W271 Assignment 1

Emanuel Mejía

Contents


```
library(tidyverse)
library(binom)
library(ggplot2)
```

1 Maximum Likelihood (30 Points)

Suppose we have a random sample of n observations, $X_1, X_2, ..., X_n$, drawn from a distribution with probability density function

$$f(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

Where:

$$x > 0, \theta > 0$$

1.1 Log-likelihood function (6 points)

Write the log-likelihood function $L(\theta|Data)$.

$$\begin{aligned} L(\theta|Data) &= L(\theta|x_1, x_2, ..., x_n) \\ &= \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_2 = x_2) \cdot ... \cdot \mathbb{P}(X_n = x_n) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}} \\ &= \frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}} \end{aligned}$$ □

To get the log-likelihood function we need to compute the log of the previous result:

$$\begin{aligned} ln[L(\theta|Data)] &= ln[\frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}}] \\ &= ln\left[\frac{1}{\theta^n}\right] + ln\left[e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}}\right] \\ &= ln\left[\theta^{-n}\right] - \frac{\sum_{i=1}^{n} x_i}{\theta} \\ &= -n(ln[\theta]) - \frac{1}{\theta}\sum_{i=1}^{n} x_i \end{aligned}$$ □

1.2 Maximum likelihood estimator (6 points)

Derive the maximum likelihood estimator of θ, denoted as $\hat{\theta}$,

We first need to differentiate our log-likelihood function:

$$\begin{aligned} \frac{\delta ln[L(\theta|Data)]}{\delta\theta} &= \frac{\delta}{\delta\theta}\left[-n(ln[\theta]) - \frac{1}{\theta}\sum_{i=1}^{n} x_i\right] \\ &= \frac{-n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \end{aligned}$$

Now we set this derivative equal to zero to find the Maximum Likelihood Estimate:

$$
\begin{aligned}
&\Rightarrow 0 = \frac{-n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \\
&\iff \frac{n}{\theta} = +\frac{1}{\theta^2}\sum_{i=1}^{n} x_i \\
&\iff \theta^2 n = \theta \sum_{i=1}^{n} x_i \\
&\iff \theta = \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}
$$

Therefore the Maximum Likelihood Estimator is $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$ □

1.3 Bias of the maximum likelihood estimator (6 points)

Show that the maximum likelihood estimator $\hat{\theta}$ is unbiased.

- **Note:** $f(x)$ is an exponential distribution with $E[X] = \theta$ and $Var[X] = \theta^2$.

We have the following:

$$
\begin{aligned}
E[\hat{\theta}] &= E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\
&= \frac{1}{n} n E[X] \\
&= \theta
\end{aligned}
$$

Since $E[\hat{\theta}] = \theta$ we can say that $\hat{\theta}$ is an unbiased estimator of θ

1.4 Mean squared error of the maximum likelihood estimator (6 points)

Prove that $MSE(\hat{\theta}) = \frac{\theta^2}{n}$.
We have the following:

$$
\begin{aligned}
MSE_{\hat{\theta}} &= E\left[(\hat{\theta}-\theta)^2\right] \\
&= E\left[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2\right] \\
&= E\left[\hat{\theta}^2\right] - E\left[2\hat{\theta}\theta\right] + E\left[\theta^2\right] \\
&= E\left[\hat{\theta}^2\right] - 2\theta E\left[\hat{\theta}\right] + \theta^2 \\
&= E\left[\hat{\theta}^2\right] - E\left[\hat{\theta}\right]^2 + E\left[\hat{\theta}\right]^2 - 2\theta E\left[\hat{\theta}\right] + \theta^2 \\
&= Var\left[\hat{\theta}\right] + E\left[\hat{\theta}\right]^2 - 2\theta E\left[\hat{\theta}\right] + \theta^2 \\
&= Var\left[\hat{\theta}\right] + \left[E\left[\hat{\theta}\right] - \theta\right]^2 \\
&= Var\left[\hat{\theta}\right] + Bias_{\hat{\theta}}^2
\end{aligned}
$$

