22-BANA 7041 Statistical Methods (4 cr.)

 Prof. Nanhua ZHANG and Prof. Yan YU

Fall Semester, 2017

**Homework Formatting Expectations**

*One of the major goals for this course is for students to learn how to* ***clearly*** *and* concisely *present technical results in the context of real world problems. As such homework problem solutions for the* ***computational******problems*** *using major statistical software should read like a report you would give to your boss. The following guidelines are to help you practice this skill. On all assignments, some points will be reserved for formatting and good presentation.*

* We will specify **page number limit**. Please follow strictly. In general, most responses to non-computational problemsdo not need to be more than one page. Keep in mind that more is not always better when writing a report. An example of a correct and concise solution is enclosed below.

* Please **clearly label the questions and answers**. **Use context for all conclusions.** SAS output or numbers with no labels or context will be counted as **wrong and lead to a formatting grade deduction**. A good rule of thumb is that **the grader should be able to tell what the question is asking without referencing the material**. Trust me your future boss is not going to look more than once at a confusingly organized or worded report.
* Please **put all of your code and output (such as tables and figures) as the last item in the very back of each problem**. The grader only really needs to see your code if your process seems correct but your numbers are wrong. Otherwise (just like your future bosses) the grader does not need (or want) to see your code at all. You will want to reference a table or chart from the back of your assignment, so please number and label them and use that number as a reference. An example on the next page will give you a good idea of how to do this. If the question specifically requests a chart you may put it next to the question. Also if you are not using part of the output then **you should NOT include** it in the report. It provides no benefit and is more likely to confuse the reader.
* Students always ask if they need to include formulas in their reports. The answer is, you must either provide a **reference to a formula from the text book, notes**, or alternatively **report all mathematical formulas using mathematical symbols**. For example, if you want to specify a sample standard deviation, then

s = sqrt(sum((Xi-Xbar)^2)/((n-1)) = 2.31should be written $s=\sqrt{\begin{array}{c}\sum\_{i}^{}\left(X\_{i}-\overbar{X}\right)^{2}/(n-1) \end{array}}=2.31$. I know this takes time but if you are going to be writing technical reports in the future you are going to need to get accustomed to this. Alternatively, to save time, just cite the text, e.g., “s = 2.31 (see formula 3.5.1 in Snedecor and Cochran)”. Usually, you can cite the place in the output in which the value of s can be found. Then you will need to cite the appropriate table at the back of the problem where this may be found. For example, “s = 2.31 (see Table 2)”. So you see, the tables can be useful and save you time.

* Please put the questions in order they are assigned. This is mostly for your benefit. If the questions are out of order the grader is more likely to miss that you completed a problem.

Finally, **the TA has been instructed by me not to accept requests for more points on a problem unless there is a grading mistake**. When the TA makes grading decisions they are implemented uniformly for everyone in the class. Change cannot be made on your grade without changing everyone else’s as well. Not only would this be a logistical nightmare but it is not a productive use of time. If a grading mistake has been made (which is bound to happen from time to time) or you really don’t understand what your mistake is, please see TA during the office hours. TA is happy to help try and explain any concepts you do not understand or grading decisions made. However, before you come to the TA’s office please make sure you have checked the solutions and have identified a specific question to ask.  **I recommend writing any questions down to save time.**

==========================================================================

Example of good format in hypothesis testing problem: HW 7 #2.

THE PROBLEM: In a study of beer preferences and gender, members of a sample of 150 potential consumers were asked to express preference for each of three types of beers, Light, Regular, or Dark. It turned out that there were 80 males and 70 females in the sample. The data are in the following table.

|  |  |  |
| --- | --- | --- |
|  |  | Beer Preference |
| **Gender** | Light | Regular | Dark |
| Male | 20 | 40 | 20 |
| Female | 30 | 30 | 10 |

Test the hypothesis of independence of gender and beer preference at the 0.05 significance level.

THE SOLUTION:

1. The hypotheses are:

H0: The two characteristics, customer gender and beer preference are independent

H1: The two characteristics, customer gender and beer preference are not independent

Perform the test at significance level α = 0.05.

