Central limit theorem
Confidence intervals

Specifications

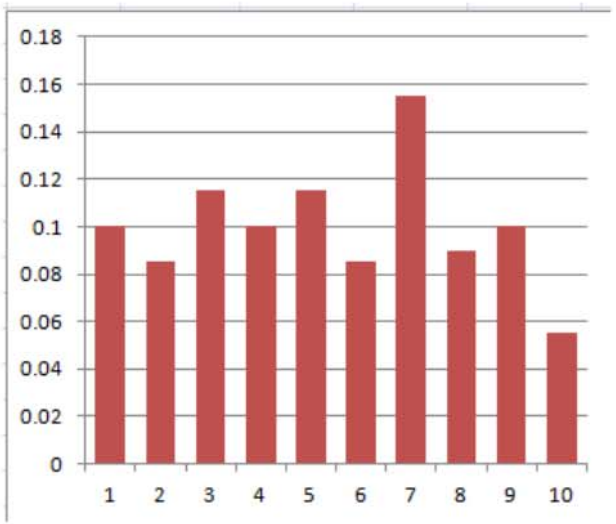
The sampling distribution of the mean of a random sample from a normal distribution.	To include the standard error of the sample mean, $\frac{\sigma}{\sqrt{n}}$, and its estimator, $\frac{S}{\sqrt{n}}$.
A normal distribution as an approximation to the sampling distribution of the mean of a large sample from any distribution.	Knowledge and application of the Central Limit Theorem.
Confidence intervals for the mean of a normal distribution with known variance.	Only confidence intervals symmetrical about the mean will be required.
Confidence intervals for the mean of a distribution using a normal approximation.	Large samples only. Known and unknown variance.
Inferences from confidence intervals.	Based on whether a calculated confidence interval includes or does not include a 'hypothesised' mean value.

Introduction :

Consider this distribution:

Outcome	1	2	3	4	5	6	7	8	9	10
Freq	20	17	23	20	23	17	31	18	20	11
Freq density	0.1	0.085	0.115	0.1	0.115	0.085	0.155	0.09	0.1	0.055

Mean	5.36
Stand dev	2.71
Variance	7.33



Out of these 200 outcomes, we select samples of 30 and we work out the mean $\bar{x}$

Sample 1:

4
7
4
4
4
3
3
4
1
3
10
3
1
7
3
8
7
7
3
8
7
1
7
3
3
2
7
3
1
3

mean:4.4

Sample 2:

2
5
5
7
10
5
4
8
4
1
2
4
5
10
10
1
3
7
5
7
7
6
4
1
4
9
2
6
1
6

mean:5.0

Sample 3:

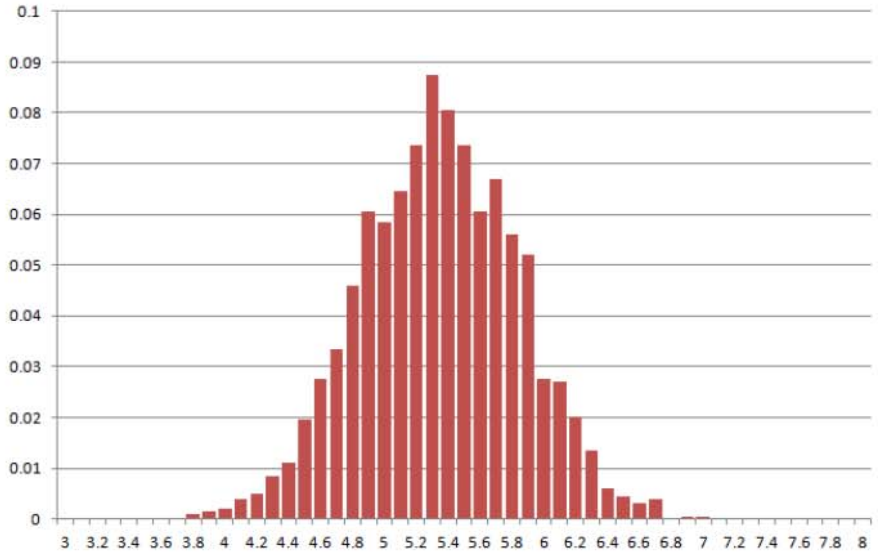
10
5
7
1
7
1
4
5
6
2
4
5
8
4
2
6
5
10
5
9
7
7
6
9
7
8
7
3
6
8

mean:5.8 *etc...*

Here we have a frequency table recording the mean of 2000 samples (rounded to 1decimal place)

