Ch 7 – Estimation with Interval Estimates

Preamble:

Rather than jump right in to this chapter, I want to go through a series of sampling experiments using the bead sampler. This population of coloured beads allows us to see an example of a physical random sample, and how the sampling variability requires our attention when we use the sample to infer something about the population. A difference between this bead sampler and much of the discussion in Ch 7 is that the feature of interest in this bead population is a proportion (of white beads, for example) rather than a mean of a quantitative variable. However, we have shown that our theory about estimating means of quantitative variables also applies to estimating proportions since a coding of population objects as 0 or 1 (not white or white, for example) allows the proportion to be the same thing as an average.

The nice thing about the bead sampler is that the proportions of beads of various colours are known. But we can imagine that this information is not known, and we selected a sample of 10,25,50 or 100 and in each case try to use the sample to inform us about the population – more precisely, to estimate the population parameters which in this case are the proportions of beads of various colours.

Note that this population can in no way be thought of as a "normal" population. Whether we code the beads as 0-1, or 0,1,2,...,8 (there are 8 colours), the population relative frequency of each code cannot be considered as a continuous frequency distribution on $(-\infty,\infty)$, and so it is not even a continuous distribution. We want to imagine that all we know about the population is that the proportions of the eight colours (white,green,red,black,violet,yellow,pink,blue) are some numbers $p_1,p_2,...,p_8$ where each $0\le p_i\le 1$ and $\sum_{i=1}^{i=8} p_i = 1$. (Actually, if we did not know which colours were in the population, we could allow many more p_i's and allow that some might be 0, but to keep things simple, we assume we know the colours represented.)

To make things a little bit realistic, lets imagine that the colours represent the intended votes of the electorate for eight candidates for president in a student society election. We will call the candidates by their color to avoid further confusion. Now lets suppose that cadidate White has been working hard on the campaign and wants to know if he/she is likely to win. An opinion poll based on a small sample of students random selected from the eligible students would produce a certain proportion of intended votes for White. The sampling paddle can simulate this process – we will do this in class but I will describe one particular outcome here – in fact, since the information based on a small sample is likely to be unconvincing, I will take samples of 10, 25, 50 and 100 to see what happens as I increase the sample size. (The 25 does not include the 10, etc – these are independent samples.)

Sampling Outcomes

Results of these samples are as follows:

Sample Size	No of white	Estimated Proportion of White
10	3	.30
25	7	.28
50	28	.56
100	48	.48

The estimates of the population proportion of Whites varies quite a bit – in the population there is a particular proportion that is the same for each of these four experiments. The fact that the sample proportions vary is a result of sampling variation.

Of course, if we take larger and larger samples, we could expect the variability to damp down. But can we be more quantitative about the variability of the sample proportion that would be expected from a particular sample size?

The SD of a sample proportion

Since a sample proportion can be deemed to be a sample mean, we can say that the variability of the sample proportion can be estimated as s/$\sqrt{n}$. But what is s in these samples? The population is very large but can be deemed to consist of a proportion p of 1s and a proportion (1-p) of 0's. Knowing this allows us to compute the SD in terms of p from the SD formula. The mean of the population is p (can you see why?). The deviations are $(x_i\text{-p})^2$ and there are only two possible values for these: $(0\text{-p})^2$ and $(1\text{-p})^2$, and we know the proportion of each in terms of p. There are a proportion p of them like $(1\text{-p})^2$ and (1-p) of them like p^2. So the expected deviation is $p*(1\text{-p})^2 + (1\text{-p})*p^2 = p(1\text{-}p)*(1\text{-p+p})=p(1\text{-p})$. This is the population variance $(=E(X\text{-}E(X))^2)$. So the population SD = $\sqrt{p(1-p)}$. In other words, if we knew p, we would not have to go through the general SD formula to get the SD. For example, if p = 0.4, pop SD = $\sqrt{.4(1-.4)}$ = .49.

Now the SD of the sample proportion = SD of the population/$\sqrt{n}$ and in this case it is

$$\sqrt{\frac{p(1-p)}{n}}$$

Estimating the population proportion p

Now when we sample a population of 0s and 1s, we do not usually know p. But the sample does allow us to estimate it. For example, from out sample size 100 selection noted above, we estimated p = 0.48. So we can use this as an approximate value of p and say that the SD of this estimate (i.e. the estimate is 0.48 in this instance) is about

$$\sqrt{.48(1-.48)}\Big/\sqrt{100} = .025$$

This means that we think our estimate of p, the 0.48, is likely within 1 or 2 SDs of the population value. That is, our population p is probably somewhere in .48 ± .025 or, to be safer, .48 ± .050.

Note what would have happened if we had relied on a sample of n=25? We would have said our true population mean was in the interval .28 ± 2* $\sqrt{.28(1-.28)}\Big/\sqrt{25}$ = .28 ±.08. This interval does not even overlap with the n=100 interval, so one of them is certainly wrong. This can happen. If we use mean ± 2SDs, we will only be correct 95% of the time. Our example for n=25 was one of those 5% instances that can happen. Usually (95% of the time) the population mean will lie in the interval computed from the sample this way.

Notice one other thing about the n=25 example compared to the n=100 example. In the n=100 example, the interval width was .05, while it was .16 in the n=25 example. Bigger samples tend to give narrower intervals, and narrower intervals are better information about the parameter we are estimating, right?

Revisiting the SD of a sample proportion

When we computed the SD formula for 0-1 populations, did it look familiar? It actually is the same as the formula for the Binomial SD when n=1, since $\sqrt{np(1-p)}$ = $\sqrt{p(1-p)}$ in this case, or the SD of the Bernoulli RV. (The n in this formula is not the same as the sample size here – every member of the population is either 0 or 1, whereas for the Binomial, the population can be 0,1,2,...,n.)

