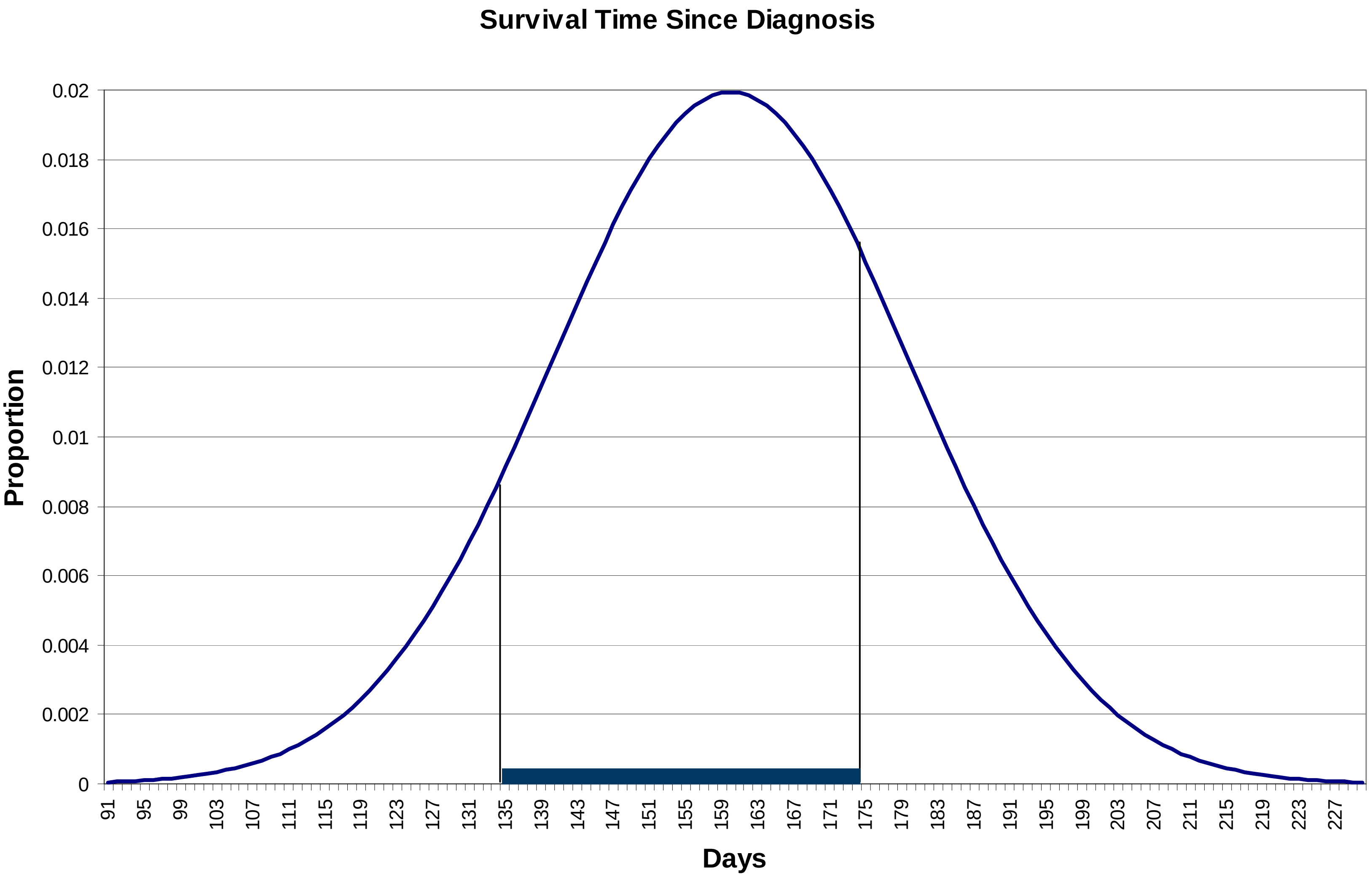
  - Areas under the standard normal curve between different z-scores are equal to areas between corresponding values on any normal distribution
  - Tables of areas have been calculated for each z-score,
    - so if you standardise your observation, you can find out the area above or below it.
- But we saw earlier that areas under density functions correspond to probabilities:
  - so if you standardise your observation, you can find out the probability of other observations lying above or below it.

b) Converting x to z values:

Example:

- Suppose that the survival time of brain tumour patients following diagnosis is found to be normally distributed. You have records on all such diagnoses (I.e. the population). The average survival time is 160 days with a standard deviation of 20 days. Suppose you want to find the ***proportion of brain tumour patients who survive between 135 and 175 days***.
  - How would you do that given:

$$\boxed{z_i = \frac{x_i - \mu}{\sigma}}$$


Survival Time Since Diagnosis
Proportion
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
91
95
99
103
107
111
115
119
123
127
131
135
139
143
147
151
155
159
163
167
171
175
179
183
187
191
195
199
203
207
211
215
219
223
227
Days

Brain Tumour Example cont'd:

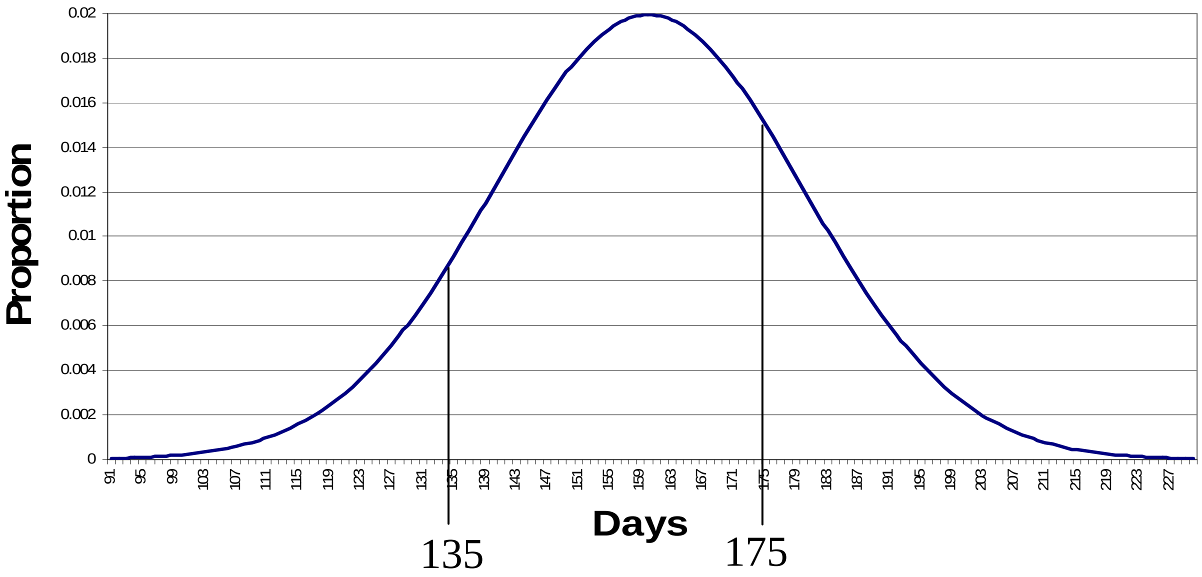
- Suppose that the survival time of brain tumour patients following diagnosis is found to be normally distributed. You have records on all such diagnoses (I.e. the population). The average survival time is 160 days with a standard deviation of 20 days. Find the ***proportion of brain tumour patients who survive between 135 and 175 days***.