3.7	3.8	3.9	4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7	7.1
0	2	3	4	8	10	17	22	39	55	67	92	121	117	129	147	175	161	147	121	134	112	104	55	54	40	27	12	9	6	8	0	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0	0	0.01	0.01	0	0	0	0	0	0	0

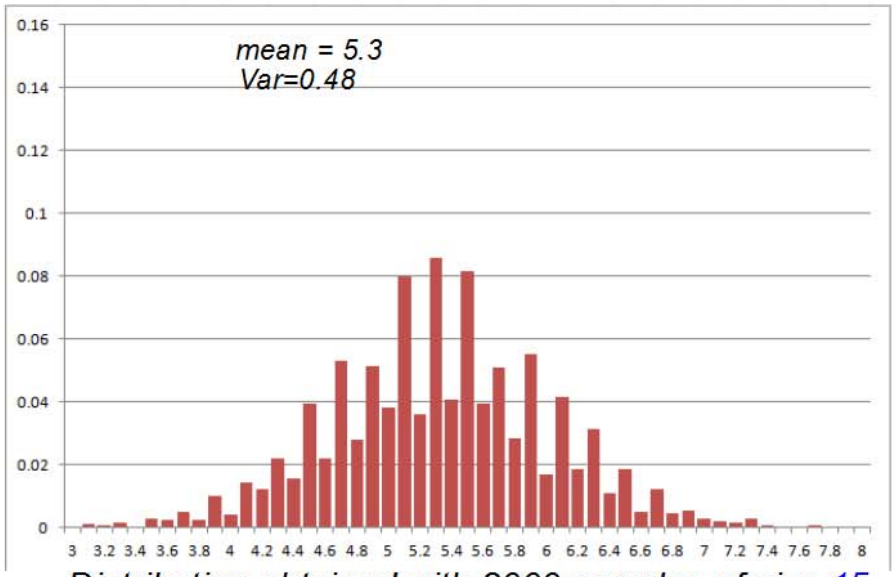
and a representation of this distribution:



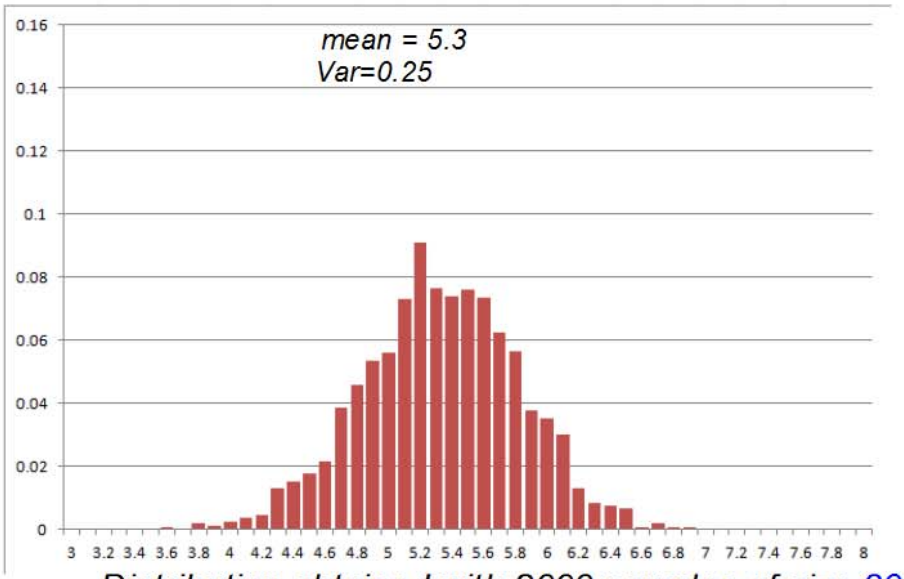
Mean	5.3
Std dev	0.50
Var	0.25

what do you notice?

