Name:

Math 130 Section 01 – Introduction to Statistics - Final Exam
Thursday, May 15th, 9 am - 12 pm in BEBU 107 (Paino)

Instructions:

1. Show all work. You may receive partial credit for partially completed problems.

2. You may use calculators and a two-sided sheet of reference notes (two one-sided pages are fine), as well as the provided normal and t tables. You may not use any other references or any texts. You may NOT use your cell phone as your calculator.

3. You may not discuss the exam with anyone but Prof. Wagaman. Uphold the honor code. Cell phones and mobile devices must be turned off and put away in your backpacks.

4. Suggestion: Skim/read all questions before beginning. Complete the ones you know best first. Point values per problem are displayed below. More space than needed is provided.

5. Carry calculations out to four decimals, unless otherwise specified. I.E. 8/10 would be .8 but 2/3 would be .6667. Expected count computations only need to be out to two decimals.

6. If you cannot solve an earlier part of a problem, and a later part depends on that part, use a reasonable value and state you are assuming that is the right value to use. This allows you to still get credit for later parts if you get stuck on something early on.

7. BE SPECIFIC when stating and checking conditions. Saying “nearly normal condition” by itself does not tell me what you think must be nearly normal. A check mark does not show me you know how to check the conditions. Demonstrate that you know what to check.

8. Interpretations/conclusions/any written statement must be in context of the problem.

9. Note: If you give multiple answers, the worst one will be graded. For example, if a question asks you to pick a specific condition out of the relevant conditions for a procedure, you cannot list ALL conditions and expect to get credit.

10. Initial in the box to acknowledge that you have read and understand the instructions, especially in regard to #8 and the honor code in #2.

11. Good luck and best wishes for your summers.

<table>
<thead>
<tr>
<th>Problem</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points Earned</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
<td>_______</td>
</tr>
<tr>
<td>Possible Points</td>
<td>5</td>
<td>13</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>16</td>
<td>11</td>
<td>100</td>
</tr>
</tbody>
</table>
1. A food processing plant receives broccoli from three different growers, labeled 1, 2, and 3. In a recent shipment from each grower, the plant foreman randomly selected 4 boxes from each grower and took 3 clusters of broccoli out of each box. The plant mandates that each cluster should weigh at least 350 units to be of use. We want to examine the broccoli clusters for differences in weight between the three growers. The boxplots show the distribution of cluster weight by grower.

a. The ______________________ for weight for grower 3 is a little bit larger than the maximum observed broccoli cluster weight for grower 1.

b. As described here, an ANOVA to compare average cluster weights between growers would be balanced unbalanced. (Circle one)

Summary statistics by grower:

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>326</td>
<td>338.75</td>
<td>351.0</td>
<td>360.50</td>
<td>378</td>
<td>352.1667</td>
<td>15.93072</td>
</tr>
<tr>
<td>2</td>
<td>340</td>
<td>360.50</td>
<td>366.5</td>
<td>371.25</td>
<td>374</td>
<td>363.5833</td>
<td>10.30850</td>
</tr>
<tr>
<td>3</td>
<td>328</td>
<td>343.75</td>
<td>357.5</td>
<td>378.50</td>
<td>388</td>
<td>358.7500</td>
<td>21.41421</td>
</tr>
</tbody>
</table>

c. The three broccoli cluster populations are independent and the clusters were randomly selected in a multistage sampling process. Two conditions for inference remain. State the remaining condition that does NOT deal with normality and determine whether or not it checks out based on the information provided, explaining your rationale.

The relevant condition is that ...

This condition does does not appear to check out, because...
The 1996 American National Election Study was taken before the presidential election between Clinton and Dole. Respondents self-reported their own political “status” – liberal, moderate, or conservative, which was given a numerical code, and stated whom they intended to vote for in the presidential election. Some results from the survey are shown in the table.

<table>
<thead>
<tr>
<th>Status</th>
<th>Liberal</th>
<th>Moderate</th>
<th>Conservative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P(Code)</td>
<td>.2818</td>
<td>.2712</td>
<td></td>
</tr>
</tbody>
</table>

a. Fill in the appropriate value in the table to create a probability distribution for political Code.

b. What is the expected value of political Code?

c. Which of the following is the most likely value of the standard deviation of political Code? (Circle one). Explain your choice.

