Contents 40

Sampling Distributions and Estimation

Learning outcomes

You will learn about the distributions which are created when a population is sampled. For example, every sample will have a mean value; this gives rise to a distribution of mean values. We shall look at the behaviour of this distribution. We shall also look at the problem of estimating the true value of a population mean (for example) from a given sample.

Sampling Distributions

40.1

Introduction

When you are dealing with large populations, for example populations created by the manufacturing processes, it is impossible, or very difficult indeed, to deal with the whole population and know the parameters of that population. Items such as car components, electronic components, aircraft components or ordinary everyday items such as light bulbs, cycle tyres and cutlery effectively form infinite populations. Hence we have to deal with samples taken from a population and estimate those population parameters that we need. This Workbook will show you how to calculate single number estimates of parameters - called point estimates - and interval estimates of parameters - called interval estimates or confidence intervals. In the latter case you will be able to calculate a range of values and state the confidence that the true value of the parameter you are estimating lies in the range you have found.

Prerequisites

Before starting this Section you should ...

- understand and be able to calculate means and variances
- be familiar with the results and concepts met in the study of probability
- be familiar with the normal distribution

Learning Outcomes

On completion you should be able to ...

- understand what is meant by the terms sample and sampling distribution
- explain the importance of sampling in the application of statistics
- explain the terms point estimate and the term interval estimate
- calculate point estimates of means and variances
- find interval estimates of population parameters for given levels of confidence

1. Sampling

Why sample?

Considering samples from a distribution enables us to obtain information about a population where we cannot, for reasons of practicality, economy, or both, inspect the whole of the population. For example, it is impossible to check the complete output of some manufacturing processes. Items such as electric light bulbs, nuts, bolts, springs and light emitting diodes (LEDs) are produced in their millions and the sheer cost of checking every item as well as the time implications of such a checking process render it impossible. In addition, testing is sometimes destructive - one would not wish to destroy the whole production of a given component!

Populations and samples

If we choose n items from a population, we say that the size of the sample is n. If we take many samples, the means of these samples will themselves have a distribution which may be different from the population from which the samples were chosen. Much of the practical application of sampling theory is based on the relationship between the 'parent' population from which samples are drawn and the summary statistics (mean and variance) of the 'offspring' population of sample means. Not surprisingly, in the case of a normal 'parent' population, the distribution of the population and the distribution of the sample means are closely related. What is surprising is that even in the case of a non-normal parent population, the 'offspring' population of sample means is usually (but not always) normally distributed provided that the samples taken are large enough. In practice the term 'large' is usually taken to mean about 30 or more. The behaviour of the distribution of sample means is based on the following result from mathematical statistics.

The central limit theorem

In what follows, we shall assume that the members of a sample are chosen at random from a population. This implies that the members of the sample are *independent*. We have already met the Central Limit Theorem. Here we will consider it in more detail and illustrate some of the properties resulting from it.

Much of the theory (and hence the practice) of sampling is based on the Central Limit Theorem. While we will not be looking at the proof of the theorem (it will be illustrated where practical) it is necessary that we understand what the theorem says and what it enables us to do. Essentially, the Central Limit Theorem says that if we take large samples of size n with mean $\bar{X}$ from a population which has a mean μ and standard deviation σ then the distribution of sample means $\bar{X}$ is normally distributed with mean μ and standard deviation $\dfrac{\sigma}{\sqrt{n}}$.

That is, the **sampling distribution of the mean** $\bar{X}$ follows the distribution

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

Strictly speaking we require $\sigma^2 < \infty$, and it is important to note that no claim is made about the way in which the original distribution behaves, **and it need not be normal**. This is why the Central Limit Theorem is so fundamental to statistical practice. One implication is that a random variable which takes the form of a sum of many components which are random but not necessarily normal will itself be normal provided that the sum is not dominated by a small number of components. This explains why many biological variables, such as human heights, are normally distributed.

In the case where the original distribution is normal, the relationship between the original distribution $X \sim N(\mu, \sigma)$ and the distribution of sample means $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ is shown below.

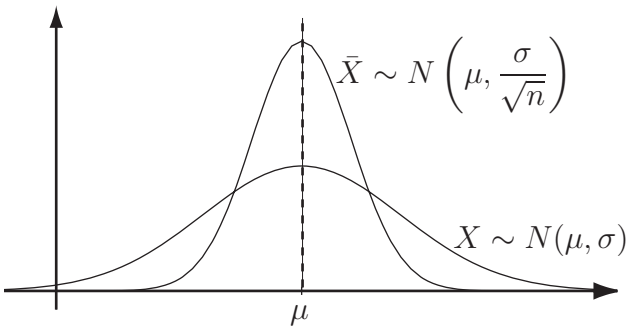


Figure 1

The distributions of X and $\bar{X}$ have the same mean μ but $\bar{X}$ has the smaller standard deviation $\frac{\sigma}{\sqrt{n}}$

The theorem says that we must take *large* samples. If we take *small* samples, **the theorem only holds if the original population is normally distributed**.

