Lecture 05 Part 02 - Interval Estimation

Data Analysis in CJ (CJUS 6103)

Outline

In this part of the lecture we will be discussing:

I. Point estimation
II. Interval Estimation

Point Estimation

Point estimate – the sample statistic used as an estimate of an unknown population parameter. We have been using the sample mean as a point estimate for the population mean, and the sample standard deviation as a point estimate for the population standard deviation.

There are two criteria that we want to meet in choosing a good estimator:

1) **Unbiasedness** – the mean of the sampling distribution is equal to the parameter being estimated. In other words, with an infinite number of samples of size n, the mean of the sample means ($\mu_{\overline{x}}$) converges to the true population mean (μ_x). The property of unbiasedness does not mean that we will get the *true* answer every time, only that we will get the true answer *on average*.

2) **Efficiency** – the sampling distribution clusters tightly about the true population parameter. In other words, the standard error of the mean ($\sigma_{\overline{x}}$) is at a minimum relative to other estimators. This means that we will be close to the true answer, *on average*. Moreover, the efficiency of an estimator improves as we increase the sample size.

Interval Estimation

Confidence interval – the interval within which a parameter has a known probability of lying. We know that there is sample-to-sample variability in the value of the sample mean, and confidence intervals take this uncertainty directly into account. Instead of estimating a single value for the unknown population characteristic, we estimate a range of values within which we believe the true population value falls. Conventional levels of confidence are 90, 95, and 99 percent.

We interpret a 95 percent confidence interval as follows: if we drew an infinite number of samples of size n from the population and constructed a confidence interval around each sample mean, 95 percent of these intervals would contain the true population mean. We often use a shorthand interpretation, however, although it is not entirely accurate: we are 95 percent confident that the true population mean lies between the upper and lower limits of the confidence interval.

The formula used to construct a confidence interval around the sample mean for a given level of confidence is:

$$\text{C.I.} = \overline{\text{x}} \pm z_{\alpha/2}(\sigma_{\overline{\text{x}}}) = \overline{\text{x}} \pm z_{\alpha/2}(\sqrt{\sigma^2/\text{n}}) = \overline{\text{x}} \pm z_{\alpha/2}(\sigma/\sqrt{\text{n}})$$

Note: $z_{\alpha/2}$ simply represents the z-score you would need for a two-tailed test at a particular α level, where the α level is equal to one minus the desired level of confidence (e.g., 1-.90 for a 90% confidence interval).

The z-score is different depending on how confident you want to be. With higher levels of confidence, you are "throwing the net" out farther away from the sample mean in order to capture the true population mean μ. To obtain the z-score for a given level of confidence, we must resort to the standard normal distribution (or the z-distribution). We must first divide our confidence in half, which gives us the probability under the normal curve that lies between the mean and the z-score that bounds the confidence interval (e.g., $90/2 = 45 = .4500$ probability on either side of the mean). We then use this probability to locate the z-score.

Confidence Level	Proportion from the Mean	Proportion in Tail	z-Score
90%	.4500	.0500	1.65
95%	.4750	.0250	1.96
99%	.4950	.0050	2.58
99.9%	.4995	.0005	3.27

Interval Estimation Example - Minneapolis Crime Hot Spots

A study of crime hot spots by Sherman, Gartin, and Buerger (1989) found that in Minneapolis, the mean number of calls to police for all addresses and intersections ("places") in 1986 was 2.82 ($\sigma = 2.31$). Treat these as population figures. In a handful of particularly unsafe neighborhoods with 3,795 places, there was an average 43.03 calls to police for service ($s = 21.24$). Construct a 90 percent confidence interval around the point estimate.

$$43.03 \pm 1.65(2.31/\sqrt{3795}) = 43.03 \pm .06 \Rightarrow [42.97, 43.09]$$

Now let's systematically increase our level of confidence and see what happens.

$$95\%\ \text{C.I.} = 43.03 \pm 1.96(2.31/\sqrt{3795}) = 43.03 \pm .07 \Rightarrow [42.96, 43.10]$$

$$99\%\ \text{C.I.} = 43.03 \pm 2.58(2.31/\sqrt{3795}) = 43.03 \pm .10 \Rightarrow [42.93, 43.13]$$

$$99.9\%\ \text{C.I.} = 43.03 \pm 3.27(2.31/\sqrt{3795}) = 43.03 \pm .12 \Rightarrow [42.91, 43.15]$$

What does it mean when the confidence interval does not contain 2.82 (μ)? From a practical standpoint, this means that the neighborhoods that make up our sample have significantly more calls to police than all of Minneapolis, leading us to conclude that these are particularly high-crime areas, different from the average "place" in Minneapolis.

