# Homework 1

**Lucien Chen**

**2024-04-09**

# Homework 1

## Problem 1

Let $p$ be the probability that a newborn baby is a girl.

$$\mathrm{H}_0 : p_1 = p_2 = \ldots = p_n$$

$$\mathrm{H}_1 : p_1 \neq p_2 \neq \ldots \neq p_n$$

Our null hypothesis is that the probability that a newborn baby is a girl is the same for all counties in California where each $i \in \{1, 2, \ldots, n\}$ represents a specific county. The alternative hypothesis is that the probability that a newborn baby is a girl is not the same for all counties in California.

```
load("natality-california-2022.rda")
head(df)
```

```
##   Gender                  County Births
## 1 Female      Alameda County, CA   7966
## 2 Female        Butte County, CA    906
## 3 Female Contra Costa County, CA   5666
## 4 Female    El Dorado County, CA    792
## 5 Female       Fresno County, CA   6932
## 6 Female     Humboldt County, CA    590
```

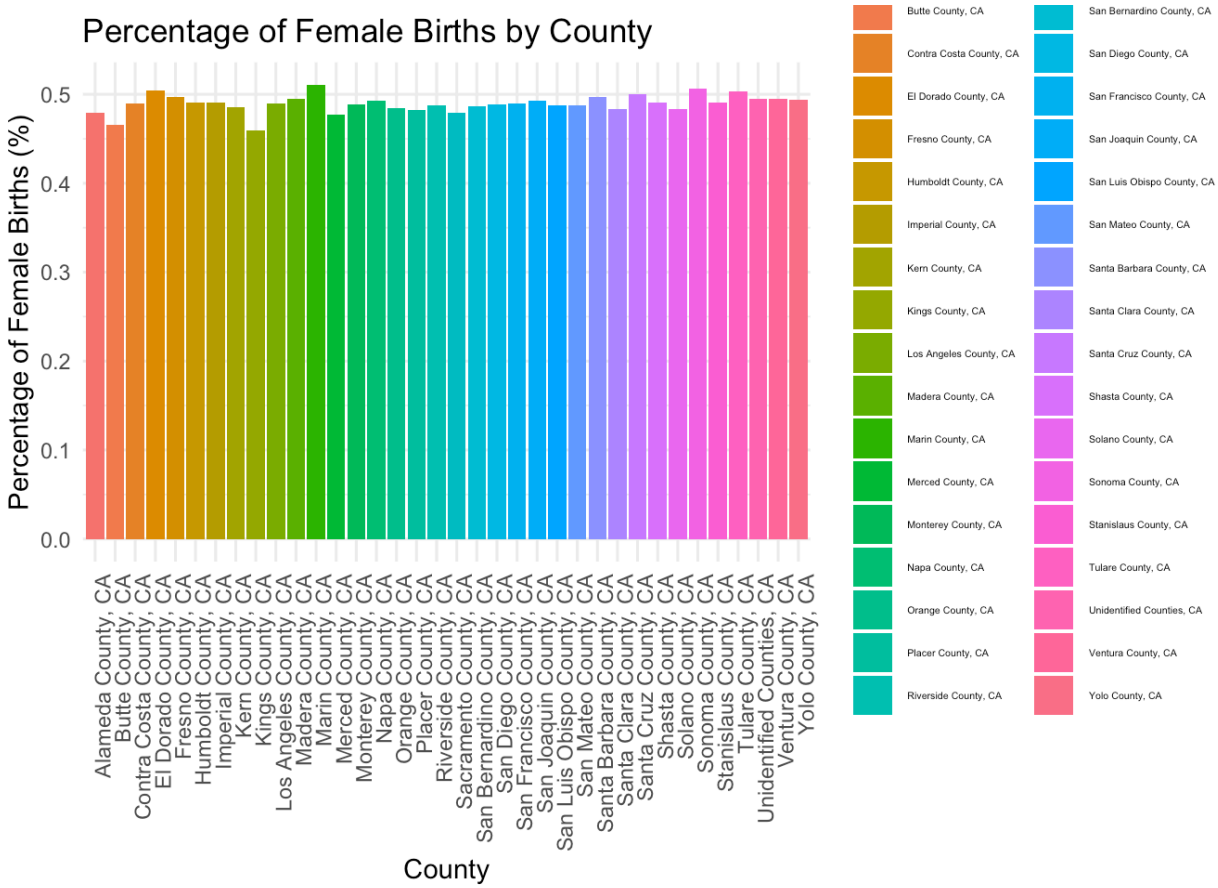
```
df_summary <- df %>% group_by(County) %>% summarise(
  Total_Births = sum(Births),
  Female_Births = sum(ifelse(Gender == "Female", Births, 0)),
  Male_Births = sum(ifelse(Gender == "Male", Births, 0))
)

df_summary <- df_summary %>% select(!Total_Births)
df_summary
```

```
## # A tibble: 36 × 3
##    County                  Female_Births Male_Births
##    <chr>                           <dbl>       <dbl>
##  1 Alameda County, CA               7966        8647
##  2 Butte County, CA                  906        1040
##  3 Contra Costa County, CA          5666        5904
##  4 El Dorado County, CA              792         778
##  5 Fresno County, CA                6932        7018
##  6 Humboldt County, CA               590         611
##  7 Imperial County, CA              1258        1305
##  8 Kern County, CA                  6071        6423
##  9 Kings County, CA                  938        1102
## 10 Los Angeles County, CA          46905       48919
## # ℹ 26 more rows
```