Standard error of the mean

You will meet this term often if you read statistical texts. It is the name given to the standard deviation of the population of sample means. The name stems from the fact that there is some uncertainty in the process of predicting the original population mean from the mean of a sample or samples.

Key Point 1

For a sample of n independent observations from a population with variance σ^2, the **standard error of the mean** is $\sigma_n = \frac{\sigma}{\sqrt{n}}$.

Remember that this quantity is simply the standard deviation of the distribution of sample means.

Finite populations

When we sample without replacement from a population which is not infinitely large, the observations are not independent. This means that we need to make an adjustment in the standard error of the mean. In this case the standard error of the sample mean is given by the related but more complicated formula

$$\sigma_{n,N} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

where $\sigma_{n,N}$ is the standard error of the sample mean, N is the population size and n is the sample size.

Note that, in cases where the size of the population N is large in comparison to the sample size n, the quantity

$$\frac{N-n}{N-1} \approx 1$$

so that the standard error of the mean is approximately $\sigma/\sqrt{n}$.

Illustration - a distribution of sample means

It is possible to illustrate some of the above results by setting up a small population of numbers and looking at the properties of small samples drawn from it. Notice that the setting up of a small population, say of size 5, and taking samples of size 2 enables us to deal with the totality of samples, there are $\binom{5}{2} = \frac{5!}{2!3!} = 10$ distinct samples possible, whereas if we take a population of 100 and draw samples of size 10, there are $\binom{100}{10} = \frac{100!}{10!90!} = 51,930,928,370,000$ possible distinct samples and from a practical point of view, we could not possibly list them all let alone work with them!

Suppose we take a population consisting of the five numbers 1, 2, 3, 4 and 5 and draw samples of size 2 to work with. The complete set of possible samples is:

$$(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5)$$

For the parent population, since we know that the mean $\mu = 3$, then we can calculate the standard deviation by

$$\sigma = \sqrt{\frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}} = \sqrt{\frac{10}{5}} = 1.4142$$

For the population of sample means,

$$1.5, 2, 2.5, 3, 2.5, 3, 3.5, 3.5, 4, 4.5$$

their mean and standard deviation are given by the calculations:

$$\frac{1.5+2+2.5+3+2.5+3+3.5+3.5+4+4.5}{10} = 3$$

and

$$\sqrt{\frac{(1.5-3)^2+(2-3)^2+\cdots+(4-3)^2+(4.5-3)^2}{10}} = \sqrt{\frac{7.5}{10}} = 0.8660$$

We can immediately conclude that the mean of the population of sample means is the same as the population mean μ.

Using the results given above the value of $\sigma_{n,N}$ should be given by the formula

$$\sigma_{n,N} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

with $\sigma = 1.4142,\ N = 5$ and $n = 2$. Using these numbers gives:

$$\sigma_{2,5} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{1.4142}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = \sqrt{\frac{3}{4}} = 0.8660 \text{ as predicted.}$$

Note that in this case the 'correction factor' $\sqrt{\frac{N-n}{N-1}} \approx 0.8660$ and is significant. If we take samples of size 10 from a population of 100, the factor becomes

$$\sqrt{\frac{N-n}{N-1}} \approx 0.9535$$

and for samples of size 10 taken from a population of 1000, the factor becomes

$$\sqrt{\frac{N-n}{N-1}} \approx 0.9955.$$

Thus as $\sqrt{\frac{N-n}{N-1}} \to 1$, its effect on the value of $\frac{\sigma}{\sqrt{n}}$ reduces to insignificance.

Two-centimetre number 10 woodscrews are manufactured in their millions but packed in boxes of 200 to be sold to the public or trade. If the length of the screws is known to be normally distributed with a mean of 2 cm and variance 0.05 cm^2, find the mean and standard deviation of the sample mean of 200 boxed screws. What is the probability that the sample mean length of the screws in a box of 200 is greater than 2.02 cm?

Your solution

Answer
Since the population is very large indeed, we are effectively sampling from an infinite population. The mean and standard deviation are given by

$$\mu = 2 \text{ cm} \quad \text{and} \quad \sigma_{200} = \frac{\sqrt{0.05}}{\sqrt{200}} = 0.016 \text{ cm}$$

Since the parent population is normally distributed the means of samples of 200 will be normally distributed as well.
Hence $\mathrm{P}(\text{sample mean length} > 2.02) = \mathrm{P}(z > \dfrac{2.02 - 2}{0.016}) = \mathrm{P}(z > 1.25) = 0.5 - 0.3944 = 0.1056$

2. Statistical estimation

When we are dealing with large populations (the production of items such as LEDs, light bulbs, piston rings etc.) it is extremely unlikely that we will be able to calculate population parameters such as the mean and variance directly from the full population.

