MSCR 509: Homework 2

Anish Shah, MD

February 3, 2020

Question 1

Describe the appropriate statistical framework for addressing the following scientific questions (statistical hypothesis testing, statistical estimation, development of a statistical model, etc). Briefly describe the procedure – this may include: specifying the hypotheses (H0, Ha) and the quantities that you will be calculating, stating how you will make a decision, etc.

To assess the effectiveness of a new drug for treating stroke, patients are randomized to the new drug or a standard treatment. Stroke severity (continuous measure) is measured at the end of the study.

1. The outcome is a linear variable of *stroke severity*
2. This is a comparison of two groups of patients, randomized to a drug or standard treatment
3. H_0: *There is no difference between the mean stroke severity between groups*
4. H_1: *The mean stroke severity in the treatment group is different than that of the control group.*
5. $\alpha = 0.05$
6. To start, we should assess distribution of the outcome, including size, and assess whether it can follow a t-distribution.
7. If so, then a two-sample t-test would be an appropriate starting test.
8. If the mean severity score is less, with an appropriately p-value less than our prescribed α, then this new drug is worth further investigation
9. Can also think about confounding variables about patient population that would interfere with our simple analysis

Subjects are recruited to determine the prevalence of chronic fatigue syndrome among the general population.

1. *Chronic Fatigue Syndrome (CFS)* is a binary variable, either present or absent. To assess prevalence, an observational study would be needed. The recruitment process would need an effective sampling method
2. As we are assessing prevalence, there isn't a necessary statistical test or hypothesis to be explored. Howeve, the study design is important. In this case, a cross-sectional study would be appropriate.
3. I would collect data in such a fashion that groups could be tested (*chi-square*) if need be later on, although this currently just assesses prevalence.

To determine the risk factors (early life stress, gender, etc) that are associated with depression, a cross sectional study is conducted.

1. In this case, *depression* is the outcome variable. There are two approaches to assess the covariates: a) we can assess them in a predictive model, or b) in a causal model.
2. The IV can be both categorical and continuous (e.g. age). The dependent variable could be assessed as a linear/continuous variable of depressive symptoms, or as a categorical variable of the presence/absence of depression. Per the question stem, it seems to be phrased for a categorical/dichotomous dependent variable.
3. A multivariable logistic regression would be the most appropriate test.
4. H_0: *The beta coefficients for each covariate is 0*

5. H_1: *The beta coefficients for at least one of the covariates is not 0*
6. $\alpha = 0.05$
7. This testing woudl allow us to examine the relationship of each variable in the model. The question becomes how to best build the model, and which variables to add at what time. I would start with variables supported by the literature and use a step-wise model.
8. I would make my decision with the final model.

Asthma cases and matched controls are enrolled to determine whether the asbestos exposure is related to asthma.

1. This describes a case control study, thus a fixed rate. We can assess the odds ratio in this case, as the exposure variable is binary. We can also assess the probability of distribution / independence of hte populations.
2. Assuming that no box in this 2x2 matrix has less than 5 observations, we can use a *Chi-square test*.
3. H_0: *There is no association between asbestos and asthma*
4. H_1: *There is an independent association between asbestos and asthma*
5. $\alpha = 0.05$
6. We would have to identify degrees of freedom, and look at how the data falls on a chi-square distribution to assess a p-value. If so, we can reject the null hypothesis.

To determine the accuracy of a new procedure for measuring renal volume (continuous measurement), each patients' renal volume is measured by both the new procedure and MRI (the gold standard).

1. There are two measurements, the new renal volume and the MRI volume. Each patient is thus being tested twice, and each measurement should be "paired" essentially.
2. H_0: *There is no difference between renal volumes by new or MRI methods*
3. H_1: *There is a difference in renal volumes by the new versus MRI method*
4. $\alpha = 0.05$
5. I would assess this by a *paired t-test*, assuming the volume distributions followed a t-distribution.
6. If there was no difference between new and MRI methods based on the p-value (e.g. I had insufficient evidence to reject the null), I would be able to use this new renal volume measurement as an alternative.

