PubHlth 540 - Introductory Biostatistics
Fall 2008
Examination III

Before you begin:
Again, this is a “take-home” exam. You are welcome to use any reference materials you wish. You are welcome to use the computer as you wish, too. However, you MUST work this exam by yourself and you may not consult with anyone.

Instructions and Checklist:
__1. Start each problem on a new page.
__2. Write your name on every page.
__3. Make a photo-copy of your exam for safekeeping prior to submission
__4. Complete the signature page

How to submit your exam:

<table>
<thead>
<tr>
<th>Online Section</th>
<th>Worcester Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) By mail (address and telephone number below) and post-marked no later than 11:59 PM Monday December 22, 2008</td>
<td>(1) By mail (address and telephone number below) and post-marked no later than 11:59 PM Monday December 22, 2008</td>
</tr>
<tr>
<td>OR</td>
<td>OR</td>
</tr>
<tr>
<td>(2) By upload to EXAM II turn in link no later than 11:59 PM Monday December 22, 2008</td>
<td>(2) Bring your exam to Linda Hollis Monday December 22, 2008</td>
</tr>
</tbody>
</table>

Faxed exams are NOT permitted - sorry                                       Faxed exams are NOT permitted – sorry

Address and telephone number for mailing

Carol Bigelow
School of Public Health/402 Arnold House
University of Massachusetts/Amherst
715 North Pleasant Street
Amherst, MA 01003-9304
Tel. 413-545-1319
Signature

This is to confirm that in completing this exam, I worked independently and did not consult with anyone.

Signature: __________________________________________________________

Printed Name: ______________________________________________________

Date: ___________________________
1. (10 points total)

Two measures of risk for drug use were made on each of 10 students in the 5th grade. The first measure (PRE) was made prior to participation in the DARE program. The second measure (POST) was made upon completion of the DARE program. A lower score for the second measure (the one taken upon completion of the program) indicates a reduced risk of drug use. Following are the data:

<table>
<thead>
<tr>
<th>Student</th>
<th>Risk Score - PRE</th>
<th>Risk Score - POST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>61</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>47</td>
<td>12</td>
</tr>
</tbody>
</table>

You may assume normality.

a. (5 points)

Construct a 95% confidence interval for the change in score.

b. (5 points)

Write a clear interpretation of the confidence interval.
2. (10 points total)

Consider again the data on risk of drug use in question #1. Is there statistically significant evidence that participation in the DARE program reduced risk of drug use? Carry out the appropriate statistical hypothesis test. In developing your answer:

a. (2 points)

What is the null hypothesis?

b. (2 points)

What is the alternative hypothesis?

c. (2 points)

What is the appropriate test statistic and what is its value for these data?

d. (2 points)

What is the p-value?

e. (2 points)

Write a clear interpretation of the results of your test.
3. (10 points total)

An investigator would like to estimate the diameter of Christmas trees in a tree farm to within plus or minus 0.1 cm. Accordingly, a 95% confidence interval is to have total width equal to 0.2 cm. You may assume that the standard deviation of tree diameters is known and is $\sigma = 0.1$ cm. How many trees $n$ should be included in the investigator’s sample?
4. (10 points total)

a. (2 points) ________________________

TRUE or FALSE.
Consider the construction of a 95% confidence interval. Suppose one repeats the sampling process indefinitely. Suppose further that, for each sample drawn, a new 95% confidence interval estimate is obtained. True or False → If for each sample, the investigator states that the population parameter value is contained in the interval, about 95% of these statements will be correct.

b. (2 points) ________________________

TRUE or FALSE.
Consider testing a hypothesis about a population mean parameter for a normal probability distribution. The alternative hypothesis is two sided. Investigators A and B each draw a sample of size n, and compute a sample mean and sample variance. Suppose the sample sizes and means are identical. Suppose further that the sample variance obtained by investigator B is larger. True or False → The achieved level of significance (the p-value) calculated by investigator B will be smaller.

c. (2 points) ________________________

TRUE or FALSE.
Many journals recognize the tendency among researchers to publish disproportionately many trials suggesting the benefit of an experimental treatment in comparison to trials for which the benefit of an experimental treatment is not established. Consider hypothesis tests for which the type I and II errors have equal probability of 0.05. True or False → The phenomenon described above suggests an excess of type II errors in the literature relative to type I errors.
d. (2 points) ________________________

**TRUE or FALSE.**
Consider the estimation of a population parameter. A *non*-probability sample is drawn and used to obtain the required estimate. *True or False ➔* Short of defining the sample as the entire population, there is no way to assess the accuracy of the resulting estimate.


e. (2 points) ________________________

**TRUE or FALSE.**
A one sample t-test is performed to test the null hypothesis that the mean of a normal distribution with unknown variance parameter is equal to a specified value. *True or False ➔* The value of the t-statistic calculated from the data will achieve smaller level of significance if the alternative is two sided than if it is one sided.
5. (10 points total)

The following table summarizes the survival in days of female and male cockroaches, *Blatella vaga* when kept without food or water.