We have to use processes which enable us to estimate these quantities. There are two basic methods used called point estimation and interval estimation. The essential difference is that point estimation gives single numbers which, in the sense defined below, are best estimates of population parameters, while interval estimates give a range of values together with a figure called the confidence that the true value of a parameter lies within the calculated range. Such ranges are usually called **confidence intervals**.

Statistically, the word 'estimate' implies a defined procedure for finding population parameters. In statistics, the word 'estimate' does not mean a guess, something which is rough-and-ready. What the word does mean is that an agreed precise process has been (or will be) used to find required values and that these values are 'best values' in some sense. Often this means that the procedure used, which is called the 'estimator', is:

(a) **consistent** in the sense that the difference between the true value and the estimate approaches zero as the sample size used to do the calculation increases;

(b) **unbiased** in the sense that the expected value of the estimator is equal to the true value;

(c) **efficient** in the sense that the variance of the estimator is small.

Expectation is covered in Workbooks 37 and 38. You should note that it is not always possible to find a 'best' estimator. You might have to decide (for example) between one which is

consistent, biased and efficient

and one which is

consistent, unbiased and inefficient

when what you really want is one which is

consistent, unbiased and **efficient.**

Point estimation

We will look at the point estimation of the mean and variance of a population and use the following notation.

Notation

	Population	Sample	Estimator
Size	N	n	
Mean	μ or $\mathsf{E}(x)$	$\bar{x}$	$\hat{\mu}$ for μ
Variance	σ^2 or $\mathsf{V}(x)$	s^2	$\hat{\sigma}^2$ for σ^2

Estimating the mean

This is straightforward.

$$\hat{\mu} = \bar{x}$$

is a sensible estimate since the difference between the population mean and the sample mean disappears with increasing sample size. We can show that this estimator is unbiased. Symbolically we have:

$$\hat{\mu} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

so that

$$\begin{aligned} \mathsf{E}(\hat{\mu}) = &= \frac{\mathsf{E}(x_1) + \mathsf{E}(x_2) + \cdots + \mathsf{E}(x_n)}{n} \\ &= \frac{\mathsf{E}(X) + \mathsf{E}(X) + \cdots + \mathsf{E}(X)}{n} \\ &= \mathsf{E}(X) \\ &= \mu \end{aligned}$$

Note that the expected value of x_1 is $\mathsf{E}(X)$, i.e. $\mathsf{E}(x_1) = \mathsf{E}(X)$. Similarly for $x_1, x_2, \cdots, x_n$.

Estimating the variance

This is a little more difficult. The true variance of the population is $\sigma^2 = \dfrac{\sum(x-\mu)^2}{N}$ which suggests the estimator, calculated from a sample, should be $\hat{\sigma}^2 = \dfrac{\sum(x-\mu)^2}{n}$.

However, we do not know the true value of μ, but we do have the estimator $\hat{\mu} = \bar{x}$.

Replacing μ by the estimator $\hat{\mu} = \bar{x}$ gives

$$\hat{\sigma}^2 = \frac{\sum(x-\bar{x})^2}{n}$$

This can be written in the form

$$\hat{\sigma}^2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{\sum x^2}{n} - (\bar{x})^2$$

Hence

$$\mathsf{E}(\hat{\sigma}^2) = \frac{\mathsf{E}(\sum x^2)}{n} - \mathsf{E}\{(\bar{X})^2\} = \mathsf{E}(X^2) - \mathsf{E}\{(\bar{X})^2\}$$

We already have the important result

$$\mathsf{E}(x) = \mathsf{E}(\bar{x}) \quad \text{and} \quad \mathsf{V}(\bar{x}) = \frac{\mathsf{V}(x)}{n}$$

Using the result $\mathsf{E}(x) = \mathsf{E}(\bar{x})$ gives us

$$\begin{aligned}
\mathsf{E}(\hat{\sigma}^2) &= \mathsf{E}(x^2) - \mathsf{E}\{(\bar{x})^2\} \\
&= \mathsf{E}(x^2) - \{\mathsf{E}(x)\}^2 - \mathsf{E}\{(\bar{x})^2\} + \{\mathsf{E}(\bar{x})\}^2 \\
&= \mathsf{E}(x^2) - \{\mathsf{E}(x)\}^2 - (\mathsf{E}\{(\bar{x})^2\} - \{\mathsf{E}(\bar{x})\}^2) \\
&= \mathsf{V}(x) - \mathsf{V}(\bar{x}) \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

This result is **biased**, for an unbiased estimator the result should be σ^2 not $\frac{n-1}{n}\sigma^2$.

Fortunately, the remedy is simple, we just multiply by the so-called Bessel's correction, namely $\frac{n}{n-1}$ and obtain the result

$$\hat{\sigma}^2 = \frac{n}{n-1}\frac{\sum(x-\bar{x})^2}{n} = \frac{\sum(x-\bar{x})^2}{n-1}$$

There are two points to note here. Firstly (and rather obviously) you should not take samples of size 1 since the variance cannot be estimated from such samples. Secondly, you should check the operation of any hand calculators (and spreadsheets!) that you use to find out exactly what you are calculating when you press the button for standard deviation. You might find that you are calculating either

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \quad \text{or} \quad \hat{\sigma}^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

It is just as well to know which, as the first formula assumes that you are calculating the variance of a population while the second assumes that you are estimating the variance of a population from a random sample of size n taken from that population.

