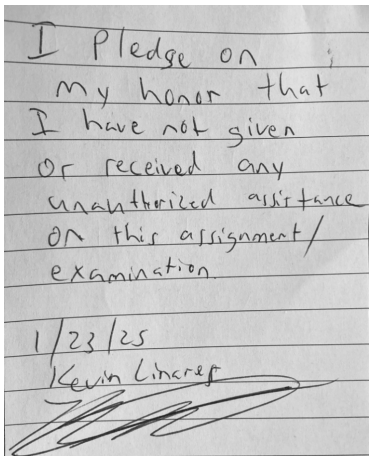
SURV616, Homework 1

Kevin Linares

2025-01-25



1. The following data are from Google Trends show the number of times that the term "film noir" was searched using Google.

week	film_noir
2022-10-02	68
2022-10-09	73
2022-10-16	58
2022-10-23	59
2022-10-30	72
2022-11-06	70
2022-11-13	77
2022-11-20	57
2022-11-27	56
2022-12-04	76
2022-12-11	63
2022-12-18	52

- There is a belief that the number of searches each week changed over that period of time

1. a) Calculate maximum likelihood estimate of **p** (i.e. the proportion of all 781 searches that occurred in each week). Graph these 12 proportions.

- The maximum likelihood estimation of a proportion p is given as, which is the success over trials:

$$\theta = g(p), \text{ is } \hat{\theta} = g(\hat{p})$$

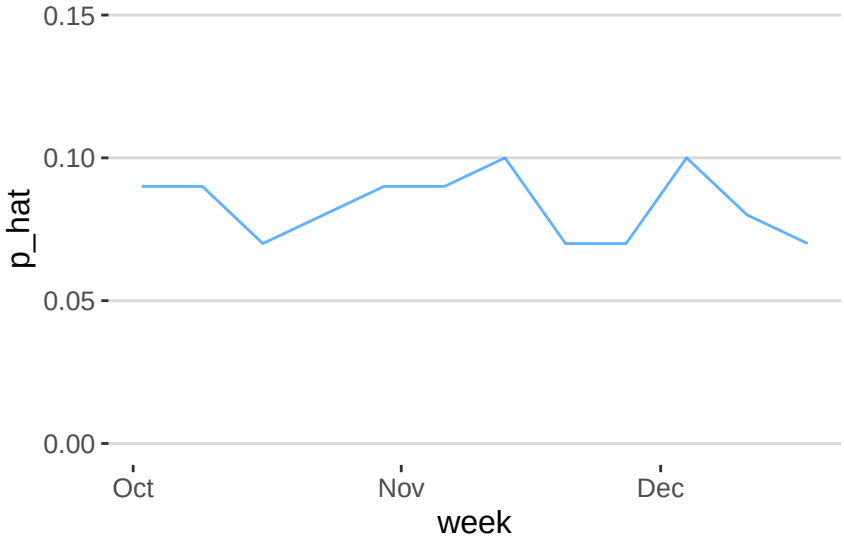
```
gtrends <- gtrends |>
  mutate(p_hat = round(film_noir / sum(film_noir), 2))

kable(gtrends)
```

week	film_noir	p_hat
2022-10-02	68	0.09
2022-10-09	73	0.09
2022-10-16	58	0.07
2022-10-23	59	0.08
2022-10-30	72	0.09
2022-11-06	70	0.09
2022-11-13	77	0.10
2022-11-20	57	0.07
2022-11-27	56	0.07
2022-12-04	76	0.10
2022-12-11	63	0.08
2022-12-18	52	0.07

- We visualize the proportions by week date.

```
gtrends |>
  ggplot(aes(x=week, y=p_hat)) +
  geom_line(color="dodgerblue", alpha=.7) +
  ylim(0, .15) +
  theme_hc()
```



1. b) Write the null hypothesis that the proportion of searches for "film noir" is the same each week. Also, write the alternative hypothesis (i.e., that there has been a change in the proportion of searches each week).

- The null hypothesis, $H_0 : p_1 = p_2 = p_3 ... = p_{12}$, is that the weekly proportion of web searches for "film_noir" is the same across weeks. The alternative hypothesis, $H_a : p_1 \neq p_2 \neq p_3 ... \neq p_{12}$, is that there is at least one proportion not equal to the other weekly proportions of web searches, in other words, there has been a change in the proportion of searches each week.

1. c) Compute the χ^2 and G^2 statistics. What do these tell us?

