Lab 8

STA 216 F19

The dataset “lab 8 – blood pressure.xslx” was used as a case study at the 2003 meeting of the Statistical Society of Canada. It contains 500 observations (subjects), 250 of which had low blood pressure and 250 had high blood pressure (i.e. hypertension). The response variable is systolic blood pressure in mm Hg, which is the higher of the two blood pressure measurements usually taken (the other is called diastolic). The variables in the dataset include:

|  |  |
| --- | --- |
| SAS variable name | Description |
| sbp | Systolic blood pressure |
| smoke | Smoking status (Y = Smoker, N = Non-smoker) |
| Overwt | Weight category (1 = Normal, 2 = Overweight, 3 = Obese) |
| Race | Categorical variable taking values 1, 2, 3, 4 (specific races not identified) |
| Exercise | Exercise level (1 = Low, 2 = Medium, 3 = High) |
| Alcohol | Alcohol use (1 = Low, 2 = Medium, 3 = High) |
| Stress | Stress level (1 = Low, 2 = Medium, 3 = High) |
| Salt | Salt (NaCl) Intake level: 1 = Low, 2 = Medium, 3 = High |

1. After reading the data into SAS, make a label for sbp and labels/formats for smoke, exercise, overwt, alcohol, salt, and stress. (If you’re feeling fancy, you can use the same format for multiple variables.)   
   Afterwards, copy and paste tables from PROC FREQ for the variables above (except sbp) to demonstrate that the labels/formats are correct (and to see how many subjects are in each category).
2. Summarize the variable systolic blood pressure.
   1. Copy and paste a histogram.
   2. Copy and paste a table containing the mean, standard deviation, and the 5-number summary, rounding each to 1 decimal place.
   3. Normal blood pressure is defined having a systolic blood pressure less than 120 mm Hg and a diastolic blood pressure less than 80 mm Hg. What percent of subjects in this dataset have systolic blood pressure less than 120?
3. It is well-known that smoking, exercise, being overweight, salt intake, and stress are related to blood pressure. For each of the following variables, copy and paste a table containing the mean and standard deviation of blood pressure separated by category. Also, briefly explain whether the means by category for each variable support the common knowledge about blood pressure.
   1. Smoke
   2. Exercise
   3. Overwt
   4. Salt
   5. Stress
4. If we assume this dataset is a random sample, we can make inferences about the effects of the variables on blood pressure for the population. For instance, we can make a confidence interval for the difference in population mean systolic blood pressure by smoking status. Since smoking status has 2 categories, we can use either PROC TTEST or PROC GLM to calculate this interval.
   1. Copy and paste the table that gives this 95% confidence interval from PROC TTEST, highlighting the numbers that this CI. (Recall that we had used Method=Satterthwaite under our two-sample model that assumed separate variances for each group; Method=Pooled assumes the variances of both groups are the same.)
   2. Interpret this interval within the context of the problem. Use words like “less” or “greater” instead of “difference”.
   3. Copy and paste the table that gives this 95% confidence interval from PROC GLM with the LSMEANS statement, highlighting the numbers that give it.
   4. Note that the interval from PROC GLM matches the one from PROC TTEST with Method=Pooled and is very slightly different from the one for Method=Satterthwaite. (The results for Method=Pooled and Method=Satterthwaite are so close because the standard deviations for the two groups are almost identical, as seen in 3a.) PROC GLM is based on the one-way ANOVA. Which one of the four assumptions of the one-way ANOVA model indicates that PROC GLM would give the same results as Method=Pooled for PROC TTEST?
5. Consider the variable exercise. Using PROC GLM, perform the global F test of equality of the population means of blood pressure for different exercise levels.
   1. State the null and alternative hypotheses in terms of Greek letters; then define these Greek letters.
   2. Report the F statistic and p-value and copy and paste the table that shows them.
   3. Write the decision about Ho.
   4. Write a conclusion in the context of the problem.
6. Continuing with the variable exercise,
   1. For the following 3 pairwise comparisons of the population mean systolic blood pressure for the different categories of exercise, state the p-value for test and whether it represents a significant difference.
      1. Low vs. Medium
      2. Low vs. High
      3. Medium vs. High
   2. Copy and paste the table that gives these p-values.
   3. Copy and paste the table titled “T Comparison Lines for Least Squares Means of exercise”. Explain how the column with the letters A and B gives the same results as in 6a.
   4. State the 95% confidence interval for the difference in population means of blood pressure for low and medium exercise level, calculated as low minus medium.
   5. Interpret this interval within the context of the problem. Use words like “less” or “greater” instead of “difference”.
7. Use Fisher’s LSD procedure for determining which (if any) of the blood pressure means are significantly different for the different levels of stress at the alpha=.05 level. Show supporting SAS output.

**[8-11 are optional]**

1. As we saw earlier in the semester, statistical significance does not always agree with practical significance. Here, we will look at two differences in means where the smaller difference is statistically significant and the larger difference is not statistically significant. (Weird, huh?)
   1. Calculate the size (absolute value) of the difference in sample means of blood pressure for subjects that have “High” and “Medium” alcohol use.
   2. Calculate the size of the difference in sample means of blood pressure for subjects in race categories 3 and 4.
      1. Copy and paste a SAS table that gives the 95% confidence interval for the difference   
         , (the difference in alcohol categories), highlighting it in the table.
      2. Does this represent a significant difference? Why or not?
      3. Copy and paste a SAS table that gives the 95% confidence interval for the difference   
         , (the difference in race categories), highlighting it in the table.
      4. Does this represent a significant difference? Why or not?
2. Let’s calculate these intervals by hand to see what’s causing this. As shown on Handout 17, the formula is . Let’s do the calculation for the difference in high and medium alcohol categories first. You will need to obtain the ANOVA table from PROC GLM.
   1. What is ?
   2. for a 95% confidence interval can be calculated from the code block  
        
      where dfError is replaced by the DF(Error) from the ANOVA table.  
      Find t\* in this case.
   3. is the root MSE. Find it.
   4. Calculate .
   5. Recall that the margin of error of the interval is also called the least significant difference (LSD). What is the LSD in this case?  
      [Also, verify for yourself that this gives the same CI.]
3. Calculate the 95% confidence interval for the difference , the difference in race categories, by hand.
   1. Find .
   2. Find s.
   3. Calculate .
   4. Calculate the LSD.
4. Which part of the confidence interval formula ( or ) is the greatest contributor to the difference in LSD between the two intervals?