From now on we will assume that we divide by $n-1$ in the sample variance and we will simply write s^2 for s^2_{n-1}.

Interval estimation

We will look at the process of finding an interval estimation of the mean and variance of a population and use the notation used above.

Interval estimation for the mean

This interval is commonly called the Confidence Interval for the Mean.

Firstly, we know that while the sample mean $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$ is a good estimator of the population mean μ. We also know that the calculated mean $\bar{x}$ of a sample of size n is unlikely to be exactly equal to μ. We will now construct an interval around $\bar{x}$ in such a way that we can quantify the confidence that the interval actually contains the population mean μ.

Secondly, we know that for sufficiently large samples taken from a large population, $\bar{x}$ follows a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Thirdly, looking at the following extract from the normal probability tables,

$Z = \dfrac{X-\mu}{\sigma}$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.9	.4713	4719	4726	4732	4738	4744	4750	4756	4762	4767

we can see that $2 \times 47.5\% = 95\%$ of the values in the standard normal distribution lie between ± 1.96 standard deviation either side of the mean.

So before we see the data we may say that

$$\mathsf{P}\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

After we see the data we say with 95% confidence that

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

which leads to

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

This interval is called a 95% confidence interval for the mean μ.

Note that while the 95% level is very commonly used, there is nothing sacrosanct about this level. If we go through the same argument but demand that we need to be 99% certain that μ lies within the confidence interval developed, we obtain the interval

$$\bar{x} - 2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}$$

since an inspection of the standard normal tables reveals that 99% of the values in a standard normal distribution lie within 2.58 standard deviations of the mean.

The above argument assumes that we know the population variance. In practice this is often not the case and we have to estimate the population variance from a sample. From the work we have seen above, we know that the best estimate of the population variance from a sample of size n is given by the formula

$$\hat{\sigma}^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

It follows that if we do not know the population variance, we must use the estimate $\hat{\sigma}$ in place of σ. Our 95% and 99% confidence intervals (for large samples) become

$$\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} \qquad \text{and} \qquad \bar{x} - 2.58\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.58\frac{\hat{\sigma}}{\sqrt{n}}$$

where

$$\hat{\sigma}^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

When we do not know the population variance, we need to estimate it. Hence we need to gauge the confidence we can have in the estimate.

In small samples, when we need to estimate the variance, the values 1.96 and 2.58 need to be replaced by values from the Student's t-distribution. See HELM 41.

Example 1

After 1000 hours of use the weight loss, in gm, due to wear in certain rollers in machines, is normally distributed with mean μ and variance σ^2. Fifty independent observations are taken. (This may be regarded as a "large" sample.) If observation i is y_i, then $\sum_{i=1}^{50} y_i = 497.2$ and $\sum_{i=1}^{50} y_i^2 = 5473.58$.

Estimate μ and σ^2 and give a 95% confidence interval for μ.

Solution

We estimate μ using the sample mean: $\bar{y} = \dfrac{\sum y_i}{n} = \dfrac{497.2}{50} = 9.944$ gm

We estimate σ^2 using the sample variance:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum (y_i - \bar{y})^2 = \frac{1}{n-1}\left\{\sum y_i^2 - \frac{1}{n}\left[\sum y_i\right]^2\right\} \\ &= \frac{1}{49}\left\{5473.58 - \frac{1}{50}497.2^2\right\} = 10.8046 \text{ gm}^2 \end{aligned}$$

The estimated standard error of the mean is $\sqrt{\dfrac{s^2}{n}} = \sqrt{\dfrac{10.8046}{50}} = 0.4649$ gm

The 95% confidence interval for μ is $\bar{y} \pm 1.96\sqrt{\dfrac{s^2}{n}}$. That is $9.479 < \mu < 10.409$

Exercises

1. The voltages of sixty nominally 10 volt cells are measured. Assuming these to be independent observations from a normal distribution with mean μ and variance σ^2, estimate μ and σ^2. Regarding this as a "large"sample, find a 99% confidence interval for μ. The data are:

10.3	10.5	9.6	9.7	10.6	9.9	10.1	10.1	9.9	10.5
10.1	10.1	9.9	9.8	10.6	10.0	9.9	10.0	10.3	10.1
10.1	10.3	10.5	9.7	10.1	9.7	9.8	10.3	10.2	10.2
10.1	10.5	10.0	10.0	10.6	10.9	10.1	10.1	9.8	10.7
10.3	10.4	10.4	10.3	10.4	9.9	9.9	10.5	10.0	10.7
10.1	10.6	10.0	10.7	9.8	10.4	10.3	10.0	10.5	10.1

2. The natural logarithms of the times in minutes taken to complete a certain task are normally distributed with mean μ and variance σ^2. Seventy-five independent observations are taken. (This may be regarded as a "large" sample.) If the natural logarithm of the time for observation i is y_i, then $\sum y_i = 147.75$ and $\sum y_i^2 = 292.8175$.

