Interval Estimates (Ch 7)

Ch 7 is basically about interval estimates of parameters (population parameters of course). The main technique for interval estimates is the Confidence Interval (CI).

Recap of CIs for population proportions

The program I was running in class on monday was CIp(), which generated confidence intervals for a proportion. Actually, it was generating approximate 95% confidence intervals. "Approximate" for two reasons – 1. it assumed the sample proportion had a normal distribution, but this is a Central Limit Theorem result which is only exact for an infinite sample size (in practice, n>30 is usually enough) – and 2. The SD of the sample proportion depends on the unknown population sd (which depends on the unknown population value of p), and we estimated this using the sample p. So obviously, both these approximations would tend to fail if the sample size were small (like 10, for example). Now this is not a serious practical problem because the point estimate of p using the sample proportion (of 1s, or of white beads), is very poorly estimated in small sample sizes, so the interval estimate is likely to be so wide as to be useless unless p is at least 50 or so (as our class experiment suggested).

Consideration of CIs for population averages

So now we need to discuss the situation of a population for a continuous random variable, and we are selecting a random sample of size n (=10,25,50, 100, or whatever) in order to estimate a population parameter (such as the population mean).

Since in this case it is possible to get reasonable quality information for small samples, we need to consider what can be done in this case (when the CLT cannot provide a good enough approximation). **We want to estimate the population mean**. We start with

Case 1: Population distribution is normal and population SD is known (Sec 7.1 in text)

This case is detailed in the text in Section 7.1 (pp 281-289). Apart from more detail about varying the confidence level, the calculation procedure is the same as we did in class for the population proportion. But this time the normality of the sample mean is an exact consequence of the normality of the population. And if the SD is known, we don't have to worry that the small sample is too small to estimate it well. See the box on p 284. The 95% CI is sample mean ± 2* popSD/sqrt(n). Here are the factors for other confidence levels – read from the Normal table in your book.

Confidence Level	Factor
90%	1.645
95%	1.960 (almost = 2)
99%	2.575
99.7%	3.000

In essence, we find a probability that $\overline{X}$ is within a certain distance from μ, and after we observe $\overline{X}=\bar{x}$, we turn the probability in to a statement about the distance of μ from $\bar{x}$.

As described on pp288-289, this strategy can be used for other statistics and parameters.

Choice of Sample Size

A common question of statisticians is "What sample size do I need?" and the statistician always answers "How much precision do you need in your parameter estimate?". For example, suppose we want to determine if a person's reaction time is lower at mid-day than it is in the evening. The measurement we would use for each individual is the difference in two reaction times. We are trying to measure the average difference (evening reaction time – midday reaction time) for a population of individuals, and lets suppose we are able to select a random sample of individuals from this population, and we will base our information about reaction times on this sample. How many individuals do we need in our sample? The question about precision requires practical information about the context of the study – how much difference in reaction times would be an interesting and/or useful finding? Now reaction times are typically around 0.2 seconds and can be measured, for an individual, to the nearest .01 seconds. So an average difference of .02 seconds might be deemed to be of interest. In other words, if the average difference is .02 or more (assuming positive), we would have an interesting result. This means we want a sample size that would ensure our CI width is less than .02, since then an average difference of .02 or more would give rise to a CI that does not include 0. 0 would not be a credible value for the average difference in that case, which would be an interesting result.

Now the width of a 95% CI is $2*\sigma/\sqrt{n}$, so to make this less than .02, we can solve for n and have $n\geq 10000\ \sigma^2$.

If we knew the variability (from one person to the next) of the difference in reaction time readings, we would have our required sample size. Lets suppose σ=.02 seconds is known so that σ^2=.0004 and $n\geq 4$ should be enough in this case to produce the required precision of estimate. If we did not know the σ we would be stuck, but in this "Case 1" we assume we do know σ.

The general strategy for sample size calculations is to equate the half-width of a confidence interval to the allowable error. The different Cases are just to outline the different ways that a CI may be computed.

Case 2. Population distribution is Normal but σ is unknown (Sec 7.3 in book)

When σ was known, $\bar{x}$ is normal (Ch 5) and so is $z=(\bar{x}-\mu)/(\sigma/\sqrt{n})$. But when σ is unknown, and must be estimated by s, $z_8=(\bar{x}-\mu)/(s/\sqrt{n})$ is only approximately normal. The variability in s induces extra variability in z_8, and it turns out even the shape of the distribution of z_8 is different from Normal. To emphasize this distinction, we will in future refer to this z_8 as T and call its distribution **the t-distribution.** The T statistic has exactly a t-distribution when the population distribution (of X) is Normal. (Theorem p 300). Note that the t-distribution has only one parameter – the so-called "degrees of freedom". In most cases, the degrees of freedom for the t-distribution is n-1, just 1 less than the sample size used to estimate σ.

