

9 ESTIMATION

Objectives

After studying this chapter you should

- appreciate the importance of random sampling;
- understand the Central Limit Theorem;
- understand the concept of estimation from samples;
- be able to determine unbiased estimates of the variance;
- be able to find a confidence interval for the mean, μ .

9.0 Introduction

How will 'first time' voters cast their votes in a general election?

How do they differ from older voters?

Which issues concern them most?

Before these questions are considered it is worth noting a few ideas about statistics itself.

Firstly, if everyone was the same, there would be no need for statistics or statisticians; you could find out everything you needed to know from one person (or one event or one result). Statistics involves the study of variability so that estimates and predictions can be made in complex situations where there is no certain answer. The quality and usefulness of these predictions depend entirely on the quality of the data upon which they are based.

Activity 1

Consider again the three questions above.

Talk with other people in your group and decide:

- Which groups of people are referred to in the questions?
- How can each target group be defined? (i.e. How can you decide whether a person belongs to either group or not?)
- How can the information be obtained?
- Is it feasible to obtain information from all members of a population?
- Why might taking a sample/samples be a good idea?

- (f) Could a sample survey possibly give better quality information than a census of the whole population?
-

9.1 Sampling methods

Methods of sampling have already been considered in Chapter 2; some of them will be revised here. You will need the 'fish' sheet from Section 7.2 (p135).

Activity 2 Finding the mean by sampling

A : non random samples

- Select a **sample** of 5 fish which you think are representative.
- Measure the length of each fish in your sample (in mm).
- Calculate the mean length of the 5 fish in your sample and record your result.
- Repeat this for two more samples.
- Collect the results for everyone in your class and record them on a stem and leaf diagram (or frequency table).

B : random samples

Note that the fish are numbered from 1 to 57.

Use 2-figure random numbers from a random number table, calculator or computer to select a sample of 5 fish from the population. The method is described here. For 3-figure random numbers from a calculator, decide in advance whether you will use the first two digits, the last two digits, or the first and last.

Some of your 2-figure numbers will be larger than 57. These can be ignored without affecting the fairness of the selection process.

Here is an example showing a line of random numbers from a table:

25	82	33	06	74	18	34	09
	↓			↓			
	ignore			ignore			

The fish selected are numbered

25 33 6 18 and 34.

Note that you must use random numbers **consecutively** from the table after making a random start. You may **not** move about at will selecting numbers from different parts of the table.

Measure the lengths of these 5 fish as before.

Find the mean length of your sample.

Collect together sample means from all the students in your group. Display your results on a stem and leaf diagram.

Comparing sets of sample means

Compare the two stem and leaf diagrams for your sample means.

What do you notice?

Activity 3 Analysing the results

Answer the following questions with reference to your two sets of results from Activity 2. Firstly, though, measure the lengths of all fish on the sheet and find the true population mean, μ .

- How close were your results to the true population mean?
- How many samples under estimated μ ?
- How many samples over estimated μ ?
- Is either of your two sets of samples biased?

Definitions

To clarify your ideas, precise definitions will now be given.

Population

A **population** is the set of all elements of interest for a particular study. Quantities such as the population mean μ are known as **population parameters**.

Sample

A **sample** is a subset of the population selected to represent the whole population. Quantities such as the sample mean \bar{x} are known as **sample statistics** and are **estimates** of the corresponding population parameters.

Random sample

A **random sample** is a sample in which each member of the population has an equal chance of being selected. Random samples generate **unbiased** estimates of the population mean, whereas non-random samples may not be unbiased. Also, the variability within random samples can be mathematically predicted (as the next section will show).

9.2 Sample size

The next experiment will consider the significance of the sample size. As the sample gets larger, so the estimate of the sample mean should become closer to the true population mean.

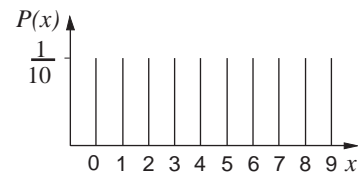
Activity 4 *Selecting your samples*

In this experiment you will need a table of random numbers or a calculator or computer to generate random numbers.