- The Pearson Chi-square test;

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- where $O_i = n_i$ is the observed count in the i^{th} category, and $E_i = np_{0i}$ is the expected count in the i^{th} category (from H_0).

- We can compute the expected frequencies by dividing the cells over the sum of observed values.

```
gtrends <- gtrends |>
  # add new column with expected frequencies
  mutate(expected = sum(film_noir)/n())

# print data table
gtrends
```

```
# A tibble: 12 x 4
   week       film_noir p_hat expected
   <date>         <dbl> <dbl>    <dbl>
 1 2022-10-02        68  0.09     65.1
 2 2022-10-09        73  0.09     65.1
 3 2022-10-16        58  0.07     65.1
 4 2022-10-23        59  0.08     65.1
 5 2022-10-30        72  0.09     65.1
 6 2022-11-06        70  0.09     65.1
 7 2022-11-13        77  0.1      65.1
 8 2022-11-20        57  0.07     65.1
 9 2022-11-27        56  0.07     65.1
10 2022-12-04        76  0.1      65.1
11 2022-12-11        63  0.08     65.1
12 2022-12-18        52  0.07     65.1
```

```
# calculate chi_square, than use function to check calculation
chi_square <- sum(((gtrends$film_noir - gtrends$expected)^2) / gtrends$expected)
print(chi_square)
```

```
[1] 12.52113
```

```
# use function to double check hand calculation
chisq.test(gtrends$film_noir)
```

```
    Chi-squared test for given probabilities

data:  gtrends$film_noir
X-squared = 12.521, df = 11, p-value = 0.3258
```

- The χ^2 goodness of fit test tells us how likely it is that our observed data is due to chance. The goodness of fit statistic is testing how well the observed distribution of the data fits with the distribution that is expected if the variables are independent. We observe a p>.05 for this test statistic, meaning that we reject the Null hypothesis that at least some of our observed frequencies are not close to the expected frequencies.
- The likelihood ratio statistic G^2 is a statistical method used to compare the goodness-of-fit of two competing statistical models. It approximates the χ^2 test statistic, and is expressed as:

$$G^2 = 2\sum_{i=1}^{k} O_i ln(\frac{O_i}{E_i})$$

```
# calculate LRT with formula above
LRT <- sum((2*gtrends$film_noir)*log(gtrends$film_noir/ (gtrends$expected)))
print(LRT)
```

```
[1] 12.58505
```

- We see that our $G^2 = 12.58$ approximates our $\chi^2 = 12.53$ closely. Overall our observations in each category does not fit our heuristic or theoretical expectations.

2. A graduate student decided to track the number of steps they took each day for a week. The student took a walk every afternoon. The student also walked to class and other places. The student wanted to know if they were walking about the same number of steps each day. Here are the data on steps tracked:

- The student wants to be walking about the same number of steps each day. Hence, the null hypothesis is that the number of steps are equally likely to be walked on each day, or that the daily proportion of each weeks total steps is the same:

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7$$

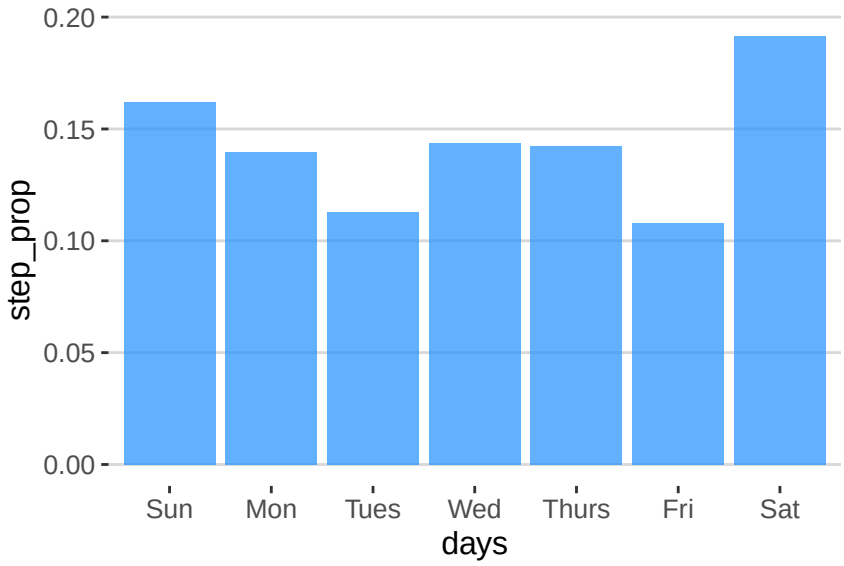
```
# enter steps data
steps <- tibble(
  days = c("Sun", "Mon", "Tues", "Wed", "Thurs", "Fri", "Sat"),
  step_counts = c(3358, 2894, 2346, 2981, 2956, 2239, 3974)
) |>
  # create days as factor, calculate daily proportions
  mutate(days = fct(days,
                    levels =
                      c("Sun", "Mon", "Tues", "Wed", "Thurs", "Fri", "Sat")),
    step_prop = step_counts/sum(step_counts))

# print total, should be 20,748
steps |> tally(step_counts)
```

```
# A tibble: 1 x 1
      n
  <dbl>
1 20748
```

