MAS1403/ACE2013

Quantitative Methods for Business Management

Statistics for Marketing and Management

Semester 2, 2008–09

Lecturer: Dr. Lee Fawcett

School of Mathematics & Statistics

Chapter 1

Estimation

1.1 Introduction

Statistics is the science that studies the collection and interpretation of data to enable us to make **inferences** from the **sample data** to the **population** from which the sample has been drawn. As the sample we draw from the population gets bigger and bigger, it becomes increasingly difficult to "picture" the data just by looking at the raw observations; the data are easier to understand if we can find some way of *summarising* them.

Recall from semester 1 that data can be summarised in two ways: **graphically** and **numerically**. Various graphical techniques available for summarising data were presented, including stem and leaf plots, bar charts, histograms, relative frequency histograms and frequency polygons. You were also introduced to two methods of summarising data numerically, through measures of **location** and measures of **spread**.

A measure of location is a quantity which is 'typical' of the data; examples of such measures include

(i) the **sample mean** ("add them up and divide by how many we have");

(ii) the **sample median** ("the one in the middle"), and

(iii) the **sample mode** ("the value which occurs most often").

A measure of spread is a value which quantifies the variability in the data (or how "spread out" the observations are); examples include

(i) the **range** ("largest minus smallest");

(ii) the **variance** ("average squared distance of observations from the mean" – the standard deviation is the square root of this value), and

(iii) the **inter–quartile range** ("upper quartile minus lower quartile" – this value represents the middle 50% of the data, as the lower quartile has one quarter of the data less than it, and the upper quartile has three–quarters of the data less than it).

1.2 Estimation

Last semester we concentrated on pure description of data, although we recognised that this might prompt us to ask pertinent questions about the population from which the sample was drawn. What exactly does the sample, often a tiny subset, tell us of the population? We can never observe the whole population, even if it is finite, except at enormous expense, and so the population mean and variance (or indeed any aspect of the population distribution) can never be known exactly. We call these unknown quantities **parameters** and use Greek letters to denote them: μ ("mu") is the symbol commonly used for the population mean and σ ("sigma") for the population standard deviation. Hopefully (and if we have a representative sample), the sample mean ($\bar{x}$) will be quite close to the true population mean μ; likewise, the sample standard deviation (s) will be a good estimator for σ. In this section, we concentrate on $\bar{x}$ as an estimator for μ.

Before we can use our sample of n observations we must ask the question: Is $\bar{x}$ a "good" estimate of μ? How do we **infer** (find something out about) the unknown μ using $\bar{x}$? So long as the sample size n is fairly large, we can hope that $\bar{x}$ is close to μ. But how close is it? To answer this question, we must make some plausible assumptions about the population.

Suppose $x_1, x_2, \ldots, x_n$ are a random sample from a $N(\mu, \sigma^2)$ distribution. Then their mean, $\bar{x}$, is an observation from a $N(\mu, \sigma^2/n)$ distribution. The standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$ and is usually called the **standard error** of $\bar{x}$ to remind ourselves that $\bar{x}$ is used as an estimate of the unknown μ.

The Central Limit Theorem
Suppose now that $x_1, x_2, \ldots, x_n$ are a random sample from *any* population, with mean μ and variance σ^2. If n is large, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{approximately,}$$

no matter what the original probability distribution.

1.3 Interval estimation

The values we calculate for sample means and variances are **point estimates**; they are single values based on a limited sample of the whole population. Suppose that we wish to estimate the mean μ of a population. The natural estimate for μ is the sample mean $\bar{x}$. However, $\bar{x}$ is never exactly equal to μ; all we really hope is that $\bar{x}$ will be close to μ. One way of improving our inference is to construct **interval estimates** or **confidence intervals**. We simply place an interval over the point estimate for μ which allows us to say (with a certain level of confidence) within what range the population mean lies. The calculation of these intervals depends on the size of our sample (n), the level of confidence we choose, and whether or not the population variance (σ^2) is known.

1.3.1 Case 1: Known variance σ^2

We know from the results above that, if our random sample is drawn from a normal distribution, or if n is large (i.e. $n \geq 30$), then

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

If we initially assume we know the population variance σ^2, we can standardise $\bar{x}$ as we did last semester; i.e.

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}.$$

Recall that the standard normal distribution is $Z \sim N(0, 1)$, i.e. Z has zero mean and variance (and so standard deviation) 1; also recall that approximately 95% of the standard normal distribution lies between -1.96 and 1.96, i.e.