- (a) Select a sample consisting of five single-digit random numbers (taking them consecutively from the random number table after making a random start). If you are using a computer or random number tables you require single-digit random numbers, so use each digit, one at a time. Treat three-figure random numbers from a calculator as three single numbers for your sample.
- (b) Record these values together with their mean.
- (c) Repeat this for four more samples (continuing to use consecutive random numbers).
- (d) Now repeat the procedure for five samples each consisting of ten single-digit random numbers. Record your results together with their mean.
- (e) Collect together the class results for means of samples of size $n = 5$ and $n = 10$ separately.
- (f) Calculate the means of your two groups of sample means, and also the variances. Enter the values of \bar{x} obtained for samples of size $n = 5$ into your calculator. Use the statistical functions to find $\bar{\bar{x}}$, the mean of the sample means, and its standard deviation $\sigma_{\bar{x}}$. Square this second result to find the variance of the \bar{x} 's.
Repeat for samples of size $n = 10$.
- (g) Which group of sample means is more variable?

Before any further analysis or discussion can be undertaken, the population mean and variance must be known.

The population of single-digit random numbers is theoretically infinite and consists of the numbers 0 to 9. These occur with equal probabilities and form a discrete uniform distribution, which you have already met in Chapter 4.



What is the value of $p(x)$ for $x = 0, 1, \dots, 9$?

Activity 5 Exploring population parameters

- (a) Use the formulae

$$\mu = \sum x p(x) \quad , \quad \sigma^2 = \sum x^2 p(x) - \mu^2$$

to find the mean and variance of the population of single-digit random numbers.

- (b) Do your class distributions for \bar{x} ($n = 5$ and $n = 10$) appear to be uniform distributions?

How would you describe them?

- (c) Do any values of \bar{x} appear to be more likely than others?

- (d) Compare the mean of \bar{x} with the population mean μ for $n = 5$ and $n = 10$.

- (e) For samples of size $n = 5$, compare the variance of \bar{x} with the population variance.

Is it close to $8.25 \div 5$?

- (f) For samples of size $n = 10$, compare the variance of \bar{x} with the population variance.

Is it close to $8.25 \div 10$?

9.3 The distribution of \bar{X}

Consider the idea of taking samples from a population. If it is a large population, it is possible (but perhaps not practical) to take a large number of samples, all of the same size from that population. For each sample, the mean \bar{x} can be calculated. The value of \bar{x} will vary from sample to sample and, as a result, is itself a random variable having its own distribution.

The value of the mean from any one sample is known as \bar{x} . If the distribution of all the possible values of the sample means is considered, this theoretical distribution is known as the **distribution of \bar{X}** .

So \bar{X} itself is a random variable which takes different values for different random samples selected from a population.

In general, sample means are usually less variable though, than individual values. This is because, within a sample of size $n = 10$, say, large and small values in the sample tend to cancel each other out when \bar{x} is calculated. In the example in the last section, even in a sample of ten random digits, \bar{x} is unlikely to take a value greater than 7 or less than 2. Larger samples will generate values of \bar{X} which are even more restricted in range (less variable).

There is an inverse relationship between the size of the samples and the variance of \bar{X} . Also, the distribution of \bar{X} tends to be a peaked distribution (with mean and mode at μ) which approaches a normal distribution for large samples.

The Central Limit Theorem

The Central Limit Theorem describes the distribution of \bar{X} if **all** possible random samples (of a given size) are selected from a population. The following results hold.

1. The mean of all possible sample means is μ the population mean; i.e.

$$E(\bar{X}) = \mu$$

2. The variance of the sample means is the population variance divided by the sample size

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

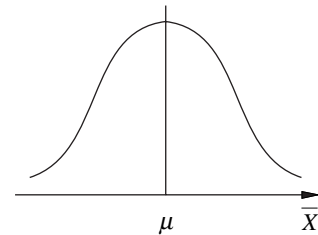
As n increases, the variance of \bar{X} decreases and as $n \rightarrow \infty$

$$V(\bar{X}) \rightarrow 0.$$

3. If all possible values of \bar{X} are calculated for a given sample size $n \geq 30$ a normal distribution is formed irrespective of the distribution of the original population: i.e. for $n \geq 30$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Note that these results are true only for **random samples**. For non-random samples you cannot make predictions in terms of mean, variance or distribution of \bar{X} .



Activity 6 Computer follow up

Use a computer package to investigate the distribution of \bar{X} for random samples

- (a) of different size, n ;
- (b) selected from different populations.

