ST 502 Final Project

Eric Warren

2024-04-18

Contents

1   Introduction

In this report we are going to use something called the NcNemar's Test. This test looks at a 2 x 2 contingency table and checks the marginal homogeneity of two opposite variables (and in the case we are going to look at two competing drugs). The test requires one categorical variable with two competing categories (in our case two different, competing drugs) and one independent variable with two dependent groups (in our case the subjects participating in the study who are split into two different groups). The NcNemar's Test is typically used in experimental design which looks at how subjects react a treatment versus some kind of control group or two competing treatments to see which one subjects react better to. The types of data being collected will be in the form of a matched pairs design where the subjects will experience both options to compare results on how well their response is. Then we will perform the analysis doing a hypothesis test saying our null hypothesis is how both treatments have the same effect on subjects and the alternative is how treatment 1 is either better, worse, or not equal to treatment 2. This report will explain more in depth how this NcNemar's Test is used in practice. Also note throughout this report there will be R code included within it to follow along and be able to replicate different things that are examined.

2   Analyzing our Data

Here we are going to consider a dataset on 250 subjects in which acid reflux is treated by a drug. The response is either success (reflux stopped) or failure (reflux still present). Half of the subjects are randomly selected to use drug A the first day they have reflux and drug B the second day they have reflux. The others were assigned to use drug B first, then drug A. We want to know if the drugs have a different probability of successful relief. The outcomes for the relief given by the drugs is given in the table below.

Drug	Success	Failure	Total
A	100	150	250

Drug	Success	Failure	Total
B	125	125	250
Total	225	275	500

However, we wouldn't want to analyze this data with a test of homogeneity because the observations in the table are not independent. In this situation, we really only have 250 subjects, which means we really do not have 500 independent observations. Instead, we can look at a table of concordant and discordant pairs.

	Drug A Success	Drug A Failure	Total
Drug B Success	85	15	100
Drug B Failure	40	110	150
Total	125	125	250

The diagonal now represents the observations that were the same for an individual (concordant pairs). The off diagonals represent observations that were not the same for individuals (discordant pairs). To test whether or not the drugs have a different effect, we can now apply McNemar's test on the table of concordant/discordant pairs.

Our null hypothesis is that there is no relationship between drug and relief with the alternative hypothesis being that the null hypothesis is not true in some way (we say that the cell probabilities are "free" – other than the sum to 1 as the constraint). Please note here to do the test we will use Pearson's Chi-Square test statistic (call it X^2), which can be found by $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$ where where n_{ij} represents the observed count in cell ij. (You'll show this later on.) Our reference distribution is a χ_1^2. This test can be done in **R** using the `mcnemar.test()` function from the `stats` package.

First, we are going to find the value of the test statistic, the rejection region, and the p-value of the test using R. Note here we will say our significance level $\alpha = 0.05$, our test statistic is $X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$, our rejection region is $\{t_{obs} : t_{obs} > \chi_{1,.05}^2\}$, and our p-value is $P(\chi_1^2 \geq X^2)$.

```
# Get the n21 and n12 values
n21 <- 40
n12 <- 15


# Get the Pearson Chi-sq test stat
chi_sq_test_stat <- ((n12 - n21) ** 2) / (n12 + n21)


# Get the critical value to compare for our rejection region
critical_value <- qchisq(0.95, 1)

# Find the p-value
p_value <- pchisq(chi_sq_test_stat, 1, lower.tail = F)
```

In the case of our experiment and if we say that $\alpha = 0.05$, we can say our test statistic $X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = 11.3636$, our rejection region of $\{t_{obs} : t_{obs} > \chi_{1,.05}^2\}$ becomes $\{t_{obs} : t_{obs} > 3.8415\}$, and our p-value shown by $P(\chi_1^2 \geq X^2)$ is 0.0007. Based on these results, we can say that we have statistically significant evidence to reject the null hypothesis which said there was no relationship between drug and relief. That means with statistically significant evidence we can say a relationship between drug and relief is present. We are going to double check this using our `mcnemar.test()` function to make sure our results are the same.

```r
# Make the matrix
df <- matrix(c(85, 40, 15, 110), nrow = 2, ncol = 2)

# Do the mcnemar test
test <- mcnemar.test(df, correct = FALSE)
```