In this specific case we know that $Bias_{\hat{\theta}} = 0$ so we can reduce the previous result to $MSE_{\hat{\theta}} = Var\left[\hat{\theta}\right]$

$$
\begin{aligned}
\Rightarrow MSE_{\hat{\theta}} &= Var\left[\hat{\theta}\right] \\
&= Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \\
&= \frac{1}{n^2} Var[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} Var[X_i] \\
&= \frac{1}{n^2} nVar[X] \\
&= \frac{\theta^2}{n}
\end{aligned}
$$

1.5 A sample estimate (6 points)

Using the provided set of observations below, compute a sample estimate of $\hat{\theta}$ and then use it to calculate the probability $P[X > 5]$.

$$x_1 = 2, x_2 = 0.5, x_3 = 1.5, x_4 = 3, x_5 = 0.5$$

We have the following estimator $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$. Using the sample, we can compute it as follows:

```
n <- 5
X <- c(2, 0.5, 1.5, 3, 0.5)
theta.hat <- sum(X)/n
```

And so, in this case, we have $\hat{\theta} = 1.5$

Now we have to compute $\mathbb{P}[X > 5]$, so we'll do it as follows:

$$\begin{aligned}
\mathbb{P}[X > 5] &= 1 - \mathbb{P}[X \leq 5] \\
&= 1 - \int_0^5 f_X(x, \theta) dx \\
&= 1 - \int_0^5 \frac{1}{\theta} e^{-\frac{x}{\theta}} dx \\
&= 1 - \frac{1}{\theta} \int_0^5 e^{-\frac{x}{\theta}} dx \\
&= 1 - \frac{1}{\theta} \left[\frac{e^{-\frac{x}{\theta}}}{-1/\theta} \right]_0^5 \\
&= 1 - \left[-e^{-\frac{x}{\theta}} \right]_0^5 \\
&= 1 - \left[1 - e^{-\frac{5}{\theta}} \right] \\
&= e^{-\frac{5}{\theta}}
\end{aligned}$$

Substituting our MLE: $\hat{\theta} = 1.5$

```
x.thresh <- 5
prob.result <- exp(-x.thresh / theta.hat)
```

And then $\mathbb{P}[X > 5] = 0.035674$

2 Customer churn study: Part-1 (70 points)

Customer churn refers to the rate at which customers discontinue using a company's products or services. It has a significant impact on a company's revenue and growth. High churn rates may indicate customer dissatisfaction or stiff competition. Therefore, it's crucial for businesses to monitor churn and devise strategies to retain customers.

This project consists of three parts. In Part 1, we examine how senior customers differ from non-senior customers in terms of churn

In Part 2 and Part 3 (HW-2, HW-3), we will employ regression techniques to build more complex models and explore further questions about customer churn behavior.

```
telcom_churn <- read.csv("./data/Telco_Customer_Churn.csv", header=T,na.strings=c("","NA"))
```

Churn dataset consists of 21 variables and 7043 observations. You can find variable definitions in the `data_description` file.

For the remainder of this section, pay particular attention to two variables: `Churn` and `SeniorCitizen`.

2.1 Data Preprocessing (5 points)

In this section, review the structure of data to ensure proper data types. Convert variables as needed, and manage any missing values.

We can first check the type of data we have in each column:

```r
sapply(telcom_churn, function(x) typeof(x))
```

```
##       customerID           gender    SeniorCitizen          Partner
##      "character"      "character"        "integer"      "character"
##       Dependents           tenure     PhoneService    MultipleLines
##      "character"        "integer"      "character"      "character"
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##      "character"      "character"      "character"      "character"
##      TechSupport      StreamingTV  StreamingMovies         Contract
##      "character"      "character"      "character"      "character"
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##      "character"      "character"         "double"         "double"
##            Churn
##      "character"
```

We could turn the SeniorCitizen column to have the same data type as Churn so it's easier to understand its meaning:

```r
telcom_churn["SeniorCitizen"][telcom_churn["SeniorCitizen"] == 0] <- "No"
telcom_churn["SeniorCitizen"][telcom_churn["SeniorCitizen"] == 1] <- "Yes"
```