$$z_i = \frac{x_i - \mu}{\sigma}$$

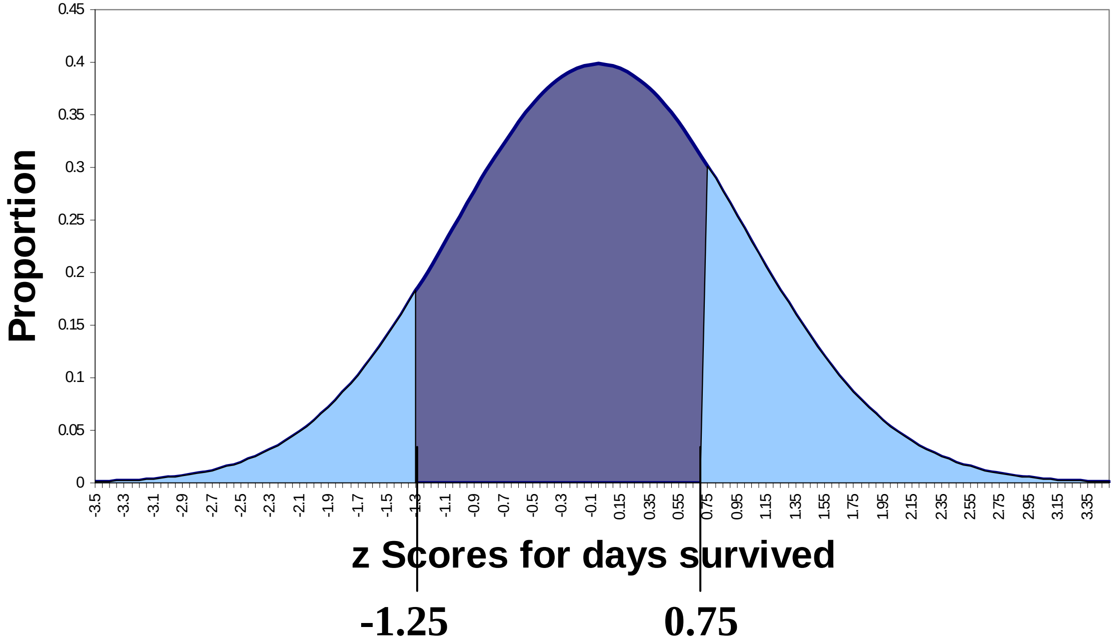
- (i) Find z scores for $x_1 = 135$ and $x_2 = 175$:
  - $z_1 = (135 - 160)/20 = \underline{\mathbf{-1.25}}$; and $z_2 = (175 - 160)/20 = \underline{\mathbf{0.75}}$
  - $P(135 < \text{days} < 175) = P(-1.25 < z < 0.75)$
- (ii) Find area A under z curve where: $A = P(z < -1.25) = 0.1056$
- (iii) Find area B under z curve where: $B = P(z < 0.75) = 0.7734$
- (iv) take area A from area B: $C = B-A = P(-1.25 < z < 0.75)$

  $C = P(135 < \text{days} < 175)$
  $= P(-1.25 < z < 0.75)$
  $= B - A$
  $= 0.7734 - 0.1056$
  $= \underline{\mathbf{0.6678}}$

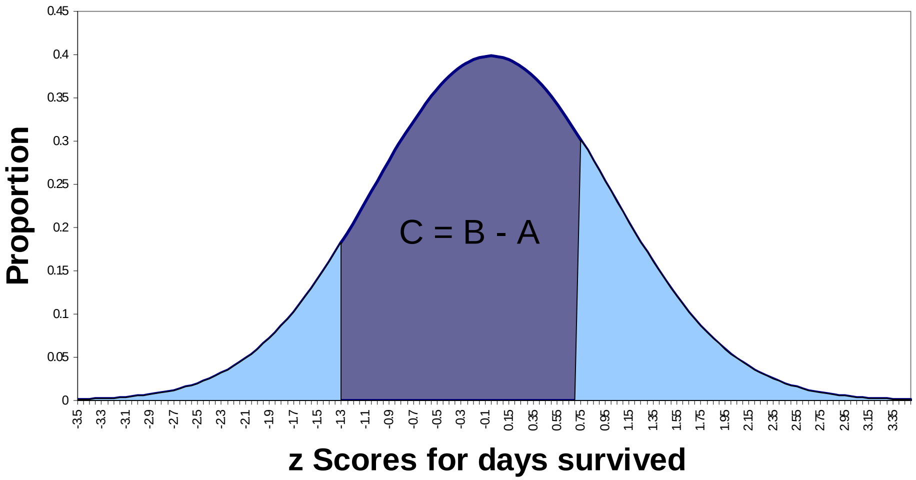
Survival Time Since Diagnosis



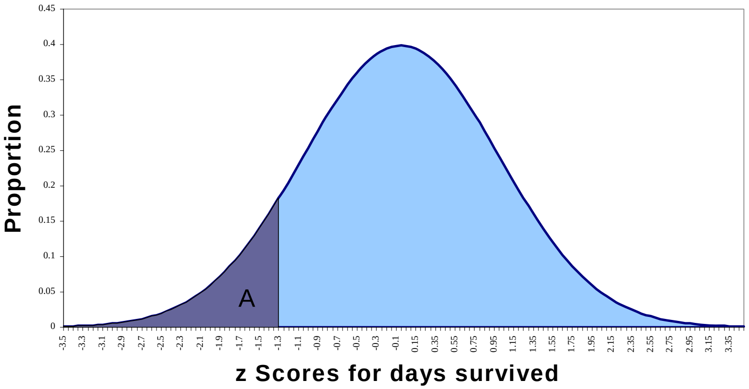
z Scores for Survival Time Since Diagnosis



