Quiz 3: Nutr 6101-Nutrition Research Methods

Your Name: Melissa Vigdor

[Please use this file to type your answers. Add your last name to the file name when you save your responses (i.e., 'Save As…"). If your name is Smith, the file name will be "Nutr6101-Quiz3_Smith". Send the file with your answers to me **by no later than** Sunday, November 17th @ 5:00pm.]

Q. You have assessed: education level achieved (i.e., H.S., B.S., M.S., PhD/MD, etc), current income, age, race, height, weight, body fat percent, LDL cholesterol, HDL cholesterol, and total cholesterol in all the active players on 3 NFL teams (60 x 3).
   1. You wish to determine the variables most associated with LDL, HDL, and Total Cholesterol in this sample. Describe the statistical procedure(s) you would use to make this determination. (Not just the name of the statistical test, but also the procedure you would follow leading up to the test that includes identifying the dependent/independent variables.)

First, I would define my variables:

- Dependent variables: LDL cholesterol, HDL cholesterol, Total Cholesterol (all continuous)
- Independent variables: education level, current income, age, race (categorical), height, weight, body fat percentage (ratio)

Then I would enter the variables into SPSS as follows:

Categorical
- Age (yrs) - 2 places, no decimals
- Education Level: HS = 1, BS = 2, MS = 3, PhD/MD = 4 (one place no decimals)
- Race: Caucasian = 1, African American = 2, Asian = 3, Latino = 4, Pacific Islander = 5, Other = 6 (one place no decimals)
- Current Income ($) – 7 places no decimals

Ratio
- Height (cm) – 3 places no decimals (eventually can add a new variable in spss to convert this value to meters by dividing by 100, which would allow a BMI calculation if desired)
- Weight (kg) – 3 places no decimals
- Body fat percent (%) – 2 places no decimals

Continuous
- LDL cholesterol (mg/dL) – 3 places no decimals
- HDL cholesterol (mg/dL) – 3 places no decimals
- Total cholesterol (mg/dL) – 3 places no decimals

Then I would write my research and null hypotheses for each correlation I wanted to investigate for example:

Quiz 3: Nutr 6101-Nutrition Research Methods

Research Hypothesis: There is a relationship between increased age and increased total cholesterol in NFL players.

Null Hypothesis: There is not a relationship between increased age and increased total cholesterol in NFL players.

I would set the alpha score at $P<0.05$. Then I would create several scatter plots to illustrate the linear relationship between the dependent and independent variables. For example, total cholesterol levels (dependent variable) on the y-axis and the age of the player (independent variable) on the x-axis. The way the points scatter would allow me to see if a linear relationship exists. This linear relationship could be positive or negative depending on the slope of the line. For example as age increases so does total cholesterol level (positive) or as age increases, total cholesterol decreases (negative). This relationship is not causal, but will allow you to infer if you know a person's age how that will influence their total cholesterol level.

This correlation test can be performed many times with different combinations of variables to determine if there is a relationship between any of the independent variables and LDL, HDL and total cholesterol in the sample population.

Additionally, the Pearson's correlation coefficient (r) value can be used to predict the strength of each relationship. The closer the value is to 1 or -1 the stronger the relationship is. If the number is negative, it is a negative correlation and if the number is positive it is a positive correlation. This value usually confirms the visual scatter plot

2. Assuming age and total cholesterol are found be strongly related, please indicate all the different ways you can think of for handling the 'age' and 'total cholesterol' values for doing multiple different Pearson correlations.

There are several ways to handle the age and total cholesterol values using the Pearson correlation test:

- Restrict the range of one of the other variables. For example, if we look at the relationship between age and total cholesterol only in players who weigh less than 200lb or only in African American players. Restricting the population in this way would cause the strength of the correlation to be smaller than if the entire population was used.
- Review the data for outliers, extremely high or low data points can skew your data and eliminate the linear relationship.
- Z-score indicates how many standard deviations the value is above or below the mean. Calculating the z-score allows you to compare scores on one distribution to scores on the other distribution and see if a correlation exists.