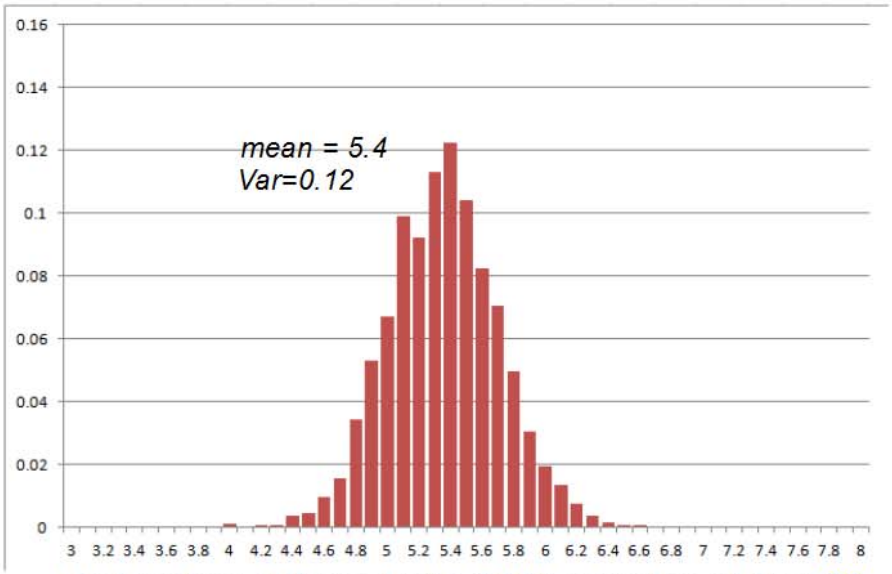
Effect of the size, n ,of the samples:



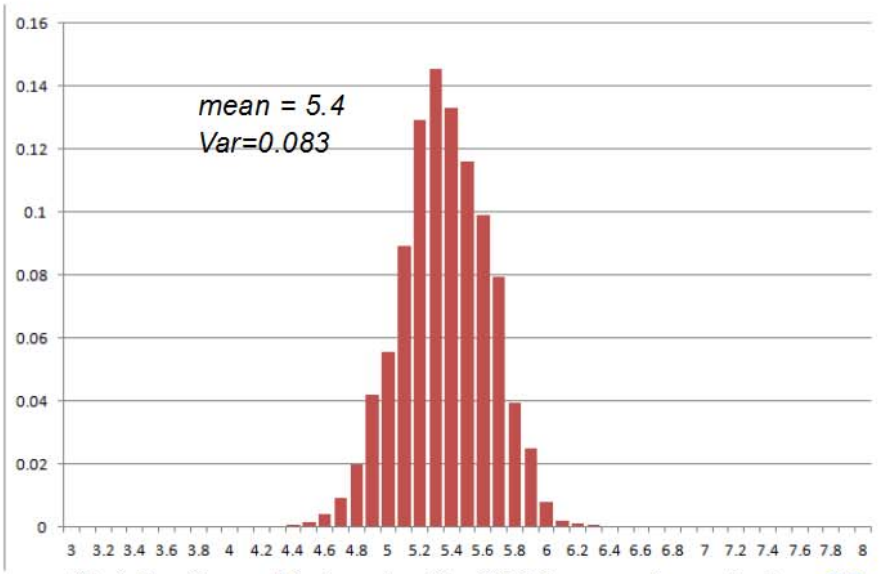
Distribution obtained with 2000 samples of size 15



Distribution obtained with 2000 samples of size 30



Distribution obtained with 2000 samples of size 60



Distribution obtained with 2000 samples of size 90

a) Compare the mean of these distributions to the mean of the original dist.

b) work out " Variance×n" in each case. what do you notice?

Central limit theorem

A bakery makes loaves of bread with a mean weight of 900 g and a standard deviation of 20 g. An inspector selected four loaves at random and weighed them. It is unlikely that the mean weight of the four loaves she chose would be exactly 900 g. In fact the mean weight was 906 g. A second inspector then chose four loaves at random and found their mean weight to be 893 g. There is no limit to how many times a sample of four can be chosen and the mean weight calculated. These means will vary and will have a distribution.

This distribution is known as **the distribution of the sample mean**.