   Estimate μ and σ^2 and give a 95% confidence interval for μ.

   Use your confidence interval to find a 95% confidence interval for the median time to complete the task.

Answers

1. $\sum y_i = 611.0$, $\sum y_i^2 = 6227.34$ and $n = 60$. We estimate μ using the sample mean:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{611.0}{60} = 10.1833 \text{ V}$$

We estimate σ^2 using the sample variance:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{1}{n-1}\left\{\sum y_i^2 - \frac{1}{n}\left[\sum y_i\right]^2\right\} \\ &= \frac{1}{59}\left\{6227.34 - \frac{1}{59}611.0^2\right\} = 0.090226 \end{aligned}$$

The estimated standard error of the mean is

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.090226}{60}} = 0.03878 \text{ V}$$

The 99% confidence interval for μ is $\bar{y} \pm 2.58\sqrt{s^2/n}$. That is

$$10.08 < \mu < 10.28$$

2. We estimate μ using the sample mean:

$$\bar{y} = \frac{\sum y_i}{n} = \frac{147.75}{75} = 1.97$$

We estimate σ^2 using the sample variance:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum(y_i - \bar{y})^2 = \frac{1}{n-1}\left\{\sum y_i^2 - \frac{1}{n}\left[\sum y_i\right]^2\right\} \\ &= \frac{1}{74}\left\{292.8175 - \frac{1}{75}147.75^2\right\} = 0.02365 \end{aligned}$$

The estimated standard error of the mean is

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.02365}{75}} = 0.01776$$

The 95% confidence interval for μ is $\bar{y} \pm 1.96\sqrt{s^2/n}$. That is

$$1.935 < \mu < 2.005$$

The 95% confidence interval for the median time, in minutes, to complete the task is

$$e^{1.935} < M < e^{2.005}$$

That is

$$6.93 < M < 7.42$$

Interval Estimation for the Variance

40.2

Introduction

In Section 40.1 we have seen that the sampling distribution of the sample mean, when the data come from a normal distribution (and even, in large samples, when they do not) is itself a normal distribution. This allowed us to find a confidence interval for the population mean. It is also often useful to find a confidence interval for the population variance. This is important, for example, in quality control. However the distribution of the sample variance is not normal. To find a confidence interval for the population variance we need to use another distribution called the "chi-squared" distribution.

Prerequisites

Before starting this Section you should ...

- understand and be able to calculate means and variances
- understand the concepts of continuous probability distributions
- understand and be able to calculate a confidence interval for the mean of a normal distribution

Learning Outcomes

On completion you should be able to ...

- find probabilities using a chi-squared distribution
- find a confidence interval for the variance of a normal distribution

1. Interval estimation for the variance

In Section 40.1 we saw how to find a confidence interval for the mean of a normal population. We can also find a confidence interval for the variance. The corresponding confidence interval for the standard deviation is found by taking square roots.

We know that if we take samples from a population, then each sample will have a mean and a variance associated with it. We can calculate the values of these quantities from first principles, that is we can use the basic definitions of the mean and the variance to find their values. Just as the means form a distribution, so do the values of the variance and it is to this distribution that we turn in order to find an interval estimate for the value of the variance of the population. Note that if the original population is normal, samples taken from this population have means which are normally distributed. When we consider the distribution of variances calculated from the samples we need the chi-squared (usually written as χ^2) distribution in order to calculate the confidence intervals. As you might expect, the values of the chi-squared distribution are tabulated for ease of use. The calculation of confidence intervals for the variance (and standard deviation) depends on the following result.

Key Point 2

If $x_1, x_2, \cdots, x_n$ is a random sample taken from a normal population with mean μ and variance σ^2 then if the sample variance is denoted by S^2, the random variable

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared (χ^2) distribution with $n-1$ degrees of freedom.

Clearly, a little explanation is required to make this understandable! Key Point 2 refers to the chi-squared distribution and the term 'degrees of freedom.' Both require some detailed explanation before the Key Point can be properly understood. We shall start by looking in a little detail at the chi-squared distribution and then consider the term 'degrees of freedom.' You are advised to read these explanations very carefully and make sure that you fully understand them.

The chi-squared random variable

The probability density function of a χ^2 random variable is somewhat complicated and involves the gamma (Γ) function. The gamma function, for positive r, is defined as

$$\Gamma(r) = \int_0^\infty x^{r-1}\mathrm{e}^{-x}dx$$

It is easily shown that $\Gamma(r) = (r-1)\Gamma(r-1)$ and that, if r is an integer, then

$$\Gamma(r) = (r-1)(r-2)(r-3)\cdots(3)(2)(1) = (r-1)!$$

The probability density function is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} \mathrm{e}^{-x/2} \qquad x > 0.$$

The plots in Figure 2 show the probability density function for various convenient values of k. We have deliberately taken even values of k so that the gamma function has a value easily calculated from the above formula for a factorial. In these graphs the vertical scaling has been chosen to ensure each graph has the same maximum value.

