# Math185_HW1

2024-04-11

# Q1

```
dataset <- read.table("natality-california-2022.txt", header = TRUE)
# Save the dataset as an RDA file
save(dataset, file = "natality-california-2022.rda")
# load the data
load('natality-california-2022.rda')
```

## Null hypothesis:

The chance of a baby being born a girl is the same across counties in California. Let A_i be the probability of the baby is a girl where i represent all possible counties in California. Such that A_i are the same for all i.

## Hypothesis:

The chance of a baby being born a girl are the same acroos counties in California.

```
# load all the girls / boys
girl <- subset(dataset, Gender.Code == "F")
boy <- subset(dataset, Gender.Code == "M")

# order them by their county
girls <- girl[order(girl$County),]
head(girls)
```

```
##   Gender Gender.Code                  County County.Code Births
## 1 Female           F      Alameda County, CA        6001   7966
## 2 Female           F        Butte County, CA        6007    906
## 3 Female           F Contra Costa County, CA        6013   5666
## 4 Female           F    El Dorado County, CA        6017    792
## 5 Female           F       Fresno County, CA        6019   6932
## 6 Female           F     Humboldt County, CA        6023    590
```

```
boys <- boy[order(girl$County),]
head(boys)
```

```
##    Gender Gender.Code                  County County.Code Births
## 37   Male           M      Alameda County, CA        6001   8647
## 38   Male           M        Butte County, CA        6007   1040
## 39   Male           M Contra Costa County, CA        6013   5904
## 40   Male           M    El Dorado County, CA        6017    778
## 41   Male           M       Fresno County, CA        6019   7018
## 42   Male           M     Humboldt County, CA        6023    611
```

```
total <- girls$Births +boys$Births
total
```

```
##  [1] 16613  1946 11570  1570 13950  1201  2563 12494  2040 95824  2150  2224
## [13]  3849  5691  1150 30929  3766 27878 18205 26212 37587  7115 10091  2411
## [25]  7497  5670 19080  2305  1776  4851  4468  7071  6801  9987  8642  1927
```

```
# calculate the probability that girl will born for each county
girlProb = girls$Births / total
# run chisq
chisq.test(girlProb)
```

```
## Warning in chisq.test(girlProb): Chi-squared approximation may be incorrect
```