If a random sample of size n is taken from any distribution with mean μ and standard deviation σ then:

- $\bar{x}$, the sample mean, will be distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, *or* variance $\frac{\sigma^2}{n}$
- the distribution will be approximately normal provided n is sufficiently large – the larger the size of n the better the approximation.

Consequence:

As the sample gets larger the standard deviation of the sample mean (sometimes called the **standard error**) gets smaller.

Remember: If a random variable X has mean μ and variance σ^2

then the SAMPLE MEAN $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ if n is large.

(n is the size of the samples)

n > 30 is the value used in the exam questions

Unless X is normally distributed
(then n could be any value)

Example/exercise:

The weights of pebbles on a beach are distributed with mean 48.6 g and standard deviation 8.5 g.

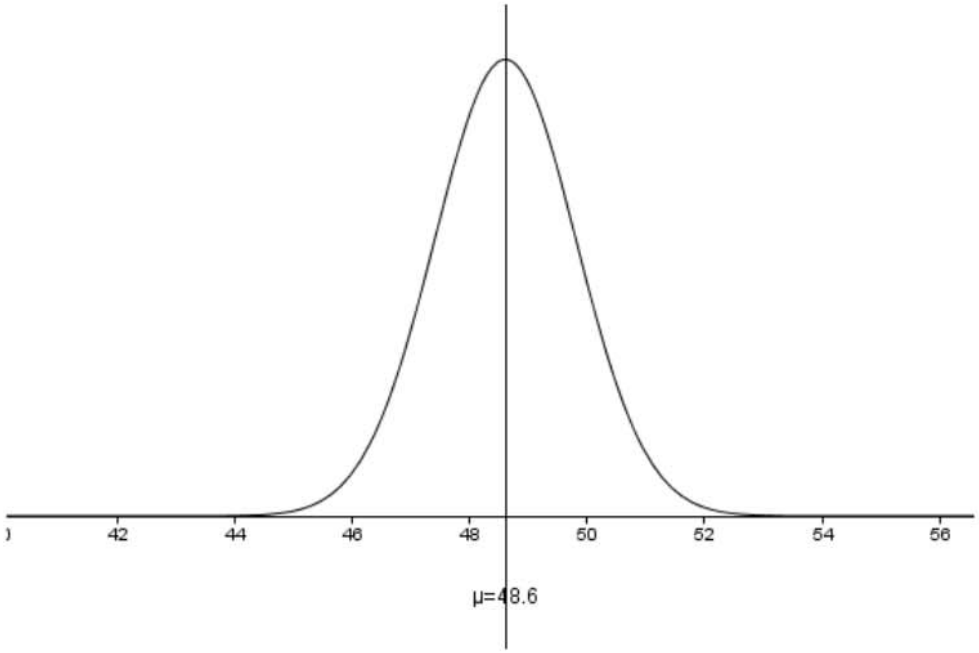
(a) A random sample of 50 pebbles is chosen. Find the probability that:

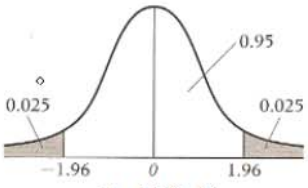
(i) the mean weight will be less than 49.0 g,

(ii) the mean weight will be 47.0 g or less.

(b) Find limits within which the central 95% of such sample means would lie.

(c) How large a sample would be needed in order that the central 95% of sample means would lie in an interval of width at most 4 g?





Z~N(0,1)

a)i)62.9% ii) 9.18%
b) $46.25 < \bar{x} < 50.95$
c) 70

Exercises:

1 A population has a mean of 57.4 kg and a standard deviation of 6.7 kg. Samples of 80 items are chosen at random from this population. Find the probability that a sample mean:

(a) will be 58.4 kg or less,

(b) will be less than 56.3 kg,

(c) will lie between 56.3 kg and 58.4 kg.

2 It is found that the mean of a population is 46.2 cm and its standard deviation is 2.3 cm. Samples of 100 items are chosen at random.

(a) Between what limits would you expect the central 95% of the means from such samples to lie?

(b) What limit would you expect to be exceeded by only 5% of the sample means?

(c) How large should the sample size be in order for the central 95% of such sample means to lie in an interval of width at most 0.8 cm?