It is possible to discern two things from the diagrams.

Firstly, as k increases, the peak of each curve occurs at values closer to k. Secondly, as k increases, the shape of the curve appears to become more and more symmetrical. In fact the mean of the χ^2 distribution is k and in the limit as $k \to \infty$ the χ^2 distribution becomes normal. One further fact, not obvious from the diagrams, is that the variance of the χ^2 distribution is $2k$.

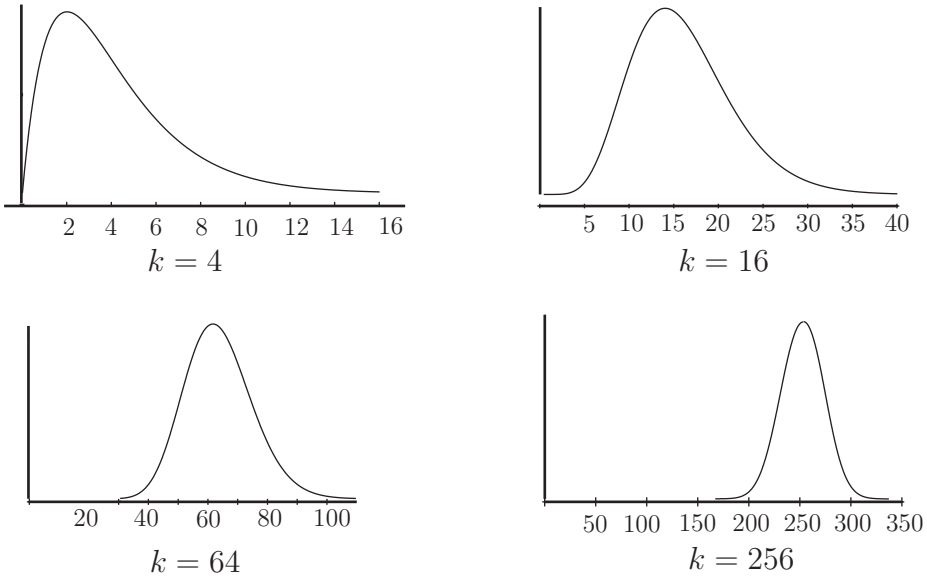


Figure 2

A summary is given in the following Key Point.

Key Point 3

The χ^2 distribution, defined by the probability density function

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} \mathrm{e}^{-x/2} \qquad x > 0.$$

has mean k and variance $2k$ and as $k \to \infty$ the limiting form of the distribution is normal.

Degrees of freedom

A formal definition of the term 'degrees of freedom' is that it is the 'number of independent comparisons that can be made among the elements of a sample.' Textbooks on statistics e.g. *Applied Statistics and Probability for Engineers* by Montgomery and Runger (Wiley) often give this formal definition. The number of degrees of freedom is usually represented by the Greek symbol ν pronounced 'nu'. The following explanations of the concept should be helpful.

Explanation 1

If we have a sample of n values say $x_1, x_2, x_3 \cdots, x_n$ chosen from a population and we are trying to calculate the mean of the sample, we know that the sum of the deviations about the mean must be zero. Hence, the following constraint must apply to the observations.

$$\sum(x - \bar{x}) = 0$$

Once we calculate the values of $(x_1 - \bar{x}),\ (x_2 - \bar{x}),\ (x_3 - \bar{x}),\ \cdots\ (x_{n-1} - \bar{x})$ we can calculate the value of $(x_n - \bar{x})$ by using the constraint $\sum(x - \bar{x}) = 0$. We say that we have $n - 1$ degrees of freedom. The term 'degrees of freedom' may be thought of as the number of independent variables minus the number of constraints imposed.

Explanation 2

A point in space which can move freely has three degrees of freedom since it can move *independently* in the x, y and z directions. If we now restrict the point so that it can only move along the straight line

$$\frac{x}{a} = \frac{y}{b} = \frac{z}{c}$$

then we have effectively imposed two constraints since the value of (say) x determines the values of y and z. In this situation, we say that the number of degrees of freedom is reduced from 3 to 1. That is, we have one degree of freedom.

A similar argument may be used to demonstrate that a point in three dimensional space which is restricted to move in a plane leads to a situation with two degrees of freedom.

Key Point 4

The term 'degrees of freedom' may be thought of as the number of independent variables involved minus the number of constraints imposed.

Figure 3 shows a typical χ^2 distribution and Table 1 at the end of this Workbook show the values of $\chi^2_{\alpha,\nu}$ for a variety of values of the area α and the number of degrees of freedom ν. Notice that Table 1 gives the area values corresponding to the right-hand tail of the distribution which is shown shaded.

