Homework Assignment 2

Due Date: Monday, October 10th

Note: There are 6 problems on this assignment. The first 4 provide examples and extra practice. They are relevant for the exams but do not need to be turned in. Solutions for them are available on the class web site. You must turn in Problems 5-6 together with the corresponding STATA/SAS printouts to receive full credit. The assignment is due Monday, October 10th. Officially it is due in class but you may turn it in to the Adam's mail folder (in CHS 51-254) any time before 3:00 with no penalty.

Note: Output from any calculations done in STATA or SAS MUST be included with your assignment for full credit. If I do not specify which way to do a problem you may chose whether to do it by hand or on the computer. All the STATA/SAS commands needed to complete this homework are given at the end of the assignment and will be reviewed in the lab. You do not need to hand in a separate lab report–simply turn in the relevant output as part of your homework.

Note: You are encouraged to work with fellow students on these problems. However, you MUST write up your solution ON YOUR OWN and IN YOUR OWN WORDS. The style of your write-up is as important as getting the correct answer. Your solutions should be easy to follow, and contain English explanations of what you are doing and why. You do not have to write an essay for each problem, but you should give enough comments so that someone who has not seen the problem statement can understand your work. You do not have to type your assignments. However, if they are too sloppy to read, too hard to understand, or give just numbers with no comments, you WILL lose points. Problems labeled (GS) are adapted from our optional text, *Primer of Applied Regression & Analysis of Variance* by Glantz and Slinker. Problems labeled "Rosner" are adapted from the 100AB text, *Fundamentals of Biostatistics, 6th ed.* by Bernard Rosner.

Warm-up Problems

(1) ANOVA Basics (Courtesy of Prof. Afifi): A partly filled in ANOVA table is shown below:

Source	df	SS	MS	F
Between (Treatment)	3		16.42	
Within (Error)		102.50		
Total	31			

(a) Complete the table and explain what each of the quantities SSB, SSW, MSB, MSW and F tells you.

(b) How many treatment groups are there? Briefly explain your reasoning.

(c) A "completely randomized design" is an approach for carrying out an experiment where the treatments are assigned to experimental units (subjects) completely at random. A "balanced design" is one in which the number of subjects in each treatment group is the same for all treatments. Is it possible that this table was derived from a balanced completely randomized design? Explain your reasoning.

(d) Obtain a p-value for a test of significance of the hypothesis that all of the treatment means are equal. In particular, would you conclude that result is significant at the $\alpha = .05$ level? (Note: This can be done approximately using an F table or exactly using STATA or SAS.)

(2) Protein Intake (Based on Rosner 12.1-12.5) Researchers compared protein intake among three groups of postmenopausal women: (1) women eating a standard American diet (SRD) (2) women eating a lacto-ovo-vegetarian diet (LAC) and (3) women eating a strict vegetarian diet (VEG). The mean and standard deviation of protein intake as well as the group sizes are presented in the table below. Use then to answer the following questions:

Group	Mean	SD	Number in group
STD	75	9	10
LAC	57	13	10
VEG	47	17	6

(a) Perform an overall F test to determine whether there is a significant difference in mean protein intake between the three groups. Be sure you an state the null and alternative hypotheses both mathematically and in words, compute the test statistic by hand, get an approximate p-value, and explain your real-world conclusions.

(b) Obtain 95% confidence intervals for each of the group means separately. Based on these, which groups appear different from one another?

(c) Now use pairwise CIs and hypothesis tests to determine whether there are differences in mean protein intake and compare your results to part (b).

(d) Suppose that you wanted to compare average protein intake on the two vegetarian diets to the protein intake on the non-vegetarian diet. Explain what contrast you would use to check whether these two means are different from one another and carry out the appropriate procedure using either a CI or hypothesis test.