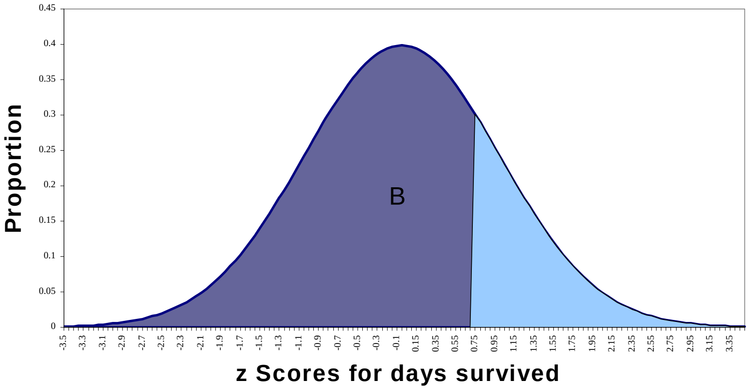
z Scores for Survival Time Since Diagnosis



z Scores for Survival Time Since Diagnosis



z Scores for Survival Time Since Diagnosis



But what about non-normal variables?

- Q/ Suppose we don’t know the shape of the population distribution of income but we want to estimate the population mean.
  - We usually can only afford to take one sample (e.g. interview 100 people).
  - But knowing something about the distribution of the sample means (I.e. the CLT) means that we can say something about how close our sample mean is likely to be to the population mean.

c) Applying z-score transformation to sampling distributions:

- The formula we learned last week for applying z scores to sampling distributions was:

$$\boxed{z_i = \frac{\bar{x}_i - \mu}{\sigma_{\bar{x}}}}$$

Now, if we rearrange this formula we get:

$$\boxed{\mu = \bar{x}_i - z_i \sigma_{\bar{x}}}$$

where:

$\mu =$ populationmean

$\bar{x}_i =$ samplemean

$\sigma_{\bar{x}} =$ standarddeviationof all thesamplemeans

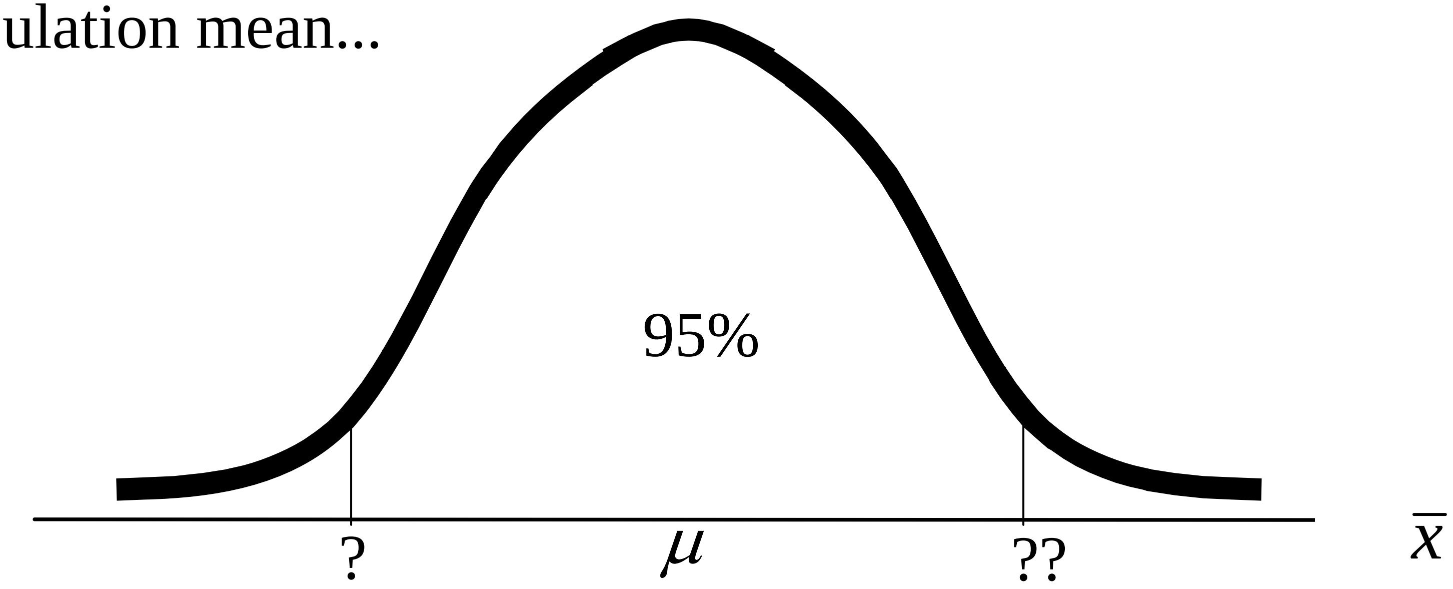
$z_i =$ z score

So if the population mean is unknown, we can then decide on the level of confidence we want, and calculate *z* to give an interval for the unknown population mean.