2. a) Graph the proportions of all steps taken on each day of the week.

```
steps |>
  ggplot(aes(x=days, y=step_prop)) +
  geom_col(fill="dodgerblue", alpha=.7) +
  theme_hc()
```



2. b) Calculate the maximum likelihood estimate of $\mathbf{p}$, as well as the maximum likelihood estimate of $\hat{V}(\hat{p})$. Note that the latter $\hat{V}(\hat{p})$ is a matrix of variances and covariances.

```
# MLE proportion
steps <-  steps |>
  mutate(p_hat = step_counts/sum(step_counts))

steps |>
  select(days, p_hat)
```

```
# A tibble: 7 x 2
  days  p_hat
  <fct> <dbl>
1 Sun   0.162
2 Mon   0.139
3 Tues  0.113
4 Wed   0.144
5 Thurs 0.142
6 Fri   0.108
7 Sat   0.192
```

- The MLE variance estimate for a Binomial Distribution is given by:

$$V(\hat{p}) = \frac{\hat{p}_i(1-\hat{p}_i)}{n}$$

```
# extract p_hats from table
p_hat <- steps |> pull(p_hat)
total_steps <- steps |> tally(step_counts)

# calculate matrix of variances on the diagonal, and covariances
var_p <- (diag(p_hat) - p_hat %*% t(p_hat))/sum(total_steps)
print(var_p)
```

```
                 [,1]             [,2]             [,3]             [,4]
[1,]  0.0000065381000 -0.0000010880541 -0.0000008820231 -0.0000011207634
[2,] -0.0000010880541  0.0000057850263 -0.0000007601474 -0.0000009658992
[3,] -0.0000008820231 -0.0000007601474  0.0000048335289 -0.0000007829991
[4,] -0.0000011207634 -0.0000009658992 -0.0000007829991  0.0000059298999
[5,] -0.0000011113642 -0.0000009577987 -0.0000007764325 -0.0000009865922
[6,] -0.0000008417945 -0.0000007254774 -0.0000005881030 -0.0000007472869
[7,] -0.0000014941006 -0.0000012876495 -0.0000010438237 -0.0000013263591
                 [,5]             [,6]             [,7]
[1,] -0.0000011113642 -0.0000008417945 -0.0000014941006
[2,] -0.0000009577987 -0.0000007254774 -0.0000012876495
[3,] -0.0000007764325 -0.0000005881030 -0.0000010438237
[4,] -0.0000009865922 -0.0000007472869 -0.0000013263591
[5,]  0.0000058884431 -0.0000007410198 -0.0000013152357
[6,] -0.0000007410198  0.0000046398969 -0.0000009962154
[7,] -0.0000013152357 -0.0000009962154  0.0000074633839
```

```
# add variances to datatable as a new column "var_p"
steps <- steps  |>
  add_column(
  var_p = round(diag(var_p), 6)
)

steps
```

```
# A tibble: 7 x 5
  days  step_counts step_prop p_hat    var_p
  <fct>       <dbl>     <dbl> <dbl>    <dbl>
1 Sun          3358     0.162 0.162 0.000007
2 Mon          2894     0.139 0.139 0.000006
3 Tues         2346     0.113 0.113 0.000005
4 Wed          2981     0.144 0.144 0.000006
5 Thurs        2956     0.142 0.142 0.000006
6 Fri          2239     0.108 0.108 0.000005
7 Sat          3974     0.192 0.192 0.000007
```

- We calculate MLE variances $\hat{V}(\hat{p})$ for each $\hat{p}$ and include it to our data table.

