MSCR 500 and 533 Final Exam

Anish Shah

December 17, 2019

Question 1

Blood-pressure measurements taken on the left and right arms of a person are assumed to be comparable. To test this assumption, 40 people are randomly sampled, and systolic blood-pressure (SBP) readings are taken simultaneously on both arms by two observers. Assume that the two observers are comparable in skill and experience, and assume that left arm SBP, right arm SBP, and the difference between left and right arm SBP are all symmetric but distinctly non-normal.

- assumption is left/right arms have same blood pressure in a person
- sample is from 40 pts at random
- simultaneously measures are taken of BP
- three variables are created: left arm, right arm, and difference between both arms
- null hypothesis would be the difference b/w arms is 0
- non-normal distribution, but n = 40
- one-sample is given

The most appropriate procedure for testing whether or not the two arms give comparable results is:

One Sample T-test based on M&M guidelines for $N \geq 40$.

Question 2

Weight loss was recorded for a random sample of 10 people who had been taking a weight-loss drug. The population of weight losses is likely to be strongly skewed.

- sample size is 10 people

- one-sample description
- true mean is known

The most appropriate procedure for testing whether the true average weight loss is greater than 10 lbs is:

Wilcoxon Signed Rank Test is most appropriate as population is skewed/non-normal ($N \leq 10$).

Question 3

Investigators measured the before-breakfast plasma citrate concentrations for 110 randomly sampled people. The average was $119 \mu mol/l$. Suppose that plasma citrate concentrations are known to follow a normal distribution.

- sample of 110 (satisfying M&M guidelines for normal as well)
- one sample testing
- SD/variance is unknown however

The most appropriate procedure for testing whether the true average before-breakfast plasma citrate concentration for the population is greater than $115 \mu mol/l$ is:

One-Sample T-test as the distribution is normal, but the sigma is unknown.

Question 4

A study was conducted to analyze the relationship between vasectomies and prostate cancer. It was found that 7 out of 2130 randomly sampled men who had not had a vasectomy had prostate cancer, while, in an independent random sample, 11 out of 2200 men with vasectomies had prostate cancer.

- comparing two populations (no vasectomies versus vasectomies)
- large sample sizes in both groups
- proportion of outcome/event is known in both
- H0: prostate cancer risk is similar without or with vasectomy

The most appropriate hypothesis test method for analyzing the relationship of interest is:

Chi-square test of homogeneity as there are two independent random samples from each population (vasectomy or not).

Question 5

A microbiologist measured the growth of two strains of a bacterium – a mutant strain and a non-mutant strain – using mouse cells in petri dishes. Nine randomly selected mice were used. Nine pairs of petri dishes were used. Within each pair of dishes, cells from the same mouse were used; in one of the dishes, the mouse's cells were exposed to the non-mutant strain, in the other dish the same mouse's cells were exposed to the mutant strain. Hence, the data were paired by mouse. The sample results are shown below. Each number in the second and third columns represents the total growth in 24 hours of the bacteria in a single dish. A priori, the researcher suspected that the mutant strain would grow faster. Test, at $\alpha = 0.05$, whether the researcher's suspicion is true. All populations involved can be assumed to be normally distributed.

Hypothesis

H_0: $\mu = 0$
H_1: $\mu > 0$
$\mu = true difference between normal and mutant strains (normal - mutant)$

Test:

Paired T-test

Justification:

Sample is "normally distributed", allowing T-distribution per M&M, also, paired as each sample is otherwise confounded by the individual mouse that hte cells came from.

Test Statistic:

The calculations were done in SAS.
t-value is 2.50

P-value:

p-value = 0.0371

Decision/conclusion:

We have sufficient evidence that we should reject the null hypothesis. The scientist hypothesized that hte mutant strain would grow faster than the normal strain. This means that the mean of the mutants would be higher than the mean of normal strains (growth rate). Our statistical test shows that the mutant and normal strain have a significant difference in mean values ($p < \alpha$), however their is a directionality issue here. The calculated mean difference is $\mu = 33.3333$, which is positive, suggesting that the normal strain actually grew faster than the mutant strain. Thus, the relationship predicted by the scientist is incorrect. There is a difference between means, however the normal strain grows faster than the mutant.