We'll also give a different level order to the values of Churning and Senior Citizen so our tables show the ones that meet these variables first (this will prove useful in the contingency table):

```r
telcom_churn$Churn <- factor(telcom_churn$Churn, levels=c("Yes", "No"))
telcom_churn$SeniorCitizen <- factor(telcom_churn$SeniorCitizen, levels=c("Yes", "No"))
```

We can also check the count of missing values for every column in our data:

```r
sapply(telcom_churn, function(x) sum(is.na(x)))
```

```
##       customerID           gender    SeniorCitizen          Partner
##                0                0                0                0
##       Dependents           tenure     PhoneService    MultipleLines
##                0                0                0                0
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##                0                0                0                0
##      TechSupport      StreamingTV  StreamingMovies         Contract
##                0                0                0                0
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##                0                0                0               11
##            Churn
##                0
```

Since we'll be focusing on the Churn and SeniorCitizen variables there's no need to deal with missing values (these columns don't have any).

2.2 Probablity of customer churn (10 points)

Calculate the probability of a customer churning, denoted as $\hat{\pi}$. Additionally, compute its confidence interval and interpret the results. Determine whether $\hat{\pi}$ is statistically different from zero.

We have the following count for churning:

```r
churn.table <- table(telcom_churn$Churn)
churn.table
```

```
##
##  Yes   No
## 1869 5174
```

```r
n <- sum(churn.table)
w <- churn.table[1]

pi.hat <- w / n
```

So we have $\hat{\pi} = 0.2653699$

And we can compute different confidence intervals as follows:

```r
alpha <- 0.05
binom.confint(x = w, n = n, conf.level = 1 - alpha, methods = "all")
```

```
##           method    x    n      mean     lower     upper
## 1  agresti-coull 1869 7043 0.2653699 0.2551873 0.2758082
## 2     asymptotic 1869 7043 0.2653699 0.2550582 0.2756816
## 3          bayes 1869 7043 0.2654032 0.2551121 0.2757317
## 4        cloglog 1869 7043 0.2653699 0.2551093 0.2757291
## 5          exact 1869 7043 0.2653699 0.2550860 0.2758483
## 6          logit 1869 7043 0.2653699 0.2551869 0.2758087
## 7         probit 1869 7043 0.2653699 0.2551609 0.2757822
## 8        profile 1869 7043 0.2653699 0.2551447 0.2757656
## 9            lrt 1869 7043 0.2653699 0.2551459 0.2757694
## 10     prop.test 1869 7043 0.2653699 0.2551180 0.2758793
## 11        wilson 1869 7043 0.2653699 0.2551881 0.2758074
```

Since our sample is large enough ($n = 7043$) we can safely use different confidence intervals.

None of them has 0 as part of the interval, so we could say that π is statistically different from zero (with $\alpha = 5\%$).