3 The times taken by people to complete a task are distributed with a mean of 18.0 s and a standard deviation 8.5 s. Samples of 50 times are chosen at random from this population.

(a) What is the probability that a randomly selected sample mean will:

(i) be at least 19.4 s,

(ii) be 17.5 s or more,

(iii) lie between 17.4 and 19.0 s?

(b) Between what limits would you expect the central 95% of such sample means to lie?

Answers

Interpolation has been used, your answers may be slightly different if you have not used interpolation.

1 **(a)** 0.909; **(b)** 0.0710; **(c)** 0.838.

2 **(a)** 45.75 – 46.65 cm;

(b) 46.58 cm;

(c) 128.

3 **(a)** **(i)** 0.122, **(ii)** 0.661, **(iii)** 0.488;

(b) 15.6 – 20.4 s.

Confidence intervals

Part 1:

Confidence interval for the mean of a normal distribution (Stand dev known)

Example:

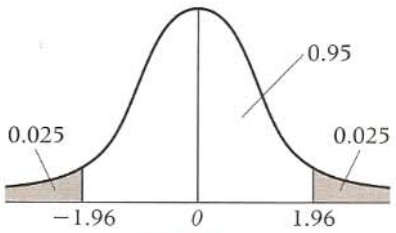
The content of a large batch of packets of baking powder are known to be normally distributed with standard deviation 7g. The mean is unknown.

- We select one packet at random

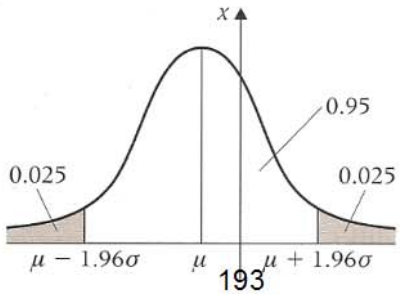
  This packet contains 193g of baking powder.
  Because it is the only information we have, we consider that 193 is the mean.
  (This is called a POINT ESTIMATE)

This is not really satisfactory as other packets would contains different amounts.
It is better to give an estimate of the mean using an interval.

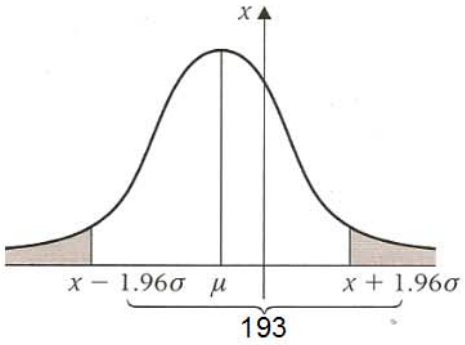
The 95% confidence interval:



Z~N(0,1)



X~N(μ,σ²)



Here,
the 95% confidence interval is:

The interval centred on μ and the interval centred on x have the same width (2x1.96σ)
and there is a 95% chance that μ will belong to the interval centred on x)

Example:

The content of a large batch of packets of baking powder are known to be normally distributed with standard deviation 7g. The mean is unknown.

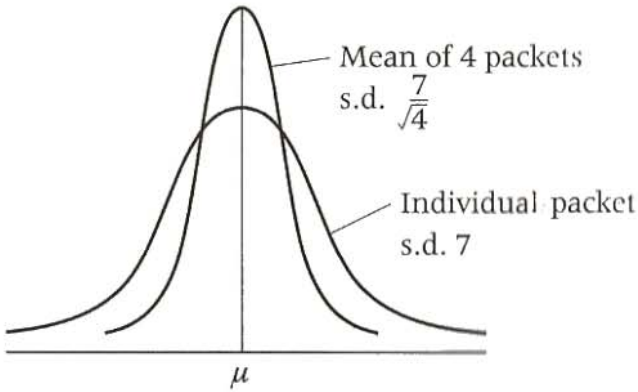
- We select four packets at random

The weights are: 193g , 197g , 212g , 184g

The mean is: 196.5g

The advantage of taking a sample of four, is that according to the central limit theorem,

the standard deviation of the sample mean is $\frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{4}} = 3.5g$



The 95% confidence interval:

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$196.5 - 1.96 \times \frac{7}{\sqrt{4}} \le \mu \le 196.5 + 1.96 \times \frac{7}{\sqrt{4}}$$

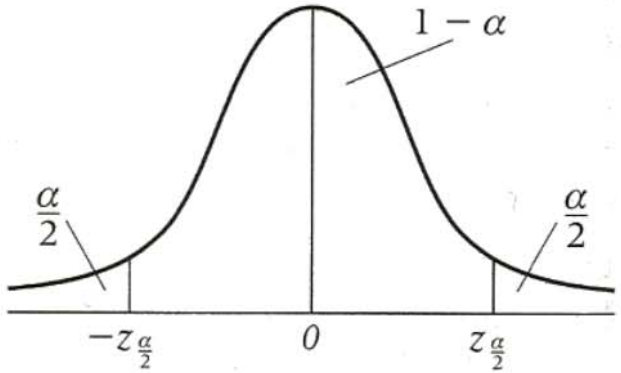
$$189.64 \le \mu \le 203.36$$

Note 1. This interval is half the width of the 95% confidence interval calculated from the weight of a single packet. This has been achieved by increasing the sample size from one to four. However there is very little advantage in increasing a sample of size 21 to one of size 24. To halve the width of the interval you need to **multiply** the sample size by four.

Note 2. If the distribution is not normal the confidence interval will be inaccurate. This could be a major problem for the single observation but would be less serious for the sample of size four. For large samples the sample mean will be approximately normally distributed. Four is not a large sample but the mean of a sample of size four will come closer to following a normal distribution than will the distribution of a single observation.

Conclusion:

If $\bar{x}$ is the mean of a random sample of size n from a normal distribution with (unknown) mean μ and (known) standard deviation σ, a $100(1-\alpha)\%$ confidence interval for μ is given by $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.



Practice exercises:

A machine fills bottles with vinegar. The volumes of vinegar contained in these bottles are normally distributed with standard deviation 6 ml.

A random sample of five bottles from a large batch filled by the machine contained the following volumes, in millilitres, of vinegar:

986 996 984 990 1002

Calculate a 90% confidence interval for the mean volume of vinegar in bottles of this batch.

A food processor produces large batches of jars of jam. In each batch the weight of jam in a jar is known to be normally distributed with standard deviation 7 g. The weights, in grams, of the jam in a random sample of jars from a particular batch were:

481 455 468 457 469 463 469 458

(a) Calculate a 95% confidence interval for the mean weight of jam in this batch of jars.

(b) Assuming the mean weight is at the upper limit of the confidence interval calculated in **(a)**, calculate the limits within which 99% of weights of jam in these jars lies.

987.2 to 996.0 ml
460.1 to 469.9
451.9 to 487.9.

Exercises

1 The potency of a particular brand of aspirin tablets is known to be normally distributed with standard deviation 0.83. A random sample of tablets of this brand was tested and found to have potencies of

58.7 58.4 59.3 60.4 59.8 59.4 57.7 60.3 61.0 58.2

Calculate:

(a) a 99% confidence interval for the mean potency of these tablets,

(b) a 95% confidence interval for the mean potency of these tablets,

(c) a 60% confidence interval for the mean potency of these tablets.

2 The diastolic blood pressures, in millimetres of mercury, of a population of healthy adults has standard deviation 12.8. The diastolic blood pressures of a random sample of members of an athletics club were measured with the following results:

79.2 64.6 86.8 73.7 74.9 62.3

(a) Assuming the sample comes from a normal distribution with standard deviation 12.8, calculate:

(i) a 90% confidence interval for the mean,

(ii) a 95% confidence interval for the mean,

(iii) a 99% confidence interval for the mean.

The diastolic blood pressures of a random sample of members of a chess club were also measured with the following results:

84.6 93.2 104.6 106.7 76.3 78.2

(b) Assuming the sample comes from a normal distribution with standard deviation 12.8, calculate:

(i) an 80% confidence interval for the mean,

(ii) a 95% confidence interval for the mean,

(iii) a 99% confidence interval for the mean.

(c) Comment on the diastolic blood pressure of members of each of the two clubs given that a population of healthy adults would have a mean of 84.8.