-0.45  0.84  2.61 is the most likely value of the standard deviation of political Code because...

d. Assume the values from the National Election Study are representative of the American populace in 1996 for the rest of the problem. If five Americans were randomly selected in 1996, find the probability that exactly 3 out of 5 would self-report their political status as “conservative”.

e. Now suppose a research group in 1996 wants to randomly select 100 Americans and have at least 35 self-report their political status as “liberal”. Find the approximate probability that in a random sample of 100, at least 35 individuals self-report as “liberal”. Be sure to justify your approximation.
2. (continued).

f. Based on the survey, we also know if individuals planned to vote for Clinton or Dole in the 1996 election. Of the self-reported liberals in 1996, 94% reported they would vote for Clinton. Of the self-reported moderates in 1996, 71% reported they would vote for Clinton. Of the self-reported conservatives in 1996, 28% reported they would vote for Clinton. If we randomly choose someone who reported they would vote for Dole, what is the probability that person is NOT a conservative?

3. As part of a survey for an introductory statistics class, all 592 students in the course at a large university reported their natural hair color and eye color restricted to the following values:

Hair color: black, brown, red, and blond;
Eye color: green, hazel, blue, and brown.

A table of their results is shown. Use the table to help address the questions that follow.

<table>
<thead>
<tr>
<th>Eye Color</th>
<th>Black</th>
<th>Brown</th>
<th>Red</th>
<th>Blond</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>5</td>
<td>29</td>
<td>14</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>Hazel</td>
<td>15</td>
<td>54</td>
<td>14</td>
<td>10</td>
<td>93</td>
</tr>
<tr>
<td>Blue</td>
<td>20</td>
<td>84</td>
<td>17</td>
<td>94</td>
<td>215</td>
</tr>
<tr>
<td>Brown</td>
<td>68</td>
<td>119</td>
<td>26</td>
<td>7</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td>108</td>
<td>286</td>
<td>71</td>
<td>127</td>
<td>592</td>
</tr>
</tbody>
</table>

a. If you randomly selected a student from this class, what is the probability he or she would have naturally blond hair?

b. If you randomly selected a student with naturally black hair from this class, what is the probability that student has blue eyes?
3. (continued).
c. If you want to know whether or not there is a relationship between eye color and natural hair color, based on the data collected, what is the appropriate analysis to perform? Be specific.

d. If we have no intent to generalize, we do not need to check conditions relating to sampling procedures. However, there is still a condition related to obtaining the distribution of the test statistic that you need to check. State that condition, and state if it checks out by computing as support ONLY the minimum expected count. (Hint: Think about which cell will have the minimum. Do not waste time computing them all.)

The condition is that ______________________________________________________________.

The minimum expected count is _______________, which allows us to state that the condition is is not met. (Circle one).

Rcmdr reports the following line of output for performing the test from c:

\textit{Pearson's Chi-squared test: } \chi^2 = 138.2898, \textit{ p-value < 2.2e-16}

e. At right, sketch and label the distribution used to find the \( \text{p-value} \), being sure to show where your test statistic and \( \text{p-value} \) are. Your picture need not be to scale, but should convey an understanding of the relevant distribution.

f. The entire class answered this survey. Describe the sampling method used here (with a statistical term) and explain possible issues (if any) that might result from using such a method in general.
4. Many Amherst students make use of the bike path, officially called the Norwottuck Rail-Trail. The Pioneer Valley Planning Commission collected data along the trail north of Chestnut Street in Florence, MA for ninety days from April 5, 2005 to November 15, 2005. One variable recorded was the volume of traffic that day along the trail. Suppose we classify a day as “high volume” if the volume is higher than 400 trail users.

a. For the 90 days for which we have data, 37 days have volume over 400. Assuming the volume each day is independent of other days, and that our 90 days are a representative sample of days, check any remaining conditions and find a 90% confidence interval for the fraction of days that have “high volume” along this stretch of the Rail Trail, if appropriate.

b. Interpret the 90% confidence level.

c. Does the confidence interval provide evidence, at a significance level of 0.05, that more than 30% of days along this stretch of the Rail Trail have high volume? State an appropriate conclusion and explain your reasoning.
5. Short Answer and True/False. It is possible to answer the short answer questions in 1-3 sentences.

a. Describe briefly why randomization is a required principle in experimental design.

b. Describe briefly how blocking and stratified sampling are different from one another.

Assess whether the following statements are True or False.

c. When individuals randomly chosen to complete a survey fail to complete it, the researcher has encountered response bias.