1. The test statistic is the chi-square statistic and its value can be computed from the contingency table (see Table 1) as

$χ^{2}=\frac{\sum\_{cells}^{}\left(f\_{ij}-F\_{ij}\right)^{2}}{F\_{ij}}=6.1224$ . Or see formula on p. 211 of Snedecor and Cochran. The chi-square value is given in Table 2. Here, *fij* is the observe frequency of cell *ij* and can be seen as the top number of each cell in Table 1. *Fij* is the expected frequency of cell *ij*. And it can be seen as the second number in each cell of Table 1.

1. Using the p-value approach, the decision rule is to reject the null hypothesis and conclude that the two characteristics are independent at the 0.05 level of significance if the p-value, namely$ P\left(χ^{2}>6.1224\right)<0.05$.
2. Conclusion: Since the p-values is 0.0468 (see p-value in SAS output Table 2 below), we can reject the null hypothesis. That is, conclude that the two characteristics, customer gender and beer preference are not independent at the 0.05 significance level.

Relevant output shows the following contingency table and chi-square information:

**Table 1**



|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2** |  |  |  |



The SAS Code

/\* Use contingency table approach with PROC FREQ\*/

data pref;

input gender $ beer $ numcell @@;

cards;

male light 20 male regular 40 male dark 20

female light 30 female regular 30 female dark 10;

proc freq data=pref;

weight numcell;

tables gender\*beer/ expected chisq; run;

/\* Chi-Square test yields df= 2 Chisquare= 6.1224 p-value= 0.0468\*/

Example of bad formatting for the same problem

=======================================================================

Solution:



**=================================================================**

**Example of good format for Problem 7 of HW 3**

Problem: Compute a 95% confidence interval for the mean salary of all trained executives using the sample of 30 trained executives you selected in HW 1 using simple random sampling from the data set pay1.xls. Use the t-method.

SOLUTIONS: We can do this problem 3 ways:

First, the formula (Snedecor and Cochran 4.10.1) for a 95% CI for the mean of a normal population is $\overbar{X}\pm t\_{29,.025}S/\sqrt{n}$ where the sample mean, sample standard deviation of the sample of size n = 30, and the .975 quantile of the t29 distribution are,

$\overbar{X}$ = **52590.80**, *S* = **4444.77,** $t\_{29,.025}$ = 2.0423, respectively. So the 95% CI for the mean salary is (50931.10, 54250.50). These are found in Table 1 using a SAS data step. Note the t-quantile was found using the SAS quantile function for this distribution.

Second, using PROC MEANS, the 95% confidence interval is given in Table 2 based on the option CLM with ALPHA =0.05.

Finally, using PROC UNIVARIATE with the option CIBASIC (TYPE=TWOSIDED ALPHA =**.05)**, we get the result is Table 3.

The meaning of this 95% CI for the mean salary is: if random samples each of 30 executives were independently drawn repeatedly an infinite number of time, then 95% of the time the true population mean would lie within the limits $\overbar{X}\pm t\_{29,.025}S/\sqrt{n}$ . Since the results on our sample of n = 30 gives (50931.1, 54250.5), we can say we have 95% confidence that the true mean salary lies within those limits.

 Relevant Output and SAS Code:

TABLE 1: Result of the Data Step analysis



TABLE 2: Result of PROC MEANS



TABLE 3: Result of PROC UNIVARIATE



The SAS code:

**PROC** **IMPORT** DATAFILE = "C:\pay1.xlsx" OUT = salaries replace dbms=xlsx;

 GETNAMES = yes;

**RUN**;

/\* Sample of size 30 from HW 1\*/

**proc** **surveyselect** data=salaries seed=**45221**

 sampsize=**30** method=srs out=final;

**PROC** **MEANS** data=final N MEAN STD CLM ALPHA=**.05**;

 VAR salary; **RUN**;

/\*Or using a SAS data step\*/

**data** CI2;

mean= **52590.80**;STD= **4444.77**;n=**30**;

t=quantile('t',**.975**,n-**1**);

t\_CI\_lower=mean-t\*std/sqrt(n);

t\_CI\_upper=mean+t\*std/sqrt(n);

**proc** **print**;var t t\_CI\_lower t\_CI\_upper; **run**;

/\* Or use proc univariate\*/

**PROC** **UNIVARIATE** data=final CIBASIC (TYPE=TWOSIDED ALPHA =**.05**);

 VAR salary; **RUN**;

========================================================================

Example of bad formatting for the same problem

Solution to Problem 7 HW3:



**===============================================================================**