Using the McNemar's Test, we get our test-statistic to be 11.3636 with 1 degree of freedom and a p-value of 0.0007. This matches up with the results we got from before our rejecting our null hypothesis and having statistically significant evidence to say a relationship between drug and relief is present.

3   Deriving the Test

Now that we saw how to use our McNemar's Test let us derive some important parts that can get us our final solution. First let us look at something we assumed in our results that $\pi_{1.} = \pi_{.1}$ and $\pi_{2.} = \pi_{.2}$ are equivalent to $\pi_{12} = \pi_{21}$. Note before we solve this that $\pi i.$ is the sum of all values in the i row and $\pi.j$ is the sum of all values in the j column.

First let us look at $\pi_{1.} = \pi_{.1}$. This is the same as saying the sum of all the π values in the first row are equal to the sum of the π values in the first column. In this case we can see that we can rewrite $\pi_{1.} = \pi_{.1}$ as $\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21} <=> \pi_{12} = \pi_{21}$. Now let us look at $\pi_{2.} = \pi_{.2}$. This is the same as saying the sum of all the π values in the second row are equal to the sum of the π values in the second column. In this case we can see that we can rewrite $\pi_{2.} = \pi_{.2}$ as $\pi_{21} + \pi_{22} = \pi_{12} + \pi_{22} <=> \pi_{21} = \pi_{12} <=> \pi_{12} = \pi_{21}$, which is the same as what we had before with our first case. Now we can see that our two cases we essentially have an equivalent procedure of $\pi_{12} = \pi_{21}$ as the test, so we have shown that $\pi_{1.} = \pi_{.1}$ and $\pi_{2.} = \pi_{.2}$ are equivalent to $\pi_{12} = \pi_{21}$.

Next under this null restriction on our multinomial, we are going to derive the maximum's for $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$. The steps for this are outlined below. **COME BACK TO THIS PART AFTER REVIEWING!!!**

Now that we have found the maximums for our $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ values, we want to show the form of our likelihood ratio test which in this case is $-2\ln(\frac{L(\tilde{\pi_{11}},\tilde{\pi_{12}},\tilde{\pi_{21}},\tilde{\pi_{12}})}{L(\hat{\pi_{11}},\hat{\pi_{12}},\hat{\pi_{21}},\hat{\pi_{12}})}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} Obs_{ij}\ln(\frac{Obs_{ij}}{Exp_{ij}})$. Please note before solving this that $\hat{\pi}_{ij,MLE} = \frac{Y_{ij}}{n}$. Below are the written steps out for how to solve this.

- Find our observed likelihood ratio test. In this case we can say that $\Lambda = \frac{L(\tilde{\pi_{11}},\tilde{\pi_{12}},\tilde{\pi_{21}},\tilde{\pi_{12}})}{L(\hat{\pi_{11}},\hat{\pi_{12}},\hat{\pi_{21}},\hat{\pi_{12}})} = \frac{\Pi_{i=1}^{2}\Pi_{j=1}^{2}\tilde{\pi}_{ij}^{y_{ij}}}{\Pi_{i=1}^{2}\Pi_{j=1}^{2}\hat{\pi}_{ij}^{y_{ij}}} = \Pi_{i=1}^{2}\Pi_{j=1}^{2}(\frac{\tilde{\pi}_{ij}}{\hat{\pi}_{ij}})^{y_{ij}}$