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

If we think about this graphically, we can see this more clearly:

Since we have an expression for Z, we can rearrange this expression:

$$\Pr\left(-1.96 < \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} < 1.96\right) = 0.95$$
$$\Pr\left(\bar{x} - 1.96\sqrt{\sigma^2/n} < \mu < \bar{x} + 1.96\sqrt{\sigma^2/n}\right) = 0.95.$$

Thus, we can say that the two values $\bar{x} - 1.96\sqrt{\sigma^2/n}$ and $\bar{x} + 1.96\sqrt{\sigma^2/n}$ are the lower and upper bounds (respectively) of the (95%) confidence interval.

This can be written as

$$\bar{x} \pm 1.96\sqrt{\sigma^2/n}.$$

Consider the following example.

Example 1
A coffee machine fills cups with hot water; the variance of the filling process is known to be $\sigma^2 = 10$ml. A sample of 100 filled cups gives a sample mean and we have calculated a sample mean of $\bar{x} = 40$ml. What is the 95% confidence interval of the population mean μ?

We already have a formula for the 95% confidence interval:

$$\bar{x} \pm 1.96\sqrt{\sigma^2/n}.$$

So, inputting our values, we get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/100}, \qquad \text{i.e.} \\ 40 &\pm 0.61. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is $(39.39, 40.61)$.

What would happen if the sample size increased to 200 and everything else remained the same? We'd get

$$\begin{aligned} 40 &\pm 1.96\sqrt{10/200}, \qquad \text{i.e.} \\ 40 &\pm 0.44. \end{aligned}$$

Hence, the 95% confidence interval for the population mean μ is $(39.56, 40.44)$. This should be intuitive, since as the sample size increases we are becoming more sure of our estimate for the population value.

What would be the 99% confidence interval in this case? From tables for the standard normal distribution, we can find that

$$\Pr(-2.58 < Z < 2.58) = 0.99;$$

hence, the 99% confidence interval is given by

$$\bar{x} \pm 2.58\sqrt{\sigma^2/n},$$

in this case giving

$$\begin{aligned} 40 &\pm 2.58\sqrt{10/200}, \qquad \text{i.e.} \\ 40 &\pm 0.58. \end{aligned}$$

Hence, the 99% confidence interval for the population mean μ is $(39.42, 40.58)$. You should note that this gives a wider range than the 95% confidence interval. This is (again) intuitive; as you increase the percentage of certainty you want, you will naturally incorporate a larger range.

1.3.2 Case 2: Unknown variance σ^2

If the population variance σ^2 is unknown, we can no longer use the normal distribution and instead have to use the t–distribution to calculate confidence intervals. We have seen that when our random sample follows a normal distribution, or indeed any distribution (if the sample size is large), then the sample mean $\bar{x} \sim N(\mu, \sigma^2/n)$. From this, it follows that

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}},$$

where Z is the standard normal distribution, i.e. $Z \sim N(0, 1)$. However, if the population variance is unknown, then the quantity

$$T = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$$

does *not* have a $N(0, 1)$ distribution (note that the *population* variance σ^2 in Z has been replaced with the *sample* variance s^2 in T). Instead it has a Student's t–distribution. This distribution is similar to the $N(0, 1)$ distribution in that it is symmetrical and bell–shaped, but it is more heavily tailed to allow for greater uncertainty in $\bar{x}$ since the true variability is now unknown. Its exact shape is determined by one parameter called the "degrees of freedom". Table 1.1 gives critical values of Student's t–distribution with various degrees of freedom. These numbers depend on two quantities: ν, the degrees of freedom, and p, a probability.

The expression for the confidence interval in this case is similar to the case where σ^2 is known:

$$\bar{x} \pm t_p \sqrt{s^2/n},$$

where σ^2 has been replaced with the sample variance s^2 and t_p is the appropriate value from the t–distribution tables. But how do we find this value?

First, we need to find p. If we are looking for the 95% confidence interval, we are looking for the value of p which satisfies the equation

$$\begin{aligned} 100(1-p)\% &= 95\%, \qquad \text{i.e.} \\ p &= 0.05. \end{aligned}$$

We would look up the value in the t tables (table 1.1) in the p column, or in this case the 5% column.

We also need to know which row to look in. The rows are given as the degrees of freedom, ν, where $\nu = n - 1$. Hence, if our sample was of size $n = 10$ and we were looking for the 95% confidence interval, we would look in the $\nu = 9$ row and the $p = 5\%$ column to give us a value of 2.262 to use in our calculation.

