More Sampling Experiments (Ch 7-8)

Underlying the techniques in Chapters 7-9 is the important concept of the sampling distribution of a statistic, and in particular the sampling distribution of a sample mean or of a sample proportion. I want to do a few more sampling experiments with the bead-bowl sampler, and then explore this model further with some R-simulations of the bead-bowl sampler.

1. Sampling variability of a proportion – dependence on the population proportion

Our bead sampler has some colours that are fairly rare – take the red ones for example. There are many less than the white ones – we know now that the white ones make up 50% of the population in the bowl, but the red ones are something much less. Lets imagine we want to know the proportion of red ones. (To add to the realism, we could think of this as the proportion of students who will get F or D on this course – hopefully a very small proportion!). Here are the results of a sampling experiment using our bead-bowl where the focus is on the red beads:

n	#red	proportion red
10	0	0.00
25	1	0.04
50	6	0.12
100	8	0.08

Now, if we had based our estimate of the population proportion "red" on one of the four experiments shown here, we would obviously incurred an error in at least three of them, and in general we expect some error of estimation. But the question is, how much error do we expect? The true SD of the sample proportion (based on a sample of some size n) will depend not only on n but also on the population proportion of "red". Why? Because the population SD is $\sqrt{p(1-p}$. But we do not know p. But our estimates of p from the samples do give, in each case, an estimate of p. In fact, by putting these sample proportions into the population SD formula, we get an estimate of the SD of the sample proportion:

n	#red	proportion red	est SD of prop. red
10	0	0.00	0
25	1	0.04	.20
50	6	0.12	.32
100	8	0.08	.27

So a 95% CI for the population proportion is essentially unknown for n= 10 or 25 since these samples are too small to use the CLT. But for n = 50 we have .12 ±.64/7 = .12 ± .9

and for n=100 it is .08 ± .54/10 = .08±.05. Note that even our sample of n=50 does not give very much info about the unknown proportion of "red", but the n=100 example is starting to be of some use.

2. Sampling variability of a sample mean:

We can use the bead sampler for this by endowing the colours with a number. Say white = 1, green=2, red=3, black=4, violet=5, yellow=6, pink =7, and blue =8.

Here is an example of what we would get in four experiments:

n	sample.avg	sample.SD
10	2.10	1.45
25	1.80	1.19
50	1.88	1.39
100	1.90	1.17

and now our 95% CI is possible to compute for all the sample sizes, but not valid for the small sample sizes since the population is not normal. So again we are limited to an approximate CI for the larger sample sizes:

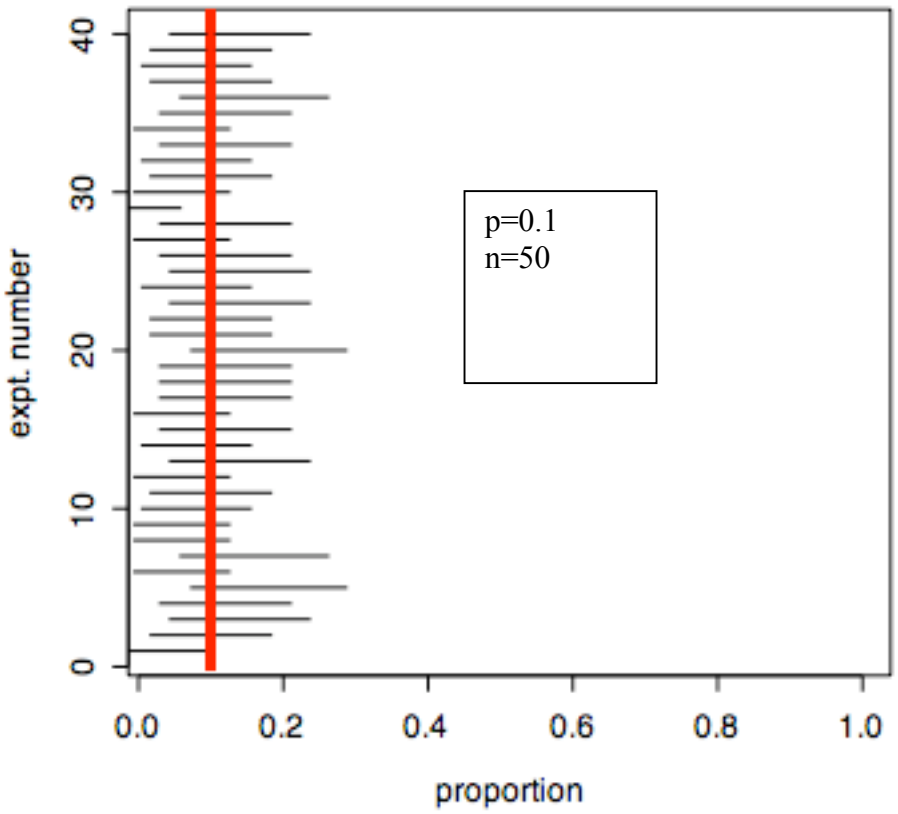
for n=50, 1.88 ± 1.39/7 or 1.88±.2
for n=100, 1.90± 1.17/10 or 1.90±.12