Now that we have calculated the percentage of female births for each county, let's visualize it.

```
ggplot(
  df_summary,
  aes(
    x=County,
    y=Female_Births/(Female_Births + Male_Births),
    fill=County
  )
) + geom_bar(
  stat="identity"
) + labs(
  x="County", y="Percentage of Female Births (%)", title="Percentage of Female Births by Count
y"
) + theme_minimal() + theme(
  axis.text.x = element_text(angle=90, hjust=1),
  legend.text = element_text(size=4)
)
```




Now, let's conduct our hypothesis test to see what we get.

```
chisq.test(df_summary %>% subset(select=c(Female_Births, Male_Births)))
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_summary %>% subset(select = c(Female_Births, Male_Births))
## X-squared = 60.435, df = 35, p-value = 0.00481
```

After running our chi-squared test, we get a p-value of approximately 0.005. Based on this result, we would reject the null hypothesis that the probability that a newborn girl is the same across all counties in California at a significance level of 0.05. It seems that the probabilities do differ across counties in California.

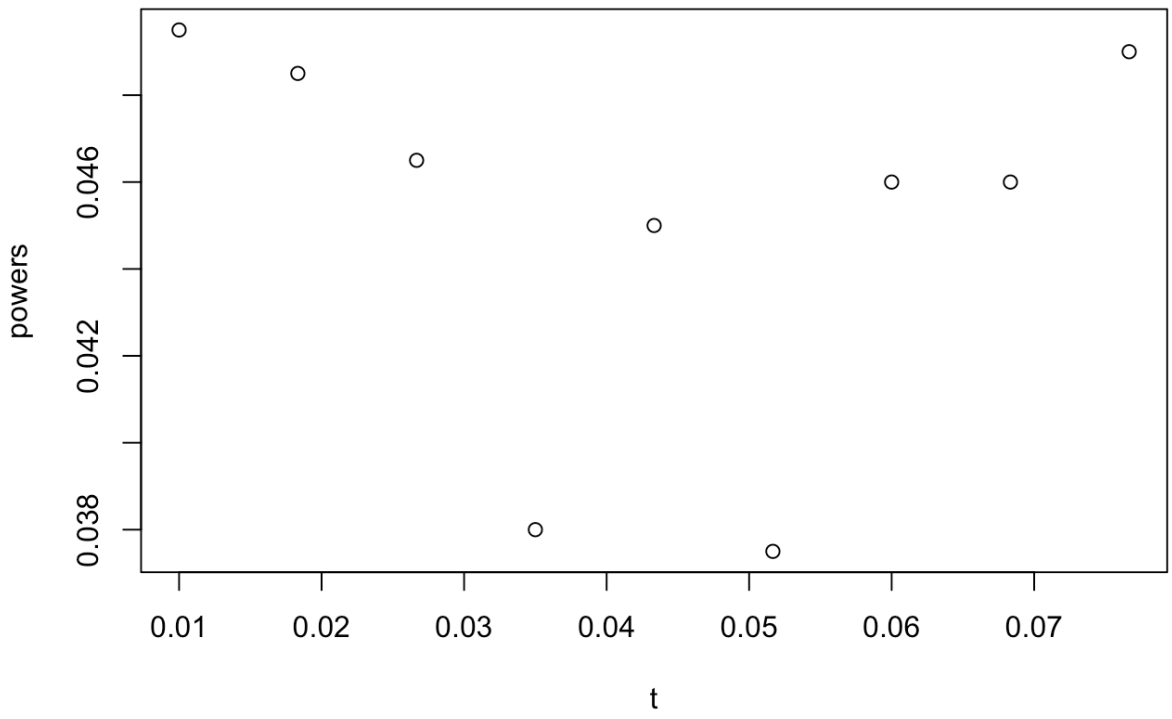
# Problem 2

I define chisq.power with the following code:

```
chisq.power <- function(k, t, n, B = 2000){
  R = vector(length=B)
  for (i in 1:B){
    alpha = 0.05
    two_k = 2*k
    pt_data = c(rep(1/two_k + t, k), rep(1/two_k - t, k))
    obs = sample(pt_data, n, replace=TRUE)
    exp = sample(1:two_k, n, replace=TRUE)
    X = chisq.test(obs, exp)
    R[i] = as.integer(X$p.value <= alpha)
  }
  return(mean(R))
}
```

Now, let's plot the curve, using our function:

```
k = 6
t = seq(0.01, 1/(2*k), 1/(2*k)/10)
