MONT 107N – Understanding Randomness
Selected Solutions for Problem Set 4
April 12, 2010

Chapter 26/6 (a) Since this part says the 350 people are chosen from a *large* population, under the null hypothesis that the selection *was random*, we can model the process of selection by thinking of drawing 350 times with replacement from the box $[0, 1]$. The expected number of 1's is 175 and the SE for the number is $\sqrt{350} \times \sqrt{(.5)(.5)} \doteq 9.35$. Then the observed z-statistic is

$$z = \frac{102 - 175}{9.35} \doteq -7.8$$

This is so small that the p-value of the test is approximately zero.

(b) Note that the problem says the selection is done *without replacement* now. This means that in computing the SE we must use the *correction factor* from pp. 367-370 in the book. The SE for the number of 1's is

$$\sqrt{\frac{350 - 100}{350 - 1}} \times \sqrt{100} \times \sqrt{(.29)(.71)} \doteq 3.84$$

Then

$$z = \frac{9 - 29}{3.84} \doteq -5.21.$$

Again, the p-value is almost zero.

(c) The conclusion to be drawn from both parts here is that the jury selection was biased against women.

8. The null hypothesis is that the average educational level in the county is also 13 (the national average), and the alternative is that the average in the county is higher. Note that we are *not comparing two independent samples here*, so this is a one-sample test, not a two-sample test. Using the SD of the sample to estimate the SD of the county population, we compute

$$z = \frac{14 - 13}{5/\sqrt{1000}} = 6.32.$$

This is very strong evidence to reject the null hypothesis ($p \doteq 0$).

9. You need to think about this one like this: On each trial, the computer is drawing 100 times with replacement from the $[0, 0, 0, 0, 1]$ box. So the EV for the sum is $100 \times \frac{1}{5} = 20$ and the SE for the sum is $\sqrt{100} \times \sqrt{(1/5)(4/5)} = 4$. Now, we are repeating that process 144 times. So this should be like drawing 144 random samples from a population with average 20 and SD = 4 and computing the average. The null hypothesis is that this is a good model for the process the computer program is performing. But then the SE for the average is $4/\sqrt{144} = 4/12 \doteq .33$. However, using the observed average 21.13, we get

$$z = \frac{21.13 - 10}{.33} \doteq 3.39.$$

The chance of observing a z value this big if the null hypothesis is true is very small. So there must be something wrong.

Chapter 27/2. (a) We want to test the alternative hypothesis that box B contains a greater percentage of positive numbers versus the null hypothesis that the two percentages are the same. There are essentially two different ways that are used to estimate standard errors in problems like this. Let us work the problem both ways and compare the results.

Method 1: Use the bootstrap to estimate the SE for the percent from each box, then combine using the formula for the SE for the difference:

$$\mathrm{SE}_A = \sqrt{\frac{(.5)(.5)}{100}} \times 100\% = 5\%$$

and

$$\mathrm{SE}_B = \sqrt{\frac{(.524)(.476)}{250}} \times 100\% \doteq 3.2\%$$

Then

$$\mathrm{SE}_{\text{diff}} = \sqrt{SE_A^2 + SE_B^2} = \sqrt{5^2 + 3.2^2} \doteq 5.94\%.$$

If we wanted to combine the calculation into one formula, the SE for the difference could be written like this:

$$\begin{aligned}\mathrm{SE}_{\text{diff}} &= \sqrt{\left(\sqrt{\frac{(.5)(.5)}{100}}\right)^2 + \left(\sqrt{\frac{(.524)(.476)}{250}}\right)^2} \times 100\% \\ &= \sqrt{\frac{(.5)(.5)}{100} + \frac{(.524)(.476)}{250}} \times 100\% \\ &\doteq 5.91\%,\end{aligned}$$

which is a more accurate approximation since we were not rounding as we went along. Note that in the simplified form on the second line above, the terms under the square root here are *not squared*. The squaring in the formula for the SE of the difference just gets rid of the square roots in the formulas for SE_A and SE_B.

Method 2: The slightly troubling part of Method 1 is that if we are thinking of an SE for computing z in a hypothesis test, we usually want to be doing the computation under the assumption the null hypothesis is *true*, and that means that we should not be using the two separate fractions .5 and .524. The way to get around this objection is to lump the two samples together and get a common value for the fraction of positive numbers like this:

$$\frac{50 + 131}{350} \doteq .517$$

Then we just use this at both places in the formula for the SE of the difference:

$$\mathrm{SE}_{\text{diff}} = \sqrt{\frac{(.517)(.483)}{100} + \frac{(.517)(.483)}{250}} \times 100\% = 5.91\%$$

Note that the two computations come out (almost) the same in this case(!) This explains why people often use Method 1 – it is a computational shortcut that gives virtually equivalent results in most cases. Still, Method 2 is the "logically correct way" to estimate SE's in this case, and I would say it is better for that reason.

To finish the problem, using either estimate for SE_{diff},

$$z = \frac{52.4 - 50}{5.9} \doteq .406$$

This means that the difference could easily be due to chance.

The same comments apply in problems 3, 5, and part (a) of 7.

4. This was a tricky question that fooled most people. So I treated it as an Extra Credit question in computing the scores for the problem set. Recall that in applying the test for difference of means (or percents), we must assume that the two random samples are independent. We cannot do that here, since it *the same group of people* that gave the ratings for doctors and for druggists. The two-sample z-test should not be applied in this situation.