3 Applicants for an assembly job are to be given a test of manual dexterity. The times, in seconds, taken by a random sample of applicants to complete the test are

63 229 165 77 49 74 67 59 66 102 81 72 59

Calculate a 90% confidence interval for the mean time taken by applicants. Assume the data comes from a normal distribution with standard deviation 57 s.

4 A rail traveller records the time she has to queue to buy a ticket. A random sample of times, in seconds, were

136 120 67 255 84 99 280 55 78

(a) Assuming the data may be regarded as a random sample from a normal distribution with standard deviation 44 s, calculate a 95% confidence interval for the mean queuing time.

(b) Assume that the mean is at the lower limit of the confidence interval calculated in **(a)**. Calculate limits within which 90% of her waiting times will lie.

(c) Comment on the station manager's claim that most passengers have to queue for less than 25 s to buy a ticket.

5 A food processor produces large batches of jars of pickles. In each batch, the gross weight of a jar is known to be normally distributed with standard deviation 7.5 g. (The gross weight is the weight of the jar plus the weight of the pickles.)
The gross weights, in grams, of a random sample from a particular batch were:

514 485 501 486 502 496 509 491 497
501 506 486 498 490 484 494 501 506
490 487 507 496 505 498 499

(a) Calculate a 90% confidence interval for the mean gross weight of this batch.

The weight of an empty jar is known to be exactly 40 g.

(b) **(i)** What is the standard deviation of the weight of the pickles in a batch of jars?

(ii) Assuming that the mean gross weight is at the upper limit of the confidence interval calculated in **(a)**, calculate limits within which 99% of the weights of the pickles would lie.

(c) The jars are claimed to contain 454 g of pickles. Comment on this claim as it relates to this batch of jars.

Answers

1 **(a)** 58.64–60.00; **(b)** 58.81–59.83; **(c)** 59.10–59.54.

2 **(a)** **(i)** 65.0–82.2, **(ii)** 63.3–83.8, **(iii)** 60.1–87.0;
(b) **(i)** 83.9–97.3, **(ii)** 80.4–100.8, **(iii)** 77.1–104.1;
(c) Athletes seem to have a lower mean diastolic blood pressure than for the population of healthy adults (84.8 is above the 95% confidence interval, although it is just inside the 99% interval). On this evidence chess club members are consistent with the population of healthy adults as 84.8 lies within the confidence intervals.

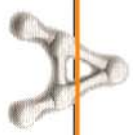
3 63.5–115.5.

4 **(a)** 101.7–159.2; **(b)** 29.3–174.1;
(c) Station manager's claim is incorrect. Even making the lowest reasonable estimate of the mean the great majority of passengers will queue for more than 25 s.

5 **(a)** 494.69–499.63; **(b)** **(i)** 7.5 g, **(ii)** 440.3–478.9;
(c) Confidence interval calculated in **(a)** suggests that the mean weight of pickles in a jar is above 454 g but interval calculated in **(b)** suggests that many individual jars will contain less than 454 g of pickles.

Part 2:

Confidence interval for mean based on a large sample.



If a large random sample is available:

- it can be used to provide a good estimate of the population standard deviation σ,
- it is safe to assume that the mean is normally distributed.

Examples:

Seventy packs of butter, selected at random from a large batch delivered to a supermarket, are weighed. The mean weight is found to be 227 g and the standard deviation is found to be 7.5 g. Calculate a 95% confidence interval for the mean weight of all packs in the batch.

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$227 - 1.96 \times \frac{7.5}{\sqrt{70}} \le \mu \le 227 + 1.96 \times \frac{7.5}{\sqrt{70}}$$

$$225.24 \le \mu \le 228.76$$

A telephone company selected a random sample of size 150 from those customers who had not paid their bills one month after they had been sent out. The mean amount owed by the customers in the sample was £97.50 and the standard deviation was £29.00.
Calculate a 90% confidence interval for the mean amount owed by all customers who had not paid their bills one month after they had been sent out.

£93.61 ≤ μ ≤ £101.39

Exercises

2 A sample of 64 fish caught in the river Mirwell had a mean weight of 848 g with a standard deviation of 146 g. Assuming these may be regarded as a random sample of all the fish caught in the Mirwell, calculate, for the mean of this population:

(a) a 95% confidence interval,

(b) a 64% confidence interval.