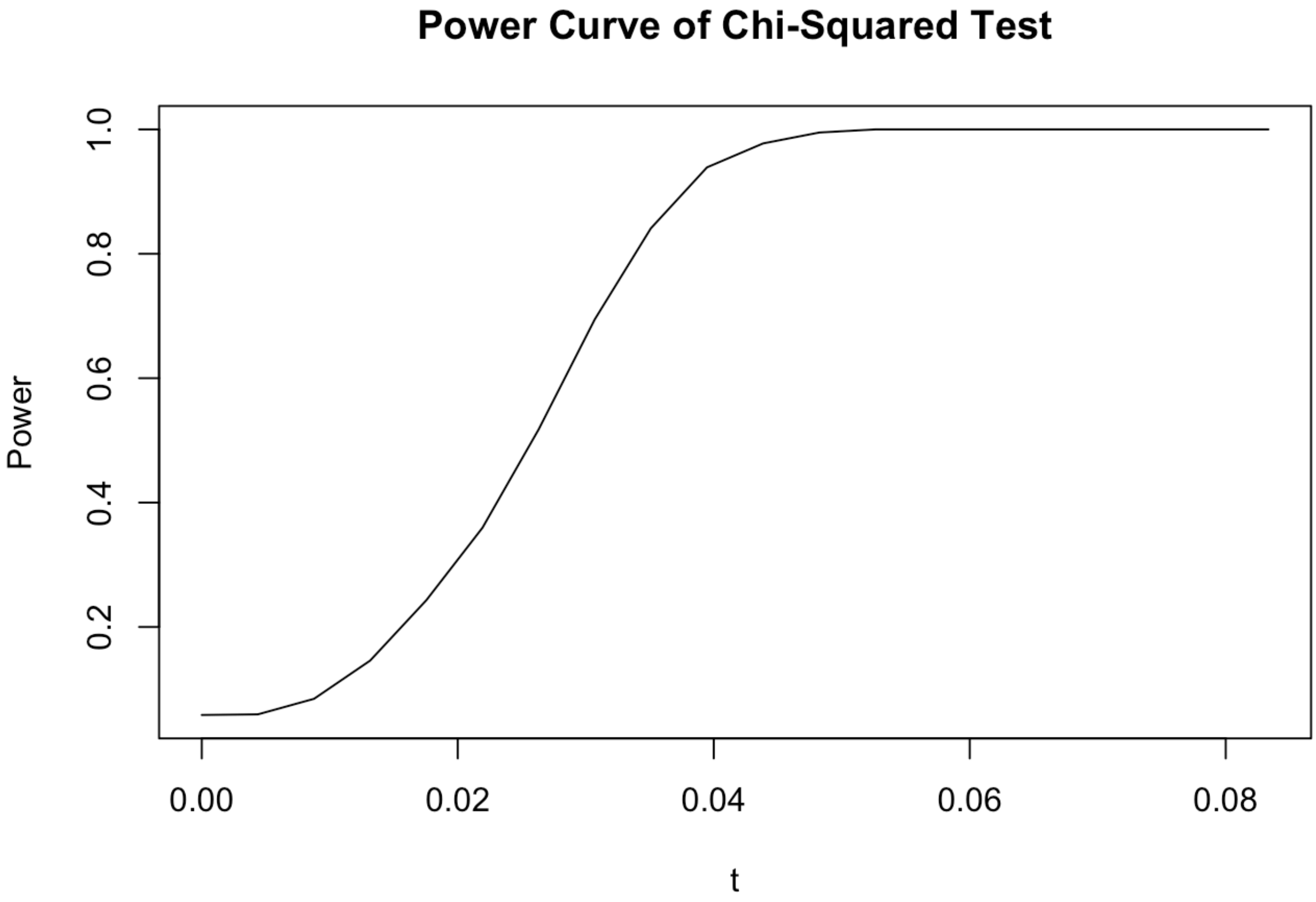
```
## 
##  Chi-squared test for given probabilities
## 
## data:  girlProb
## X-squared = 0.0071707, df = 35, p-value = 1
```

## Conclusion:

Since p-value = 1 > 0.05, we does not reject null hypothesis. Hence the chance of a baby being born a girl are the same across counties in California.