The t-distribution is bell shaped but not Normal! See the figure on p 300. The t has fatter tails than the standard normal, as you might expect from that extra variability in s. But we still use a similar formula for the CI. See bos p 302. As you will see, the only effect on the CI of estimating σ with s is that the normal multiple (1.96 for a 95% CI) is replaced by a larger number from the tables of the t-distribution. This number depends on the sample size since the degrees of freedom of t = n-1. For example, if n=4, the degrees of freedom n-1=3 and the multiple for the 95% CI is (from Table A.8) 3.23. In the Case 1 formula for the 95% CI, instead of σ one uses s, and instead of 1.96 one uses 3.23.

So a wider CI will usually result when the σ has to estimated by s.

This theory about the t-distribution does depend on the population distribution (of X) being Normal. However, this assumption is not critical in practice reasonable intervals will result from this method even if the population is a bit different from Normal.

Case 3. Population distribution not Normal, and σ is unknown (Sec 7.2 in book).

This is the case where we need to rely on the Central Limit Theorem saving the day. The formula is very similar to what we did in Case 1 except that we replace σ by its estimate s, and do not worry about the extra variability! Clearly we need a large sample for this to work – for two reasons. We need s to be close to σ, and we need the Normal approximatio to be good enough. See the box on p 292**. Our rule of thumb is that the CLT is good enough when $n \geq 30$.** Actually, the sample size for an adequate approximation depends on how non-normal the population is – if it is nearly normal, a sample as small as 10 may be enough, and if it is violently skewed, we may need n=50.

Related Topics

1. **Bootstrap Confidence Intervals**
2. **Confidence Intervals for Other Parameters than the Pop. Mean**
3. **Prediction Intervals**

1. Bootstrap Confidence Intervals

In the lectures before the midterm, we showed how the bootstrap could produce an estimate of the SD of any statistic based only on the data in the random sample. Note that the values of the statistic from each bootstrap sample will also estimate the whole distribution of the statistic. Of course, the centre of this distribution will be sensitive to the particular original sample we have, but the shape of this distribution may still be used to estimate a confidence interval for the statistic. An example of this approach is described on pp 289-290. Just be aware that there is a bootstrap approach that can be used for moderate sample sizes when the population distribution is unknown.

2. Confidence Intervals for Other Parameters than the Pop. Mean (Sec 7.4)

We only discussed in detail the CI for a population mean. But CI for other parameters are possible when certain model assumptions are satisfied. p 288-289 gives the details of one example. A more important example is the content of Sec 7.4 in the book – CI for σ.

First a simple fact that will help in future statistics courses as well as understanding this course. If X ~N(0,1) then X^2 ~ Chi-squared distribution on 1 degree of freedom (df). Moreover, if X_i~N(0,1) and independent, then $\sum_{i=1}^{k} X_i^2$ ~ Chi-squared distribution on k df. Now look at the box on p 308. It says that the sample standard deviation S^2, when multiplied by a suitable constant, has a chi-squared distribution on n-1 df, where n is the sample size. Think of df as the number of independent observations – that should make the term "degrees of freedom" seem plausible. Note that while the summation in the boxed Theorem has n terms, it apparently only has n-1 degrees of freedom. We lose one degree of freedom in estimating μ with $\bar{x}$. In fact, if we did happen to know μ, and used it in the theorem instead of $\bar{x}$, the statistic shown would have a chi-squared distribution with n df.

Anyway, we can use the theorem to say how close s is likely to be to σ, since the theorem describes (exactly) the distribution of the ratio S^2/σ^2. The result is given in the box on p 310.

3. Prediction Intervals

Note that a confidence interval is a description of a reasonable range for a population parameter. This is not the same as a reasonable range for the data itself. Of course, the reasonable range for the data itself **is** the data itself! But another way to ask the question about the range of the data is "What interval would contain the next observation with high probability (say probability 0.95)?" The answer is called a **prediction interval.** See the box on p 304. Note that it is much wider than a CI for the population mean.

Tolerance Intervals: you can ignore this extension of the prediction interval idea (for this course).