3 A boat returns from a fishing trip holding 145 cod. The mean length of these cod is 74 cm and their standard deviation is 9 cm. The cod in the boat may be regarded as a random sample from a large shoal. The normal distribution may be regarded as an adequate model for the lengths of the cod in the shoal.

(a) Calculate a 95% confidence interval for the mean length of cod in the shoal.

(b) It is known that the normal distribution is not a good model for the weights of cod in a shoal. If the cod had been weighed, what difficulties, if any, would arise in calculating a confidence interval for the mean weight of cod in the shoal? Justify your answer. [A]

4 A sweet shop sells chocolates which appear, at first sight, to be identical. Of a random sample of 80 chocolates, 61 had hard centres and the rest soft centres. The chocolates are in the shape of circular discs and the diameters, in centimetres, of the 19 soft-centred chocolates were:

2.79 2.63 2.84 2.77 2.81 2.69 2.66 2.71 2.62 2.75
2.77 2.72 2.81 2.74 2.79 2.77 2.67 2.69 2.75

The mean diameter of the 61 hard-centred chocolates was 2.690 cm.

(a) If the diameters of both hard-centred and soft-centred chocolates are known to be normally distributed with standard deviation 0.042 cm, calculate a 95% confidence interval for the mean diameter of:

(i) the soft-centred chocolates,

(ii) the hard-centred chocolates.

(b) Calculate an interval within which approximately 95% of the diameters of hard-centred chocolates will lie.

(c) Discuss, briefly, how useful knowledge of the diameter of a chocolate is in determining whether it is hard- or soft-centred. [A]

1 £93.6–£101.4.

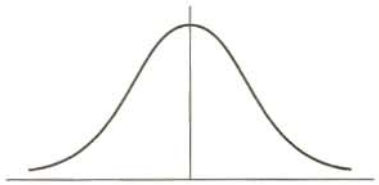
2 **(a)** 812.2–883.8 g; **(b)** 831.3–864.7 g.

3 **(a)** 72.54–75.46 cm;
(b) No difficulty as sample is large so mean will be approximately normally distributed.

4 **(a)** **(i)** 2.717–2.755 cm, **(ii)** 2.679–2.701 cm;
(b) 2.608–2.772 cm;
(c) Confidence intervals do not overlap so mean for soft centres clearly greater than mean for hard centres. However interval calculated in **(b)** shows that many hard-centred chocolates are bigger than the mean of the soft-centred chocolates. Diameter not a great deal of use because of large amount of overlap.

Key points to remember

1 The main features of the normal distribution are that it:
- is bell shaped and continuous,
- is symmetrical about the mean (and median and mode),
- has total area under the curve equal to 1.



2 $z = \dfrac{(x - \mu)}{\sigma}$ is the **standard normal variable** or z score, with mean 0 and standard deviation 1, where x is an observation from a Normal distribution with mean μ and standard deviation σ.

3 Table 3 in the AQA Formulae Book gives the probability p, that a normally distributed variable Z, with mean 0 and standard deviation 1, is less than or equal to a particular value z.

Such probabilities may also be obtained from the normal cumulative distribution function on graphical calculators.

4 Table 4 in the AQA Formulae Book gives the z score for a given probability p where $P(Z \leqslant z) = p$.

Some graphical calculators have an inverse normal function where z scores can be obtained from given probabilities.

5 If a random sample of size n is taken from any distribution with mean μ and standard deviation σ, then $\bar{x}$, the sample mean, will be distributed with mean μ, and standard deviation $\dfrac{\sigma}{\sqrt{n}}$ (which is sometimes called the **standard error**).

6 The **central limit theorem** states that the distribution of $\bar{x}$ will be approximately normal, provided n is sufficiently large (at least 30).

Key points to remember

1 If $\bar{x}$ is the mean of a random sample of size n from a Normal distribution with (unknown) mean μ, and (known) standard deviation σ, a $100(1 - a)\%$ confidence interval for μ, is given by $\bar{x} \pm z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$

2 If a large random sample is available:
- it can be used to provide a good estimate of the population standard deviation σ,
- it is safe to assume that the mean is normally distributed.