Contents 40

Sampling Distributions and Estimation

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

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

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

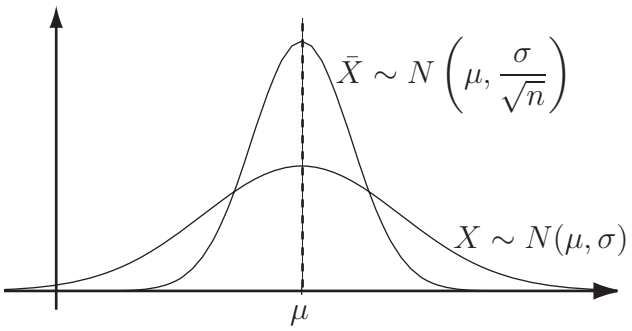

Figure 1

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

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

$$\sqrt{\frac{N-n}{\quad}} \quad 1$$

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

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

Since the parent population is normally distributed the means of samples of 200 will be normally distributed as well.

Hence $P(\text{sample mean length} > 2.02) = P\left(z > \frac{2.02 - 2}{0.016}\right) = P(z > 1.25) = 0.5 - 0.3944 = 0.1056$

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 $\mathrm{E}(x)$	$\bar{x}$	$\hat{\mu}$ for μ
Variance	σ^2 or $\mathrm{V}(x)$	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

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

Note that the expected value of x_1 is $\mathrm{E}(X)$, i.e. $\mathrm{E}(x_1) = \mathrm{E}(X)$. 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

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

We already have the important result

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

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

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

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

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

$Z = \frac{X - \mu}{\sigma}$	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 2×47.

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.

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:

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

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:

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

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

The probability density function is

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

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

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

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

Degrees of freedom

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.

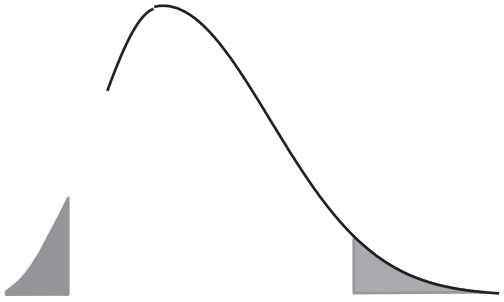

Figure 4

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 o 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

$$X^2 = (n$$

Key Point 5

If $x_1, x_2, x_3, \cdots, x_n$ is a random sample with variance S^2

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

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 $U = 5.64$.