- Now plug in $\Lambda = \Pi_{i=1}^{2}\Pi_{j=1}^{2}(\frac{\tilde{\pi}_{ij}}{\hat{\pi}_{ij}})^{y_{ij}}$ into $-2\ln(\Lambda)$. $-2\ln(\Lambda) = -2\ln((\frac{\tilde{\pi}_{ij}}{\hat{\pi}_{ij}})^{y_{ij}}) = -2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}(\ln(\frac{\tilde{\pi}_{ij}}{\hat{\pi}_{ij}})) = -2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}(\ln(\tilde{\pi}_{ij})-\ln(\hat{\pi}_{ij})) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}*-1(\ln(\tilde{\pi}_{ij})-\ln(\hat{\pi}_{ij})) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}(\ln(\hat{\pi}_{ij})-\ln(\tilde{\pi}_{ij})) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}(\ln(\frac{\hat{\pi}_{ij}}{\tilde{\pi}_{ij}}))$

- Here note that $Obs_{ij} = y_{ij}$ and $Exp_{ij} = n\tilde{\pi}_{ij}$. We can also plug in that $\hat{\pi}_{ij} = \frac{y_{ij}}{n}$ since $\hat{\pi}_{ij,MLE} = \frac{Y_{ij}}{n}$. Let us do that with what we have of our last part getting us $-2\ln(\frac{L(\tilde{\pi_{11}},\tilde{\pi_{12}},\tilde{\pi_{21}},\tilde{\pi_{12}})}{L(\hat{\pi_{11}},\hat{\pi_{12}},\hat{\pi_{21}},\hat{\pi_{12}})}) = -2\ln(\Lambda) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}\ln(\frac{\hat{\pi}_{ij}}{\tilde{\pi}_{ij}}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}\ln(\frac{\frac{y_{ij}}{n}}{\tilde{\pi}_{ij}}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}y_{ij}\ln(\frac{y_{ij}}{n\tilde{\pi}_{ij}}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}Obs_{ij}\ln(\frac{Obs_{ij}}{Exp_{ij}})$.

From all of these steps we have shown that $-2\ln(\frac{L(\tilde{\pi_{11}},\tilde{\pi_{12}},\tilde{\pi_{21}},\tilde{\pi_{12}})}{L(\hat{\pi_{11}},\hat{\pi_{12}},\hat{\pi_{21}},\hat{\pi_{12}})}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2}Obs_{ij}\ln(\frac{Obs_{ij}}{Exp_{ij}})$. From getting this result, we also need to prove that this gives us a χ_1^2 distribution. How did we get the one degree of freedom? Note that we calculate the number of degrees of freedom by the formula $df = dim(\Omega) - dim(\omega_0)$ which is just saying the number of free parameters in the alternative distribution minus the number of free

parameters in the null distribution. For the alternative distribution, we can select any value for each of the π_{ij} values and since we have 4 different parameters with no constraints our $dim(\Omega) = 4$. For our null distribution, we have a similar situation except we assumed that $\pi_{12} = \pi_{21}$. This means our π_{11} and π_{22} are free parameters but only one of π_{12} and π_{21} is free. This means we only truly have 3 free parameters under the null distribution and our $dim(\omega_0) = 3$. So this is why we have a χ_1^2 distribution because our degrees of freedom of $dim(\Omega) - dim(\omega_0) = 4 - 3 = 1$.

Lastly for our tests to hold we said we could use the Pearson's chi-square test statistic instead of this likelihood ratio test statistic since they are asymptotically equivalent. However, we made the assumption that the Pearson's chi-square test statistic is $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$, so we need to prove this. To do this proof there are some things we need to know. First, we must remember that $Obs_{ij} = y_{ij} = n_{ij}$ in our case. We also need to know that $\tilde{\pi}_{ij} = \frac{n_{ij}}{n}$ for $\tilde{\pi}_{11}$ and $\tilde{\pi}_{22}$. For $\tilde{\pi}_{12} = \tilde{\pi}_{21} = \frac{n_{12} + n_{21}}{2n}$ under the null distribution. Lastly, we must state that $Exp_{ij} = n * \tilde{\pi}_{ij}$. Knowing all of this we can prove that $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$.