3. You wish to see if different age players have different Total Cholesterol values, and you are limited to doing a t-test. Describe how you would do this.

The t-test is used to compare the means of two independent groups to find out if they are statistically different from each other. Before the test was run, I would confirm that the data was normally distributed since this test is used for normally distributed data. Creating a histogram in SPSS will allow me to see the way the data is distributed. I would also set the alpha score at $P<0.05$.

My objective would be to see if total cholesterol is significantly higher in players who are 25 years old or older than in players who are under 25 years old. My hypothesis is that total cholesterol will be higher in players 25 or older. The null hypothesis is that total cholesterol will not be higher in players that are 25 years old or older. The dependent variable is total cholesterol (continuous) and the independent variable is age (categorical).

Next I would perform the t-test in SPSS. The test would be run using the entire population of 180 players. I would use a cut point of 25 to define my age groups: participants who are less than 25 years old will be in group one and those who are greater than or equal to 25 years old will make up group two. This way, I would be able to compare total cholesterol level in each age group. After the test is run, I would review the results. The one-tailed number explains the significance on one side of the distribution and it is usually easier to show significance by using this value as opposed to the two-tailed number. The null hypothesis will be rejected if the one-tailed t-value is less than the observed T and the means are in the right order. Descriptive statistics on mean and standard deviation should also be available in the SPSS output report.

4. Same as question "C", but for education level instead of age.

As in the previous answer, I would first confirm that the data was normally distributed using a histogram, since this test is used for normally distributed data. I would also set the alpha score at $P<0.05$.

My objective would be to see if total cholesterol is significantly lower in players who have a graduate level education or higher (MS, Phd/MD) than in players with undergraduate or lower education (BS, HS). My hypothesis is that total cholesterol will be lower in players with a graduate level education or higher. The null hypothesis is that total cholesterol will not be lower in players that have a graduate level education or higher. The dependent variable is total cholesterol (continuous) and the independent variable is education level (categorical).

Next I would perform the t-test in SPSS. The test would be run on the entire population of 180 players. I would input the four education levels associating them with the numbers indicated in question 1 (HS =1, BS =2, MS =3, Phd/MD=4). I would

use a cut point of MS education and above (value of 3+) to define my education groups: participants who have an education level below MS will be in group one and players with an education level that is MS or higher will comprise group two. This will allow for the comparison of total cholesterol level within the "higher education" and "lower education" groups. Then, I would run the test and analyze the results to see if significance exists (if the value is less than 0.05). The one-tailed number explains the significance on one side of the distribution and is usually easier to show significance by using this value as opposed to the two- tailed number.

5. Same as question "C", but for income instead of age.

As in the previous two answers, I would first confirm that the data was normally distributed using a histogram, since this test is used for normally distributed data. I would also set the alpha score at $P<0.05$.

My objective would be to see if total cholesterol is significantly lower in players who have an income level of $3,000,000 per year or more than it is in players with an annual income below $3,000,000. My hypothesis is that total cholesterol will be lower in players with a $3,000,000 annual income or higher. The null hypothesis is that total cholesterol will not be lower in players that have a $3,000,000 annual income or higher. The dependent variable is total cholesterol (continuous) and the independent variable is income level (categorical).

Next I would perform the t-test in SPSS. The test would be run on the entire population of 180 players. I would use a cut point of $3,000,000 annual income and above to define my two income groups: participants who have an income level below $3,000,000 per year will be in group one (low income) and players with an income level that is $3,000,000 or higher will comprise group two (high income). This will allow for the comparison of total cholesterol level between the high income ($3,000,000+) and low income (below $3,000,000) groups. Then, I would run the test and analyze the results to see if significance exists. The one-tailed number explains the significance on one side of the distribution and is usually easier to show significance by using this value as opposed to the two-tailed number.

6. You wish to predict total cholesterol from the variables you have collected. Please describe the statistical procedure you will follow to do this, and indicate the values you will look for to determine the strength of your prediction.

