Lab 4: Paired t-test, Categorical Inference (10 pts. + 1 pts. Bonus)

Objectives
Part 1: Confidence Intervals/Hypothesis tests
  1.1) Paired Design
Part 2: Categorical Data
  2.1) Chi-Square Distribution
  2.2) Goodness of Fit Test
  2.3) Test of Independence

Remember:
a) Please put your name, STAT 503 and the lab # on the front of the lab
b) Label each part and put them in logical order.
c) ALWAYS include your SAS code for each problem.

1. Confidence Intervals/Hypothesis tests:

1.1 Paired - Design
For paired data, we will again use “proc ttest” like we did in Lab 3; however, we just use a “paired” command for both variables in place of the “var” command for a single variable. The difference will be the first variable in the paired command minus the second variable. In addition, we now have a separate column for each of the two variables instead of using a grouping variable. The file d1.dat was taken from Example 8.2.4 in the book for the following learning code. In addition, this example uses a directional alternative hypothesis. When you are performing directional hypothesis, be sure to know which variable is which so the direction is appropriate. In the example in the book (which was nondirectional), I am choosing the direction of the drug decreases how hungry the women are.

SAS Learning code: (d1.sas)

```sas
data hunger;
infile 'H:\d1.dat';
input subject drug placebo;
run;

proc print data=hunger; run;

Title 'Two Sample Paired Inference';
proc ttest data=hunger alpha=0.01 SIDE=L;
paired drug*placebo;
/*generates the hypothesis test with alpha = 0.01 (99% CI) for drug - placebo
  with Ha = mu1 < mu2 (SIDE = L). If SIDE=U, it would be Ha: mu1 > mu2*/
run;
```
SAS Learning output:

The TTEST Procedure

```
N    Mean    Std Dev     Std Err     Minimum     Maximum
9    -29.5556    32.8219    10.9406    -90.0000     8.0000

Mean  99% CL Mean    Std Dev    99% CL Std Dev
-29.5556   -Infty    2.1336    32.8219    19.8127    80.0650

DF   t Value    Pr < t
8    -2.70    0.0135
```
SAS Learning code: (d2.sas)

data chisquared;
*chis is the calculated test statistic, df is the degrees of freedom for
  the test, alpha is the significance level, for the critical value, we want
  the percentile to be alpha;
chis = 43.2;
df = 3;
Pvalue = 1 - PROBCHI(chis,df);
alpha=0.05;
CritVal = QUANTILE('CHISQ',alpha,df);

proc print data=chisquared; run;
quit;

SAS Learning output:

<table>
<thead>
<tr>
<th>Obs</th>
<th>chis</th>
<th>df</th>
<th>Pvalue</th>
<th>alpha</th>
<th>CritVal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.2</td>
<td>3</td>
<td>2.2318E-9</td>
<td>0.05</td>
<td>0.35185</td>
</tr>
</tbody>
</table>

P-value = 2.2318 x 10^-9, $\chi^2$ = 0.35185

Problem 2 (1 pt.)
Calculate the following which refers to Example 9.4.6 (p.353).
  a) The P-value ($\chi^2_s = 7.71$, df = 5).
  b) The critical value for $\alpha = 0.005$.

Your submission should consist of the answers to parts a and b, your code (clearly indicating
which part is coming from which line) and the appropriate parts of the output file.

2.2. Goodness of Fit Test
The Goodness-of-Fit test uses proc freq. The difficulty is the output does not include the
expected value ($e_i$) only the percent. Therefore, to convert this back to what we use in class,
you will have to manually compute the $e_i$'s for each of the observations. However, the program
does calculate $\chi^2_s$ and the P-value for you.

The learning code is based on Example 9.4.6 (p. 353). SAS does not like 0's, it considers that
missing data, so I inputted a value of 0.001 for the 0 for Variegated Low.
**SAS Learning code:** (d3.sas)

```sas
data flaxseed;
  infile 'H:\d3.dat';
  input coloracid $ count;
  *I converted the two inputs Color and Acid Level into one variable;
run;

proc print data=flaxseed; run;

Title 'Goodness of Fit';
proc freq data=flaxseed order=data;
tables coloracid/nocum chisq
  testp = (0.1875,0.375,0.1875,0.0625,0.125,0.0625);
  *the qualitative variable is listed in the Table statement;
  *nocum: prevents printout of the cumulative percentages;
  *testp: the percentages for the expected, these need to be calculated out;
  weight count; *the number of observations or 'count' is listed as the weight;
run;
quit;
```