2. c) Calculate the maximum likelihood estimate of the proportion of steps taken on the weekend (Sunday and Saturday, p1+p7) and the maximum likelihood estimate of the variance of the proportion of steps taken on the weekend.

```
steps |>
 # keep days needed
 filter(days %in% c("Sun", "Sat")) |>
 summarise(
   # take sum of p_hats
   p_hat = sum(p_hat),
   # compute variance for p_hats
           var_p = (1-p_hat) * p_hat / sum(total_steps))
```

```
# A tibble: 1 x 2
  p_hat     var_p
  <dbl>     <dbl>
1 0.353 0.0000110
```

- The MLE proportion of steps was .35 with an estimated MLE variance of .00001.

2. d) Test the H that, by computing both the χ^2 and G^2 statistics. What do you conclude?

$$H_0 : (p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7) \text{ vs } H_A : (p_1 \neq p_2 \neq p_3 \neq p_4 \neq p_5 \neq p_6 \neq p_7)$$

```r
# add expectations to the dataframe
steps <- steps |>
  mutate(expected = sum(step_counts) / n())

# calculate chi-square using formula above
chi_square <- sum(((steps$step_counts - steps$expected)^2) / steps$expected)
print(chi_square)
```

```
[1] 704.5
```

```r
# check calculation that was done by hand
chisq.test(steps$step_counts)
```

```
    Chi-squared test for given probabilities

data:  steps$step_counts
X-squared = 704.5, df = 6, p-value < 2.2e-16
```

```r
# calculate LRT
LRT <- sum((2*steps$step_counts)*log(steps$step_counts/ (steps$expected)))
print(LRT)
```

```
[1] 695.3545
```

- We calculate χ^2 and G^2 goodness of fit statistics and we reject the Null hypothesis and state is no statistical difference between the observed distribution of steps versus the theoretical distribution we would expect.

3. The following table is based on a study of aspirin use and myocardial infarction. The data are similar to actual data.

```
# construct 2x2 table
myo_table <- matrix(c(173, 83, 9879, 9970), nrow= 2, ncol = 2, byrow = FALSE)
rownames(myo_table) <- c("Placebo", "Aspirin")
colnames(myo_table) <- c("Yes", "No")

# add marginal counts
myo_table <- addmargins(myo_table)

# calculate proportion of disease; disease=yes / N
disease_prop <- myo_table[[3, 1]] / myo_table[[3, 3]]

myo_table
```

```
        Yes    No   Sum
Placebo 173  9879 10052
Aspirin  83  9970 10053
Sum     256 19849 20105
```

3. a) About 1.27% (n11+n21)/(n11+n21+n12+22) had myocardial infarction. Since this was a designed experiment, 50% were assigned to take a placebo. If the use of aspirin or placebo was independent of risk of myocardial infarction (i.e. if the risk of myocardial infarction was no different whether you took placebo or aspirin), what would the expected counts be in each cell (n11, n12, n21, and n22)?

- It appears that the overall probability of having a Myocardial Infarction is .0127 and the total sample is $N = 20,105$ which the design experiment split the random sample evenly between the two conditions. We can assume that the risk of Myocardial infarction is the same regardless of condition.

- We calculate the expected frequencies for the expected frequencies for having disease by both conditions, and not having the disease by both conditions. The expected frequencies for having the disease regardless of condition is 128, and not having the disease in the placebo group the expected frequency is 9,924, and the expected frequency for not having the disease in the aspirin group is 9,925.

```r
# calculate placebo expectation, disease proportion * totaL placebo group
disease_placebo_expect <- disease_prop * myo_table[[1, 3]]

# calculate aspirin expectation, disease proportion * totaL aspirin group
disease_aspirin_expect <- disease_prop * myo_table[[2, 3]]

# calculate placebo expectation, no disease placebo counts - p(disease placebo)
no_disease_placebo_expect <-  myo_table[[1, 3]] - disease_placebo_expect

# calculate aspirin expectation, no disease aspirin counts - p(disease aspirin)
no_disease_aspirin_expect <-  myo_table[[2, 3]] - disease_aspirin_expect

# save expected freq in 2x2 table
expected_freq <- matrix(c(disease_placebo_expect, disease_aspirin_expect,
                          no_disease_placebo_expect, no_disease_aspirin_expect),
                          ncol=2, byrow=FALSE) |>
  round(1)

rownames(expected_freq) <- c("Placebo", "Aspirin")
colnames(expected_freq) <- c("Yes", "No")

addmargins(expected_freq)
```

```
        Yes    No   Sum
Placebo 128  9924 10052
Aspirin 128  9925 10053
Sum     256 19849 20105
```