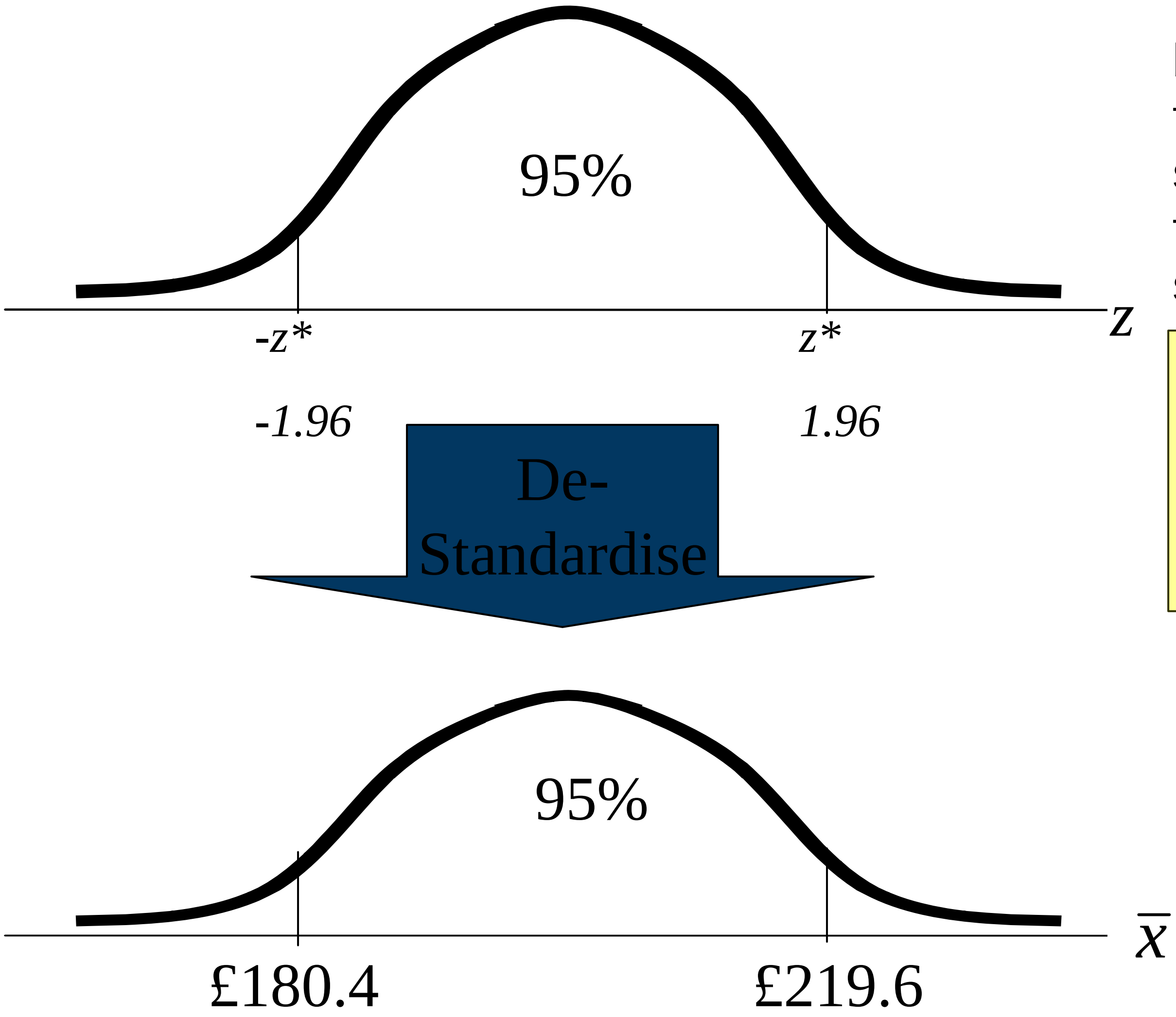
We want to know where 95% of sample means lie:

we can then say that we are 95% sure the population mean will lie between £? and £??

We can find out where 95% of sample means lie because we know that the sample mean is normally distributed around the population mean...



… and this means we can use *z*



Find the z-scores that bound the central 95%, then convert these z-scores to sample means to find the central 95% of sample means.

$$\mu = \bar{x}_i - z_i \sigma_{\bar{x}}$$

$$\mu = £200 \pm 1.96 \times 10$$

$$\mu = £200 \pm 19.6$$

I.e. 95% of samples will have means that lie between £180.4 and £219.6

d) Confidence Intervals are based on 5 steps of logic:

- (1) **CLT says that:** sample mean is normally distributed. We call the standard deviation of the this distribution the "SE of the mean"
- (2) **95% Rule:** for any normally distributed variable, 95% of observations lie within 2 standard deviations of the mean.
- (3) **Statements (1) & (2) imply that:**
  - 95% of samples will have means that lie within 2 SEs of μ
- (4) **$\Rightarrow$ μ is within 2 SEs of the sample mean**
  - to say that we are 95% sure that the sample mean lies within 2 SEs of μ is the same as saying that we are 95% sure that μ is within 2 SEs of the sample mean.
- (5) **So 95% of all samples will capture the true population mean in the interval:**

$$\bar{x} - 2\mathrm{SE} \quad \text{to} \quad \bar{x} + 2\mathrm{SE}$$

- (5) **So 95% of all samples will capture the true population mean in the interval:**

$$\bar{x} - 2\mathrm{SE} \quad \text{to} \quad \bar{x} + 2\mathrm{SE}$$

- Put another way, there are only 2 possibilities:
  - Either the interval sample mean ± 2SE contains μ
  - Or our sample was one of the few samples (I.e. one of the 5%) for which the sample mean is not within 2SE of μ

- E.g. Suppose SE of the mean = £10 for repeated samples of income.
- Because the sampling distribution of mean income is normal (assuming large sample sizes) this means 95% of mean incomes lie between ± 2x£10 of the population mean.
- So if the **population** mean income is £200, we know that in 95% of samples, the sample mean will lie between...

… £180 and £220.

- We also know that in 95% of samples, the **population** mean will lie within £20 of the sample mean.

Algebraic proof that the statement:

the sample mean lies within 2 SEs of μ

is the same as saying that

μ is within 2 SEs of the sample mean.

$$\mu - £20 \leq \bar{x} \leq \mu + £20$$

Sample mean lies within: μ±£20

$$-£20 \leq \bar{x} - \mu \leq £20$$

$$-\bar{x} - £20 \leq -\mu \leq -\bar{x} + £20$$

$$\bar{x} + £20 \geq \mu \geq \bar{x} - £20$$

$$\bar{x} - £20 \leq \mu \leq \bar{x} + £20$$

μ lies within: sample mean ±£20

2. Three steps of Interval estimation for μ: the *large* sample case

- **1.** Choose the appropriate formula and decide on the level of confidence (e.g. 95%):

$$\boxed{\mu = \bar{x}_i \pm z_i \sigma_{\bar{x}}}$$

- **2.** Find the value for z* such that:
  - Prob(-z^* ≤ z ≤ z^*) = Confidence level (e.g. 95%)
- **3. Calculate the confidence interval**
  - substitute your values for the sample mean, *z** and the standard error of the mean into the formula.

Amazing as the Central Limit Theorem is, it has at least 2 problems:

- 1. We need to know the standard error of the mean
  - i.e. the average deviation of the sample mean from sample to sample
- 2. CLT depends on the sample size being large

Let's look at the first problem in the context of sampling distributions:

- 1. We need to know the standard error of the mean
  - i.e. the average deviation of the sample mean from sample to sample, denoted as:

$$\sigma_{\overline{x}}$$

  - But we only have one sample. How might we find a way of estimating SE(mean)?