powers <- vector(length=length(t))
for (i in 1:length(t)){
  powers[i] = chisq.power(k, t[i], 100)
}
plot(x=t, y=powers)
```




# Problem 3

```
load("school-improvement.rda")
head(d)
```

```
##                         School.Name     City State                              District.Name
## 1 HOGARTH KINGEEKUK MEMORIAL SCHOOL SAVOONGA    AK             BERING STRAIT SCHOOL DISTRICT
## 2                   AKIACHAK SCHOOL AKIACHAK    AK                    YUPIIT SCHOOL DISTRICT
## 3                    GAMBELL SCHOOL  GAMBELL    AK             BERING STRAIT SCHOOL DISTRICT
## 4              BURCHELL HIGH SCHOOL  WASILLA    AK MATANUSKA-SUSITNA BOROUGH SCHOOL DISTRICT
## 5                     AKIAK SCHOOL    AKIAK    AK                    YUPIIT SCHOOL DISTRICT
## 6                    MIDVALLEY HIGH  WASILLA    AK MATANUSKA-SUSITNA BOROUGH SCHOOL DISTRICT
##   X2010.11.Award.Amount Model.Selected
## 1            $471014.00 Transformation
## 2            $520579.00 Transformation
## 3            $449592.00 Transformation
## 4            $641184.00 Transformation
## 5            $399686.00 Transformation
## 6            $697703.00        Restart
##                                                        Location
## 1            200 MAIN ST\nSAVOONGA, AK 99769\n(63.6687, -170.603)
## 2         AKIACHAK 51100\nAKIACHAK, AK 99551\n(60.8911, -161.376)
## 3             169 MAIN ST\nGAMBELL, AK 99742\n(63.7413, -171.689)
## 4     1775 WEST PARKS HWY\nWASILLA, AK 99654\n(61.5794, -149.495)
## 5                  AKIAK 5227\nAKIAK, AK 99552\n(60.8879, -161.2)
## 6 7362 WEST PARKS HWY 725\nWASILLA, AK 99654\n(61.5023, -149.796)
```