Here we can say that $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} = \frac{(Obs_{11} - Exp_{11})^2}{Exp_{11}} + \frac{(Obs_{12} - Exp_{12})^2}{Exp_{12}} + \frac{(Obs_{21} - Exp_{21})^2}{Exp_{21}} + \frac{(Obs_{22} - Exp_{22})^2}{Exp_{22}} = \frac{(n_{11} - n(\frac{n_{11}}{n}))^2}{n(\frac{n_{11}}{n})} + \frac{(n_{12} - n(\frac{n_{12} + n_{21}}{2n}))^2}{n(\frac{n_{12} + n_{21}}{2n})} + \frac{(n_{21} - n(\frac{n_{12} + n_{21}}{2n}))^2}{n(\frac{n_{12} + n_{21}}{2n})} + \frac{(n_{22} - n(\frac{n_{22}}{n}))^2}{n(\frac{n_{22}}{n})} = \frac{(n_{11} - n_{11})^2}{n_{11}} + \frac{(n_{12} - \frac{n_{12} + n_{21}}{2})^2}{\frac{n_{12} + n_{21}}{2}} + \frac{(n_{21} - \frac{n_{12} + n_{21}}{2})^2}{\frac{n_{12} + n_{21}}{2}} + \frac{(n_{22} - n_{22})^2}{n_{22}} = 0 + \frac{(\frac{n_{12} - n_{21}}{2})^2}{\frac{n_{12} + n_{21}}{2}} + \frac{(\frac{n_{21} - n_{12}}{2})^2}{\frac{n_{12} + n_{21}}{2}} + 0 = \frac{\frac{n_{12}^2}{4} - \frac{2n_{12}n_{21}}{4} + \frac{n_{21}^2}{4}}{\frac{n_{12} + n_{21}}{2}} + \frac{\frac{n_{12}^2}{4} - \frac{2n_{12}n_{21}}{4} + \frac{n_{21}^2}{4}}{\frac{n_{12} + n_{21}}{2}} = \frac{\frac{n_{12}^2}{4} - \frac{n_{12}n_{21}}{2} + \frac{n_{21}^2}{4}}{\frac{n_{12} + n_{21}}{2}} + \frac{\frac{n_{12}^2}{4} - \frac{n_{12}n_{21}}{2} + \frac{n_{21}^2}{4}}{\frac{n_{12} + n_{21}}{2}} = \frac{\frac{n_{12}^2}{2} - n_{12}n_{21} + \frac{n_{21}^2}{2}}{n_{12} + n_{21}} + \frac{\frac{n_{12}^2}{2} - n_{12}n_{21} + \frac{n_{21}^2}{2}}{n_{12} + n_{21}} = \frac{\frac{2n_{12}^2}{2} - 2n_{12}n_{21} + \frac{2n_{21}^2}{2}}{n_{12} + n_{21}} = \frac{n_{12}^2 - 2n_{12}n_{21} + n_{21}^2}{n_{12} + n_{21}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$. Thus, we have shown that $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$ and by showing this we have now shown that all of our assumptions for our statistical test using the McNemar's Test prior are valid with accurate test results.

4   Simulation Study

Sometimes we might like to know how well our statistical testing does at detecting certain alternatives (which is what we would call its power). While we can try to derive these things, there are times that the derivations are too difficult so instead we use simulation based methods which will provide a nice method for finding approximate results. In this section, we are going to investigate if we control for α, how well does our Pearson chi-square test identify alternative hypotheses when they are present and should be selected (or really reject our null hypothesis). This is what we refer to as the test's power. In this case of the simulation, we are going to set $\alpha = 0.05$ since this is a standard control for Type I Error.

