Question 1

Below I will describe a number of different study designs. You will be asked to

(1) Identify the null hypothesis
(2) Identify the correct statistical test for this hypothesis

Part A: For this study, I am interested in determining if a student's major (Humanities/STEM/Social Sciences) is associated with their final exam score in STA-209.

Major: categorical variable

Score: continuous

Null H0: no association/no difference in mean exam score

Test: ANOVA or t-test with Bonferonni correction

Part B: A two-day workshop for learning basic R has been created, where attendees are tested in their R skills both prior to the workshop and after the workshop has been completed. For each of these tests, a numeric score is given. We wish to determine whether or not the workshop has been effective in improving the R skill of the attendees.

Score: continuous

Category: before and after

H0: No difference before and after session

Test: paired t-test

Part C: Binge drinking is defined as a pattern of drinking that involves consuming 5 or more standard drinks within 2 hours. Respondents of a survey were asked for their sex and whether or not they have engaged in binge drinking more than twice in the previous week. We wish to determine whether or not there is a difference in binge drinking patterns between men and women.

Sex: Categorical

Binge Drinking (yes/no): Categorical (2 categories)

Test: Difference in proportion

H0: No difference between men and women

or

Test: χ^2 test of independence

H0: no association between sex and binge drinking

Question 2

Cocaine addicts have been reported to have a significant depletion of stimulating neurotransmitters and thus continue to use cocaine to avoid feelings of depression and anxiety. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo (lithium is the standard treatment for cocaine addiction). One third of the subjects were randomly assigned to each treatment group with the following results:

	Relapse	No Relapse
Desipramine	10	14
Lithium	18	6
Placebo	20	4

Part A: What type of plot would you use to visually display these results

Stacked or conditional bar chart (treatment on x axis, number on y axis, color for relapse)

Part B: Describe the null hypothesis of this study and construct a table of expected counts under the assumption of the null hypothesis.

χ^2 test of independence, null is that there is no association between treatment and outcome

```r
m <- matrix(c(10,14,18,6,20,4), nrow = 3, byrow = TRUE)
rownames(m) <- c("Desipramine", "Lithium", "Placebo")
colnames(m) <- c("Relapse", "No Relapse")
m %>% addmargins()
```

```
##             Relapse No Relapse Sum
## Desipramine      10         14  24
## Lithium          18          6  24
## Placebo          20          4  24
## Sum              48         24  72
```

```r
# Under the null
# expectred count of desipramine and relpase is product of probabilities
(24/72)*(48/72) * 72
```

```
## [1] 16
```

```r
(24*48)/72
```

```
## [1] 16
```

```r
cc <- chisq.test(m)
## Observed counts
m
```

```
##             Relapse No Relapse
## Desipramine      10         14
## Lithium          18          6
## Placebo          20          4
```

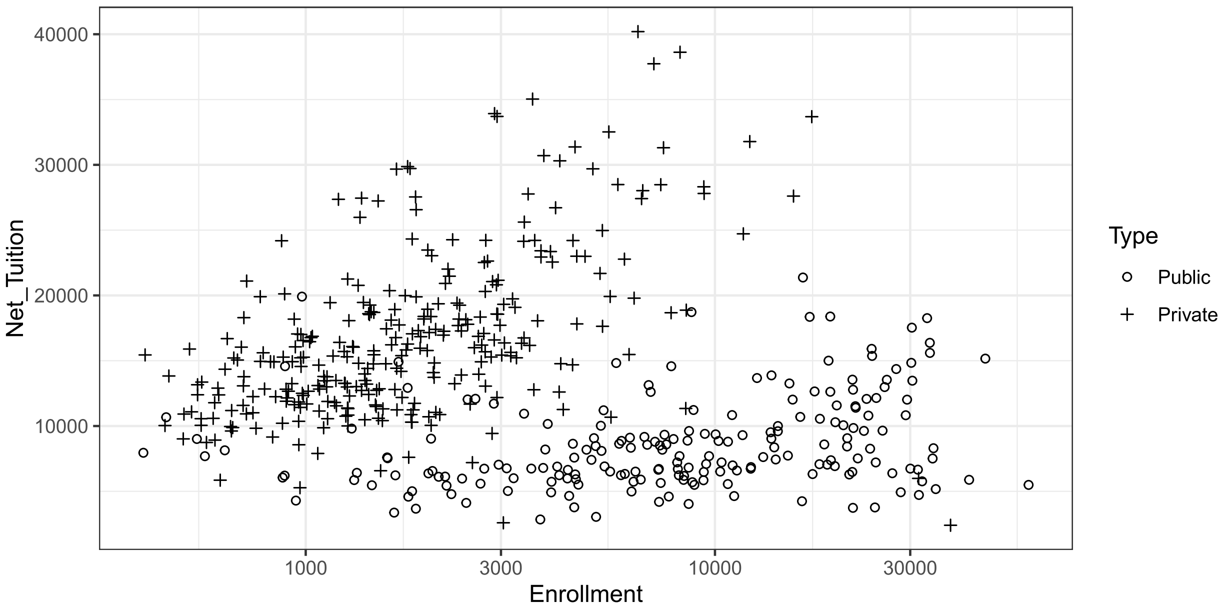
```r
# expected counts
cc$expected
```

```
##             Relapse No Relapse
## Desipramine      16          8
## Lithium          16          8
## Placebo          16          8
```

Part C: The χ^2 test statistic has a value of $\chi^2 = 10.5$ with a p-value of p-val $= 0.0052$. Based on this, what conclusion would you reach if testing at the $\alpha = 0.05$ level?

We would reject since $p < \alpha = 0.05$

Question 3



Model 1:

```
lm(formula = Net_Tuition ~ Enrollment, data = college)

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept) 14225.3137   272.8034    52.1 <0.0000000000000002 ***
Enrollment     -0.0820     0.0265    -3.1               0.002 **

Residual standard error: 7180 on 1093 degrees of freedom
Multiple R-squared:  0.00869,   Adjusted R-squared:  0.00779
F-statistic: 9.58 on 1 and 1093 DF,  p-value: 0.00201
```

Model 2:

```
lm(formula = Net_Tuition ~ Enrollment + Type, data = college)

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  5746.1019   377.2481    15.2 <0.0000000000000002 ***
Enrollment      0.2533     0.0239    10.6 <0.0000000000000002 ***
TypePrivate 10808.5970   398.6370    27.1 <0.0000000000000002 ***

Residual standard error: 5550 on 1092 degrees of freedom
Multiple R-squared:  0.408, Adjusted R-squared:  0.406
F-statistic:  376 on 2 and 1092 DF,  p-value: <0.0000000000000002
```

Part A: For this part, consider **Model 1** from above. What is the null hypothesis in linear regression? Based on the summary output, how would you describe the relationship between enrollment and tuition?

H0: no linear relationship between enrollment and net tuition

Summary: Based on summary, evidence to suggest that there *is* a linear relationship so we reject H0

Describe: seems that negative linear relationship

Part B: Now consider **Model 2**, which includes an indicator for whether or not a college is private. How would you interpret the intercept in this model? Is this a meaningful value in this model?

Intercept: reference var is public school

Interpretation: 5746 is estimated tuition for public school when enrollment = 0

Meaningful?: No.

Part C: Compare the coefficient for Enrollment between **Model 1** and **Model 2**. What has changed? In other words, what impact has adding an indicator for Private had on this value, and why did it result in such a drastic change?

Big change: from negative to positive

Why: because now considering within public/private groups

Name for phenomena (not on test): Simpson's Paradox