Question 2

To assess the effectiveness of a new drug for treating an infectious disease, patients with the disease at the baseline are randomized to a new drug or a placebo treatment. The subjects are tested for disease improvement based on culture result in blood/urine at the end of study (at 8 week). Negative culture suggests improvement. The SAS dataset diseaseX can be found on Canvas.

Variables:

- Outcome (week8_result): 1=Culture Negative, 0=Culture Positive
- Covariates: treat (test or control), age, gender (1=Male 2=Female), BMI, antibiotic resistance (1=resistant, 0=not resistant).

```
# Data intake
disease <- read_sas("diseasex.sas7bdat")

# Preview data for assignment
head(disease)
```

```
## # A tibble: 6 x 7
##   PatientId treat   week8_result   age   BMI Gender antibiotic
##       <dbl> <chr>          <dbl> <dbl> <dbl>  <dbl>      <dbl>
```

```
## 1      1001 Test              1  25.8    25      2          1
## 2      1002 Control           1  33.1    17      1          0
## 3      1003 Test              0  19.4    20      1          0
## 4      1005 Test              1  24.2    25      1          0
## 5      1006 Test              1  19.9    15      2          0
## 6      1007 Control           1  49      25      1          0
```

The goal of the experiment is to assess the treatment effectiveness by culture conversion. *Using descriptive results, describe the baseline characteristics for those who received the treatment vs those who did not. In your Table, you need to describe age and BMI as continuous variables as well as binary variables by creating two new variables such Age_cat: Age < 30/Age >= 30 ; and BMI_cat: BMI < 18.5 / BMI >= 18.5. Please provide an informative Table by including mean, SD etc. (See Table 1 style and format from article by Royster et al located on the canvas).*

```
# Data
df <- disease

# Create age variables
df$age_cat <- 0
df$age_cat[df$age >= 30] <- 1

# Create BMI variable
df$BMI_cat <- 0
df$BMI_cat[df$BMI >= 18.5] <- 1

# Identify which variables are factors
attr(df$age, "label") <- "Age"
df$age_cat %<>% factor(., levels = c(0,1), labels = c("<30", ">= 30"))
attr(df$age_cat, "label") <- "Age Category"
df$BMI_cat %<>% factor(., levels = c(0,1), labels = c("<18.5", ">= 18.5"))
attr(df$BMI_cat, "label") <- "BMI Category"
df$Gender %<>% factor(., levels = c(1,2), labels = c("Male", "Female"))
df$antibiotic %<>% factor(., levels = c(0,1), labels = c("Not Resistant", "Resistant"))
attr(df$antibiotic, "label") <- "Abx Resistance"

# Table based on those that were treated versus not treated
compareGroups(treat ~ age + age_cat + Gender + BMI + BMI_cat + antibiotic,
              data = df) %>%
    createTable(., show.p.overall = TRUE) %>%
    export2md(., size = 8,
                caption = "Characteristics in Test versus Control Groups for Disease")
```

Table 1: Characteristics in Test versus Control Groups for Disease

	Control N=86	Test N=83	p.overall
Age	35.1 (12.4)	32.1 (10.7)	0.098
Age Category:			0.400
<30	39 (45.3%)	44 (53.0%)	
>= 30	47 (54.7%)	39 (47.0%)	
Gender:			0.983
Male	54 (62.8%)	51 (61.4%)	
Female	32 (37.2%)	32 (38.6%)	
BMI	20.8 (3.22)	21.0 (3.69)	0.726
BMI Category:			0.255
<18.5	24 (27.9%)	16 (19.3%)	
>= 18.5	62 (72.1%)	67 (80.7%)	

Table 1: Characteristics in Test versus Control Groups for Disease *(continued)*

	Control	Test	p.overall
Abx Resistance:			0.865
Not Resistant	76 (88.4%)	75 (90.4%)	
Resistant	10 (11.6%)	8 (9.64%)	