The first thing we are going to do in our simulation is to make a function that will take in a value of sample size (n), our π_1 value which is the mean of success for drug A, our π_2 value which is the mean of success for drug B, and the correlation between our data will be generated on. The function will then go through and make 1000 different 2x2 data sets that will have a McNemar's Test conducted on each. We then will look at the proportion of tests rejected, which is what this function will return for whatever combination of values inputted.

```
# Import library needed to run simulation
library(MultiRNG)

# Here we are going to write a function where we can plug in our sample size,
# pi1, pi2, and correlation values to get the percentage of tests that are
# correctly rejected (find the power of the test)
get_results <- function(n, pi1, pi2, correlation) {

  # Set seed for reproducibility
  set.seed(9)
```

```r
  # Replicate results 1000 times
  test_results <- replicate(1000, {

    # Make our correlation matrix
    cmat <- matrix(c(1, correlation, correlation, 1), nrow = 2, ncol = 2)

    # Make our binary draw random data based on the information inputted in function
    cor_df <- draw.correlated.binary(no.row = n,
                                     d = 2,
                                     prop.vec = c(pi1, pi2),
                                     corr.mat = cmat)

    # Count the number of times certain situations occur
    # Off diagonals are differences
    count_vector <- NULL
    for (i in 1:nrow(cor_df)) {
      if (cor_df[i , 1] == 0) {
        if (cor_df[i , 2] == 0) {
          count_vector <- append(count_vector, "a")
        } else {
          count_vector <- append(count_vector, "b")
        }
      } else {
        if (cor_df[i , 2] == 0) {
          count_vector <- append(count_vector, "c")
        } else {
          count_vector <- append(count_vector, "d")
        }
      }
    }
    # Make a matrix of each type of situation
    data <- matrix(
      c(
        sum(count_vector == "a"),
        sum(count_vector == "c"),
        sum(count_vector == "b"),
        sum(count_vector == "d")
      ),
      nrow = 2,
      ncol = 2
    )

    # Run our McNemar Test to accept or reject this particular data
    # TRUE is rejected and FALSE is fail to reject
    mcnemar.test(data, correct = FALSE)$p.value <= 0.05
  })
  # After getting 1000 simulated results get a proportion rejections
  mean(test_results, na.rm = TRUE)
}
```

Now after making our function, we are going to get our proportion of McNemar's Tests that were rejected for each combination of $n, \pi_1, \pi_2,$ and our correlation. The loop here will get each proportion and our data frame made will have each combination with their proportion result.

```r
# Get different possible values for each
n <- c(25, 40, 80, 200)
pi1 <- c(0.1, 0.4, 0.8)
added <- c(0, 0.02, 0.05, 0.1)
correlation <- c(0, 0.2, 0.5)

# Make data frame of all different possible combinations
diff_combos <- expand.grid(n = n, pi1 = pi1, added = added, correlation = correlation)

# Make for loop to get values for all different combinations
temp <- NULL
simulated_data <- NULL
for (i in correlation) {
  for (j in added) {
    for (k in pi1) {
      for (l in n) {
        # Get temp result
        temp <- get_results(n = l,
                            pi1 = k,
                            pi2 = j + k,
                            correlation = i)

        # Combine all results into a data frame to use for later
        simulated_data <- rbind(simulated_data, temp)
      }
    }
  }
}

# Get rid of row names
row.names(simulated_data) <- NULL

# Combine with expanded grid data frame for full results
final_data <- cbind(diff_combos, simulated_data)
```