(3) Birthweight and Smoking (Based on Rosner 12.14-12.10) Birthweight of an infant has been hypothesized to be associated with smoking status of the mother during the first trimester of pregnancy. This hypothesis is tested by recording the birthweights of infants and smoking status of the mother during pregnancy for all mothers who register at the prenatal clinic at a particular hospital within a 1-month period. The mothers are divided into four groups according to smoking habit: (1) Mother is a non-smoker, (2) Mother is an ex-smoker (i.e. smoked at some point during her life but not during pregnancy), (3) Mother is a current smoker but smokes less than 1 pack per day, and (4) Mother is a current smoker and smokers more than 1 pack per day. The sample birthweights in pounds within each group are as follows and are also in the honmework 2 data files as the variables "birthweight" and "group":

Group 1: 7.5, 6.2, 6.9, 7.4, 9.2, 8.3, 7.6
Group 2: 5.8, 7.3, 8.2, 7.1, 7.8
Group 3: 5.9, 6.2, 5.8, 4.7, 8.3, 7.2, 6.2
Group 4: 6.2, 6.8, 5.7, 4.9, 6.2, 7.1, 5.8, 5.4

(a) Are the mean birthweights different in the four groups? Perform an appropriate hypothesis test to answer this question using classical ANOVA methods. State the null and alternative hypotheses mathematically and in words, compute an ANOVA table and the corresponding the test statistic **by hand**, use STATA or SAS to verify the results of your calculations and obtain a p-value, and explain your real-world conclusions. Use $\alpha = .05$.

(b) Test for differences among each pair of group means by calculating appropriate t-statistics and corresponding p-values. You should do one of these calculations by hand but the rest may be obtained from STATA or SAS. Based on these tests, which means are different from one another?

(c) Suppose we want to compare the average birthweight for mothers who are not currently smoking (groups 1 and 2) to that for those who are currently smoking (groups 3 and 4). Write down an appropriate contrast, compute a confidence interval for it, and test whether it is significantly different from 0 by hand. Verify your hypothesis test in STATA. What are your practical conclusions?

(4) Health is Where the HAART Is (An Old Midterm Problem): Professor U. R. Helpful is an AIDS researcher at the University of Calculationally Literate Adults. (Did he perversely choose his field based on his name?) He is interested in how different forms of "highly active antiretroviral therapy" or HAART affect patients' viral loads. He has selected n=124 HIV positive individuals and randomly assigned them to four treatment groups: nonHAART (N), HAART A (A), HAART B (B), and a combination of HAART A and B (AB). (Ignore the ethical implications of this design for now! I am also skipping medication names to keep this simpler.) The accompanying data file contains the grouping variable, "haart", and his outcome variable, "vload", which is the $\log_{10}$ viral load. (Viral loads are often skewed and are log transformed to make them closer to normally distributed). Use the data to answer the following questions.

(a) Is there evidence that mean $\log_{10}$ viral load differs across the treatment regimens? Justify your answer by performing an appropriate **overall** hypothesis test. State the mathematical hypotheses in the classical ANOVA framework, give the p-value, and your real-world conclusions. You may use SAS or STATA to do all the computations.

(b) Test for differences among each pair of group means by calculating appropriate t-statistics and corresponding p-values. You should do one of these calculations by hand but the rest may be obtained from STATA or SAS. Based on these tests, which means are different from one another? Explain as carefully as you can what this suggests about the relative merits of the treatment regimens.

(c) Suppose Professor Helpful wants to test whether patients taking HAART A have lower viral loads than patients not taking HAART A (regardless of whether or not they are taking HAART B). Write down an appropriate linear combination, LC, for the comparison he wishes to make, give your best estimate of the effect of HAART A, and perform the corresponding hypothesis test. You should give the null and alternative hypotheses mathematically and in words with a justification of your choice, and then use STATA or SAS to find the p-value and explain your real world conclusions using $\alpha = .05$. Perform a similar calculation for HAART B. You do not need to write out all the details. Just give your best estimate for the effect of HAART B and explain whether it has a significant effect.

(d) Suppose that Professor Helpful had wanted to see whether there was a difference in viral load between patients on HAART (any of A, B or the combination) and patients not on HAART. Give the linear combination appropriate to this problem and the corresponding conclusions.

