```
---
title: "Chi-Sq Tests vs Proportion Tests vs Logistic Regression: Binary Responses and
Binary/Categorical Predictors"
author: [Your Names Here]
output: html_document
---

```{r, include=F}
library(Stat2Data); library(tidyverse); library(ggformula); library(mosaic)
data("MathPlacement")
```


PART I. PREDICTORS WITH 2 CATEGORIES (BINARY PREDICTORS)

Students at a small liberal arts college took a placement exam prior to entry in order to provide them guidance when selecting their first math course. The dataset **MathPlacement** (in Stat2Data) contains the placement scores for 2696 students, along with whether they took the recommended course, and several admissions variables (GPA, SAT score, etc.). We want to use this data to decide how well the math placement process is working. If they take the recommended course, do they succeed (where "success" is defined as a grade of "B" or above)?

We already investigated this question using Logistic regression in the "Multiple Logistic Regression" activity. But **there are other statistical methods/tests** that we could use to answer this question.

Some data cleaning: Before we go further, I'm going to remove the students who are missing values for "CourseSuccess".
```{r}
MathPlacement <- filter(MathPlacement, !is.na(CourseSuccess))
```


A. Logistic Regression

Summarize your results from part D of "Multiple Logistic Regression". If students take the recommended course, are they signficiantly more likely to succeed than those who don't take the recommendation? *How much* more likely to succeed?

B. 2-Sample Proportion Test

We have binary response variable (CourseSuccess) and a binary explanatory variable (RecTaken). Thus, we could investigate the relationship between them using a 2-sample proportion test.

1. Hypotheses. If $\pi_{RecTaken}$ is the probability of success for those who took the recommended course and $\pi_{NotRecTaken}$ is the probability of success for those who didn't take the recommended course, the hypotheses are (write in words and in symbols):

2. Assumptions/Conditions. A 2-sample proportion test has the same conditions as the one-sample proportion test, applied to both samples.

- number of successes for both groups is at least 10

- number of failures for both groups is at least 10

- samples are independent of each other

Are the conditions met in this case? Make sure you discuss all 3 conditions. (In this case, our two groups are: 1) those who took the recommended course; and 2) those who didn't take the recommended course.)
```

3. Using R for Inference. Just as in the one-sample case, `prop.test()` will find the p-value and confidence interval.

Syntax:

`prop.test(response var ~ explanatory var)`

Use prop.test to test the the hypotheses in #1.
```{r}

```

Notice that the "sample estimates" at the bottom of the output are the estimates of the probability of course *failure* as opposed to course *success*. This is because R is choosing "CourseSuccess=0" to be "success". Of course, we want "CourseSuccess=1" to be the success, so we can change the code like this:
```{r}
prop.test(CourseSuccess~RecTaken, data=MathPlacement, success=1) #the "success" argument tells R what the value of the response variable should be considered a success
```

If you have a **summary 2-way table** (as opposed to untabulated data), you use this alternative syntax

`prop.test(x=c(successes_in_group1,successes_in_group2),
n=c(total_#_in_group1,total_#_in_group1))`
```{r}
prop.test(x=c(396,1045), n=c((396+247),(1045+441))) #396 successes among those (396+247) who didn't take the recommended course; 1045 successes among those (1045+441) who did take the recommended course
```

4. Conclusion Make a conclusion about H_0 in the context of the problem.

5. Confidence intervals. `prop.test()` will give you the CI for the difference between the two proportions, as long as the "alternative" is two-sided (which is the default). Report and interpret the 95% confidence interval for the difference in success rates between those who did and didn't take the recommended course.

C. Chi-Squared Test for Association

Another way of testing whether the two groups have significantly different probabilities of success is to conduct a Chi-square Test for a 2x2 table. You may have seen this in your Intro Stats class, where it was called the "Test for Independence", the "Test for Association", or the "Test for Homogeneity".

1. Hypotheses. 2 ways to state the hypotheses...

H_0: the two variables (CourseSuccess and RecTaken) are independent vs. H_a: not independent

OR

H_0: the probability of success is the same for both populations (those who took the recommendation and those who didn't) vs. H_a: the proportions are not the same

2. Assumptions/Conditions. The only conditions for the Chi-square Test are that:

- all expected cell counts are at least 5

- the sample (or samples) are randomly selected, or representative of the population of interest

You've already thought about the second condition; we'll discuss whether the first condition is met later.

3. Test Statistic & Distribution. Consider the table below.

```{r}
tally(CourseSuccess ~ RecTaken, data=MathPlacement)
```

Under the null hypothesis of the Chi-square Test, we'd expect our two groups (those who took the recommendation and those who didn't) to success at about the same rate. Thus, the frequencies in each cell should correspond to the overall proportions of successes and failures for the entire sample. The Chi-square test statistic, χ^2, compares the observed cell counts to the expected cell counts under this assumption.

$$\chi^2 = \sum{\frac{(Observed - Expected)^2}{Expected}}$$

where the sum is taken over all the cells in the table (in this case, four).

Under the null hypothesis, we'd expect χ^2 to be quite small, since Observed and Expected should be nearly equal. How far apart these numbers are tells us how unusual our data is; thus, big values of χ^2 are an indication that H_0 is false. Under H_0, χ^2 follows a chi-squared distribution with 1 degree of freedom.

4. Using R to find the test stat & p-value.

Syntax: First, you must create a 2-way table of successes and failures. Then you run `chisq.test` on that table.