It turns out that the true population mean underlying these results (simulations really, since I discovered that the red beads are barely discernable from the pink and purple!) is 1.885 and the SD is 1.24 (based on white,green,red,black,violet,yellow,pink,blue with probabilties .5,.3,.1,.05,.03,.01,.005,.005 respectively – which you can check).
So the two 95% CIs do in fact include the true pop. mean of 1.885, just as you would expect.

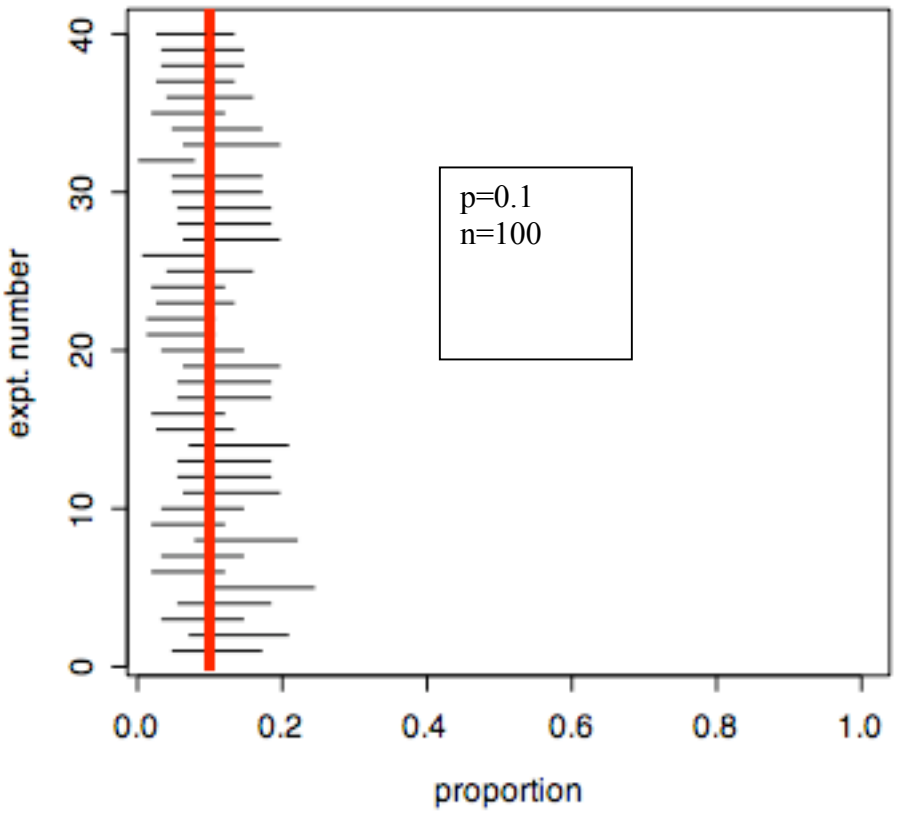
3. Simulating many CIs for proportions and means

Simulation allows one to show what would happen in a large number of sampling instances – for example, how often does the 95% CI actually include the true population parameter value. Lets run that CI program for the situations in 1. and 2. above.

If we run the simulation 40 times, for n=50, we get 38 of 40 CIs (luckily, exactly 95%) that include the true value 0.1 of the population proportion. See figure next page.



For n=100, we get even tighter CIs:



The notes were originally intended to include the CI simulation for the population 1,2,...8 with the various frequencies, but this is quite similar to the posting for Friday March 9, and so I will omit here.

You should check the notes about intro to Ch 8 posted