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample size, n</th>
<th>Sample mean</th>
<th>Sample standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>(n_{\text{FEMALES}} = 10)</td>
<td>(\bar{X}_{\text{FEMALES}} = 8.5) days</td>
<td>(S_{\text{FEMALES}} = 3.6) days</td>
</tr>
<tr>
<td>Males</td>
<td>(n_{\text{MALES}} = 10)</td>
<td>(\bar{X}_{\text{MALES}} = 4.8) days</td>
<td>(S_{\text{MALES}} = 0.9) days</td>
</tr>
</tbody>
</table>

Is the apparent greater variability in length of survival that is observed for females, relative to males, statistically significant? Carry out the appropriate statistical hypothesis test. In developing your answer, please provide the following.

a. **(2 points)**

What is the null hypothesis?

b. **(2 points)**

What is the alternative hypothesis?

c. **(2 points)**

What is the appropriate test statistic and what is its value for these data?

d. **(2 points)**

What is the p-value?

e. **(2 points)**

Write a clear interpretation of the results of your test.
6. (10 points total)

a. (2 points)

Consider the construction of a 98% confidence interval of the mean for the setting of a simple random sample from a Normal distribution where the variance parameter $\sigma^2$ is known. What is the value of $z_{1-\alpha/2}$ that is used in this confidence interval construction?

Questions #6b, #6c, #6d, and #6e all pertain to the following:
Suppose next that you are told that (262.09, 374.11) is the result of calculating a 95% confidence interval for the mean of a Normal distribution for the setting of a simple random sample of size n=80 from a normal distribution where the variance parameter $\sigma^2$ is known.

b. (2 points)

What is the point estimate of the population mean $\mu$?

c. (2 points)

What is the value of the standard error of the mean?

d. (2 points)

What is the value of the population variance?

e. (2 points)

Write a clear interpretation of the confidence interval.
7. (10 points total)

A study was investigated of length of hospital stay associated with seat belt use among children hospitalized following motor vehicle crashes. The following are the observed sample mean and sample standard deviations for two groups of children: 290 children who were not wearing a seat belt at the time of the accident plus 123 children who were wearing a seat belt at the time of the accident.

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample size, n</th>
<th>Sample mean</th>
<th>Sample standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seat belt = no</td>
<td>( n_{NO} = 290 )</td>
<td>( \bar{X}_{NO} = 1.39 ) days</td>
<td>( S_{NO} = 3.06 ) days</td>
</tr>
<tr>
<td>Seat belt = yes</td>
<td>( n_{YES} = 123 )</td>
<td>( \bar{X}_{YES} = 0.83 ) days</td>
<td>( S_{YES} = 2.77 ) days</td>
</tr>
</tbody>
</table>

You may assume normality. You may also assume that the unknown variances are equal. Construct a 95% confidence interval estimate of the difference between the two population means. In developing your answer, please provide the following.

(a) (2 points)
What is the value of the point estimate?

(b) (2 points)
What is the value of the estimated standard error of the point estimate?

(c) (2 points)
What is the value of the confidence coefficient?

d. (2 points)
What are values of the lower and upper limits of the confidence interval?

e. (2 points)
Write a clear interpretation of the confidence interval.
8. (10 points total)

Suppose it is known that the distribution of serum cholesterol for healthy adult males is normal with mean $\mu = 200$ mg/dl and standard deviation $\sigma = 16.67$ mg/dl.

(a) (5 points)

An investigator measures the serum cholesterol for a sample of $n=49$ overweight males obtains an average value of $\bar{X}_{n=49} = 225$ mg/dl. Construct a 95% confidence interval estimate of the mean cholesterol in the population of overweight males. You may assume that the distribution of serum cholesterol for overweight males is also normal and that its standard deviation is known and equal to 16.67 mg/dl.