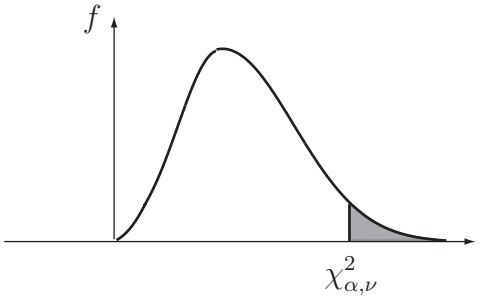


Figure 3

The $\chi^2_{\alpha,\nu}$ values for (say) right-hand area values of 5% are given by the column headed 0.05 while the $\chi^2_{\alpha,\nu}$ values for (say) left-hand area values of 5% are given by the column headed 0.95. Figure 4 shows the values of $\chi^2_{\alpha,\nu}$ for the two 5% tails when there are 5 degrees of freedom.

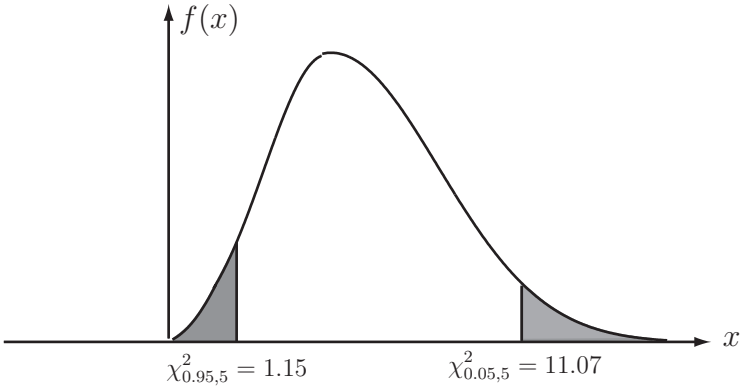


Figure 4

Use the percentage points of the χ^2 distribution to find the appropriate values of $\chi^2_{\alpha,\nu}$ in the following cases.

(a) Right-hand tail of 10% and 7 degrees of freedom.

(b) Left-hand tail of 2.5% and 9 degrees of freedom.

(c) Both tails of 5% and 10 degrees of freedom.

(d) Both tails of 2.5% and 20 degrees of freedom.

Your solution

Answer
Using Table 1 and reading off the values directly gives:
(a) 12.02 (b) 2.70 (c) 3.94 and 18.31 (d) 9.59 and 34.17

Constructing a confidence interval for the variance

We know that if $x_1, x_2, x_3, \cdots, x_n$ is a random sample taken from a normal population with mean μ and variance σ^2 and if the sample variance is denoted by S^2, the random variable

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared distribution with $n-1$ degrees of freedom. This knowledge enables us to construct a confidence interval as follows.

Firstly, we decide on a level of confidence, say, for the sake of illustration, 95%. This means that we need two 2.5% tails.

Secondly, we know that we have $n-1$ degrees of freedom so that the value of X^2 will lie between the left-tail value of $\chi^2_{0.975,n-1}$ and the right-tail value of $\chi^2_{0.025,n-1}$. If we know the value of n then we can easily read off these values from the χ^2 tables.

The confidence interval is developed as shown below.
We have

$$\chi^2_{0.025,n-1} \leq X^2 \leq \chi^2_{0.975,n-1}$$

so that

$$\chi^2_{0.025,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{0.975,n-1}$$

hence

$$\frac{1}{\chi^2_{0.975,n-1}} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi^2_{0.025,n-1}}$$

so that

$$\frac{(n-1)S^2}{\chi^2_{0.975,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{0.025,n-1}}$$

Another way of stating the same result using probability directly is to say that

$$\mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{0.975,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{0.025,n-1}}\right) = 0.95$$

Noting that $0.95 = 100(1-0.05)$ and that we are working with the right-hand tail values of the χ^2 distribution, it is usual to generalize the above result as follows. Taking a general confidence level as $100(1-\alpha)\%$, (a 95% interval gives $\alpha = 0.05$), our confidence interval becomes

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

Note that the confidence interval for the standard deviation σ is obtained by taking the appropriate square roots.

The following Key Point summarizes the development of this confidence interval.

Key Point 5

If $x_1, x_2, x_3, \cdots, x_n$ is a random sample with variance S^2 taken from a normal population with variance σ^2 then a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the appropriate right-hand and left-hand values respectively of a chi-squared distribution with $n-1$ degrees of freedom.

Example 2

A random sample of 20 nominally measured 2mm diameter steel ball bearings is taken and the diameters are measured precisely. The measurements, in mm, are as follows:

2.02 1.94 2.09 1.95 1.98 2.00 2.03 2.04 2.08 2.07
1.99 1.96 1.99 1.95 1.99 1.99 2.03 2.05 2.01 2.03

Assuming that the diameters are normally distributed with unknown mean, μ, and unknown variance σ^2,

(a) find a two-sided 95% confidence interval for the variance, σ^2;

(b) find a two-sided confidence interval for the standard deviation, σ.

