Homework #4

Dustin Leatherman

October 7, 2018

(1) Dependence Root Cause Analysis

For each example, explain the source of dependence and whether the dependence is between or within groups.

Traffic

Researchers interested in learning the effects of speed limits on traffic accidents recorded the number of accidents per year for each of 10 consecutive years on roads in a state with speed limits of 90 km/h. They also recorded the number of accidents for the next 7 years on the same roads after the speed limit had been increased to 110 km/hr. The two groups of measurements are the number of accidents per year for those years under study. (Notice that there is also a potential confounding variable here!)

Car accidents are likely to beget other car accidents so there is a serial correlation between accidents for both the 90 km/h and 110 km/h groups. The number of accidents in the 90 km/h group may also affect the number of accidents in the 110 km/h group. It is reasonable to guess that a roadway which is accident-prone may give drivers cause to drive in a different manner than normal.

Twin Intelligence

Researchers collected intelligence test scores on twins, one of whom was raised by the natural parents and one of whom was raised by foster parents. The data set consists of test scores for the two groups, boys raised by their natural parents and boys raised by foster parents.

There is a clustered dependence between both groups of twins. Since each twin came from the same parents, it is more likely that they will have similar intelligence levels.

Indooor Pollution

Researchers interested in investigating the effect of indoor pollution on respiratory health randomly select houses in a particular city. Each house is monitored for nitrogen dioxide concentration and categorized as being either high or low on the nitrogen dioxide scale. Each member of the household is measured for respiratory health in terms of breathing capacity. The data set consists of these measures of respiratory health for all individuals from houses with low nitrogen dioxide levels and all individuals from houses with high levels.

The source of the dependence in this study is within the low nitrogen dioxide groups and high nitrogen dioxide groups since the breathing capacity should be similar between households.

(2) T-test Validation Analysis

Desired Weightloss

Examine the data and discuss the validity of the using a two sample t-test to compare the mean desired weight loss between those who exercised in the last month and those who didn't (i.e. was the t-test you conducted in the last homework valid?)

The BFRSS survey by the CDC annually surveys 300,000 americans about health habits. Survey data is particularly tricky in that there are a lot of correlations between responses. This can appear within and between various groups. For example, it is likely that members in the same neighborhood will give similar responses about desired weight loss as self-image can be influenced by culture. This means that without any massaging, traditional model-based approaches such as the t-test are not valid. To that effect, researchers can apply transformations to the results in order to make the results more fit for testing. For BFRSS in

particular, the researchers attempt to normalize the sample through post-stratification, which weights results according to a known population. The CDC dataset we are working with is a sample of 20,000 responses from this initial dataset; the CDC small dataset is a sample of 1000 responses from those 20,000 responses. The source of the initial survey data has been transformed so that it is appropriate to do model-based testing with. Given that the CDC-small dataset is a sample of a transformed sample, the t-test is still valid as long as the transformed sample follows a Student T distribution.

Additionally, the Shapiro-Wilks test yields a low p-value which indicates that this data is non-normal. This is further confirmed by histograms and boxplots which show that the data is very much right skewed. A randomization test with 200,000 combinations indicates a p-value of approximately 0.001 which correlates to the p-value produced by the Two-sampled t-test. Even though the data is non-normal, the t-test is still valid since this sample comes from a transformed sample and a randomization test yielding approximately the same result as a two-tailed t-test.

What are you Zincing about?

Exercise 1.25 in The Statistical Sleuth contains data on zinc concentrations in the blood of rats that received a dietary supplement and rats that did not receive the supplement. Examine the data and discuss the validity of the using a two sample t-test to compare the mean zinc concentration between the two groups of rats.

Rats from the same group are randomly assigned whether or not they receive a supplement. The rats themselves are not randomly sampled as they are clustered within the same group. It is likely that rats in both groups may share charactertistics since they were part of the same initial group. This makes the t-test not valid without any transformation. The Shapiro-Wilks test produces a p-value of 0.0554 for the Zinc data and 0.1002 for a log transformation of Zinc data. These values indicate that there is not enough disprove the normality of the data. This means that the log transformation of Zinc may be a more appropriate two-tailed t-test but without transformation, the two-tailed t-test is not valid.

Statistical Trauma

Exercise 3.18 on page 78 of The Statistical Sleuth gives metabolic expenditures for seven patients admitted to a hospital for multiple fractures (Trauma) and eight patients admitted for other reasons (Nontrauma). Examine the data and discuss the validity of using a two sample t-test to compare the mean metabolic rate between trauma and non-trauma patients.

The patients are sampled from two groups, one where patients have multiple fractures and one where they have not. The wording suggests that these statistics were gathered at the same hospital. It is possible that there is clustered sampling within groups as hospitals follow Standard Operating Procedures after an initial diagnosis has been taken. This would mean that each group from clustered dependence from within. The Shapiro-Wilk test indicates that there is enough convincing evidence that metabolic expenditure is not normally distributed (p-value=0.0016). When comparing box plots, the Trauma group has a larger spread. The Shapiro-Wilk test using the log transformation of metabolic expenditure still indicates that there is convincing evidence that the log metabolic expenditure is not normally distributed (p-value = 0.0102). Because the data is non-normal, the two-tailed t-test is not valid.