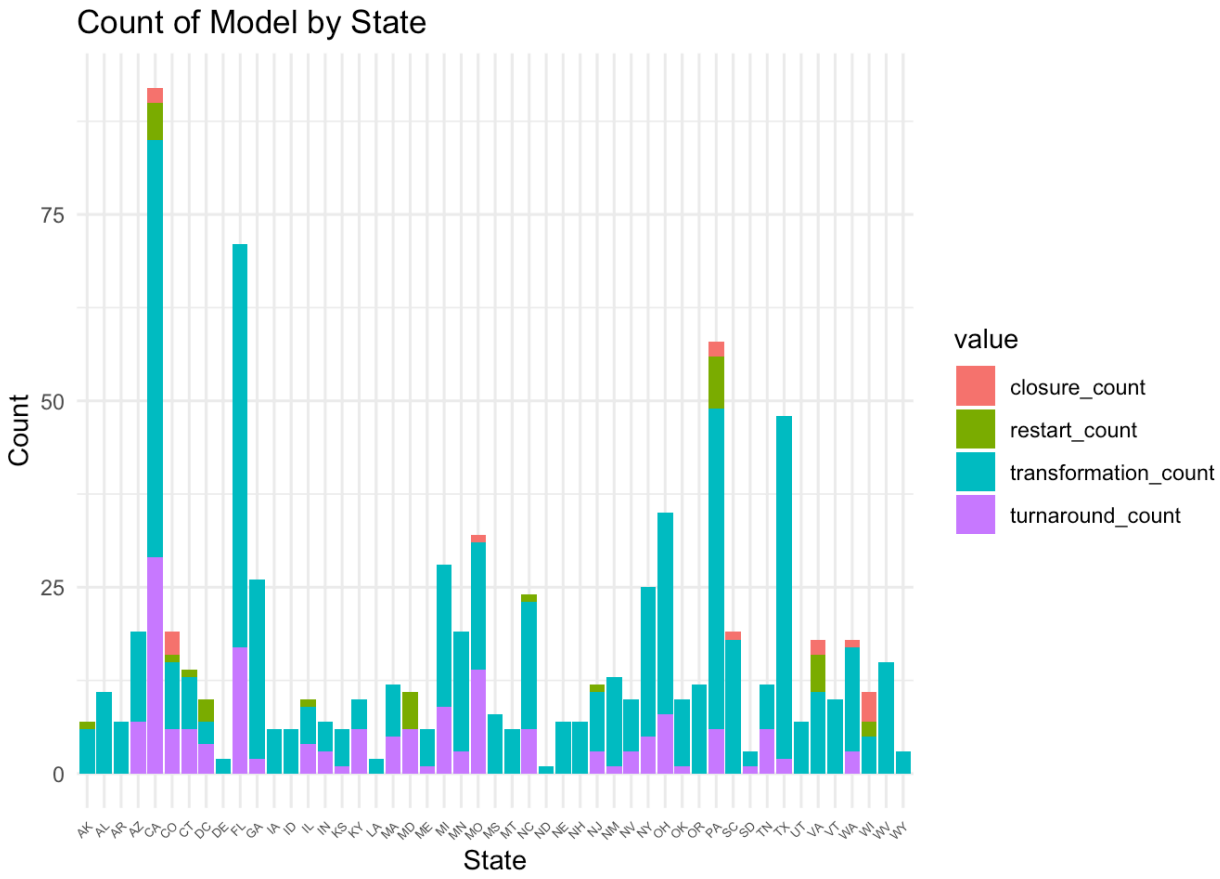
Like before, we need to do some data manipulation to get it to be in the format we want.

```
model_summary <- d %>% group_by(State) %>% summarize(
  transformation_count = length(which(Model.Selected == "Transformation")),
  restart_count = length(which(Model.Selected == "Restart")),
  turnaround_count = length(which(Model.Selected == "Turnaround")),
  closure_count = length(which(Model.Selected == "Closure"))
  )
head(model_summary)
```

```
## # A tibble: 6 × 5
##   State transformation_count restart_count turnaround_count closure_count
##   <chr>                <int>         <int>            <int>         <int>
## 1 AK                       6             1                0             0
## 2 AL                      11             0                0             0
## 3 AR                       7             0                0             0
## 4 AZ                      12             0                7             0
## 5 CA                      56             5               29             2
## 6 CO                       9             1                6             3
```

Here's what our data looks like. Now let's visualize it using a stacked bar plot.

```
ggplot(model_summary %>% gather(value, variable, -State), aes(x=State, y=variable, fill=value))
+
geom_bar(stat="identity") +
labs(
  x="State",
  y="Count",
  title="Count of Model by State"
) + theme_minimal() + theme(
  axis.text.x = element_text(size=5, angle=45, hjust=1)
)
```




Finally, as before we are going to conduct a chi-squared test. Our hypotheses are

$H_0 =$ There is no association between the models that each school selected and the state where the school is located.

versus

$H_1 =$ There is an association between the models that each school selected and the state where the school is located.

```
chisq.test(model_summary %>% subset(select=c(closure_count, restart_count, transformation_coun
t, turnaround_count)))
```

```
##
##  Pearson's Chi-squared test
##
## data:  model_summary %>% subset(select = c(closure_count, restart_count,     transformation_
count, turnaround_count))
## X-squared = 378.37, df = 144, p-value < 2.2e-16
```

Based on our test, we would reject the null that the data are independent, and hence that there is no association between the state the school is from and the model they selected at the 0.05 significance level. It seems that there is an association between the state that a school is located in and the model they picked.