Applied Statistics

MS Comprehensive Examination

April 17, 2010

Instructions:

Please answer all questions. Each question is worth 20 points.

Record your answers in your blue books.

You may use your books, notes and calculator for this exam.

You have three hours.
1. Suppose that $n$ measurements are to be taken under a treatment condition ($X$) and $2n$ measurements are to be taken independently under a control condition ($Y$). It is thought that the standard deviation of a single observation is about 10 under both conditions.
   a. How large should $n$ be so that a 95% confidence interval for $\mu_X - \mu_Y$ has a width of 4?
   b. How large should $n$ be so that the test of $H_0: \mu_X = \mu_Y$ against the one-sided alternative $H_a: \mu_X > \mu_Y$ has a power of 0.5 if $\mu_X - \mu_Y = 2$ and $\alpha = 0.05$?
   c. Suppose a 95% confidence interval for $\mu_X - \mu_Y$ is (-1,3). Explain the meaning of the interval in such a way that someone who had not ever studied statistics would understand.

2. We have 92 measurements of weight on a random sample of college students. We want to test whether or not Weight is normally distributed.
   a. Before even testing this, do you expect Weight to be normally distributed? Why or why not?
   b. Do a chi-square test for goodness of fit with estimated expected cell counts using the summary statistics given below. I recommend using 6, 8 or 10 equally likely cells. Use level of significance $\alpha = .05$. Give all of the key elements of the test, such as the hypotheses, the test statistic and its distribution under the null hypothesis (including the degrees of freedom), the critical region, bounds on the P-value, the decision, and an interpretation of the result.

   **Descriptive Statistics: Weight**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>92</td>
<td>145.15</td>
<td>2.48</td>
<td>23.74</td>
<td>95.00</td>
<td>125.00</td>
<td>145.00</td>
<td>156.50</td>
<td>215.00</td>
</tr>
</tbody>
</table>

   Weight Data (ordered):

   95 102 108 108 110 110 112 115 115 116 116 116 118 118 120 120 120
   121 122 123 125 125 125 125 125 130 130 130 130 130 131 133 135
   135 135 136 138 138 140 140 140 142 145 145 145 145 145 145 145
   150 150 150 150 150 150 150 150 150 153 155 155 155 155 155 155
   155 155 155 155 155 157 160 160 160 160 164 165 170 170 170 170
   175 175 180 180 180 185 190 190 190 190 195 215

   3. Let $(x_i,Y_i)$ be observed for $i = 1, \ldots, n$. Suppose that the $x_i$ are constants, and that $Y_i = \frac{\beta}{x_i} + \epsilon_i$, where $\beta$ is an unknown parameter, and the $\epsilon_i$ are independent, each with the N(0,$\sigma^2$) distribution.
   a. Show that the least squares estimator of $\beta$ is $\hat{\beta} = (\sum_{i=1}^{n} Y_i x_i^{-1})/(\sum_{i=1}^{n} x_i^{-2})$. Do NOT use matrix or vector space methods.
   b. Show that $\hat{\beta}$ is an unbiased estimator of $\beta$.
   c. Find $\text{Var}(\hat{\beta})$.
   d. For the following $(x_i,Y_i)$ pairs find a 95% confidence interval on $\beta$: (1,5), (2,-3), (2,-4), (2/3,1).
4. Here we will explore stratified sampling. Say that we have a workplace where the owner is considering a complete ban on smoking (indoors and out). Part of making a good decision, of course, is to survey the employees and learn what they desire. Two choices for sampling present themselves: simple random sampling or stratified sampling, both without replacement. The strata are the different job classifications; for simplicity call them A, B and C. At UT, they could be administrators, teachers and support staff. At the company say that we have the following situation:

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number (known)</th>
<th>Number in Support of a Complete Ban (not known)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>B</td>
<td>30</td>
<td>16</td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>16</td>
</tr>
</tbody>
</table>

The goal is to estimate the proportion of the entire workforce in favor of the ban. Time and energy are available to sample 40% of the population.

a. What is the actual proportion in favor of the ban?
b. Using simple random sampling, assume that we found 22 of the 40 in favor of the ban. Give the estimate, its standard error, and approximate 95% confidence interval for the proportion in favor of the ban. Give the appropriate formulas and the numerical answers. Is the true value of the parameter inside the interval?
c. Using stratified sampling where (non-optimally) 40% of each stratum is sampled, assume that we found 8, 9 and 5 in favor of the ban from classes A, B and C, respectively. Give the estimate, its standard error, and an approximate 95% confidence interval for the proportion in favor of the ban. Give the appropriate formulas and the numerical answers. Is the true value of the parameter inside the interval?

5. In an early study of the effects of a strong magnetic field on the development of mice, 7 cages, each containing 3 albino female mice were subjected for a period of ten days to a magnetic field. 21 other mice housed in 7 similar cages were not placed in a magnetic field and served as controls. The following table shows the weight gains, in grams, for each of the cages.

<table>
<thead>
<tr>
<th>Magnetic Field Present:</th>
<th>22.8</th>
<th>10.2</th>
<th>20.8</th>
<th>27.0</th>
<th>19.2</th>
<th>10.4</th>
<th>14.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic Field Absent:</td>
<td>23.6</td>
<td>31.0</td>
<td>19.5</td>
<td>26.2</td>
<td>26.5</td>
<td>25.2</td>
<td>24.5</td>
</tr>
</tbody>
</table>

a. State a nonparametric model, define hypotheses, and carry out a test at level $\alpha = 0.10$ which will enable you to decide whether there is a significant difference in weight gain between these two groups. Find the exact $p$-value and use it to make your decision.
b. Repeat part a using the normal approximation. Do you get the same conclusion from (a)?