Solution

From the data, we calculate $\sum x_i = 40.19$ and $\sum x_i^2 = 80.7977$. Hence

$$(n-1)S^2 = 80.7977 - \frac{40.19^2}{20} = 0.035895$$

There are 19 degrees of freedom and the critical values of the χ^2_{19}-distribution are

$$\chi^2_{0.975,19} = 8.91 \qquad \text{and} \qquad \chi^2_{0.025,19} = 32.85$$

(a) the confidence interval for σ^2 is

$$\frac{0.035895}{32.85} < \sigma^2 < \frac{0.035895}{8.91} \equiv 1.0927 \times 10^{-3}\text{mm} < \sigma^2 \leq 4.0286 \times 10^{-3}\text{mm}$$

(b) the confidence interval for σ is

$$\sqrt{1.0927 \times 10^{-3}} < \sigma \leq \sqrt{4.0286 \times 10^{-3}} \equiv 0.033\text{mm} < \sigma < 0.063 \text{ mm}$$

In a typical car, bell housings are bolted to crankcase castings by means of a series of 13 mm bolts. A random sample of 12 bolt-hole diameters is checked as part of a quality control process and found to have a variance of 0.0013 mm^2.

(a) Construct the 95% confidence interval for the variance of the holes.

(b) Find the 95% confidence interval for the standard deviation of the holes.

State clearly any assumptions you make.

Your solution

Answer

Using the confidence interval formula developed, we know that the 95% confidence interval is

$$\frac{11 \times 0.0013}{\chi^2_{0.025,11}} \leq \sigma^2 \leq \frac{11 \times 0.0013}{\chi^2_{0.975,11}} \qquad \text{i.e.} \qquad \frac{11 \times 0.0013}{21.92} \leq \sigma^2 \leq \frac{11 \times 0.0013}{3.82}$$

(a) The 95% confidence interval for the variance is $0.0007 \leq \sigma^2 \leq 0.0037$ mm^2.

(b) The 95% confidence interval for the standard deviation is $0.0265 \leq \sigma \leq 0.0608$ mm.

We have assumed that the hole diameters are normally distributed.

Exercises

1. Measurements are made on the lengths, in mm, of a sample of twenty wooden components for self-assembly furniture. Assume that these may be regarded as twenty independent observations from a normal distribution with unknown mean μ and unknown variance σ^2. The data are as follows.

581	580	581	577	580	581	577	579	579	578
581	583	577	578	582	581	582	580	582	579

   Find a 95% confidence interval for the variance σ^2 and hence find a 95% confidence interval for the standard deviation σ.

2. A machine fills packets with powder. At intervals a sample of ten packets is taken and the packets are weighed. The ten weights may be regarded as a sample of ten independent observations from a normal distribution with unknown mean. Find limits L, U such that the probability that $L < S^2 < U$ is 0.9 when the population variance is $\sigma^2 = 3.0$ and S^2 is the sample variance.

Answers

1. From the data we calculate $\sum y_i = 11598$ and $\sum y_i^2 = 6725744$ and we have $n = 20$. Hence

$$(n-1)s^2 = \sum (y_i - \bar{y})^2 = 6725744 - \frac{11598^2}{20} = 63.8$$

   The number of degrees of freedom is $n - 1 = 19$. We know that

$$\chi^2_{0.975,19} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{0.025,19}$$

   with probability 0.95. So a 95% confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{0.025,19}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{0.975,19}}$$

   That is $\dfrac{63.8}{32.85} < \sigma^2 < \dfrac{63.8}{8.91}$ so $1.942 < \sigma^2 < 7.160$

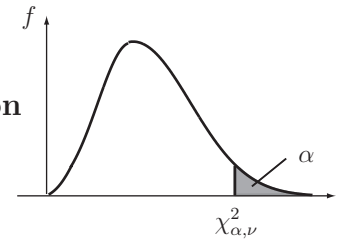
   This gives a 95% confidence interval for σ: $1.394 < \sigma < 2.676$

2. There are $n - 1 = 9$ degrees of freedom. Now

$$\begin{aligned} 0.9 &= \mathrm{P}\left(\chi^2_{0.05,9} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{0.95,9}\right) \\ &= \mathrm{P}\left(\frac{\chi^2_{0.05,9}\sigma^2}{n-1} < S^2 < \frac{\chi^2_{0.95,9}\sigma^2}{n-1}\right) \\ &= \mathrm{P}\left(\frac{3.33 \times 3.0}{9} < S^2 < \frac{16.92 \times 3.0}{9}\right) = \mathrm{P}(1.11 < S^2 < 5.64) \end{aligned}$$

   Hence $L = 1.11$ and $U = 5.64$.

Table 1: Percentage Points $\chi^2_{\alpha,\nu}$ of the χ^2 distribution



α	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
v											
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.01	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.28
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	31.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17