   True       False

d. The most important property of a simple random sample is that every individual in the population has an equal chance at being selected for the sample.

   True       False

e. For a confidence interval for a population proportion, for a specified confidence level, in order to cut your margin of error in half, you need to collect four times as many observations as your original sample.

   True       False

f. The command `pchisq(14,3,lower.tail=FALSE)` would provide the p-value for a chi-square test with a test statistic of 3 and degrees of freedom of 14.

   True       False
6. The Current Population Survey (CPS) is used to supplement census information between census years. We have access to a random sample of persons from the CPS in 1985, with information on wages, sex (M/F), number of years of education, occupational work sector (8 job sectors), race (White vs. Non-white), marriage status (Single vs. Married) and union membership (Union vs. Not) of the surveyed workers.

For each research question below, determine the appropriate analysis to pursue from among our inferential procedures this semester. If you choose hypothesis test or CI, you must specify the parameter of interest. Similarly, if you choose a chi-square test, you must specify which test.

a. Research question: Suppose number of years of education is transformed to be maximum level of education: No High School Degree, High School Degree, Some College, College Degree, or Advanced Degree, and we want to determine how average wages compare across these education designations for 1985.

b. Research question: In 1985, did a higher percentage of white workers belong to a union compared to non-white workers?

c. Research question: Would the work force in 1985 be reasonably modeled by assuming it was 20% single whites, 50% married whites, 8% single non-whites, and 22% married non-whites?

d. Research question: How much more money per hour did married workers make compared to single workers in 1985?

e. Research question: Is there a significant linear relationship between number of years of work experience and wage in 1985?
7. Investigators wanted to explore the relationship between SAT performance (an achievement test) and first year college GPA. They recorded SAT Verbal and Math scores for 219 students entering college, and waited for their first years in college to end to study their GPA. In the meantime, perhaps you can answer a related question. Do students tend to have higher scores on the Math component of the SAT relative to the Verbal component? Both tests are scored from 0-800.

a. Clearly state the hypotheses you should test, and define your parameter(s).

Assume all conditions necessary for your procedure check out. The following output was generated by someone who was not sure which procedure to run. Use the appropriate output to address the questions that follow.

```
t.test(GPA$SATM,GPA$SATV)             Welch Two Sample t-test
data:  GPA$SATM and GPA$SATV
t = 3.8505, df = 431.46, p-value = 0.0001357
alternative hypothesis: true difference in means is not equal to 0
sample estimates:
mean of SATM   mean of SATV
634.2922       605.0685
```

```
t.test(GPA$SATdiff)                        Note: SATdiff=SATM-SATV
One Sample t-test                       data:  GPA$SATdiff
t = 5.582, df = 218, p-value = 7.026e-08
alternative hypothesis: true mean is not equal to 0
sample estimates:
mean of SATdiff
29.22374
```

b. Clearly state the numeric value of the test statistic for your procedure in a. Then interpret the test statistic.

c. Clearly state the numeric value of the p-value for your procedure in a. Then interpret the p-value.
A credit union serving a large university is studying the distribution of account balances in checking accounts for users who are not using the interest-earning version of the account. A histogram and some descriptive statistics are provided of a random sample of 1000 such balances. Use the output to address the questions that follow about the variable: balance.

<table>
<thead>
<tr>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.27</td>
<td>86.5</td>
<td>215.55</td>
<td>421.45</td>
<td>2518.01</td>
<td>300.58</td>
</tr>
</tbody>
</table>

a. An intern examining the data says the range for balance is $0.27 to $2518.01, based on this sample. Is the intern correct? If not, provide a correction.

b. Describe the observed distribution of balance based on the data provided.

c. Provide two shape descriptors (which are not synonyms of each other!) that you would feel comfortable applying to the population distribution of balance.

d. If you repeatedly took random samples of 1000 balances from the credit union as described, computed the sample average balance and stored it, what shape would you expect the distribution of sample average balances to have?

e. What result allows you to provide your answer for part d.?

f. An executive for the credit union has information that, nationally, similar checking accounts have a mean balance of $275 and a standard deviation of $300. Is your random sample of 1000 balances from this credit union consistent with the national numbers? Explain and provide appropriate numeric support for your response.
9. Scholars can learn a lot about a cathedral by studying its construction. The data set ‘cathedral’ contains cathedral lengths and nave heights from England, measured in feet. Building a tall nave could be a challenge depending on when the construction took place. Suppose we want to investigate **whether or not the height of the nave can provide useful information about the length of the corresponding cathedral**. Consider ALL of the output provided before addressing the questions that follow. Not all provided output may be relevant.