2.3 Comparison between senior and non-senior customers (5 points)

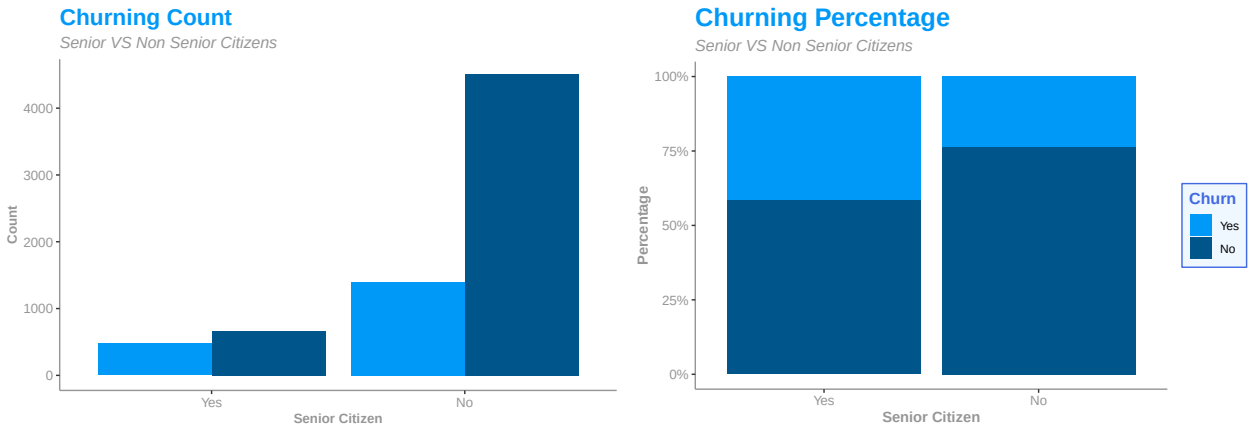
Generate a bar plot comparing seniority with churn. Are there differences between senior and non-senior customers in terms of churn?

```r
churn.cit.group <- ggplot(telcom_churn, aes(x = SeniorCitizen, fill = Churn)) +
  geom_histogram(
    position = "dodge",
    stat = "count",
    show.legend = FALSE
    ) +
  labs(
    title = "Churning Count",
    subtitle = "Senior VS Non Senior Citizens",
    x = "Senior Citizen",
```

```
    y = "Count"
    ) +
  theme_classic() +
  theme(
    plot.title = element_text(color = "#0099F8",
                              size = 17,
                              face = "bold"),
    plot.subtitle = element_text(color="#969696",
                                 size = 12,
                                 face = "italic"),
    axis.title = element_text(color = "#969696",
                              size = 10,
                              face = "bold"),
    axis.text = element_text(color = "#969696", size = 10),
    axis.line = element_line(color = "#969696"),
    ) +
  scale_fill_manual(
    values=c("#0099F8", "#00558A")
    )


churn.cit.fill <- ggplot(telcom_churn, aes(x = SeniorCitizen, fill = Churn)) +
  geom_histogram(
    position="fill",
    stat="count"
    ) +
  labs(
    title = "Churning Percentage",
    subtitle = "Senior VS Non Senior Citizens",
    x = "Senior Citizen",
    y = "Percentage"
    ) +
  theme_classic() +
  theme(
    plot.title = element_text(color = "#0099F8",
                              size = 19,
                              face = "bold"),
    plot.subtitle = element_text(color="#969696",
                                 size = 12,
                                 face = "italic"),
    axis.title = element_text(color = "#969696",
                              size = 11,
                              face = "bold"),
    axis.text = element_text(color = "#969696", size = 10),
    axis.line = element_line(color = "#969696"),
    legend.title = element_text(color = "royalblue",
                                size = 12,
                                face = "bold"),
    legend.background = element_rect(fill ="aliceblue",
                                     color ="royalblue",
                                     )
  ) + scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values=c("#0099F8", "#00558A"))
```

```
churn.cit.group
churn.cit.fill
```



By just taking a look at the charts we can see differences in terms of churn for senior and non-senior citizens.

2.4 Contingency table (10 points)

Construct a contingency table for the `Senior Citizen` and `Churn` variables and calculate the probabilities of churn for senior and non-senior customers. Determine if there is a significant difference between these probabilities.

2.4.1 Contingency table

```
cont.tbl <- xtabs(formula = ~ SeniorCitizen + Churn, data = telcom_churn)
cont.tbl
```

```
##               Churn
## SeniorCitizen  Yes   No
##           Yes  476  666
##           No  1393 4508
```

2.4.2 $\hat{\pi}$ Table

```
pi.hat.tbl <- cont.tbl / rowSums(cont.tbl)
pi.hat.tbl
```

```
##               Churn
## SeniorCitizen       Yes        No
##           Yes 0.4168126 0.5831874
##           No  0.2360617 0.7639383
```

Since we're interested in getting the churn probability we'll define churning as our table's "success"

```r
w1 <- cont.tbl[1,1]
w2 <- cont.tbl[2,1]
pi.hat1 <- pi.hat.tbl[1,1]
pi.hat2 <- pi.hat.tbl[2,1]
```

Therefore the number of successes are $w_1 = 476$ and $w_2 = 1393$ for seniors and non-seniors respectively.