Lastly, we want to go through and plot our data which will be grouped by π_1 and then subgrouped within our main group by sample size. This will allow us to see a trend in each π_1 of how correlation, difference in π_2 and π_1, and sample size can affect our proportion of rejected tests (which in this case is what we are calling power of the test).

```r
# Load library for plotting
library(tidyverse)

# Make the plots in a 1 row by number of pi1 values for plots
par(mfrow = c(1, length(pi1)))

# Loop through for pi1 values to get plots
for (i in pi1) {
  # Make plot
  plot <- final_data %>%
    filter(pi1 == i) %>%
    ggplot(aes(x = added, y = simulated_data, color = factor(correlation))) +
    geom_line() +
```
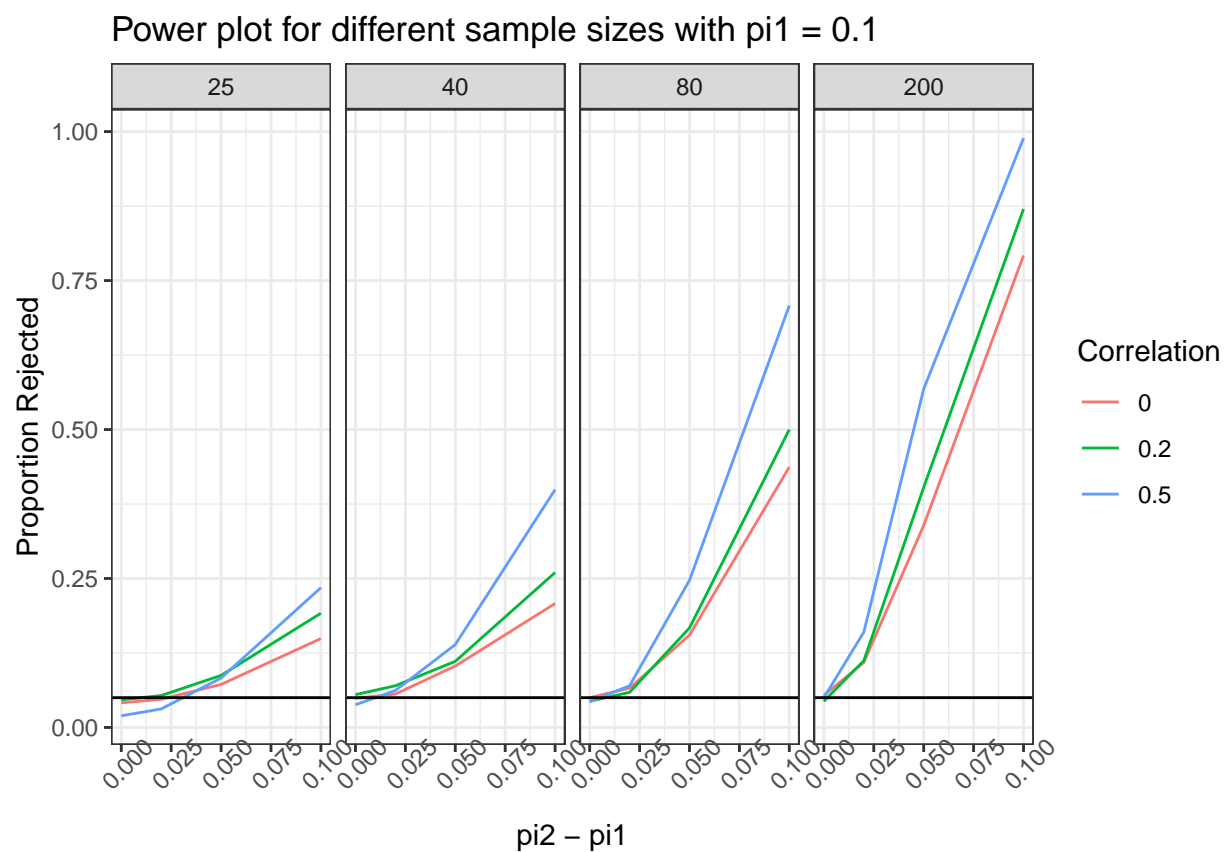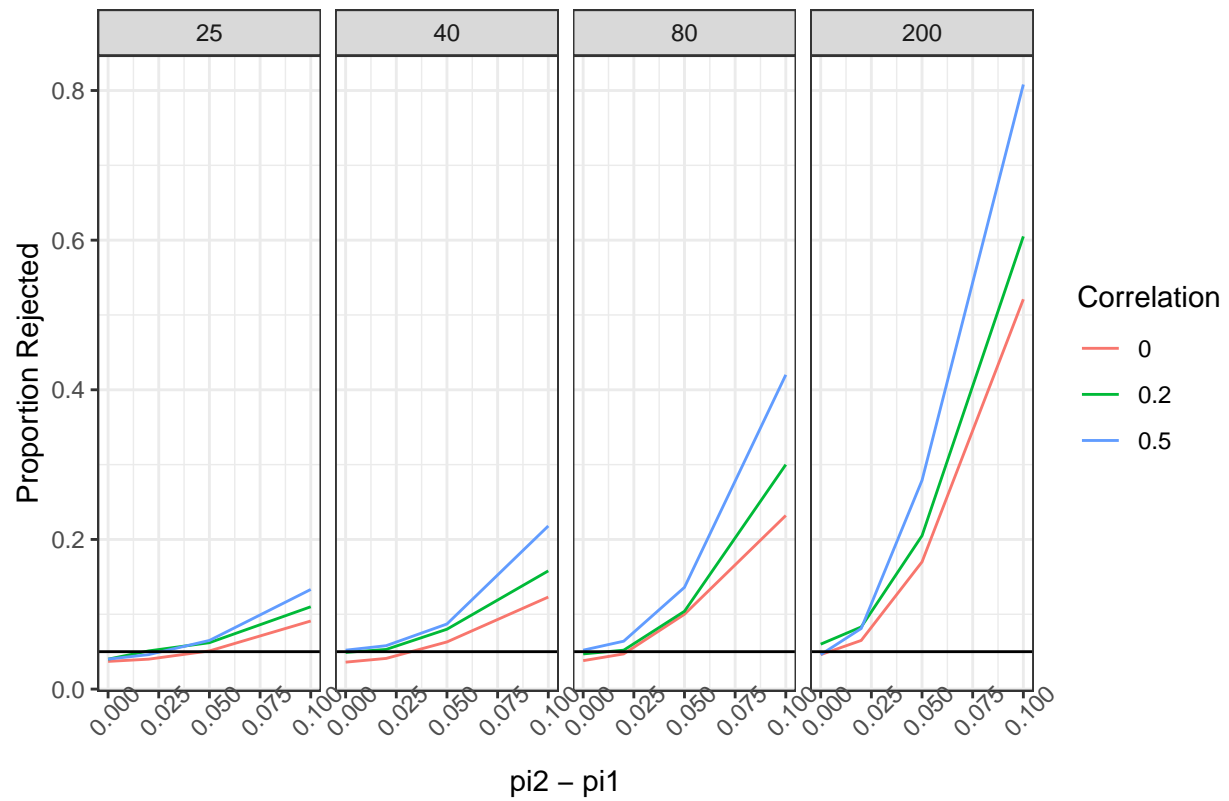
```
    geom_hline(yintercept = 0.05) +
    labs(x = "pi2 - pi1",
         y = "Proportion Rejected",
         color = "Correlation",
         title = paste0("Power plot for different sample sizes with pi1 = ", i)
         ) +
    facet_grid(~ n) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 45))

  # Print plot
  print(plot)
}
```
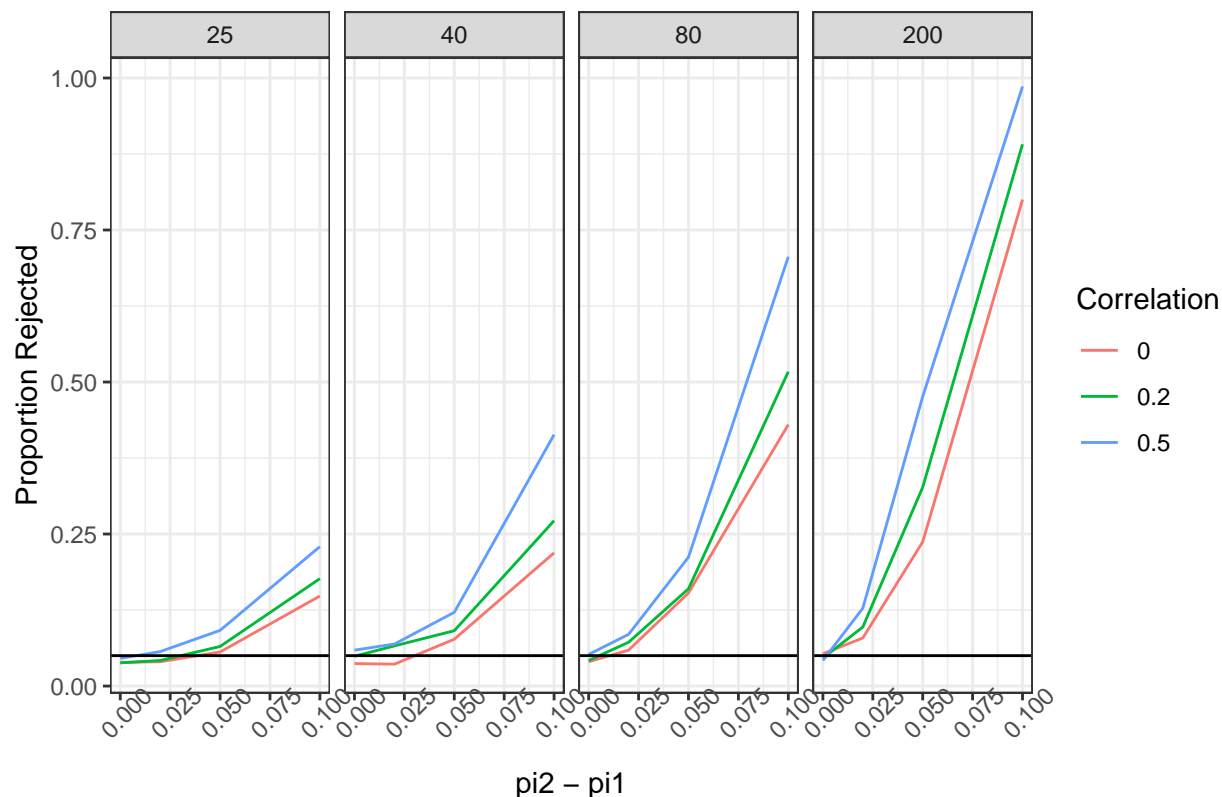