**Model 1**: Call:`lm(formula = cathlength ~ naveheight, data = cathedral)`

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 76.420   | 89.258     | 0.856   | 0.400739 |
| naveheight     | 4.669    | 1.172      | 3.985   | 0.000584 *** |

Residual standard error: 85.83 on 23 degrees of freedom
Multiple R-squared: 0.4084,  Adjusted R-squared: 0.3827
F-statistic: 15.88 on 1 and 23 DF,  p-value: 0.0005838

**Model 2**: Call:`lm(formula = naveheight ~ cathlength, data = cathedral)`

Coefficients:

|                | Estimate  | Std. Error  | t value | Pr(>|t|) |
|----------------|-----------|-------------|---------|----------|
| (Intercept)    | 37.53933  | 9.63107     | 3.898   | 0.000725 *** |
| cathlength     | 0.08748   | 0.02195     | 3.985   | 0.000584 *** |

Residual standard error: 11.75 on 23 degrees of freedom
Multiple R-squared: 0.4084,  Adjusted R-squared: 0.3827
F-statistic: 15.88 on 1 and 23 DF,  p-value: 0.0005838

`favstats(~naveheight,data=cathedral)`

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
</table>
|favstats(~cathlength,data=cathedral)`

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>Q1</th>
<th>median</th>
<th>Q3</th>
<th>max</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
</table>

This final plot provided is from the “correct” model, and is referenced in part g of the problem.
9a. Recall that we want to investigate whether or not the height of the nave can provide useful information about the length of the corresponding cathedral. Identify which variable should be the response and which should be the predictor.

Response: __________________________
Predictor: ____________________________

d. A researcher has found a cathedral with a cathedral length of 500 feet and wants an estimate of its nave height. Using your regression line from b, provide an estimate or explain why such an estimate is not reasonable to provide.

e. Explain how accurate you believe predictions using your regression line will be. Use at least two numeric values as support for your response.

I believe that predictions using this regression will be

spot on    very accurate    moderately accurate    not very accurate    terribly inaccurate

because ...
Based on previous studies in other European countries, the researchers believe the slope for predicting cathedral length from nave height should be roughly 4. Is there evidence to support that their slope is different than expected at a 0.10 significance level? Perform an appropriate procedure, assuming that appropriate conditions hold (do not need to list), and complete the conclusion.

**g.** The final plot provided above is a _____________ plot of the residuals which is used to check that

______________________________________________________________________________

This particular plot shows that the condition appears to check out with the exception of a few points in the lower tail, including the cathedrals at Bath and Ripon. These points would be called

________________________________________, and should be dealt with by...
10. As part of an agricultural study in Wisconsin, the amount of nitrogen fertilizer (pounds per acre) and corn yield (bushels per acre) were recorded for 44 fields. The fields were divided into those that had a high amount of fertilizer applied (over 50 pounds per acre) and those with a low amount. For the fields with a high amount of fertilizer applied, the mean yield was 139.69 bushels per acre with a standard deviation of 11.61. For the fields with a low amount of fertilizer applied, the mean yield was 105.67 bushels per acre with a standard deviation of 28.32.

a. A researcher is focused on two fields – one with a high amount of fertilizer with a yield of 162 bushels, and one with a low amount of fertilizer with a yield of 136 bushels. Which field is more unusual compared to its fertilizer group? Provide appropriate numeric support for your answer.

b. It turns out that the corn yield for the low fertilizer group is approximately normally distributed, while the corn yield for the high fertilizer group is fairly right-skewed. Taking the values provided from the study as the population values rather than sample values, estimate the probability that for the low fertilizer group, a field will have a yield of 75 bushels of corn or less per acre.

c. Taking the values provided from the study as the population values rather than sample values, either estimate the probability that for the high fertilizer group, a field will have a yield of 120 bushels of corn or less per acre or explain why that probability cannot be calculated with the information at hand.

d. What numeric value is the upper 12% cutoff of corn yields in bushels for the low fertilizer group?

e. Estimate the IQR of the corn yield in bushels for the low fertilizer group. Provide appropriate support for your answer via calculations.