(b) (5 points)

What sample size $n$ of overweight males would you need to have 95% confidence that the estimate of 225 is within 10 mg/dl of the true mean for overweight males? In developing your answer, your total confidence interval width is therefore $(2)(10) = 20$ mg/dl.
9. (10 points total)

(a) (2 points)
Write a clear interpretation of the meaning of a p-value = 0.032

(b) (2 points)
What does type II error tell you?

c) (3 points) ______________________________

Multiple Choice: Consider the setting of statistical hypothesis testing. If, in a given problem, you decide to decrease type I error setting from $\alpha = 0.05$ to $\alpha = 0.01$, what would happen to the type II error? Increase? Decrease? Stay the same?

d) (3 points)
What is the basis for being able to use confidence intervals to perform a statistical hypothesis test?
10. (10 points total)

For each of the following single sample hypothesis test situations, what is the critical region? In developing your answer you will be providing either a range of values of a z-score that will prompt rejection of the null hypothesis or a range of values of a t-score that will prompt rejection of the null hypothesis.

(a) (2 points) Critical Region is: _________________________________

\[ H_0: \mu = 220 \]
\[ H_A: \mu \neq 220 \]
\[ \alpha = 0.05, n = 20, \sigma \text{ is known} \]

(b) (2 points) Critical Region is: _________________________________

\[ H_0: \mu \leq 15 \]
\[ H_A: \mu > 15 \]
\[ \alpha = 0.01, n = 35, \sigma \text{ is known} \]

(c) (2 points) Critical Region is: _________________________________

\[ H_0: \mu = 120 \]
\[ H_A: \mu \neq 120 \]
\[ \alpha = 0.05, n = 25, \sigma \text{ is unknown} \]

(d) (2 points) Critical Region is: _________________________________

\[ H_0: \mu \geq 100 \]
\[ H_A: \mu < 100 \]
\[ \alpha = 0.01, n = 16, \sigma \text{ is unknown} \]

(e) (2 points) Critical Region is: _________________________________

\[ H_0: \mu \leq 55 \]
\[ H_A: \mu > 55 \]
\[ \alpha = 0.05, n = 857, \sigma \text{ is unknown} \]
optional EXTRA CREDIT I (10 points total)
Points earned here will be added to your exam I score.

(a) (2 points) ____________________________

**Multiple Choice.** A cumulative distribution
(a) expresses the same information as a probability or density function but in a different way.
(b) states the probability that a random variable is less than or equal to a particular value.
(c) always takes on a value between 0 and 1.
(d) all of the above.
(e) none of the above.

(b) (2 points) ____________________________

**Multiple Choice.**
A large study of bladder cancer and smoking produced the following findings

<table>
<thead>
<tr>
<th></th>
<th>Incidence of bladder cancer per 100,000 males/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cigarette smokers</td>
<td>48.0</td>
</tr>
<tr>
<td>NON smokers</td>
<td>25.4</td>
</tr>
</tbody>
</table>

The estimated relative risk of developing bladder cancer for male cigarette smokers compared to male non smokers is

(a) $48.0/25.4 = 1.89$.
(b) unknown.
(c) $48.0 - 25.4 = 22.6$
(d) $48.0$.
(e) $(48.0 - 25.4)/48.0 = 0.47$.

(c) (3 points) ____________________________

**Multiple Choice:** A data set consists of $n=20$ values that are fairly close to each other in value. Another value is added to the data set but this new value is an outlier. How is the standard deviation affected by its inclusion?

(a) No effect.
(b) New standard deviation is smaller.
(c) New standard deviation is larger
(d) Unknown.
(d) **(3 points)**

Both parents of a child have the brown/blue pair of eye color genes and each parent contributes one gene
to a child. Assume that if the child has at least one brown gene, that color will dominate and the eyes
will be brown. (As we now know, the actual determination of eye color is a bit more complicated).
What is the probability that the child will have brown eyes?
optional EXTRA CREDIT II (10 points total)
Points earned here will be added to your exam II score.

(a) (2 points) ____________________________
A test consists of multiple choice questions, each having four possible answers, one of which is correct. What is the probability of getting exactly four correct answers when six guesses are made?

(b) (3 points) ____________________________
After being rejected for employment, woman “A” learns that company “X” has hired only 2 women among the last 20 new employees. She also learns that the pool of applicants is very large, with an approximately equal number of qualified men and women. Help her address the charge of gender discrimination by finding the probability of getting 2 or fewer women when 20 people hired under the assumption that there is no discrimination based on gender. Does the resulting probability really support such a charge?

(c) (2 points) ____________________________
Suppose the length of newborn infants is distributed normal with mean 52.5 cm and standard deviation 4.5 cm. What is the probability that the mean of a sample of size 15 is greater than 56 cm?

(d) (3 points)
Suppose that 25 year old males have a remaining life expectancy of an additional 55 years with a standard deviation of 6 years. Suppose further that this distribution of additional years life is normal. What proportion of 25 year old males will live past 65 years of age?