9.4 Identifying unusual samples

Afzal believes that the packets of crisps in the school tuck-shop are underweight. He takes a sample of ten packets of salt and vinegar crisps and finds their mean weight is 24.6 g. As the weight stated on the packets is 25 g, he writes to the manufacturer to complain. He receives the following reply :

Dear Sir,

Thank you for your letter of 5th July. We do share your concern over the weight of crisps in our packets of salt and vinegar crisps.

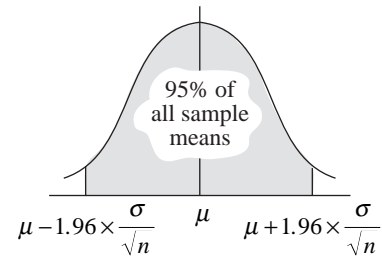
Over a period of time, we have found that the standard deviation of the weights of individual packets is a little below 1 g. For this reason we believe that your sample mean weight of 24.6 g comes well within the normal limits of acceptability.

Yours faithfully,

Does this reply give a valid argument?

The Central Limit Theorem can be used in practical situations like this to identify **unusual** samples, which are not typical of the population from which they have been selected.

For a given size of sample, the distribution of all possible sample means forms a normal distribution. The mean of this distribution is μ (the overall population mean) and the variance is $\frac{\sigma^2}{n}$



Distribution of all possible sample means for random samples of size n

Referring to normal distribution tables, 95% of any normal distribution lies between $z = -1.96$ and $z = +1.96$. So for a particular sample size n , 95% of all sample means should lie within 1.96 times the standard error each side of μ .

A sample mean outside this range may, in general, be :

- a genuine 'freak' result; after all, 5% of random samples do give means outside these limits;
- a result from a random sample selected from a different population;
- a sample selected from the population specified but not a random sample (e.g. a high proportion of children with above average IQs due to school selection procedures).

Example

A survey of adults aged 16-64 living in Great Britain, by the Office of Population Censuses and Surveys (OPCS), found that adult females had a mean height of 160.9 cm with standard deviation of 6 cm.

A sample of fifty female students is found to have a mean height of 162 cm. Are their heights typical of the general population?

Solution

The population mean is given by

$$\mu = 160.9 \text{ cm.}$$

Since the sample size is $n = 50$, the standard error is given by

$$\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.849.$$

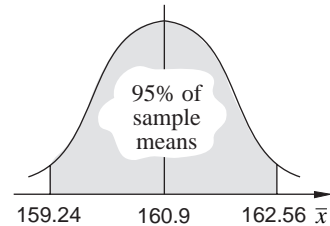
Thus the range of values for 95% of all sample means ($n = 50$) is

$$160.9 - (1.96 \times 0.849) \leq \bar{x} \leq 160.9 + (1.96 \times 0.849)$$

$$160.9 - 1.66 \leq \bar{x} \leq 160.9 + 1.66$$

So 95% of all \bar{x} should lie in the range $159.24 \leq \bar{x} \leq 162.56$.

You can see that the sample mean of 162 cm obtained from the fifty students is within the range of typical values for \bar{x} . So there is no evidence to suggest that this sample is not typical of the population in terms of height.



These ideas can be used to identify unusual sample means for large ($n \geq 30$) random samples selected from any population, or for small samples selected from a normal population provided the value of the population variance is known.

Exercise 9A

1. IQ (Intelligence Quotient) scores are measured on a test which is constructed to give individual scores forming a normal distribution with a mean of 100 points and standard deviation of 15 points. A random sample of 10 students achieves a mean IQ score of 110 points. Is this sample typical of the general population?
2. A large group of female students is found to have a mean pulse rate (resting) of 75 beats per minute and standard deviation of 12 beats. Later, a class of 30 students is found to have a mean pulse rate of 82 beats per minute. What are your conclusions?

3. Over the summer months, samples of adult specimens of freshwater shrimps are taken from a slow moving stream. Their lengths are measured and found to have a mean of 39 mm and standard deviation of 5.3 mm. During the winter, a small sample of 10 shrimps is found to have a mean length of 41 mm.
 - (a) Have the shrimps continued growing in the colder weather?
 - (b) What assumptions have you had to make in order to answer the question?
4. Re-read the crisps problem at the beginning of this section. Do you agree with Afzal or do you agree with the manufacturers? Would it help Afzal to take a larger sample?

