STAT 503

class: July 20, 2012
due: July 24, 2012

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)

```
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

                      Difference:  drug - placebo
 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
```

Problem 1 (3 pts.)

A biologist wishes to determine whether soaking in a solution of benzamil reduces the amount of healing in amputated limbs of salamanders. To study this, she amputated both hindlimbs of 17 salamanders, then put one limb from each pair in a benzamil solution and the other in a control solution. After 4 hours, she measured the amount of healing (new skin area) for each limb. Use a paired-sample t-test on these data to test whether the benzamil limb has a lower mean healing time. In the infile (benzamil.dat), the columns are the salamander number, the control data, then the benzmil data. Note: Be sure that the directionality is correct.

Your submission should consist of the code and relevant parts of the output in addition to the complete 9(8) steps of the hypothesis test. Remember that step 1 is optional because it is given in the question.

2. Categorical Data:

3.1 χ^2 Distribution

Similar to the t distribution, the χ^2 distribution depends on the "degrees of freedom". Unlike the t distribution, the χ^2 distribution is skewed and only positive values can occur.

When performing a test that involves a chi-square statistic, large values suggest a departure from the Null hypothesis. As a result, p-values of hypothesis tests involve determining an upper tail area.

SAS provides two functions involving the chi-square distribution. PROBCHI(χ^2_s, df) returns the upper tail area for a specific test statistic which is the P-value. The Quantile function which we have seen before will calculate x for a specific quantile p, or QUANTILE('CHISQ',p,df) = Pr(χ^2>x) = p for df =df. This function can be used to determine the critical value of a test. The sample code is based on Example 9.4.5 where χ^2_s = 43.2, df = 3 using an alpha of 0.05.

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:

Obs	chis	df	Pvalue	alpha	CritVal
1	43.2	3	2.2318E-9	0.05	0.35185

P-value = 2.2318×10^{-9}, $\chi^2_c = 0.35185$

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 χ^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)

```
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:

```
                    Goodness of Fit          10:31 Monday, July 16, 2012  80

                         The FREQ Procedure
```

coloracid	Frequency	Percent	Test Percent
BrownLow	15	20.83	18.75
BrownInt	26	36.11	37.50
BrownHig	15	20.83	18.75
VarLow	0.001	0.00	6.25
VarInter	8	11.11	12.50
VarHigh	8	11.11	6.25

Chi-Square Test
for Specified Proportions

Chi-Square	7.7016
DF	5
Pr > ChiSq	0.1735

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

              Sample Size = 72.001
```

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 H_0?

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 H_a is nondirectional instead of directional.

SAS Learning code: (d4.sas)

```
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*sugery/nocum chisq expected nopercent;
*expected: provides the expected values of each cell;
*nopercent: doesn't print out the percentages, just the frequencies;
weight count;
run;

quit;
```

SAS Learning output:

```
                 The FREQ Procedure

             Table of success by sugery

          success     sugery

          Frequency|
          Expected |
          Row Pct  |
          Col Pct  |real    |sham    |  Total
          ---------+--------+--------+
          success  |     41 |     15 |     56
                   | 36.587 | 19.413 |
                   |  73.21 |  26.79 |
                   |  83.67 |  57.69 |
          ---------+--------+--------+
          no       |      8 |     11 |     19
                   | 12.413 | 6.5867 |
                   |  42.11 |  57.89 |
                   |  16.33 |  42.31 |
          ---------+--------+--------+
          Total          49       26       75


       Statistics for Table of success by sugery

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      6.0619    0.0138
Likelihood Ratio Chi-Square    1      5.8549    0.0155
Continuity Adj. Chi-Square     1      4.7661    0.0290
Mantel-Haenszel Chi-Square     1      5.9810    0.0145
Phi Coefficient                       0.2843
Contingency Coefficient               0.2735
Cramer's V                            0.2843


                Fisher's Exact Test
          ----------------------------------
          Cell (1,1) Frequency (F)        41
          Left-sided Pr <= F          0.9965
          Right-sided Pr >= F         0.0156

          Table Probability (P)       0.0121
          Two-sided Pr <= P           0.0241

                 Sample Size = 75
```

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!