Statistics 13V
Fall 2002
Final Exam

Name:____________________________________

Last six digits of Student ID#:_________________

Open book and notes. If you need more space for any question you may use the back of the page. You should have 6 questions on 4 pages. Please check to make sure you have all pages.

1. (5 pts each, 20 total). Write a few sentences to answer each of the following questions.
   a. Under what circumstances can the results of a study be extended to a population *and* a cause-and-effect conclusion be made?

To extend results to a larger population, a representative sample must be chosen. To make a cause-and-effect conclusion, a randomized experiment must be done. So, for both, a randomized experiment must be done with a representative sample.

   b. Explain why is it important to consider the size of the sample when interpreting the results of a significance test

If a sample is very large, the results may be statistically significant but have no practical importance. If the sample is too small, a real and important effect may not be statistically significant.

   c. Explain the difference between sampling error and nonsampling error in a survey.

Sampling error can be quantified as the margin of error, and is based on the variability of different random samples. Nonsampling errors cannot be quantified and are the result of how questions are asked, who asks them and so on.

   d. Suppose you wanted to know the mean weight of all players on a rival school's football team. If you knew the individual weights of all of the players, would you need to construct a confidence interval to estimate the mean weight? Explain.

No. You know the weights of everyone in the population, so you can compute the population mean exactly.(A confidence interval is used to make an inference from a sample to a larger population.)

2. (10 pts total) Based on data from nine randomly selected counties near nuclear plants, a regression equation was determined for predicting y = cancer mortality rates (out of 100,000 people) from x = "exposure index." The exposure index was based on factors such as location of the plant and proximity to potentially contaminated water. The results were:

$$\hat{y} = 115 + 9.2x \quad \text{and} \quad r^2 = 85.8\%$$

   a. (3 pts) What is the correlation between exposure index and cancer mortality rate?

$$\sqrt{.858} = .9263$$

   b. (7pts) What is the predicted cancer mortality rate for a county with an exposure index of 3.0? Show your work, then write your answer in a sentence describing how many cancer deaths would be predicted per 100,000 people.

$$\hat{y} = 115 + 9.2x = 115 + 9.2(3) = 142.6$$

With an exposure index of 3.0, it is predicted that about 142.6 out of every 100,000 people will die of cancer.

3. (15 pts total) A waste removal company in one community charges customers $1 extra a week for curbside recycling pickup. They know that 40% (.40) of their customers buy this service. A nearby community includes recycling pickup in the basic cost of waste removal, so there is no way to know how many households use the service. The city manager wants to know if the proportion of households who do so is similar to the .40 found in the community that charges customers. He asks a random sample of 900 households, and 45% say they use the service.

   a. (7 pts) Give the shape, mean and standard deviation of the sampling distribution of the sample proportion for a survey of 900 households if in fact 40% of the customers use this service. Show your work.

The sampling distribution is approximately bell-shaped with mean = p = .4, and standard deviation

$$s.d.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.4(1-.4)}{900}} = .0163.$$

   b. (5 pts) Based on the sampling distribution you found in part (a), assuming the true proportion is .4, how unlikely would a sample proportion of .45 or higher be in this survey of 900 households? Show your work.

A sample proportion of .45 corresponds with a z-score of $\frac{.45-.4}{.0163} = 3.06.$

$$P(\hat{p} \geq .45) = P(z \geq 3.06) = .0011.$$

   c. (3 pts) Based on your answer in part (b), do you think the proportion of customers who use the service is higher in the community where it is included in the basic cost? Explain.

Yes. If the proportion is the same, .40, the sample proportion of .45 or something higher would only occur with probability .0011, so it's more logical that the actual population proportion is higher than .40.