Test the hypothesis that the conversion rate is the same between test and placebo group. This should include: specifying the hypotheses (H0, Ha) and statistical method, results in a Table and conclusion.

```
# Data from above
df <- df

# THis is likely a chi-square test as the DV is either culture + or -
# The IV is treatment or not
# Question becomes if treatment changes rate of cure
tConversion <- chisq.test(df$week8_result, df$treat)

print(tConversion)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$week8_result and df$treat
## X-squared = 0.7, df = 1, p-value = 0.4
```

The H_0 is that there is no difference between treatment groups for culture changing to negative. The H_1 is that the treatment group will have a different / higher probability of having cure. The appropriate test is a *chi-square test*. We will set the $\alpha = 0.05$ for this test.

In this case, the p-value is 0.398. This is above the set α, and we thus have insufficient evidence to reject H_0. We can conclude there is no significant difference in the rate of cure/conversion in those that are treated.

Test the hypothesis that mean BMI is different for treatments (variable treat). This should include: specifying the hypotheses (H0, Ha) and statistical method, results and conclusion.

```
# Asked to assess if the BMI is different between treatment groups

# Data
df <- df

# Simple t-test
tBMI <- t.test(df$BMI ~ df$treat)

# Show output
print(tBMI)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$BMI by df$treat
## t = -0.4, df = 200, p-value = 0.7
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.243  0.868
## sample estimates:
## mean in group Control    mean in group Test
```

```
##                 20.8                 21.0
```

The H_0 is that there is no difference in means between treatment groups for BMI. The H_1 is that the mean BMI between treatment groups is different. The appropriate test is a *two-sample t-test*. We will set the $\alpha = 0.05$ for this test.

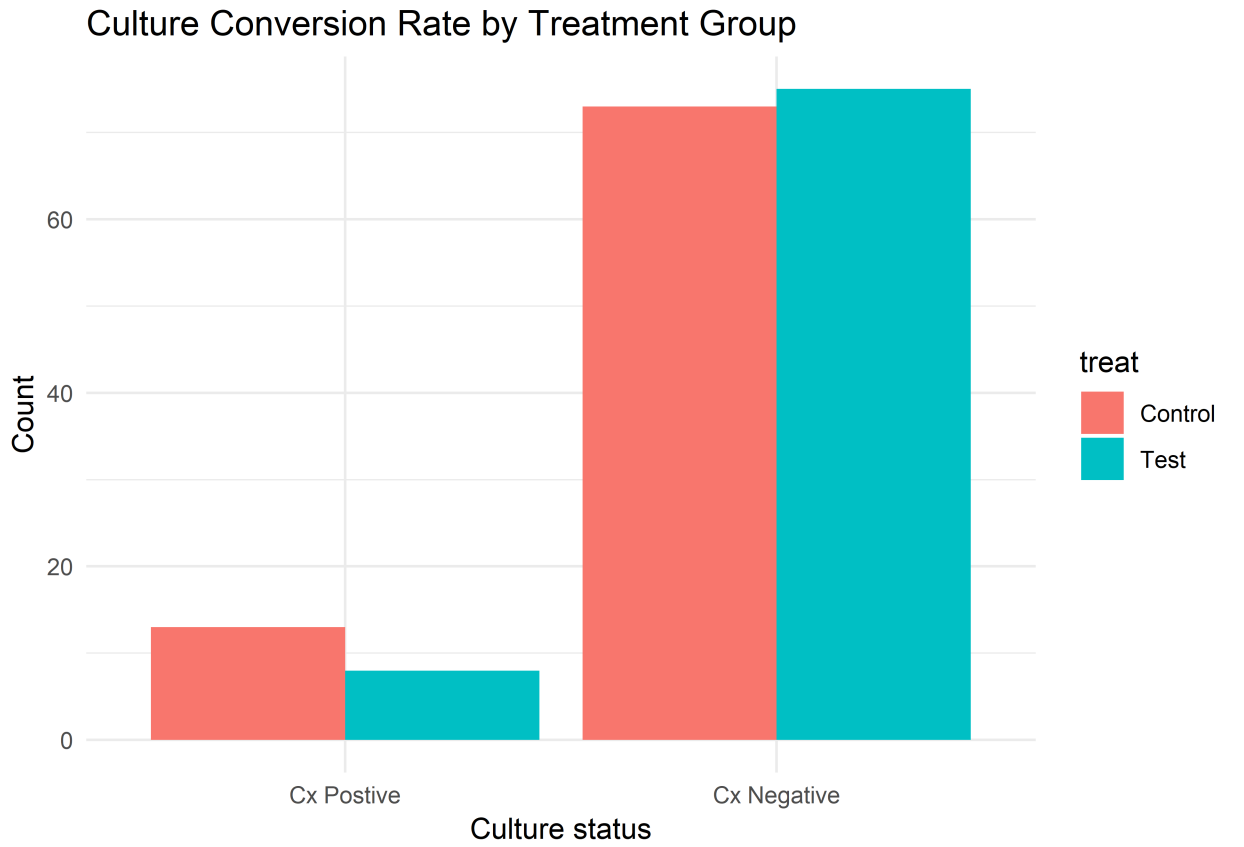
In this case, the p-value is 0.726. This is above the set α, and we thus have insufficient evidence to reject H_0. We can conclude there is no significant difference in BMI between treatment groups.