(e) In part (c) Professor Helpful implicitly assumed that the effect of HAART B did not depend on whether the subject was taking HAART A or not. Suppose he believed that the treatments were synergistic so that the effect of HAART B was **greater** for subjects who were taking HAART A than those who were not. Find an appropriate linear combination for this scenario and perform the corresponding hypothesis test. You may do all calculations in STATA or SAS but be sure to state your hypotheses both mathematically and in words with an explanation of your reasoning and give your real-world conclusions.

Problems To Turn In

(5) On the Statistical Treadmill (Data Courtesy of Dr. Belin):
In a study of "self-efficacy" (confidence in one's capability to perform a task) pertaining to exercise, subjects were randomly assigned to one of three groups ("exercise group"). Group 1 received a one-time coaching session, treadmill exercise testing, and a personal trainer three times a week for 4 weeks. Group 2 received only the coaching session and treadmill exercise testing. Group 3 received an information brochure only. Self-efficacy was measured based on the responses to a series of questionnaire items. The following self-efficacy scores, also given in the accompanying data set as the variable "efficacy," were observed after four weeks. Higher scores are better.

Group 1: 156, 119, 100, 170, 130, 154
Group 2: 132, 105, 144, 136, 132, 159
Group 3: 110, 101, 124, 106, 113, 94

(a) Use STATA or SAS to obtain descriptive statistics and 95% confidence intervals separately for each of the three groups. Sketch your CIs on a common graph (either by hand or using a computer program is fine) and use this to explain which groups you think will be different.

(b) Use the descriptive statistics from part (a) to calculate the ANOVA table for this problem by hand.

(c) Is there overall evidence of a difference in mean self-efficacy scores between the three groups? Perform an appropriate hypothesis test to answer this question using classical ANOVA methods. State the null and alternative hypotheses mathematically and in words, use STATA or SAS to verify the computation of the ANOVA table from part (a) and obtain the p-value, and explain your real-world conclusions. Use $\alpha = .05$.

(d) Compare the treatment regimens by computing 95% pairwise confidence intervals for the differences in group means. You should do one of these calculations by hand but the rest may be obtained from STATA or SAS. Based on these intervals, which means are different from one another?

(e) The confidence intervals from part (d) are not simply the differences of the confidence intevrals from part (a). Explain why this is the case and why you could even potentially get different results from parts (a) and (d) about the treatment differences. Which calculation is more appropriate and why?

(f) Suppose Groups 1 and 2 are thought of as "active treatment" and Group 3 is thought of as a "control" treatment. Provide an estimate of the mean difference between active treatment and control treatment using an appropriate linear combination. Determine whether the active treatments result in higher efficacy scores than the control treatment (i) by finding a 95% confidence interval for the difference and carefully interpreting it and (ii) by performing an appropriate hypothesis test using $\alpha = .05$. (Make sure you state your null and alternative hypotheses with a justification of your choice.) Do the calculations first by hand and then verify your answers in STATA or SAS. What are your real-world conclusions?

(g) Suppose the treatment groups had not been randomly assigned, but instead the investigator had simply found six subjects who had followed the specified regimens over the past month. Could you still infer that differences in self efficacy were attributable to the training programs? Discuss.

(6) Location, Location, Location (From Last Year's Midterm):
Prof. Urtha Green, an environmental health scientist at my favorite school, the University of Calculationally Literate Adults, is studying fine particle pollutants at elementary schools in the city of Los Seraphim. She is particularrly interested in differences in pollutant levels at different locations and in different room conditions around the schools. Specifically, she decides to take measurements in the parking (P) lot at morning drop-off, in the play yard at afternoon recess (Y), in the cafeteria (C) at lunch time, and in two

classrooms right before lunch, one with the windows closed and air conditioner running (A) and one with the windows open (W). The data are given in the files accompanying the homework as two columns, one labeled "particles" (in thousands of particles per cm^3) and one labeled "location." Use STATA to perform an ANOVA for this data, with the Bonferroni option to obtain all the follow-up pairwise comparisons of the means (see the command section at the end of the assignment for details.)