Approximating the S.E. of the mean:

- Q/ Do you think that the standard deviation within the sample you have selected will tell us anything about the SE of the mean?
  - I.e. is the spread of any one sample and the spread of all sample means related?
- A/ Yes, we would expect the variability of the possible sample means to be related to the variability of the population, which in turn is estimated by our sample s.d.

Large sample is “better” than small sample

- This is because the mean and s.d. will be closer to mean and s.d. of population the larger *n*
- So the variability of the sample mean decreases as the sample size increases
- more specifically,

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \rightarrow \frac{s}{\sqrt{n}} \text{ as } n \rightarrow \infty}$$

- I.e. provided n > 30, we can use:
  - sample standard deviation ÷ *square root of sample size*

as an approximation for SE(mean)

- So:
  - Usually we do not know the standard error of the mean.
  - A simple approximation of the standard error of the mean can be found by dividing the sample standard deviation by the square root of the sample size:

$$\boxed{\frac{s}{\sqrt{n}} \approx \sigma_{\overline{\mathrm{x}}}}$$

  - So, for large samples, we can create confidence intervals for the population mean from the sample mean and s.d. using the following formula:

$$\boxed{\mu = \bar{x}_i \pm z^* \frac{s}{\sqrt{n}}}$$

3. Three steps of Interval estimation for μ: the *large* sample case

- **1.** Choose the appropriate test statistic and decide on the level of confidence (e.g. 95%):

$$\boxed{\mu = \bar{x}_i \pm z^* \frac{s}{\sqrt{n}}}$$

- **2.** Find the value for z^* such that
  - Prob($-z^* \leq z \leq z^*$) = Confidence level (e.g. 95%)
- **3.** Calculate the confidence interval by substituting your values for the sample mean, z* and your approximation for the standard error of the mean ($s/\sqrt{n}$).

Example:

- Suppose your area of research is the disappearance of thousands of civil servants and other workers during Joseph Stalin’s *Great Purge* in Soviet Russia 1936-38. One of the questions you are interested in is the average age of the workers when they disappeared. Your thesis is that Stalin felt most threatened by older, more established ‘enemies’, and so you anticipate their average age to be over 50. Unfortunately, you only have access to 506 records on the age of individuals when they disappeared.
- You have calculated the **average age** in this sample to be **56.2** years, which would appear to confirm your thesis. The **standard deviation** of your sample was found to be **14.7** years. Assuming that your **506** records constitute a random sample from the population of those who disappeared (a questionable assumption?), calculate the **95% confidence interval** for the **population mean age**.
- Does your expected value for the population average age fall below the 95% confidence interval? If so, what does this imply?

n	=	506
xbar	=	56.2
s	=	14.7

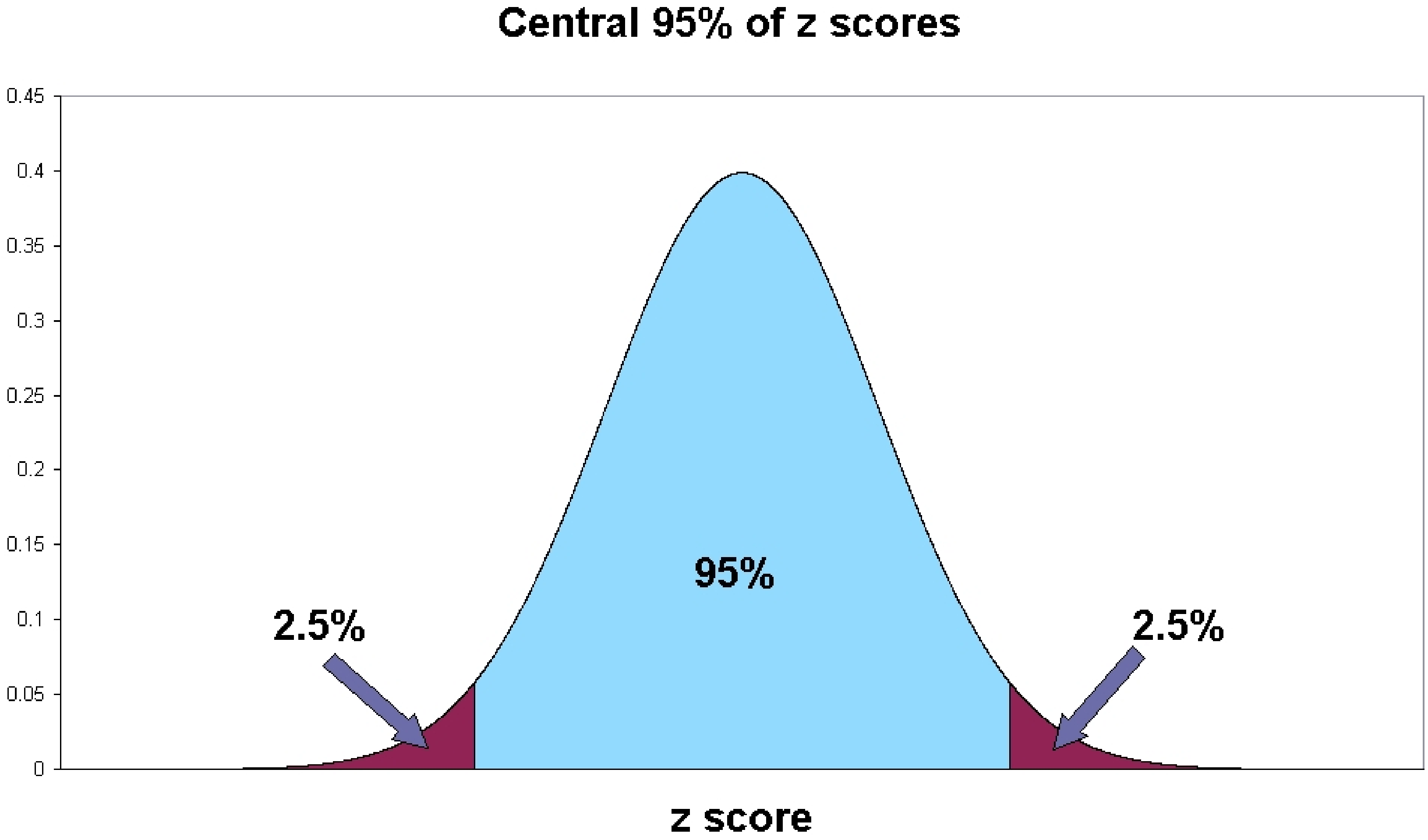
- **1.** Choose the appropriate formula and decide on the level of confidence:

$$\mu = \bar{x}_i \pm z^* \frac{s}{\sqrt{n}}$$

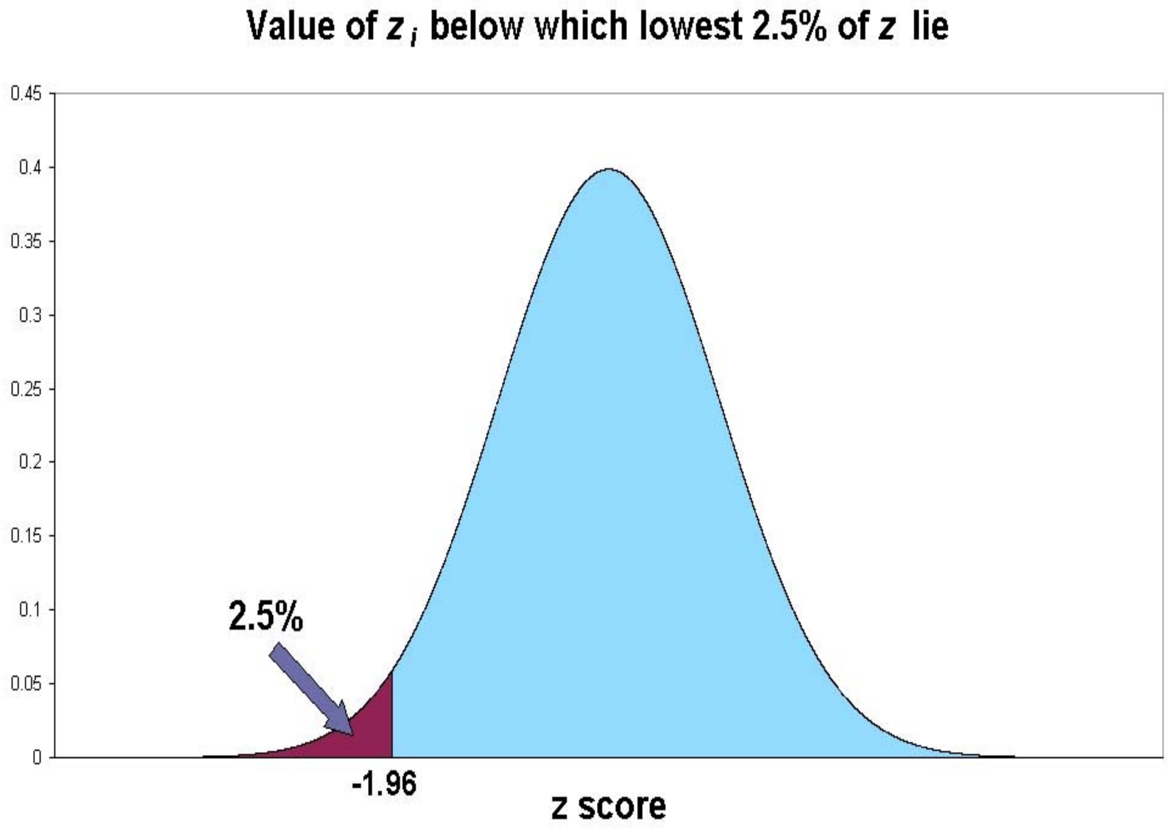
$c = 0.95$

- **2.** Find the value for z* such that:

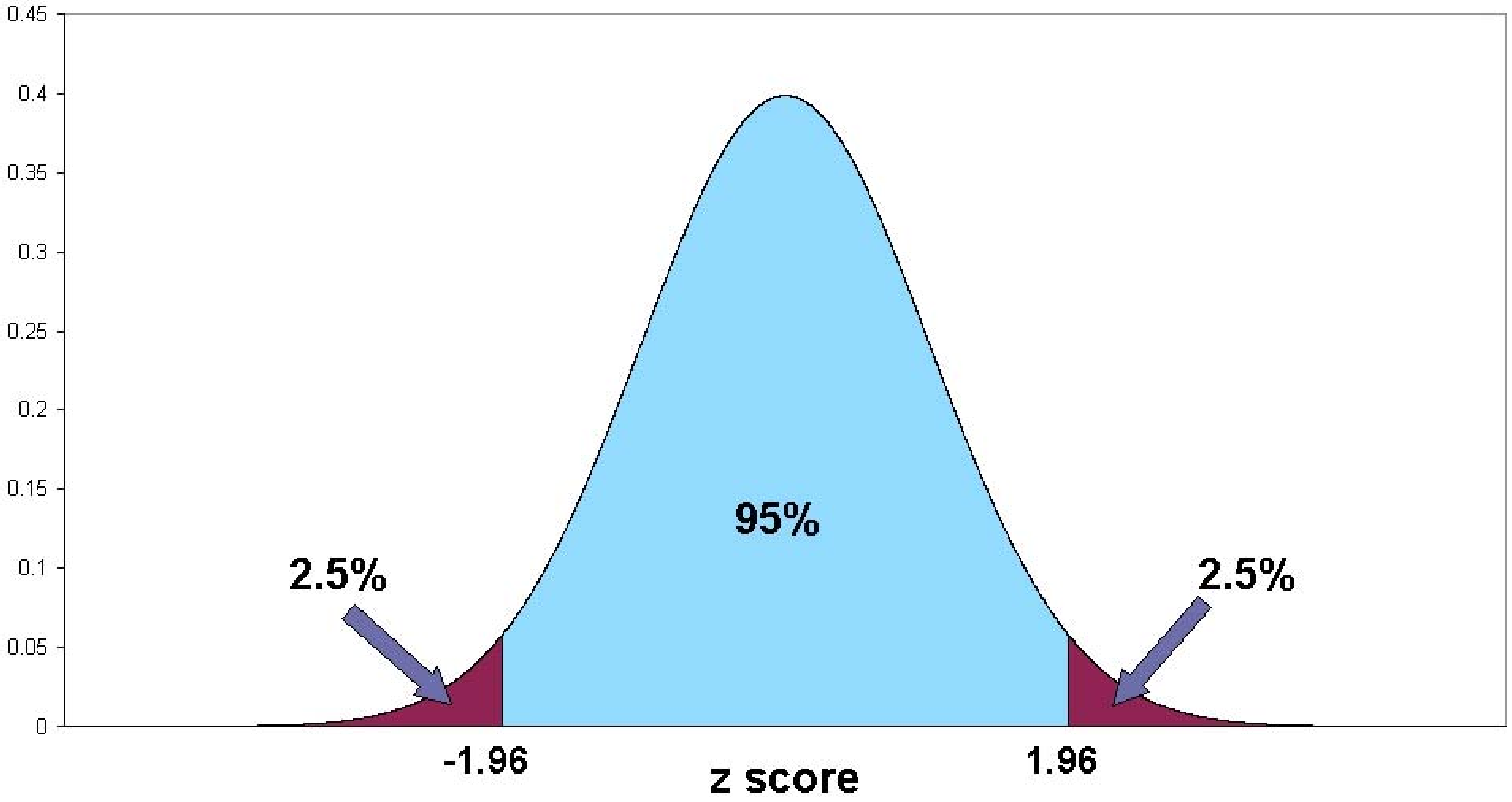
  Prob(-z* < z < z*) = 95%


Central 95% of z scores
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
95%
2.5%
2.5%
z score

look up 0.0250 in the body of the z table which tells us that the value for –z* is 1.96



The z values that partition the central 95%



Alternatively we could use the **zi_gl_zp** syntax for finding the central 95%:

```
zi_gl_zp p = (0.95).
 Value of zi such that Prob(-zi < z < zi)
  = PROB, when PROB is given
         ZIL        ZIU       PROB
    -1.95996    1.95996     .95000
```

3. Calculate the CI by substituting your values into the formula:

$$\mu = \overline{x}_i \pm z^* \frac{s}{\sqrt{n}}$$

$$= 56.2 \pm 1.96 \times \frac{14.7}{\sqrt{506}}$$

$$= 56.2 \pm 1.281$$

- error associated with using the sample mean as an estimate of the population mean =1.281 years.
- I.e. we are 95% certain that the population age of missing workers was between 54.92 years and 57.481 years.
- Note that this range is clearly above our guesstimate of the population mean of 50 years.

CI_L1M **Large sample CI for one mean (M&M pp.417-424) .**

- We could alternatively use the macro:

CI_L1M n=(506) x_bar=(56.2) s=(14.7) c=(0.95).

```
Large sample confidence interval for the population mean
          N      X_BAR        ZIL         SE        ERR      LOWER      UPPER
  506.00000   56.20000   -1.95996     .65349    1.28083   54.91917   57.48083
```

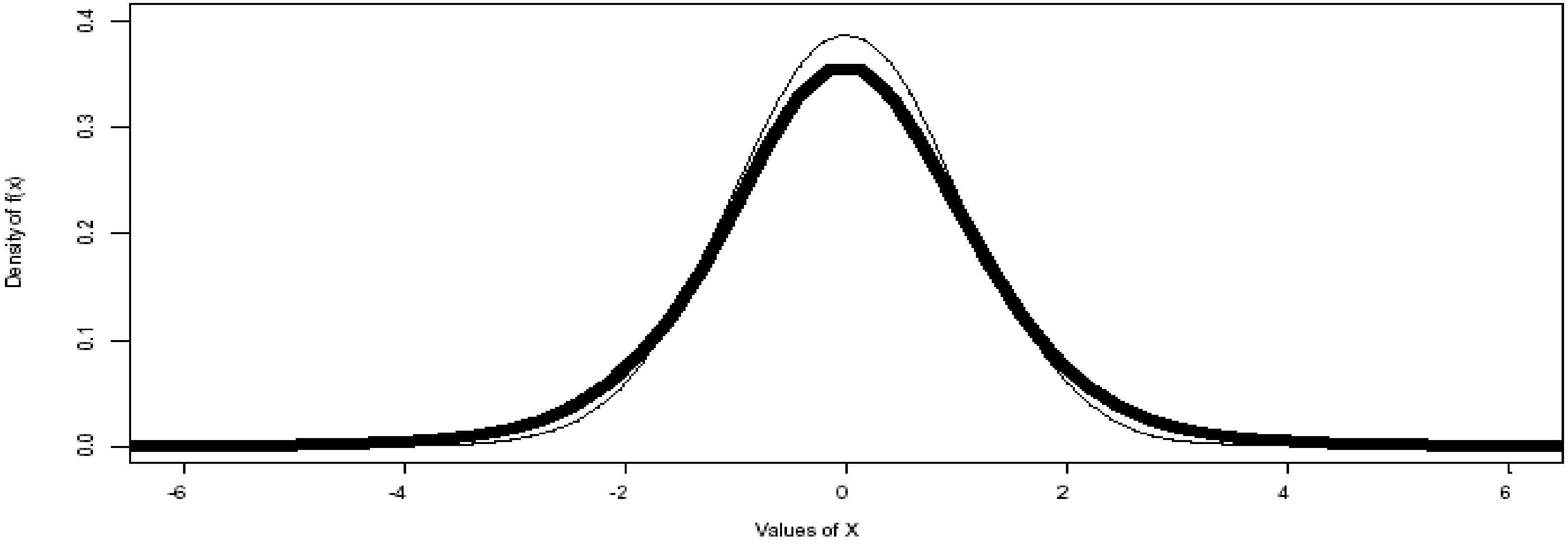
4. Small Sample CIs

- Now let’s look at the second problem of the CLT:
- It depends on the sample being large
- What if our sample is small?

Student’s t-distribution

- We mentioned earlier that we can approximate the standard error of the mean using $s \, / \, \sqrt{n}$
- However, strictly speaking, when we substitute for the SE of the mean in this way, the statistic does *not* have a normal distribution:
  - its distribution is slightly different to the normal distribution and is called the ‘t-distribution’

- Student’s t-distribution varies according to sample size
  - I.e. a different distribution for each sample size
- The spread is slightly larger than the normal distribution due to the substitution of s for σ.
  - but because $s \rightarrow \sigma$ as $n\uparrow$, the t-distribution → normal as $n\uparrow$



Assumption and implication:

- The t-distribution assumes that the variable in question is normally distributed.
- In reality, few variables are normal, but the effect of non-normality in the original variable lessens as the sample size increases
  - as *n* increases, the Central Limit Theorem kicks in.