Draw a bar graph to display your results for task 2b. (See SAS code for an example of barchart on canvas).

```r
# Data
df <- df

# Make pretty the outcome variable
df$week8_result %<>% factor(., levels = c(0,1), labels = c("Cx Postive", "Cx Negative"))
attr(df$week8_result, "label") <- "Cx Conversion"

# Barchart
ggplot(df, aes(x = week8_result, fill = treat)) +
    geom_bar(position = "dodge") +
    theme_minimal() +
    labs(
        title = "Culture Conversion Rate by Treatment Group",
        x = "Culture status",
        y = "Count"
    )
```



Perform a univariate analysis to determine the factors that may be associated with the response variable

(negative culture). A possible way of presenting the results is as follows.

```
# Data
df <- df

# Make pretty table
attr(df$treat, "label") <- "Treatment Group"

# Table based on those that were treated versus not treated
compareGroups(week8_result ~ treat + age_cat + Gender + BMI_cat + antibiotic,data = df) %>%
    createTable(., show.p.overall = TRUE, show.n = TRUE, show.ratio = TRUE, show.p.mul = TRUE, show.des
    export2md(., size = 8, caption = "Analysis of Culture Conversion by Covariates")
```

Table 2: Analysis of Culture Conversion by Covariates

	Cx Postive N=21	Cx Negative N=148	OR	p.ratio	p.overall	N
Treatment Group:					0.398	169
Control	13 (61.9%)	73 (49.3%)	Ref.	Ref.		
Test	8 (38.1%)	75 (50.7%)	1.65 [0.65;4.46]	0.293		
Age Category:					0.704	169
<30	9 (42.9%)	74 (50.0%)	Ref.	Ref.		
>= 30	12 (57.1%)	74 (50.0%)	0.75 [0.29;1.91]	0.552		
Gender:					0.238	169
Male	16 (76.2%)	89 (60.1%)	Ref.	Ref.		
Female	5 (23.8%)	59 (39.9%)	2.07 [0.76;6.76]	0.162		
BMI Category:					0.785	169
<18.5	4 (19.0%)	36 (24.3%)	Ref.	Ref.		
>= 18.5	17 (81.0%)	112 (75.7%)	0.75 [0.20;2.22]	0.627		
Abx Resistance:					0.002	169
Not Resistant	14 (66.7%)	137 (92.6%)	Ref.	Ref.		
Resistant	7 (33.3%)	11 (7.43%)	0.16 [0.05;0.51]	0.003		