Power plot for different sample sizes with pi1 = 0.1

Power plot for different sample sizes with pi1 = 0.4

Power plot for different sample sizes with pi1 = 0.8

In conclusion with this simulation, we can see that there are some similar trends in our plots. First, as our correlation increases, our power tends to increase holding everything else constant. We can also see the same trend of increasing power as our sample size grows. This makes sense as we have learned that an increase in sample size is how people try to get to a certain power level or goal for their statistical tests and experiments. We can also see that as the difference between π_2 and π_1 increases, so does our power for the test. This obviously makes sense as we would expect more tests to be rejected (and increased power) as we deviate farther from the difference being zero (as this is what our null hypothesis states that the difference is zero). The plot tends to make sense from things we know about how to control power in our statistical testing. One thing to note that I expected, but might surprise others, is that the π_1 probability that we set does make a difference. Why is that? As we get closer to 0.5 being the null probability we set, the larger our variance (standard error) is, which means our error intervals are wider. This means we could fail to reject more tests with these wider intervals, which decreases our power. So if we want to increase our power choosing a null probability then we should look at more extreme values (closer to zero or one). So in conclusion, our major takeaways are that increasing correlation, sample size, and difference between π_2 and π_1 (or mean success rates between drugs) causes an increase in power of a statistical test. Meanwhile choosing a null (baseline) π_1 for the success of drug A that is closer to 0.5 actually decreases our power so more extreme success rates (closer to zero or one) will increase our power of statistical tests as well.