9.5 Confidence intervals

In many situations the value of μ , the population mean, may not be known for the variable being measured.

Is it possible to estimate the value of μ in such cases?

The best estimate of μ is the value of \bar{x} obtained from a random sample. As the estimate consists of a single value, \bar{x} , it is referred to as a **point estimate**. (Other less reliable point estimates can be obtained from the sample median or mid-range.) The sample mean \bar{x} is an **unbiased estimator** for μ , but even so, the value of \bar{x} obtained from any particular random sample is unlikely to give the exact value of μ . In fact, as an unbiased estimator, half the values of \bar{x} will under estimate μ , while half will give over estimates.

In order to 'hedge our bets' a range of values may be given which should include the value of μ . This is called an interval estimate or **confidence interval**.

The ideas introduced in earlier sections can be used to construct such an interval estimate.

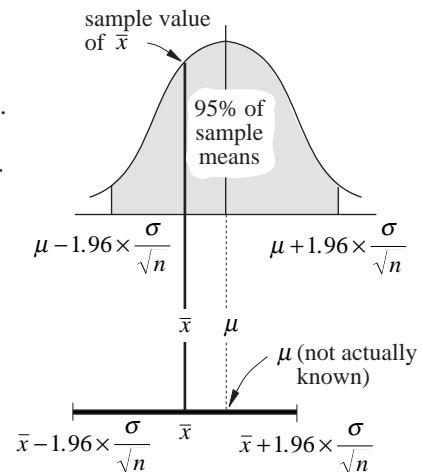
Population variance known

The distribution of all possible sample means, \bar{X} , forms a normal distribution, with a mean μ , at the true population mean.

(The variance of this distribution is $\frac{\sigma^2}{n}$ and decreases for larger sized samples.)

In reality, you are unlikely to know μ and all you have is one sample result \bar{x} . (Now \bar{x} could lie anywhere in the distribution as shown in the diagram opposite.)

Distribution of all possible values of \bar{x} obtained from random samples of size n .



In order to estimate μ , a range of values can be taken around \bar{x} which hopefully will include the true value of μ .

The **95% confidence interval** for μ is found by taking a range of 1.96 times the standard error either side of \bar{x} ; that is

$$\left(\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}} \right).$$

Providing \bar{x} lies within the central 95% of the distribution of all possible sample means, the confidence interval will include μ . This will happen for 95 random samples out of 100. If the sample is a 'freak' sample and the sample value of \bar{x} lies at one of the extreme ends of the distribution, the confidence interval will not include μ . This will happen for only 5 random samples (roughly) out of every 100.

If this margin of error is to be reduced a wider interval (which will include μ for, say, 99 samples out of 100) can be constructed. This is called a **99% confidence interval**.

What values of the standard variable z trap 99% of the distributions?

Tables of the normal distribution give $z = \pm 2.58$ to give 0.5% of the distribution in each tail.

Since the standard error is $\frac{\sigma}{\sqrt{n}}$, the range of values for which 99% of all sample means should lie is

$$\left(\bar{x} - 2.58 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \times \frac{\sigma}{\sqrt{n}} \right).$$

This gives the required confidence interval.

For 1 sample in every 100 the confidence interval will **not** include μ .

For other confidence intervals, e.g. 90%, 98%, you can look up the appropriate z value in the normal distribution tables.

Example

A random sample of 100 men is taken and their mean height is found to be 180 cm. The population variance $\sigma^2 = 49 \text{ cm}^2$. Find the 95% confidence interval for μ , the mean height of the population.

Solution

$$\text{Lower limit} = \bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$$\text{upper limit} = \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

when the standard error is given by

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{49}{100}} = 0.7$$

Hence the 95% confidence interval for μ is given by

$$\begin{aligned} &= (180 \pm 1.96 \times 0.7) \text{ cm} \\ &= (180 \pm 1.37) \text{ cm.} \end{aligned}$$

So μ should lie between 178.63 and 181.37.

It should be noted at this stage that a **parameter** is a measure of a population, e.g. the population mean μ , or population variances, etc.; whilst a **statistic** is a similar measure taken from a sample, e.g. the sample mean \bar{x} .