Example 2
A sample of size 15 is taken from a larger population; the sample mean is calculated as 12 and the sample variance as 25. What is the 95% confidence interval for the population mean μ?

We know that the confidence interval is given by

$$\bar{x} \pm t_p \sqrt{s^2/n},$$

where

$$\begin{aligned} n &= 15, \\ \nu &= n - 1 = 15 - 1 = 14, \\ p &= 5\%, \\ \bar{x} &= 12 \qquad \text{and} \\ s^2 &= 25. \end{aligned}$$

We can find our t value by looking in the $p = 5\%$ column and the $\nu = 14$ row, giving a value of 2.145. Putting what we know into our expression, we get

$$\begin{aligned} 12 &\pm t_{5\%} \sqrt{\frac{25}{15}} \\ 12 &\pm 2.145 \sqrt{\frac{25}{15}} \qquad \text{i.e.} \\ 12 &\pm 2.77. \end{aligned}$$

Hence, the confidence interval is $(9.23, 14.77)$.

1.3.3 Confidence intervals: a general approach

In this section, we summarise the general procedure for calculating a confidence interval for the population mean μ.

Case 1: Known population variance σ^2

(i) Calculate the sample mean $\bar{x}$ from the data;

(ii) Calculate your interval! For example,

- for a 90% confidence interval, use the formula

$$\bar{x} \pm 1.64 \times \sqrt{\sigma^2/n};$$

- for a 95% confidence interval, use the formula

$$\bar{x} \pm 1.96 \times \sqrt{\sigma^2/n};$$

- for a 99% confidence interval, use the formula

$$\bar{x} \pm 2.58 \times \sqrt{\sigma^2/n}.$$

Case 2: Unknown population variance σ^2

(i) Calculate the sample mean $\bar{x}$ *and* the sample variance s^2 from the data;

(ii) For a $100(1-p)\%$ confidence interval, look up the value of t under column p, row ν of table 1.1, remembering that $\nu = n - 1$. Note that, for a 90% confidence interval, $p = 10\%$, for a 95% confidence interval, $p = 5\%$ and for a 99% confidence interval, $p = 1\%$;

(iii) Calculate your interval, using

$$\bar{x} \pm t_p \times \sqrt{s^2/n}.$$

1.4 Application of Confidence Intervals

You might be asking: "why do we bother calculating confidence intervals?". Firstly, by calculating a confidence interval for the population mean, it allows us to see how confident we are of the point estimate we have calculated. The wider the range, the less precise we can be about the population value.

Secondly, it allows us to start looking at differences between groups. If the confidence intervals for two samples do not overlap, this could suggest that they are from separate populations. Or if we have a known value for a population and this does not fall within the confidence interval of our sample, this could suggest that there is something different about this sample. For example, if the average daily taking for all shops in a chain was £12550, and we calculated the confidence interval for the population mean of one particular branch as (£11000, £11500), we can say that this branch is out of line with the company as a whole, and it appears that its daily takings are lower.

Example 1.4.1
A credit card company wants to determine the mean income of its card holders. It also wants to find out if there are any differences in mean income between males and females. A random sample of 225 male card holders and 190 female card holders was drawn, and the following results obtained:

	Mean	Standard deviation
Males	£16 450	£3675
Females	£13 220	£3050

Calculate 95% confidence intervals for the mean income for males and females. Is there any evidence to suggest that, on average, males' and females' incomes differ? If so, describe this difference.

95% confidence interval for male income
The true population variance, σ^2, is unknown, so we can't use the approach of section 1.3.1. Instead, we follow that of section 1.3.2 which uses the t–distribution, i.e.

$$\bar{x} \pm t_p \times \sqrt{s^2/n}.$$

Here,

$$\begin{aligned} \bar{x} &= 16450, \\ s^2 &= 3675^2 = 13505625 \qquad \text{and} \\ n &= 225. \end{aligned}$$

The value t_p must be found from table 1.1. Recall that the degrees of freedom, $\nu = n-1$, and so here we have $\nu = 225 - 1 = 224$. Notice that table 1.1 only gives value of ν up to 29; for higher values, we use the ∞ row. Since we require a 95% confidence interval, we read down the 5% column, giving a t value of 1.96 (recall that this is the same as the

value used if σ^2 were *known* and we used the normal distribution – that's because the t–distribution converges to the normal distribution as the sample size increases). Thus, the 95% confidence interval for μ is found as

$$\begin{aligned} 16450 &\pm 1.96 \times \sqrt{13505625/225}, \qquad \text{i.e.} \\ 16450 &\pm 480.2. \end{aligned}$$

So, the 95% confidence interval is (£15969.80, £16930.20).

