Research Methods 2

Week 7: Document 1

How to use the standard error: Confidence Intervals

The problem

In fourth question of the question sheet from Week 6, you computed the mean, SD and SE of the following set of data (which are the number of polyps in nine patients with familial adenomatous polyposis coli.

26, 16, 40, 14, 16, 11, 26, 45, 32

The results are that the sample mean is 25.1, the SD is 12.0 and the SE is 4.0.

Therefore our estimate of the population mean is 25.1 and how good that is as an estimate is measured by the SE, which has value 4.0. What does this mean? How can we use the value of the SE to inform our inference?

It is the purpose of the unit this week to address this problem. The answer to the problem is to use a *confidence interval* and what this is and how to compute it will now be presented.

The Confidence Interval

Basis of the Confidence interval: Normal variables

For variables with a Normal distribution the solution to the problem of how best to use the information in the standard error is to link together the following four facts.

1. The sample mean follows a Normal distribution.

2. The population mean of that Normal distribution is the same as the population mean of the distribution of the individual observations.

3. The population SD of the distribution of sample means is estimated by the *standard error* of the sample.

4. For any Normal distribution, the 68-95-99.7 rule tells us that 95% of observations lie with two standard deviations of the mean.

Given a sample of size n with mean m and standard error $s/\sqrt{n}$, we know from 1 that m has a Normal distribution. We know from 2 that the Normal distribution has mean m, the parameter we wish to estimate. We also know from 4 that there is a 95% chance that m is within 2 standard deviations of m, i.e. there is a 95% chance that m is between $m - 2s/\sqrt{n}$ and $m + 2s/\sqrt{n}$. This formula uses $s/\sqrt{n}$, the *standard error*, because this is the *standard deviation* of the distribution of sample means: -item 3.

Computing a confidence interval[†]

Another way of saying this is that there is a 95% chance that m is located between $m - 2s/\sqrt{n}$ and $m + 2s/\sqrt{n}$. The interval $m - 2s/\sqrt{n}$ to $m + 2s/\sqrt{n}$ can be calculated from the sample and is known as the 95% confidence interval for the population mean.

So to return to our example of the polyps, the mean, m, is 25.1 and the SE, $s/\sqrt{n}$, is 4.0. Thus the 95% confidence interval is $25.1 - 2\times 4.0$ to $25.1 + 2\times 4.0 = 17.1$ to 33.1. The way this would be used in practice is as follows.

If we were to estimate the mean using just a single number, then the number we would use is 25.1. However, we know the population mean will not be exactly this value and we can acknowledge this uncertainty by estimating the mean by an interval, rather than a single value. If this is done we would use the 95% confidence interval and give our estimate of m as 17.1 to 33.1. We would be 95% certain that this interval would include m.

It might seem that a confidence interval from 17.1 to 33.1 is rather wide and that we do not have a very good idea about the value of m. If this is so, it is likely to be the case because we are only using a sample of size 9. If the sample had been of size 36, rather than 9 (i.e. four times the size) then the SE would have been half as big, i.e. it would have been 2.0. The confidence interval would then have been $25.1 \pm 2\times 2.0$, i.e. 21.1 to 29.1. This is a narrower interval and therefore our estimate of m is more precise. A confidence interval can made be arbitrarily narrow simply by collecting a sufficiently large sample. Of course, in practice, arbitrarily large samples are not readily available. However samples should be large enough to produce an estimate that is sufficiently precise for our purposes. More will be said of this later in the course.

Different confidence levels

The confidence interval is calculated by adding and subtracting a multiple of the SE from the sample mean. We use '2' as the multiplier because we want a 95% confidence interval. The '95' part of the 68-95-99.7 rule for the Normal distribution explains why 2 times the SE corresponds to a 95% interval. We can have a narrower interval if we are prepared to accept a lower level of confidence. For example, the 68 part of the 68-95-99.7 rule tells you that the interval $25.1 \pm 4.0 = 21.1, 29.1$, i.e. 1 times the SE, give a 68% confidence interval. However, there is little value in using an interval where you have a 32% chance that your claim it contains m is wrong. Occasionally you might use a multiplier larger than '2', e.g. if you wanted a 99% confidence interval. However, 95% intervals have come to be widely used, as they represent a sensible compromise between a confidence level high enough to be worthwhile and intervals narrow enough to be useful.

The Confidence Intervals: non-Normal data

Data that do not have a Normal distribution can have a very wide variety of distributions. For some types of non-Normal data special methods need to be used. For example if a variable can take only two values, say 'in remission' or 'not in remission', then the methods described in Bland, section 8.4 must be applied. However, for many continuous variables the observation made in question 5 of the example sheet for Week 6, is relevant. This was that sample means tend to have a Normal distribution, even when the underlying individual observations do not have a Normal distribution. Moreover, the larger the sample, the closer to Normal is the distribution of the sample mean. So even for continuous variables with a distribution that is not all that close to Normal, the methods outlined above can be applied.

[†] For reasons to be explained later, the intervals described here are approximate

Confidence intervals: a technicality

The footnote to the section 'Computing a confidence interval' stated that the 95% confidence interval given there was an approximation. Both in the example sheet and when reading the literature, you may find that if you compute a confidence interval in the way described above, then it will not agree with the interval in the paper or produced by the program. We will now expand on this, but if you do not wish to absorb the extra-complexity involved and are prepared to leave the minor discrepancies as a continuing minor mystery, you will not be disadvantaged in the remainder of the course.

The 95% confidence interval ($m - 2 \times s/\sqrt{n}$, $m + 2 \times s/\sqrt{n}$) does not, in fact, have a 95% chance of containing μ. The reason, which need not concern us deeply, is to do with s being an imperfect estimator of σ, in the same way that m is an imperfect estimator of μ. The solution is to amend the multiplier '2': a different multiplier is chosen so that the chance that the 'correct' confidence interval contains μ is 95%. However, as the problem is due to s being an imperfect estimator of σ, and as s will become a better and better estimator of σ as the sample size increases, the amended multiplier will change as the sample size changes. There is, in fact a different multiplier for every sample size. These numbers do not change all that much. A selection of multipliers is shown in the table below

Sample size	Multiplier
4	3.18
25	2.06
50	2.01
100	1.98

There is no need to know how these numbers are obtained, although full details are in section 10.2 of Bland. If you compute a confidence interval in a program, it will automatically use the correct multiplier.

This is a convenient place to note that the '95' part of the 68-95-99.7 rule is itself an approximation. Two SDs about the mean does, in fact, cut off about 95.4% of the population. The mean plus and minus 1.96 times the SD is the interval that cuts off exactly 95%. You will occasionally see the multiplier 1.96 in statistics texts and articles where we have used 2.