So a statistic is an **estimator** of a parameter. When investigating a practical problem, it is unlikely that information concerning all the items in a given population will be available. Knowledge will normally be limited to one sample, from which tentative conclusions may be drawn concerning the whole population from which the sample is taken. The larger the sample, the greater the confidence in the estimation.

Activity 7

The lifetimes of 10 light bulbs were observed (in hours) as

1052 1271 836 962 1019 1051 512 1027 1219 1040

Assuming that the standard deviation for light bulbs of this type is 80 hours,

- find the 95% confidence interval for the mean lifetime of this type of bulb;
 - find the % of the confidence interval that has a total range of 80 hours;
 - determine the sample size, n , needed to restrict the range of the 95% confidence interval to 50 hours.
-

Population variance unknown

If the population variance is unknown and a large sample is taken, then the variance must be established from the sample itself. If s^2 is the **sample** variance, the best estimate for the **population** variance is given by

$$\begin{aligned}\hat{\sigma}^2 &= \frac{ns^2}{(n-1)} && \text{(this is derived in the text } \textit{Further Statistics}) \\ &= \frac{n}{(n-1)} \left(\frac{1}{n} \sum x^2 - \bar{x}^2 \right) \\ &= \frac{1}{(n-1)} (\sum x^2 - n\bar{x}^2).\end{aligned}$$

[This quantity is shown on calculators as σ_{n-1} or s_{n-1} .]

This result is used in the next example.

Example

A user of a certain gauge of steel wire suspects that its breaking strength, in newtons (N), is different from that specified by the manufacturer. Consequently the user tests the breaking strength, x N, of each of a random sample of nine lengths of wire and obtains the following *ordered* results.

72.2 72.9 73.4 73.8 74.1 74.5 74.8 75.3 75.9

$$[\sum x = 666.9 \quad \sum x^2 = 49\,428.25].$$

Calculate the mean and the variance of the sample values.

Hence calculate a 95% confidence interval for the mean breaking strength.

Comment upon the manufacturer's claims that the breaking strength of the wire has a mean of 75. (AEB)

Solution

For the sample,

$$\begin{aligned}\text{mean} &= \frac{\sum x_i}{n} \\ &= \frac{666.9}{9} \\ &= 74.1\end{aligned}$$

$$\begin{aligned}\text{variance} &= \frac{1}{n} \sum x^2 - \bar{x}^2 \\ &= \frac{49428.25}{9} - (74.1)^2 \\ &\approx 1.218.\end{aligned}$$

The estimate of the population variance is given by

$$\begin{aligned}\hat{\sigma}^2 &= \frac{n}{(n-1)} s^2 \\ &= \frac{9}{8} \times 1.218 \\ &= 1.370.\end{aligned}$$

$$\Rightarrow \sigma \approx 1.170$$

The 95% confidence interval is now given by

$$\begin{aligned}&\left(\bar{x} - 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\hat{\sigma}}{\sqrt{n}} \right) \\ \Rightarrow &\left(74.1 - 1.96 \times \frac{1.170}{\sqrt{9}}, 74.1 + 1.96 \times \frac{1.170}{\sqrt{9}} \right) \\ \Rightarrow &(73.34, 74.86).\end{aligned}$$

So the manufacturer's claim of a mean of 75 N is unlikely to be true since it is not included in the 95% confidence interval.

9.6 Miscellaneous Exercises

- A sample of size 250 has mean 57.1 and standard deviation 11.8.
 - Find the standard error of the mean.
 - Give 95% confidence limits for the mean of the population.
- A company making cans for lemonade wishes to print 'Average contents x ml' on their cans, and to be 99% confident that the true mean volume is greater than x ml. The volume of lemonade in a can is known to have a standard deviation of 3.2 ml, and a random sample of 50 cans contained a mean volume of 503.6 ml. What volume of x should be stated?
- Butter is sold in packs marked as salted or unsalted and the masses of the packs of both types of butter are known to be normally distributed. The mean mass of the salted packs of butter is 225.38 g and the standard deviation for both packs is 8.45 g.

A sample of 12 of the unsalted packs of butter had masses, measured to the nearest gram, as follows.

219 226 217 224 223 216 221 228 215 229 225 229

Find a 95% confidence interval for the mean mass of unsalted packs of butter.

Calculate limits between which 90% of the masses of salted packs of butter will lie.