Three steps of Interval estimation for μ: the *small* sample case

- 1. Choose the appropriate formula and decide on the level of confidence (e.g. 95%):

$$\mu = \bar{x} \pm t^{*} \frac{s}{\sqrt{n}}$$

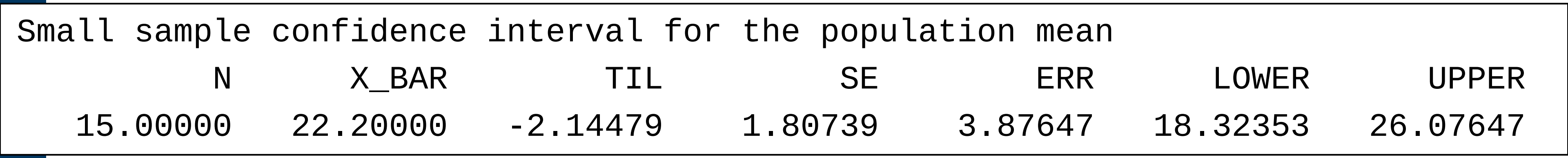
- 2. Find the value for t* such that:
  - Prob($-t^{*} \leq t \leq t^{*}$) = Confidence level (e.g. 95%)
- 3. Calculate the confidence interval by substituting your values for the sample mean, *t** and your approximation for the standard error of the mean ($s/\sqrt{n}$).

- So when the sample size is *small*, the variable is normal:
  - **we always use the Student t-distribution.**
- when the sample size is *large* and the variable is *non-normal* :
  - **we can use the z or t distributions.**
- But when the sample size is *small*, and the variable is *non-normal*:
  - **we can't use the t-distrubution** (or we do so with caution!)
    - => Resort to non-parametric methods (not covered in this course).

Macro syntax for Small Sample CI:

e.g. 95% CI for average age of graduation (*n* = 15, *s* = 7years)

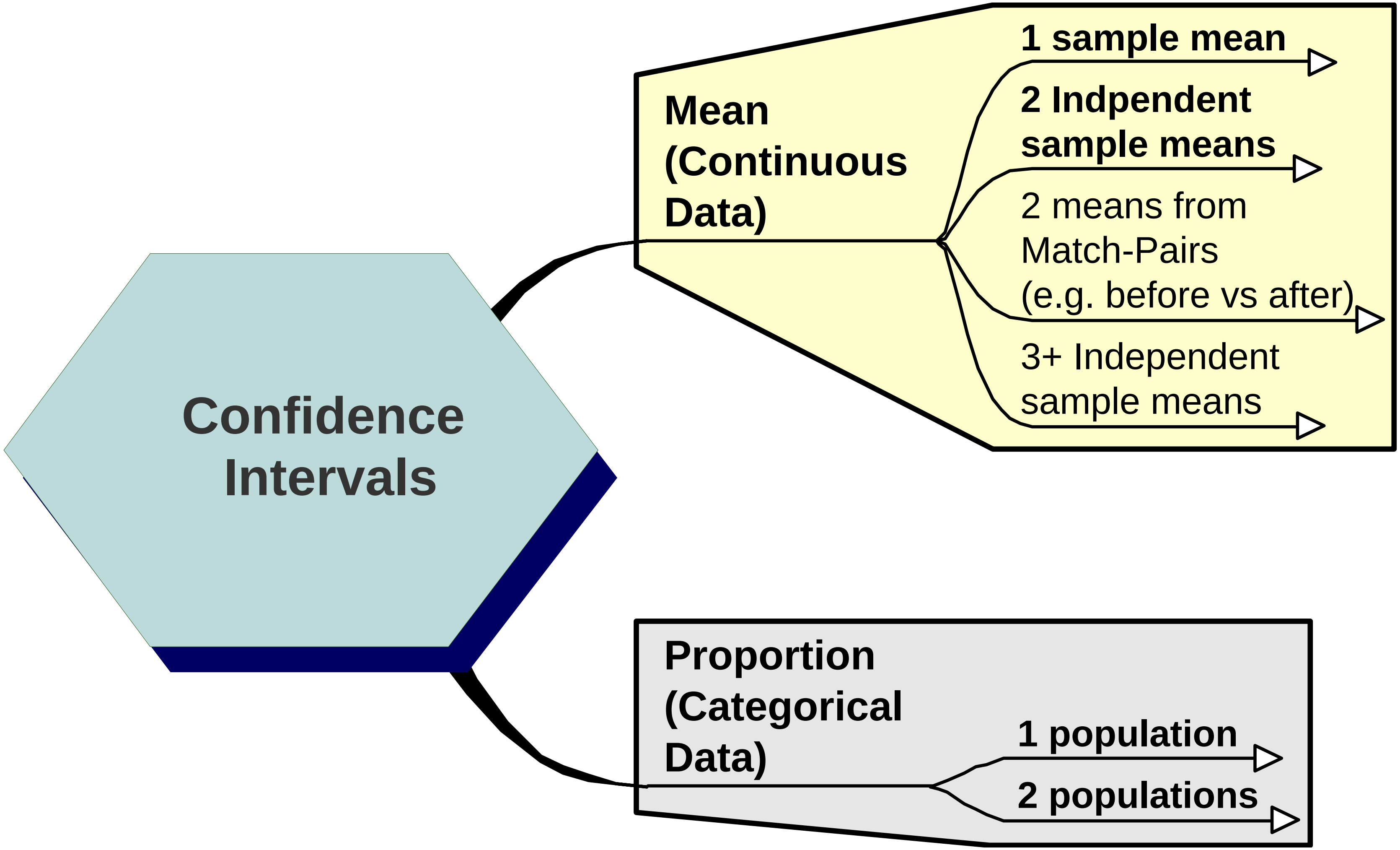
CI_S1M n=(15) x_bar=(22.2) s=(7) c=(0.95).



```
Small sample confidence interval for the population mean
         N     X_BAR       TIL        SE       ERR     LOWER     UPPER
  15.00000  22.20000  -2.14479   1.80739   3.87647  18.32353  26.07647
```

Summary:

- 1. Introduction-
  - Material covered so far
  - Intuition behind CIs
- 2. Three steps of CI Estimation
- 3. Large Sample CI for the mean
  - **CI_L1M** n=(?) x_bar=(?) s=(?) c=(?).
- 4. Small Sample CI for the mean
  - **CI_S1M** n=(?) x_bar=(?) s=(?) c=(?).



Confidence Intervals

Mean (Continuous Data)

- 1 sample mean
- 2 Indpendent sample means
- 2 means from Match-Pairs (e.g. before vs after)
- 3+ Independent sample means

Proportion (Categorical Data)

- 1 population
- 2 populations