95% confidence interval for female income

Again, the true population variance, σ^2, is unknown, so we can't use the approach of section 1.3.1, and so again we use the t–distribution as in section 1.3.2:

$$\bar{x} \pm t_p \times \sqrt{s^2/n}.$$

Now,

$$\begin{aligned} \bar{x} &= 13220, \\ s^2 &= 3050^2 \\ &= 9302500, \qquad \text{and} \\ n &= 190. \end{aligned}$$

Again, since the sample size is large, we use the ∞ row of table 1.1 to obtain the value of t_p, and so the 95% confidence interval for μ is found as

$$\begin{aligned} 13220 &\pm 1.96 \times \sqrt{9302500/190}, \qquad \text{i.e.} \\ 13220 &\pm 1.96 \times 221.27, \qquad \text{i.e.} \\ 13220 &\pm 433.69. \end{aligned}$$

So, the 95% confidence interval is (£12786.31, £13653.69).

Since the 95% confidence intervals for males and females *do not overlap*, there *is* evidence to suggest that males' and females' incomes, on average, are different. Further, it appears that male card holders earn more than women.

1.5 Exercises

1. A company packs sacks of flour. The variance of the filling process is 100g. A sample of 50 bags is taken and weighed and the resulting sample mean is 750g. Compute a 95% and 99% confidence interval for the mean weight of a bag of flour.

2. A company manufactures bolts with a process variance of 50mm. A sample of 100 bolts is taken and measured and their average length is calculated as 98mm. What is the 95% confidence interval for the mean length of bolts? If the bolts are designed to be 100mm long, is the process satisfactory?

3. A class of students has sat an exam. A sample of 40 students is taken and their marks produced first. This sample has a mean of 55% and a sample variance of 100. Calculate the 95% confidence interval for the mean mark of the class as a whole.

4. A sample of 12 students is taken and their mean IQ calculated as 110 (with a sample variance of 220). What is the 95% and 99% confidence intervals for the population value based on this sample? What do you notice about the calculated interval as the confidence level increases? Do either of these two confidence intervals contain the known population mean IQ of 100?

5. The following are the number of cars caught speeding each day on one speed camera over a two week period.

10	12	15	9	8	12	11
6	15	17	12	10	9	7

   What is the 95% confidence interval for this sample? How does this compare with the whole Northumbria police average of 8 per day per camera?

		50%	20%	10%	5%	1%
	1	1.00	3.078	6.314	12.706	63.657
	2	0.816	1.886	2.920	4.303	9.925
	3	0.765	1.638	2.353	3.182	5.841
	4	0.741	1.533	2.132	2.776	4.604
	5	0.727	1.476	2.015	2.571	4.032
	6	0.718	1.440	1.943	2.447	3.707
	7	0.711	1.415	1.895	2.365	3.449
ν	8	0.706	1.397	1.860	2.306	3.355
	9	0.703	1.383	1.833	2.262	3.250
	10	0.700	1.372	1.812	2.228	3.169
	11	0.697	1.363	1.796	2.201	3.106
	12	0.695	1.356	1.782	2.179	3.055
	13	0.694	1.350	1.771	2.160	3.012
	14	0.692	1.345	1.761	2.145	2.977
	15	0.691	1.341	1.753	2.131	2.947
	16	0.690	1.337	1.746	2.120	2.921
	17	0.689	1.333	1.740	2.110	2.898
	18	0.688	1.330	1.734	2.101	2.878
	19	0.688	1.328	1.729	2.093	2.861
	20	0.687	1.325	1.725	2.086	2.845
	21	0.686	1.323	1.721	2.080	2.831
	22	0.686	1.321	1.717	2.074	2.819
	23	0.685	1.319	1.714	2.069	2.807
	24	0.685	1.318	1.711	2.064	2.797
	25	0.684	1.316	1.708	2.060	2.787
	26	0.684	1.315	1.706	2.056	2.779
	27	0.684	1.314	1.703	2.052	2.771
	28	0.683	1.313	1.701	2.048	2.763
	29	0.683	1.311	1.699	2.045	2.756
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	∞	0.674	1.282	1.645	1.960	2.576

Table 1.1: Tabulated values of t for which $\Pr(|T| > t) = p$, where T has a t–distribution with ν degrees of freedom