Estimate the size of sample which should be taken in order to be 95% sure that the sample mean of the masses of salted packs does not differ from the true mean by more than 3 g.

State, giving a reason, whether or not you would use the same sample size to be 95% sure of the same accuracy when sampling unsalted packs of butter. (AEB)

4. The lengths of a sample of 100 rods produced by a machine are given below.

Length (cm)	5.60-5.62	5.62-5.64	5.64-5.66	5.66-5.68	5.68-5.70	5.70-5.72	5.72-5.74	5.74-5.76	5.76-5.78	5.78-5.80
Number of rods	1	3	5	5	8	20	24	16	12	6

Find the mean and standard deviation of the lengths in this sample.

Estimate the standard error of the mean, and give 95% confidence limits for the true mean length, μ , of rods produced by the machine.

Explain carefully the meaning of these confidence limits.

By taking a larger sample, the manufacturers wish to find 95% confidence limits for μ which differ by less than 0.004 cm. Find the smallest sample size needed to do this.

5. A piece of apparatus used by a chemist to determine the weight of impurity in a chemical is known to give readings that are approximately normally distributed with a standard deviation of 3.2 mg per 100 g of chemical.
- (a) In order to estimate the amount of impurity in a certain batch of the chemical, the chemist takes 12 samples, each of 100 g, from the batch and measures the weight of impurity in each sample. The results obtained in mg/100 g are as follows:
- 7.6 3.4 13.7 8.6 5.3 6.4
11.6 8.9 7.8 4.2 7.1 8.4
- (i) Find 95% central confidence limits for the mean weight of impurity present in a 100 g unit from the batch.
- (ii) The chemist calculated a 95% confidence interval for the mean weight of impurity in 100 g units from the batch. The interval was of the form $-\infty < \text{mean} \leq \alpha$. Find the value of α .
- Suggest why the chemist might prefer to use the value α rather than the limits in (i).
- (iii) Calculate an interval within which approximately 90% of the measured weights of impurity of 100 g units from the batch will lie.
- (b) Estimate how many samples of 100 g the scientist should take in order to be 95% confident that an estimate of the mean weight of impurity per 100 g is within 1.5 mg of the true value. (AEB)

6. Experimental components for use in aircraft engines were tested to destruction under extreme conditions. The survival times, X days, of ten components were as follows:

207 381 111 673 234 294 897 144 418 554

- (a) Calculate the arithmetic mean and the standard deviation of the data.
- (b) Assuming that the survival time, under these conditions, for all the experimental components is normally distributed with standard deviation 240 days, calculate a 90% confidence interval for the mean of X . (AEB)
7. A company manufactures bars of soap. The bars of soap are either pink or white in colour and differently shaped according to colour. The masses of both types of soap are known to be normally distributed, the mean mass of the white bars being 176.2 g. The standard deviation for both bars is 6.46 g. A sample of 12 of the pink bars of soap had masses, measured to the nearest gram, as follows.
- 174 164 182 169 171 187
176 177 168 171 180 175
- Find a 95% confidence interval for the mean mass of pink bars of soap.
- Calculate also an interval within which approximately 90% of the masses of the white bars of soap will lie.
- The cost of manufacturing a pink bar of soap of mass x g is $(15 + 0.065x)p$ and it is sold for 32p. If the company manufactures 9000 bars of pink soap per week, derive a 95% confidence interval for its weekly expected profit from pink bars of soap. (AEB)
8. Sugar produced by a company is classified as granulated or caster and the masses of the bags of both types are known to be normally distributed. The mean of the masses of bags of granulated sugar is 1022.51 g and the standard deviation for both types of sugar is 8.21 g.
- Calculate an interval within which 90% of the masses of bags of granulated sugar will lie.
- A sample of 10 bags of caster sugar had masses, measured to the nearest gram, as follows.

1062 1008 1027 1031 1011
1007 1072 1036 1029 1041

Find a 99% confidence interval for the mean mass of bags of caster sugar. Find a 99% confidence interval for the mean mass of bags of caster sugar.

To produce a bag of caster sugar of mass x g costs, in pence,

$$(32 + 0.023x)$$

and it is sold for 65p.

If the company produces 10 000 bags of caster sugar per day, derive a 99% confidence interval for its daily profit from caster sugar.

(AEB)