Variable to predict: dependent – total cholesterol (continuous)
Variables to predict total cholesterol: independent – other variables collected (listed in question 1 either categorical or ratio)

Quiz 3: Nutr 6101-Nutrition Research Methods

Regression analysis is used to predict the value of the dependent variable (total cholesterol) based on an independent variable. In order to predict total cholesterol from the variables I have collected, I will need to do a multiple regression analysis since there are multiple independent variables. This test is best performed without outliers because they can alter the predictive outcome or influence the statistical significance.

To preform the test, I would select total cholesterol as my dependent variable and select education level achieved, current income, age, race, height, weight, body fat percent, LDL cholesterol, and HDL cholesterol as the independent variables. I would set the alpha score at $P<0.05$. Once in SPSS, I can choose to run a forward step wise or backward step wise analysis. If I select forward step wise, it will tell me if my independent variables explain or do not explain a significant amount of variance for total cholesterol. If it doesn't explain a significant amount of variance, SPSS will eliminate it since it will only keep it if it is significant.

Next, I would develop a research hypothesis to state my belief on how each of the independent variables listed in question one would influence total cholesterol. I would also create a null hypothesis for each, for example (listing includes some not all independent variables):

Research hypothesis 1: As age increases, total cholesterol will increase.
Null hypothesis 1: There is no relationship between age and total cholesterol.

Research hypothesis 2: As income level increases, total cholesterol will decrease.
Null hypothesis 2: There is no relationship between income level and total cholesterol.

Research hypothesis 3: As weight increases, total cholesterol will increase.
Null hypothesis 3: There is no relationship between weight and total cholesterol.

Research hypothesis 4: As body fat percentage increases, total cholesterol will increase.
Null hypothesis 4: There is no relationship between body fat percentage and total cholesterol.

The relationship that exists between each independent variable and total cholesterol (dependent variable) can be analyzed by creating a scatter plot. This will show if there is a linear relationship between the two variables.

After the test is run, I would analyze the results. R is the multiple correlation coefficient it determines how strong the predictive value is. The closer this value is to 1 or -1 the more predictive the relationship between the independent variables and total cholesterol level. If the number is negative, it is a negative relationship and if the number is positive it is a positive relationship. R^2 is the coefficient of determination. It illustrates the amount of variance in the dependent variable that is

explained by the independent variables. For instance if the R^2 value is 0.727, that means that the independent variables account for 72.7% of the variability in total cholesterol level. You might also mention SEE.

The significance column on the SPSS report shows the p value for each independent variable. If the value is less than 0.05 then the relationship is significant. Based on the significance, the null hypotheses and research hypotheses can be rejected or accepted based on the findings.

1. You wish to see if the players on the 3 different teams have different LDL, HDL, and Total Cholesterol values. Describe the statistical procedure you would follow to do this.

Prior to statistical testing, I would check the data to make sure there were no outliers and that the data was normally distributed. Creating a histogram will allow you to determine the distribution of the data. I would set the alpha score at $P<0.05$. If the data were normally distributed, I would use an ANOVA to evaluate the LDL, HDL and Total Cholesterol values on the three NFL teams. The ANOVA test is used to determine if there is a significant difference between three or more groups on the same variable. My three groups are NFL team 1, NFL team 2, and NFL team 3. I would preform the ANOVA three times comparing each of the three team's means for LDL, HDL and total cholesterol. My research hypothesis is that the three NFL teams will not have the same LDL, HDL and Total Cholesterol values. The null hypothesis to test is that the three NFL groups will have the same LDL, HDL and Total Cholesterol values. The data output for this test involves the F statistic. When there is a difference between at least two groups the F statistic is significant.

If the data were skewed, I would run the non-parametric equivalent to the ANOVA, Kruskal-Wallis analysis of variance by ranks. It is a test used for three or more independent groups as well.

When you run the test, there are also different post hoc tests you may run along with it. Some apply to parametric data such as Bonferroni and Tukey and others are appropriate for non-parametric data for example Dunnett's T3.

Wow Melissa. Great Test! Score = 100 (A)