`chisq.test(table)`

Create the 2-way table below (using tally, xtabs, or table), then run `chisq.test` on the table, and save the test as `rec.chisq`.

```{r}
table2 <- #make your 2-way table here
rec.chisq <- chisq.test(table2); rec.chisq
```

Report the test statistic and p-value here:

5. Back to the assumptions. The `chisq.test()` object `rec.chisq` contains the expected cell counts in `rec.chisq$expected`: you can check that they are all at least 5. As a nice bonus, however, `chisq.test()` will actually warn you during the test if one or more of the expected values is less than 5.

Is this condition met in this case?

6. Conclusion. Make a conclusion about H_0 in the context of the problem.

D. Synthesis

We've seen three methods of looking at the relationship between a binary response (CourseSuccess) and a binary predictor (RecTaken).

1. What are the similarities/difference in the conclusions between these three methods?

2. Discuss the pros and cons of each method. Think about what information the different

methods provide, what conclusions we can make with each, and the differences between assumptions/conditions. Are there situations where you feel one or the other method would be best?

PART II. CATEGORICAL PREDICTORS (with > 2 categories)

We saw that those who didn't take the recommended course are *significantly* less likely to succeed than those who do take the recommended course. But of course, these students are being placed into different courses depending on their placement tests. Is the actual course they end up taking a significant factor in whether they succeed? Let's use the chi-sq test for association and logistic regression to investigate this.

A. Data Collection
---

1. Make a two-way table of CourseSuccess by Course.

2. The success rate in Math109 is:

The success rate in Math120 is:

The success rate in Math220 is:

3. Based only on the proportions above, does it seem that there is a relationship between success and course?

4. You see that very few students took Math114 or Math398. I'm going to filter those students out for the rest of the analysis...

```{r}
MathPlacement <- filter(MathPlacement, Course %in% c(109,117,120,122,126,128,210,220))
```

B. ~~2-Sample Proportion Test~~

We can't do a 2-sample proportion test here because we have more than 2 groups!

C. Chi-Squared Test: Test for Association
---
1. Hypotheses

2. Conduct the Chi-sq test and save it as `course.chisq`. Report the Test Stat and p-value:

3. Assumptions of the Chi-sq test Is this condition that all expected cell counts are at least 5 met in this case?

4. Conclusion. Make a conclusion about H_0 in the context of the problem.

5. ...Can we say more?

Since we have rejected the null hypothesis, it would nice to pinpoint which cells are

significantly different than expected. In this way, we can say something more than just "there is an association".

We can do this by finding the "contribution to the chi-sq" for each cell. That is, we calculate the amount that each cell "contributed" to the chi-sq test statistic you found above. The larger the contribution, the farther that cell's observed value was from what was expected (the closer to 0, the closer to what was expected).

The contributions are equivalent to the "residuals" from each cell, which you can find in `course.chisq$residuals`. Cells with larger absolute residuals have larger differences between what's observed and what's expected, and thus have "contributed" most to the rejection of the null.

a. In this case, what three courses have the largest contribution?

b. A negative "residual" means that cell had fewer successes than what was expected under the null; a positive "residual" means that cell had more successes than expected. Using this information, what can you say about the success/failure rate in those three courses with the largest contribution, and how it compares to what was expected under the null?

D. Logistic Regression

You can absolutely use a logistic regression model to come to a similar conclusion as we did in Part C above. However, the output looks a little different than it did in the case of a binary predictor variable (or a numerical predictor)...

First, note that the values of `Course` are numerical. So `glm()` will think that `Course` is a numerical variable unless you force it to be a factor (categorical variable). We've done this before: just use `as.factor(Course)` in the glm call.

```{r}
course.log <- glm(CourseSuccess~as.factor(Course), family=binomial, data=MathPlacement)
```

1. The Fitted Logistic Regression Model. Write down the fitted model here:

2. Interpretation of Intercept. What does the intercept of -0.0408 mean? Interpret this in context.

3. All the course have positive coefficient estimates. What does this mean? How do we interpret a positive coefficient?

4. Interpretation of slope/odds ratio.

The coefficient of Course=120 is _____, which means the odds ratio for that group is _____.
Interpret this odds ratio:

The coefficient of Course=220 is _____, which means the odds ratio for that group is _____.
Interpret this odds ratio:

5. What does it mean that only some courses are labelled as statistically significant? Whom are these groups significantly different from? (That is, who are they being compared to?)

6. Confidence interval for the slope/odds ratio.

Use confint() to calculate a 90% confidence interval for the slope of Course=220:

Interpret the CI in terms of odds ratios:

7. Hypotheses: drop-in-deviance test for model utility

Ho: Course is NOT a useful variable in predicting CourseSuccess vs. Ha: Course is useful in predicting CourseSuccess

8. Test statistic & p-value. The test statistic is as usual, the difference between the null deviance and the residual deviance, which follows a chi-sq distribution.

```{r}
anova(course.log, test="Chisq")
```

Report the test stat and the p-value:

9. Conclusion.

E. Synthesis

We've seen two methods of looking at the relationship between a binary response (CourseSuccess) and a categorical predictor (Course).

1. How does your p-value and conclusion in part D (logistic regression) compare to your p-value and conclusion in part C (chi-sq test)?

2. Discuss the pros and cons of the chi-sq test for association vs. the logistic model when using a categorical predictor. Think about what information the different methods provide, what conclusions we can make with each, and the differences between assumptions/conditions. Are there situations where you feel one or the other method would be best?