Choosing a multiple of SDs in estimating a population proportion

But why do I use ±2 SDs? Here is where the CLT is useful because it tells us that the sample mean (or sample proportion of 1s) has a Normal distribution, approximately. In a normal distribution, approximately 95% of the values are within 2 SDs of the mean. That is why I used 2SDs for the intervals above. If I wanted to reduce the chance of being misled by a chance result, I could ask for an interval of values around the sample proportion (like .48) that was ±3SDs. It turns out that such an interval will contain the population mean 99.7% of the time. But the catch is that the interval is wider than a 2SD interval, and so there is a balance between how sure you need to be and how close the estimate should be. It is quite common to compute 95% interval estimates.

Interpreting Confidence Intervals

There is a name for these intervals: Confidence Intervals. Note that when you say a confidence interval for the population proportion is 0.48±0.050, you need to be careful about what probability you are implying. The population mean does not vary – no matter how much we sample the population mean stays constant at the unknown p. If p=0.5, then our confidence interval is a good one, but if p=0.4 it is not. But the probability implied in the 95% concerns the sequence of intervals that one might get in repeated sampling of size n samples. Usually, the interval will contain the unknown but true population mean, but sometimes it will not. See the figure on p 285.

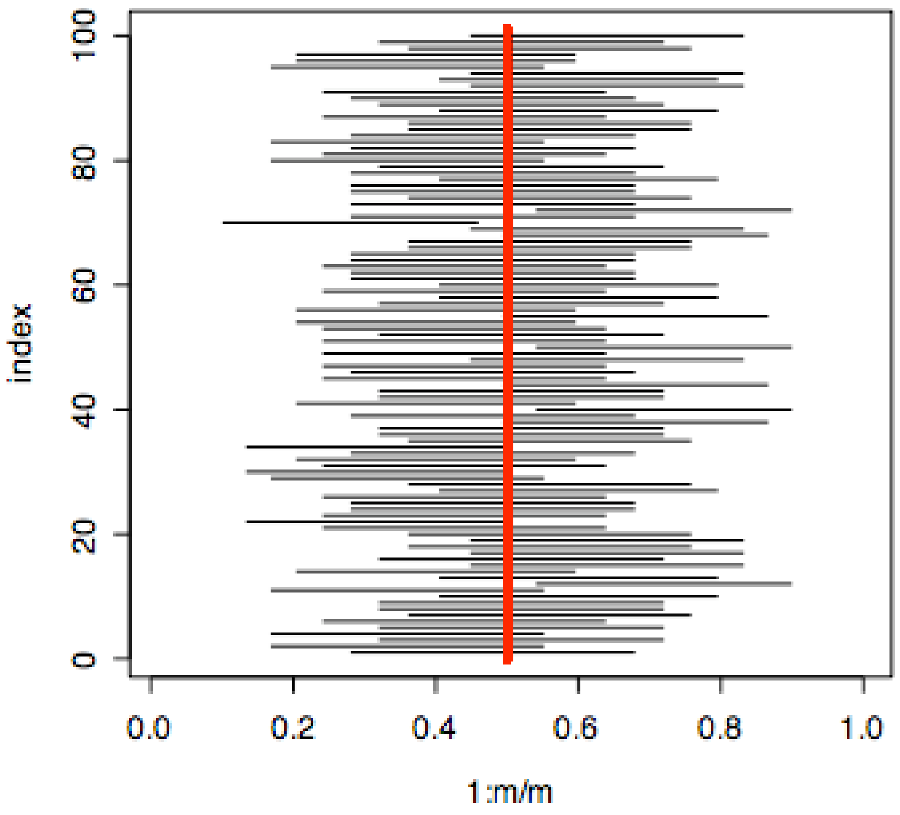
Extension of the CI estimation method to continuous data

When we had a population of 0s and 1s, we had a short-cut method of computing the sample SD based on the sample proportion of 1's. But when the population is a quantity, like Sales or GPA, we need to go back to the long formula for the sample SD. However, the rest of the procedure is unchanged, as long as we have large enough samples for the CLT to give good approximations. (Rule of thumb n>30). See pp 281-286 in your text.

Appendix

Is it really true that a 95% confidence interval for a population mean, $\bar{x} \pm 2\ {}^{s}\!/\!_{\sqrt{n}}$, will contain the population mean in 95% of the samples?

Here are some CIs based on 100 samples of size 25. In this case we were lucky and we have exactly 95 that contain the true mean (p=.5) in this case.



R Program pthat produces this graph is given on next page ...

```
> CIp
function (m=40,n=25,p=.5)
{
        index=1:m
        count=0
        pr=rbinom(m,n,p)/n
        sd=(pr*(1-pr)/n)^.5
        p.up=pr+2*sd
        p.down=pr-2*sd
        for (i in 1:m) {
        #       print(c(p.down[i],p.up[i]))
          if (is.between(p.down[i],p,p.up[i])) {count=count+1}
        }
        plot(1:m/m,index,type="n")
        for (j in 1:m) {
                lines(c(p.down[j],p.up[j]),c(index[j],index[j]))
                lines(c(p,p),c(0,101),lw=3,col="red")
                }
                print("number of intervals containing true p=")
        print(count)
        print("out of")
        print(m)
}
```

To run it, just use

```
CIp()
```

while in R
and to change parameters use

```
CIp(m=100,n=50,p=0.2)
```