(a) Is there overall evidence of a difference in pollution levels at the various locations? Justify your answer with an by an appropriate test, giving the mathematical hypotheses and your real-world conclusions using $\alpha = .05$.

(b) According to your printout, after using the Bonferroni adjustment for multiple comparisons, which pairs of locations do **not** have significantly different particle concentrations? Use this information to describe as precisely as you can the ordering of the locations in terms of air quality for the students.

(c) Which pairs of locations would have been significantly different **without adjusting for multiple testing.** Briefly explain your reasoning. (Note: There is a slow way to do this using lots of contrast statements and a fast way using your initial ANOVA printout!)

(d) Dr. Green is interested in comparing inside air quality to outside air quality.

(i) Write down the linear combination in which she is interested, briefly explaining your reasoning.
(ii) Give your best estimate for the linear combination and its standard error based on this data.
(iii) Perform the hypothesis test Dr. Green would use to show that the average air quality inside the buildings is at least 10,000 particles/cm^3 better than outside.

STATA and SAS Commands

For this assignment you need to know how to do an ANOVA in STATA and SAS, and how to obtain follow-up contrasts. The instructions for these plus some additional commands about calculating probabilities for the t and F distribution are given below. You may also want to use some of the commands from HW1. I have included commands for STATA versions 9 and 10 which are a little different from version 11 which is what is in the labs in case anyone is using one of the older versions.

(1) STATA COMMANDS:
. oneway particle location, tabulate bonferroni You can get the basic ANOVA output in STATA using the **oneway** command. Simply type **oneway** followed by your Y variable and the group variable. For example, for warm-up problem 3 the command would be:

oneway birthweight group

This produces the ANOVA table and F test but does not offer many extensions to more complicated models. To obtain somewhat more detailed output, including the group means, group standard deviations, and group sizes, you can follow the oneway command with the **tabulate** option:

oneway birthweight group, tabulate

To add pairwise comparisons of means using the bonferroni adjustment for multiple comparisons (useful for turn-in problem 6) you add the option **bonferroni**:

oneway birthweight group, bonferroni

Of course you can use multiple follow-up options at the same time. Another way to obtain ANOVA results is to use the **anova** command which can be extended to more complex models and also used to examine contrasts between group means. It has the same basic format:

anova birthweight group

This gives you the ANOVA table (it has an extra line aside from the usual Model and Residual lines labeled "group"–don't worry about this-for what we are doing it is just a duplicate of the "model" line) as well as the overall sample size, RMSE, R-squared and R-squared adjusted as if this were a regression model (which of course it is as we will see soon!) You can not use the tabulate command after the anova command. However **once you have used the anova command** you can test contrasts between group means using the commands **test** or **lincom**. Note that the way you use these commands differs between STATA version 11 (which is what is in the computer labs) and earlier versions. I have now included both versions for your reference. Note also that STATA applies the test and lincom commands to the most recent ANOVA you have fit so if you want to go back to a previous problem you will first need to restate the anova command.

First, to test whether the mean of group 1 is significantly different from the mean of group 2 in STATA Version 11 we can type either of the following:

test 1.group = 2.group
test 1.group -2.group = 0

The equivalent commands in STATA version 9 are :

test _b[group[1]] = _b[group[2]]
test _b[group[1]] - _b[group[2]] = 0

You can test all the pairs of group means this way. Note that the test statistic STATA gives is an F statistic–which is simply the square of our usual t statistic.

You can also test more complicated contrasts (relationships among the means) by simply writing the expression of interest. For example if we wanted to test whether the mean of group 1 was equal to the average of the second and third groups we would type:

test 1.group = .5*(2.group + 3.group) [STATA version 11]
test _b[group[1]] = (_b[group[2]]+ _b[group[3]])/2 [STATA version 9/10].

You can use this same approach for the tests in warm-up problems 2(c), 3(d) and 4(c)-(e) as well as for turn-in problems 5(f) and 6(d).

