Prof. Green
Intro Stats

Homework Assignment 5:
Politics Section Supplement

1. *Sampling variability: one proportion. Suppose we are interested in learning about the proportion of high school students who use illegal drugs each year in the town of Bigcity. Unbeknownst to us, the true proportion (p) is .35. The total number of high school students is 25,000. Suppose we conduct a survey in which 400 high school students are sampled at random. Suppose that they are entirely truthful about whether they have used illegal drugs within the past year. What is the probability that our sample will show that 40% or more of the students use illegal drugs?*

The standard error of proportion based on a sample of 400 is

$$SE = \sqrt{\frac{.35(1-.35)}{400}} = .0238$$

The gap between .40 and .35 is .05, and .05 represents a distance of .05/.0238 = 2.1 standard errors. The probability of obtaining a sample proportion as large .40 is therefore .0179.

B. If the sample size were 1000, what is the probability that our sample will show that 40% or more of the students use illegal drugs?

$$SE = \sqrt{\frac{.35(1-.35)}{1000}} = .015$$

Now, .4 is 3.31 standard errors above .35, and the p-value is < .001.

C. If we gathered a very large number of samples of size 400, 95% of them would fall within what range?

The interval is centered at 35% plus or minus 4.66%.

D. *Use Minitab to generate 25,000 observations in which 35% have scores of 1 and 65% have scores of 0. Imagine that you did not know that the true population proportion were .35. Draw a random sample of size 400 and construct a 95% confidence interval. Describe the formula used to construct this interval.*

The formula +/- 1.96 times the standard error formula above uses the estimated proportion, p.

Test and Confidence Interval for One Proportion

```
Test of p = 0.5 vs p not = 0.5

Success = 1

Variable          X      N  Sample p        95.0 % CI       Z-Value  P-Value
C1              154    400  0.385000  (0.337315, 0.432685)    -4.60    0.000
```

This test was based on the normal approximation. It does bracket the true mean.

E. Now examine the sampling variability of these confidence intervals. Draw 20 random samples of size 400. For each one, construct a 95% interval. (Clicking on the Edit Last Dialogue button returns you to the most recently used menu and saves time.) How many of these 20 confidence intervals captured the true p?

```
Variable          X      N  Sample p        95.0 % CI       Z-Value  P-Value
C1              124    400  0.310000  (0.264676, 0.355324)    -7.60    0.000
C2              142    400  0.355000  (0.308107, 0.401893)    -5.80    0.000
C3              131    400  0.327500  (0.281509, 0.373491)    -6.90    0.000
C4              148    400  0.370000  (0.322686, 0.417314)    -5.20    0.000
C5              129    400  0.322500  (0.276692, 0.368308)    -7.10    0.000
C6              135    400  0.337500  (0.291161, 0.383839)    -6.50    0.000
C7              132    400  0.330000  (0.283920, 0.376080)    -6.80    0.000
C8              141    400  0.352500  (0.305681, 0.399319)    -5.90    0.000
C9              158    400  0.395000  (0.347094, 0.442906)    -4.20    0.000
C10             152    400  0.380000  (0.332433, 0.427567)    -4.80    0.000
C11             136    400  0.340000  (0.293577, 0.386423)    -6.40    0.000
C12             139    400  0.347500  (0.300836, 0.394164)    -6.10    0.000
C13             128    400  0.320000  (0.274286, 0.365714)    -7.20    0.000
C14             142    400  0.355000  (0.308107, 0.401893)    -5.80    0.000
C15             142    400  0.355000  (0.308107, 0.401893)    -5.80    0.000
C16             150    400  0.375000  (0.327557, 0.422443)    -5.00    0.000
C17             143    400  0.357500  (0.310533, 0.404467)    -5.70    0.000
C18             147    400  0.367500  (0.320253, 0.414747)    -5.30    0.000
C19             144    400  0.360000  (0.312961, 0.407039)    -5.60    0.000
C20             130    400  0.325000  (0.279100, 0.370900)    -7.00    0.000
```

Answer: 20/20.

F. *Now suppose that the instructor inspects 100 students' answers to the previous problem and writes down the number of times each student reportedly captured the true p. Five students report that they captured the true p 20/20 times, and 95 report that they captured the true p 19/20 times. Should the instructor suspect that the students somehow botched or faked the results? Why or why not?*

This is not the distribution we would expect from 100 student experiments. The instructor should treat the exercise as 100 draws from a binomial distribution in which a coin with probability .95 is tossed 20 times. A binomial table shows that 36% of the time, such a process will produce 20/20 heads, and 38% of the time, it will produce 19/20 heads. This is nowhere near the observed distribution.

(After the midterm, apply a chi-square test to these data!)

2. *Sampling variability: two proportions. Now consider drug use in another town, Smallcity (with a school population of 5,000). Here, the true rate of drug use (p) is .30. Suppose we don't know the true rates of drug use, however, in Bigcity or Smallcity. We're interested in learning which city has a higher proportion of high school drug users.*

A. If we draw simple random samples of 400 students from each city, what is the probability that this study will (incorrectly) show drug use to be more frequent in Smallcity?

$$SE\ of\ difference = \sqrt{\frac{.35(1-.35)}{400} + \frac{.3(1-.3)}{400}} = .033$$

The distribution of differences is centered at .05 and has a standard deviation of .033. Thus zero lies .05/.033 = 1.52 standard errors below the mean. Thus the probability is .066.

What if each sample had contained 1000 students?

$$SE\ of\ difference = \sqrt{\frac{.35(1-.35)}{1000} + \frac{.3(1-.3)}{1000}} = .021$$

.05/.021 = 2.38. Thus, p=.009.

B. *Use Minitab to simulate the data for both cities. Draw random samples of 400 students in each city. Compare the two sample proportions and construct a 95% confidence interval around the observed difference in proportions. Did your interval bracket the true difference in proportions?*

Test and Confidence Interval for Two Proportions

```
Variable          X      N  Sample p
C1 (bigcity)    138    400  0.345000
C2 (smallcity)  129    400  0.322500

Estimate for p(C1) - p(C2):  0.0225
95% CI for p(C1) - p(C2):  (-0.0428339, 0.0878339)
```

Yes, the true difference is .05 and the interval extends from -.04 to .09.

C. *A skeptic claims that the difference you just observed between Bigcity and Smallcity samples could arise by chance if the drug-use rates in the two cities were in fact identical. If one were to assume that this skeptic is correct, what is the probability that one would observe a difference (Bigcity rate minus Smallcity rate) as far away from zero as what you just observed? [Hint: assume that there is one p for both cities and that it is equal to observed rate of drug use in both samples combined.]*

Two-tailed test, because we do not know a priori to expect that Bigcity has a higher rate. I clicked the option to pool the variances (so that the same proportion is assumed for both samples in the calculation of the standard errors):

```
Estimate for p(C1) - p(C2):  0.0225
Test for p(C1) - p(C2) = 0 (vs not = 0):  Z = 0.67  P-Value = 0.500
```

In my sample, therefore, I cannot reject the null hypothesis that the data are drawn from the same population. You may have found a more striking difference between Smallcity and Bigcity.