4. (25 pts total) As part of the 2001 Youth Risk Behavior Surveillance System done biannually by the U.S. Government, a random sample of 12th graders was asked how often they wear a seat belt while driving. Of 1302 females, 964 said "most times or always." Of 1443 males, 924 said "most times or always." Test whether there is a difference in the proportions of 12th grade males and females in the population who would answer that they wear seat belts most times or always. Go through the 5 steps of hypothesis testing.
(*Data source:* http://www.cdc.gov/nccdphp/dash/yrbs/)

Step 1 (5 pts): *Specify the hypotheses.*

Define the population proportions to be:
p_1 = proportion of 12th grade females who most times or always wear their seat belts
p_2 = proportion of 12th grade males who most times or always wear their seat belts

Then the hypotheses are:
H_0: $p_1 - p_2 = 0$
H_a: $p_1 - p_2 \neq 0$

Step 2 (10 pts): *Verify necessary conditions and compute the test statistic.*

The conditions require that the sample sizes be large enough, which clearly holds, and that the samples are independent, which also holds based on the fact that random samples were used.

The test statistic requires these pieces:

$$\hat{p}_1 = \frac{964}{1302} = .74, \quad \hat{p}_2 = \frac{924}{1443} = .64, \quad \hat{p} = \frac{964+924}{1302+1443} = \frac{1888}{2745} = .69$$

The test statistic is $z = \dfrac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} = \dfrac{.10}{\sqrt{.69(1-.69)(\frac{1}{1302}+\frac{1}{1443})}} = 5.65$

Step 3 (5 pts): *Find the p-value.*
The test is two-sided, so find the area above 5.65 on a standard normal curve and multiply by 2. From Table A.1, the area above 5.65 is about .00000001 (that's the area above 5.61), so the p-value is about .00000002.

Step 4 (2 pts): *Make a conclusion about statistical significance.*
The p-value is much less than 0.05, so the result is statistically significant. Reject the null hypothesis.

Step 5 (3 pts): *Make a conclusion in context.*
There is a significant difference in the proportions of male and female 12th graders who most times or always wear their seatbelts.

5. (15 pts total) Exercise 2.17 in the textbook gives the ages of the CEOs of the 60 top-ranked small businesses in the US in 1993 (by *Forbes* magazine). Here is a stem-and-leaf plot of the ages, with the number of leaves in each stem counted for you already:

```
Stem Leaves            Number of leaves       Five-number summary
  |3| 23                         2
  |3| 678                        3
  |4| 013344                     6
  |4| 55556677788889            14
  |5| 000000112333             12
  |5| 555666677889             12
  |6| 0111223                    7
  |6| 99                         2
  |7| 04                         2
                              ---
                              60 Total
```

a. (10 pts) In the space to the right of the stem-and-leaf plot and counts, provide a five-number summary of the ages.

The five number summary is 32, 45.5, 50, 57, 74

b. (5 pts) Based on the stem-and-leaf plot and the five-number summary, write a few sentences describing the ages, as if you were writing a short news story about the ages of these CEOs. Someone with no training in statistics should be able to understand it.

The ages of these CEOs range from 32 to 74. The median age is 50, which means that about half of the CEOs are 50 or younger and the other half are 50 or older. There are no extremes at either end of the ages.

6. (15 pts total) In a survey of 75 Penn State students one question asked was "How many minutes do you spend talking on the phone in a typical week?" The mean of the responses was 132.6 minutes and the standard deviation was 140.5 minutes.

a. (5 pts) Explain how you can tell that the data are not bell-shaped.

The standard deviation is very large compared to the mean. One standard deviation below the mean is negative, and negative values are not possible.

b. (10 pts) Assuming these students are equivalent to a random sample of all college students, find a 95% confidence interval for the mean number of minutes college students spend talking on the phone in a typical week.

Sample mean ± multiplier × standard error, where the sample mean is 132.6, the multiplier is from the t-disribution with df = 74, so t=1.99, and the standard error is $\frac{s}{\sqrt{n}} = \frac{140.5}{\sqrt{75}} = 16.22$*. So, the confidence interval is 132.6 ± (1.99)(16.22) or 132.6 ± 32.28 or 100.32 to 164.89.*

NOTE: *It's okay to use 2 as the multiplier because n is large enough that it's very close.*