**SAS Learning output:**

```
The FREQ Procedure

<table>
<thead>
<tr>
<th>coloracid</th>
<th>Frequency</th>
<th>Percent</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BrownLow</td>
<td>15</td>
<td>20.83</td>
<td>18.75</td>
</tr>
<tr>
<td>BrownInt</td>
<td>26</td>
<td>36.11</td>
<td>37.50</td>
</tr>
<tr>
<td>BrownHig</td>
<td>15</td>
<td>20.83</td>
<td>18.75</td>
</tr>
<tr>
<td>VarLow</td>
<td>0.001</td>
<td>0.00</td>
<td>6.25</td>
</tr>
<tr>
<td>VarInter</td>
<td>8</td>
<td>11.11</td>
<td>12.50</td>
</tr>
<tr>
<td>VarHigh</td>
<td>8</td>
<td>11.11</td>
<td>6.25</td>
</tr>
</tbody>
</table>

Chi-Square Test for Specified Proportions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>7.7016</td>
</tr>
<tr>
<td>DF</td>
<td>5</td>
</tr>
<tr>
<td>Pr &gt; ChiSq</td>
<td>0.1735</td>
</tr>
</tbody>
</table>

WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.
```

The information that is required from this output is: test statistic is in red, the df is in green and the P-value is in blue.

If you look in the book, the expected values of 2/6 = 1/3 of the entries are less than 5. This is what the warning is for.
Problem 3 (3 pts.)
Suppose you do not feel comfortable with SAS’s random number generator and decide to
develop your own discrete integer algorithm. This algorithm is to generate the digits 0-9 with
equal probability. After writing the macro, you decide to test it out by generating 1000 digits
and checking to see if there is any evidence that the algorithm is not performing properly.
That number of times that you generated each digit is in rannum.dat. The infile contains the
digit in the first column and the number of times that it appears in the second. Perform the
goodness of fit test for this data. Remember, you need to decide what percentage to use for
each of the digits.
Your submission should consist of the code, the relevant parts of the output, the value of the
test statistic, df and the P-value. In addition, explain why you chose the expected
percentages that you did. Does your macro work (that is, are each of the digits from 0 – 9
random?)? Why or why not? Hint: What is H₀?

2.3. Test of Independence:
The test of independence (2 x 2 contingency table) also uses the proc freq. The example I am
using is Migraine Headaches in section 10.2 (p. 365 – 369) except that that the Hₐ is
nondirectional instead of directional.

SAS Learning code: (d4.sas)
\begin{verbatim}
data migraine;
infile 'H:\d4.dat';
input success $ sugery $ count;
run;

proc print data=migraine; run;

Title 'Test of Independence';
proc freq data=migraine order=data;
tables success*surgery/nocum chisq expected nopercent;
*expected: provides the expected
* nopercent: doesn't print out the percentages, just the frequencies;
weight count;
run;

quit;
\end{verbatim}
SAS Learning output:

The FREQ Procedure

Table of success by surgery

<table>
<thead>
<tr>
<th>success</th>
<th>sugery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Expected</td>
</tr>
<tr>
<td>success</td>
<td>real</td>
</tr>
<tr>
<td>no</td>
<td>sham</td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
</tr>
</tbody>
</table>

Statistics for Table of success by surgery

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>6.0619</td>
<td>0.0138</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>5.8549</td>
<td>0.0155</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>4.7661</td>
<td>0.0290</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>5.9810</td>
<td>0.0145</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td>0.2843</td>
<td></td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td>0.2735</td>
<td></td>
</tr>
<tr>
<td>Cramer's V</td>
<td></td>
<td>0.2843</td>
<td></td>
</tr>
</tbody>
</table>

Fisher's Exact Test

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (1,1) Frequency (F)</td>
<td>41</td>
<td>0.9965</td>
<td>0.0156</td>
</tr>
<tr>
<td>Left-sided Pr &lt;= F</td>
<td>0.0121</td>
<td>0.0121</td>
<td>0.0241</td>
</tr>
<tr>
<td>Right-sided Pr &gt;= F</td>
<td></td>
<td>0.0121</td>
<td>0.0241</td>
</tr>
<tr>
<td>Table Probability (P)</td>
<td>0.0121</td>
<td>0.0121</td>
<td>0.0241</td>
</tr>
<tr>
<td>Sample Size = 75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: We are not covering the Fisher's Exact Test in the class though it is discussed in section 10.4

The contingency table that you would report in your homework is in red. The test statistic, df and P-value are in blue. We will not be using anything else in this output.
Problem 4 (3 pts.)
The file plover.dat includes the data from example 10.5.1 (p.385). The first column is the location, the second is the year and then the appropriate counts. Perform a chi-squared test of independence to determine whether nesting location is independent of year. Your submission should consist of the code and relevant parts of the output, the test statistic and the P-value. In addition, please state whether nest location is independent of year. Why or why not? Hint: What is $H_0$?

Bonus B1 (1 pt.):
Repeat Problem 4 if the count of Agricultural Field in 2004 was 28 instead of 21. Your submission should consist of the infile, code and relevant parts of the output, the test statistic and the P-value. In addition, please state whether nest location is independent of year. change problem so that it is a fail to reject.

Length good!