Interval Estimation Example - Homicide Rate

The mean homicide rate in the U.S. from 1950 to 1999 is 7.11 with a standard deviation of 2.03. In 2001, the mean homicide rate among all 50 states was 5.6 per 100,000. Let's construct a 95% confidence interval using this as our point estimate.

$95\%\ \text{C.I.} = 5.3 \pm 1.96(1.97/\sqrt{50}) = 5.3 \pm 0.55 \Rightarrow [4.75, 5.85]$

What does it mean that the true population mean is not contained in this interval? We know that the homicide rate has reached a record low during the latter half of the 1990's. This suggests that 2001 was a particularly low-homicide year relative to earlier years.

Interval Estimation without a Population Standard Deviation

Unfortunately, we rarely know the standard deviation of the population (μ). Just like we use our sample mean as an estimate of the population mean (or the mean of the sampling distribution), we can use our sample standard deviation as an estimate of the standard deviation of the sampling distribution. However, the sample standard deviation is a biased estimate of the population standard deviation. This requires us to change the formula slightly to account for this bias:

$$\text{C.I. } = \bar{x} \pm t^{n-1}_{\alpha/2}(s_{\bar{x}}) = \bar{x} \pm t^{n-1}_{\alpha/2}(\sqrt{s^2/n-1}) = \bar{x} \pm t^{n-1}_{\alpha/2}(s/\sqrt{n-1})$$

Notice that we also use a different probability distribution than the standard normal. This distribution is very similar to the z-distribution, with the exception that it is flatter and has wider tails. When we use the t-distribution, we have to determine the degrees of freedom, which is easily computed, $df = n\breve{}1$. An important feature of the t-distribution is that, as n becomes large (i.e., over 100), it is virtually identical to the z-distribution. The only difference between the formula for a small sample confidence interval and the formula for a large sample confidence interval is that we replace z with t for a given level of confidence. Notice also that when we have a small sample (and thus few degrees of freedom), the t-score is considerably larger than the z-score. This is because there is greater sample-to-sample variability in the sampling distribution.

Let's work through some examples where σ is unknown.

Juvenile Bootcamp Example

In a sample of 1500 youths sent to juvenile boot camps, you find that the mean number of prosocial activities is 1.53 ($s = 1.21$). Construct a 99 percent confidence around this point estimate.

$$95\% \text{ C.I. } = 1.53 \pm 2.576(1.21/\sqrt{1499}) = 1.53 \pm 0.081 \Rightarrow [1.449, 1.611]$$

Now, assume that you have only 10 youths in your sample. Recalculate the confidence interval ($df = 9$).

$$1.53 \pm 3.25(1.21/\sqrt{9}) = 1.53 \pm 1.311 \Rightarrow [0.219, 2.841]$$

Child Maltreatment Example

There is reason to suspect that child maltreatment is associated with later arrest. We collect data from a sample of 61 young adults with a history of family violence, and find that the mean number of juvenile arrests is 2.57 ($s^2 = 7.90$). Construct a 95% and 99.9% confidence interval around the point estimate.

$$95\% \text{ C.I. } = 2.57 \pm 2(\sqrt{7.90/60}) = 2.57 \pm 0.726 \Rightarrow [1.844, 3.296]$$

$$99.9\% \text{ C.I. } = 2.57 \pm 3.46(\sqrt{7.90/60}) = 2.57 \pm 1.256 \Rightarrow [1.314, 3.826]$$

Let's compare these two confidence intervals. The range of the 95% C.I. is 3.296 – 1.844 = 1.452, and the range of the 99.9% C.I. is 3.826 – 1.314 = 2.512. So, being more confident comes at a price; the range of the confidence interval becomes wider. Thus, there is a trade-off between degree of confidence and precision. Now, suppose the sample size is 250. What changes?

$$95\% \text{ C.I. } = 2.57 \pm 1.97(\sqrt{7.90/249}) = 2.57 \pm 0.351 \Rightarrow [2.219, 2.921]$$

$$99.9\%\text{ C.I. } = 2.57 \pm 3.33(\sqrt{7.90/249}) = 2.57 \pm 0.593 \Rightarrow [1.977, 3.163]$$

Now the range for the 95% C.I. is 2.921 – 2.219 = 0.702, and the range of the 99.9% C.I. is 3.163 – 1.977 = 1.186. By increasing the sample size, we have substantially reduced the width of our confidence intervals.

Confidence Intervals - Confidence v. Precision

As a general rule, there is a trade-off between confidence and precision. If we want to be more precise - that is, to reduce the range around where we think the population mean is - then we will be less confident in the assumption that our interval contains the population mean. By contrast, if we want to be more confident that the interval contains the population mean, we must reduce the precision of our interval and widen it.

$$\Uparrow \text{confidence, } \Downarrow \text{ precision}$$

$$\Uparrow \text{ precision, } \Downarrow \text{ confidence}$$