# Q2 A & B

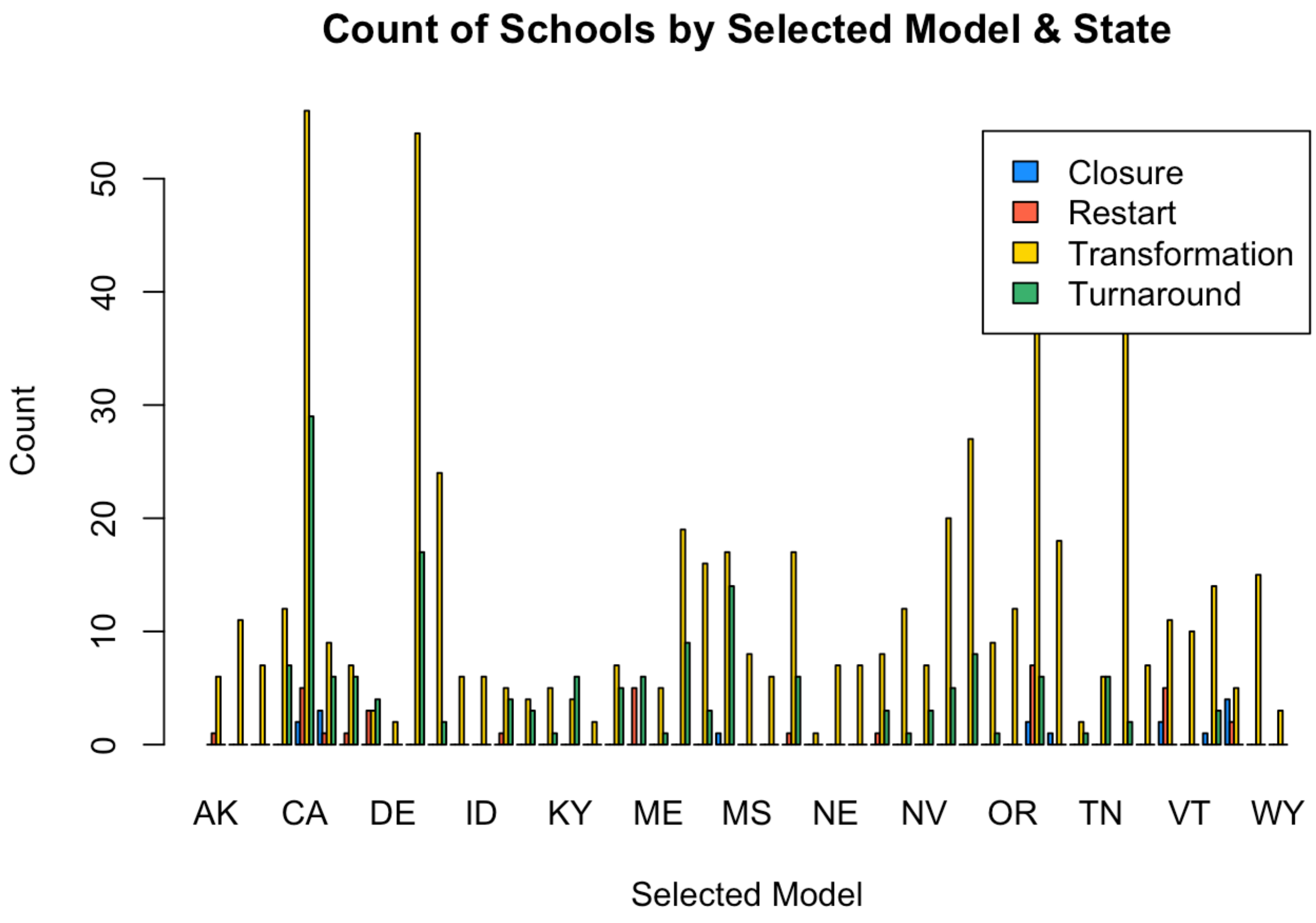
```
chisq.power <- function(k,t,n,B=2000) {
  R <- numeric(B)
  # Simulate data from Pt
  Pt <- c(rep(1/(2*k)+t, k), rep(1/(2*k)-t, k))

  for (b in 1:B) {
      # Random get n value from 1 to 2k with prob = Pt
      X <- sample(1: (2*k), n, replace = TRUE, prob = Pt)

      # Perform chi-squared test
      chisq <-chisq.test(table(factor(X, levels=1:(2*k))))
      # Check if test rejects null hypothesis
      R[b] <- as.numeric(chisq$p.value < 0.05)
  }

  # Compute proportion of rejections
  power <- mean(R)
  return(power)

}
# Part B: Plot the power curve
# Fix k = 6
# Define a range of t values
t_values <- seq(0, 1/12, length.out=20)

# Compute power for each t value
powers <- sapply(t_values, function(t) chisq.power(k=6, t=t, n=100))

# Plot
plot(t_values, powers, type = "l", xlab = "t", ylab = "Power", main = "Power Curve of Chi-Squared Test")
```