Alternatively, if you want to get a confidence interval for a particular linear combination (and also a two-sided test of whether that combination is equal to 0) you can use the command **lincom**. The syntax for the example above is

lincom 1.group - .5*(2.group + 3.group) [STATA version 11]
lincom _b[group[1]] - (_b[group[2]]+ _b[group[3]])/2 [STATA version 9/10]

Note that when using this command you have to put all the terms on one side of the expression and you don't include the "=0" in the command line.

Finally, to get the exact p-value associated with an F value from STATA if you don't have the original data you use the Ftail command. Suppose your degrees of freedom are df1 and df2 and Fobs is the value of F. Then to get the probability of being greater than that value of F type

display Ftail(df1, df2, Fobs)

Ftail tells STATA to give you the probability of being BIGGER than the specified value (if you want the probability of LESS than the specified value just type F(df1, df2, Fobs) though it is rare you would want that). The equivalent command for the t distribution is

display ttail(df, tobs)

Note that most of the t-tests we are doing in ANOVA are 2-sided so you may need to double the one-sided probability given by STATA in the ttail command. Finally if you have a probability and you want the value of t associated with it (say for doing CIs or tests with weird alpha* values) you type

display invttail(df, prob)

This tells STATA you are giving it a probability and you want the value of t such that the probability of being BIGGER than that value is prob. For instance for a t-distribution with 23 degrees of freedom, $t_{.025} = 2.069$ Thus if you type display invttail(23, .025) STATA will give back 2.069.

IN SAS:

Note about libraries: If you import the data from the Excel file I posted then your data will be in the work directory and your data statements will have the form **data = work.hw2**. This is what I have used below. However if you directly open the SAS version of the file it will be stored in a library called tmp1 and your statements would have the form **data = tmp1.hw2**.

In SAS you do basic analysis of variance using **proc ANOVA**. The syntax, using the example in warm-up problem 3 where the outcome variable is birthweight and the grouping variable is group is

```
proc ANOVA data = work.hw2;
class group;
model birthweight = group;
means group;
run;
```

First I specified the data set. Then I used the "class" statement to tell SAS that "group" was a categorical variable so it could be used as the grouping variable in the ANOVA. Then I specified the "model" which said that the outcome birthweight depended on what group you were in. Model statements in SAS always look like Y = combination of X variables. The "means" statement will give me the group means, standard deviations, and so on for birthweight within each of the groups (like the "tabulate" option in "oneway" or the "summarize" command with "by" from homework 1 in STATA or "prov univariate" in SAS.)

If you want to use contrasts after fitting the model you need to use **proc glm** (for "general linear model") instead of proc anova but everything else works the same way. The syntax is:

```
proc glm data = work.hw2;
class group;
```

```
model birthweight = group;
contrast ‘‘Group 1 vs Group 2’’ group 1 -1 0 0;
contrast ‘‘Group 1 and 2 vs 3 and 4’’ group .5 .5 -.5 -5;
contrast ‘‘Group 1 vs rest’’ group 3 -1 -1 -1;
run;
```

The “contrast” statements give tests of whether the linear combinations specified by the coefficients are equal to 0. The first contrast line in my example tests $\mu_1 - \mu_2 + 0\mu_3 + 0\mu_4 = \mu_1 - \mu_2 = 0$. In other words, it is a pairwise comparison of whether the non-smokers and ex-smokers have babies with the same birthweights. The second contrast compares birthweights of babies for those who currently don't vs currently do smoke. The third contrast compares non-smokers to those who have ever smoked (now or in the past). Note that I could have used 1 1 -1 -1 for the coefficients in the second contrast and similarly I could divide all the terms in the third contrast by 3. This doesn't change the outcome of the hypothesis test but it does change the linear combination that is being estimated in terms of the scale. The phrases in quotation markes are just labels for the output so that when SAS prints the results you will be able to remember which contrast was which. You can include as many contrasts as you like on a single model statement.