Question 6

A major court case on the health effects of drinking contaminated water took place in the town of Woburn, MA. A town well in Woburn was contaminated by industrial chemicals. During the period that residents drank water from this well, there were 16 birth defects among 414 births. In years when the contaminated well was shut off and water was supplied from other wells, there were 5 birth defects among 228 births. The plaintiffs suing the firm responsible for the contamination claimed that these data showed that the proportion of birth defects was higher when the contaminated well was in use. Was the true proportion of birth defects higher when the contaminated well was in use? Perform the appropriate hypothesis test at the $\alpha = 0.05$ level.

Hypothesis

H_0: $p_1 - p_2 = 0$
H_A: $p_1 - p_2 \neq 0$

Test:

Chi-square test of homogeneity

Justification:

There are two independent samples that meet a normal distribution by CTL and M&M. There is a categorical variable (birth defects) that occur in different frequencies (potentially) in each population. All of this makes the above test the most appropriate. All cells in the contingency table are also 5 or more.

Test Statistic:

Calculated by SAS.
$\chi = 1.2987$

P-value:

p-value $= 0.2545$

Decision/conclusion:

We do not have sufficient evidence to reject the null hypothesis. Because $p > \alpha$, we do not have enough evidence to say that the difference in frequency of birth defects based on well-water usage is significantly different.

Question 7

Researchers studied the effect of dietary supplementation of calcium on blood zinc levels. Blood zinc concentrations (mg/ml) were measured in pairs of rats. In each pair: (i) the rats came from the same litter, and (ii) one rat received a dietary supplement of calcium, and the other rat did not. The data (blood zinc

levels) for the pairs of rats is stored in the Excel file 'zinc.xls'. Your tasks: use the SAS import wizard and SAS programming statements to:

1. import data into zinc
2. DIFF variable (calcium-treated versus untreated rats) for difference in zinc concentrations
3. PAIR variable for each pair of rats
4. SEX variable (first 10 pairs are 0, 1 for the remaining pairs)
5. Save data into a permanent data set named zinc_Anish
6. Perform descriptive/inferential statistics
7. Save or attach SAS code

a. Do there appear to be extreme outliers or implausible data?

NO

b. Provide following information for the variable DIFF?

Mean: 0.076
Median: 0.055
Skewness: 0.1272
Standard Deviation: 0.3210
Interquartile Range: 0.4650
Kurtosis: -0.6228

The histogram created to assess normality of DIFF appears to show a bell-shaped distribution, and is visually normal. The probability plot of DIFF shows a mostly linear/normal pattern as well.

90% Confidence Interval for true average DIFF: [-0.0481, 0.2001]
90% Confidence Interval for true average DIFF for female rats: [-0.0887, 0.2707]
90% Confidence Interval for true average DIFF for male rats: [-0.1406, 0.2626]

c. Describe the distribution of DIFF by overall and sex?

The distribution of hte overall DIFF appears to be normal in pattern by histogram and probability plot. However, when analyzing by sex subsets, the sample size is much smaller and the histograms do not appear to be normal visually. For female rats, teh distribution is ever so slightly left skewed, while hte data almost appears bimodal in the distribution of just male rats (however this is all likely from random error as the sample is just 10 in each subgroup).

Question 9

In an exercise physiology study, the relationship between oxygen consumption (the dependent variable) and several potential predictors of oxygen consumption was to be studied. In a random sample of 31 subjects, the following variables were measured:

- Oxygen consumption (% of available oxygen consumed, as measured during 1.5 mile treadmill run. The more fit a person is, the greater their oxygen consumption will be during physical activity)
- Sex (M, F)
- Age (years)
- Weight (kg)
- Run time (time, in minutes, taken to run 1.5 miles on treadmill)
- Maximum pulse (highest pulse rate during the run)
- Performance (a subjective measure of the subject's fitness, assigned by an interviewer; larger values mean the subject was assessed as having a higher fitness level)

The data is in a file named fitness2019, make a report of this data ($\alpha = 0.05$) with the following attributes:

- table of descriptive for each variable
- state the regression model
- briefly describe methods used to arrive at your model
- state estimate of model
- report p-value for overall test for model
- report a measure of fit of the model
- report p-values for the partial tests for the model
- report key results of any diagnostic work performed for the model
- include short/clear/well-written summary describing any significant relationships
- cut/paste objects must be discussed in your report

Question 10

"I, *Anish Sanjay Shah*, will complete all the work for question 10 before the Spring 2020 semester begins."