# Q3

Read csv file

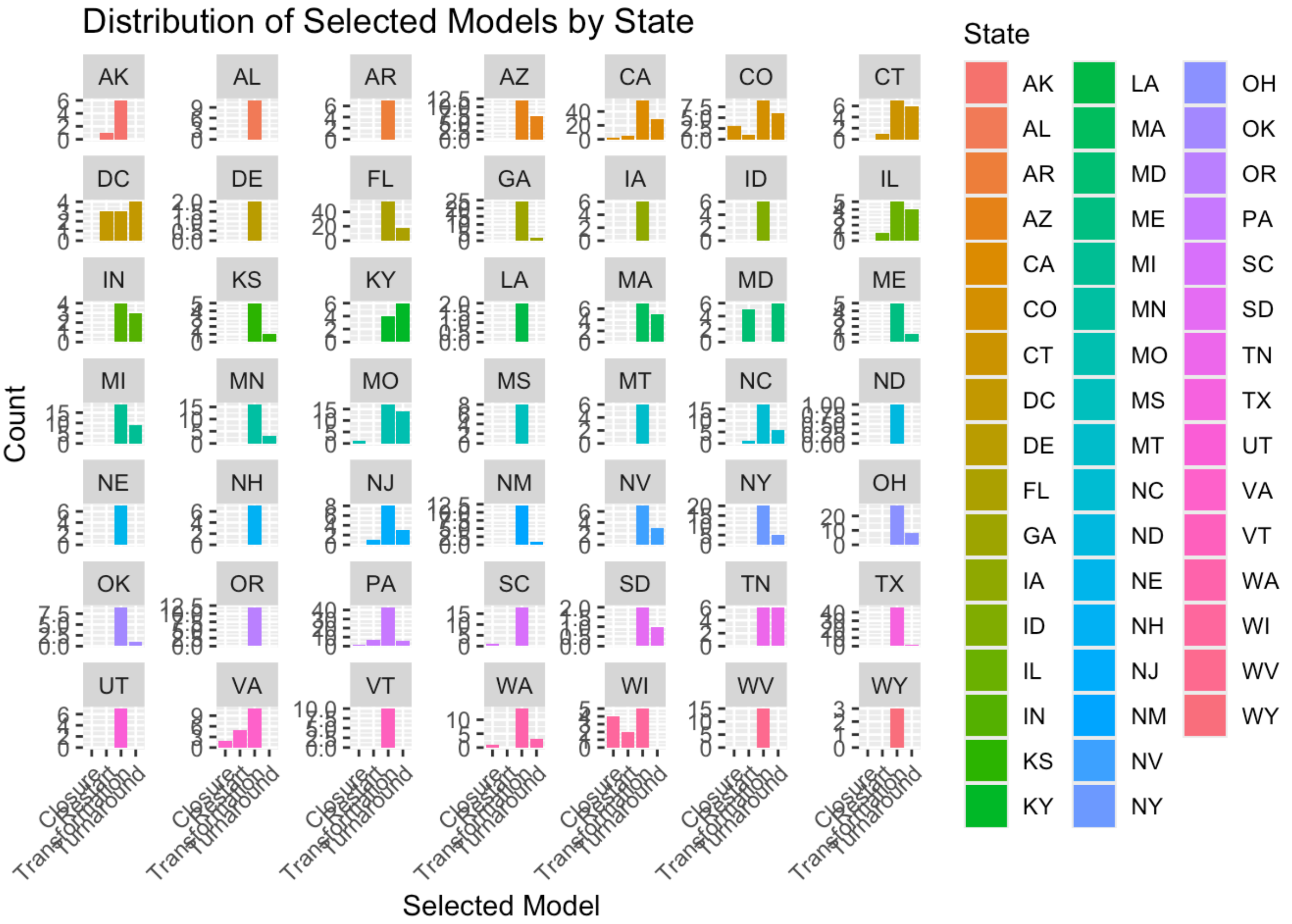
```
dataset3 <- read.csv("improvement-2010.csv", header = TRUE)
save(dataset3, file = "improvement-2010.rda")
load('improvement-2010.rda')

# remove school:
dataset3 <- subset(dataset3, State != "RI")

# create a table of counts
model = dataset3$Model.Selected
state = dataset3$State
counts <- table(model, state)

# get a barplot
custom_colors <- c("dodgerblue", "tomato", "gold", "mediumseagreen")
barplot(counts, beside = TRUE, legend = TRUE, col = custom_colors,
        main = "Count of Schools by Selected Model & State",
        xlab = "Selected Model", ylab = "Count")
```



```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
# Create a horizontal bar plot
ggplot(dataset3, aes(x = Model.Selected, fill = State)) +
  geom_bar() +
  facet_wrap(~ State, scales = "free_y", ncol = 7) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Selected Models by State", x = "Selected Model", y =
"Count")
```



# See pattern:

Thus, we can see that each state tend to have a different pattern in terms of school selection model. There seems to have some association between the model that each school selected and the state where the school was located.

# Null hypothesis:

There is no association between the model that each school selected and the state of the school.

# Hypothesis:

There is association between the model that each school selected and the state of the school.

```
# chisq test
chisq.test(counts)
```

```
## Warning in chisq.test(counts): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  counts
## X-squared = 378.37, df = 144, p-value < 2.2e-16
```

# Conclusion:

Since p value < 2.2e-16, which means p-value < 0.05, we reject the null hypothesis. Therefore, there is association between the model that each school selected and the states of the schools.