And so the probability of churning for seniors is $\hat{\pi_1} = 0.4168126$ while for non-seniors is $\hat{\pi_2} = 0.2360617$.

We could say that there seems to be a significant difference between these two probabilities.

2.5 Confidence intervals for the difference of two probabilities (10 points)

Is there a difference between the proportion of senior and non-senior customers who churned? Use both `Wald` and `Agresti-Caffo` confidence intervals. Comment on their differences or similarities.

```r
# Variables to compute CI
alpha <- 0.05
n1 <- sum(cont.tbl[1,])
n2 <- sum(cont.tbl[2,])
```

2.5.1 Wald CI

```r
var.wald <- pi.hat1 * (1 - pi.hat1) / n1 + pi.hat2 * (1 - pi.hat2) / n2
wald.ci <- pi.hat1 - pi.hat2 + qnorm(p = c(alpha/2, 1 - alpha/2)) * sqrt(var.wald)
wald.ci
```

```
## [1] 0.1501720 0.2113298
```

2.5.2 Agresti-Caffo CI

```r
pi.tld1 <- (w1 + 1) / (n1 + 2)
pi.tld2 <- (w2 + 1) / (n2 + 2)
var.ac <- pi.tld1 * (1 - pi.tld1) / (n1 + 2) + pi.tld2 * (1 - pi.tld2) / (n2 + 2)
ac.ci <- pi.tld1 - pi.tld2 + qnorm(p = c(alpha/2, 1 - alpha/2)) * sqrt(var.ac)
ac.ci
```

```
## [1] 0.1502503 0.2113636
```

2.5.3 CI Results

By taking a look at both confidence intervals, there seems to be a difference indeed between the proportions (zero is not part of any interval).

Their size is similar, although AC CI slides a little bit to the right from Wald IC, but overall we can say that we have similar results with both confidence interval methods, mainly because we have a large sample.

2.6 Test for the difference of two probabilities (10 points)

To confirm the results in part `c`, conduct a hypothesis test to assess whether the probability of churn differs between senior and non-senior customers.

- The hypothesis is given below:

$$H_0 : \pi_{senior} - \pi_{non-senior} = 0$$
$$H_a : \pi_{senior} - \pi_{non-senior} \neq 0$$

For this purpose we can use the following statistic:

$$Z_0 = \frac{\hat{\pi_1} - \hat{\pi_2}}{\sqrt{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $\bar{\pi} = w_+/n_+$

So we'll first compute $\bar{\pi}$

```
w.plus <- sum(cont.tbl[,1])
n.plus <- sum(cont.tbl)
pi.bar <- w.plus / n.plus
```

Now we just need to compute the test statistic Z_0

```
z.zero <- (pi.hat1-pi.hat2)/(sqrt(pi.bar * (1-pi.bar) * (1/n1 + 1/n2)))
```

So we get $|Z_0| = 12.663$, and having $\alpha = 5\%$ then $Z_{1-\alpha/2} \approx 1.96$.

Therefore we have that $|Z_0| > Z_{1-\alpha/2}$ and we REJECT $H_0 : \pi_{senior} - \pi_{non-senior} = 0$, confirming the results we previously observed.

2.7 Relative risks (10 points)

Calculate the relative risk of churn for a senior customer to non-senior customer and its confidence interval, and then interpret the results. Are they consistent with your findings in the previous sections?

2.7.1 RR MLE

```
rr <- pi.hat1 / pi.hat2
```

We have that $\widehat{RR} = 1.7656936$ so we can estimate that senior citizens are about 1.77 times as likely to churn.

2.7.2 RR CI

Now we can compute a confidence interval as follows:

```r
var.log.rr <- (1- pi.hat1) / (n1 * pi.hat1) + (1- pi.hat2) / (n2 * pi.hat2)
rr.ci <- exp(log(rr) + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.log.rr))
rr.ci
```

```
## [1] 1.625802 1.917622
```

And therefore we can say, with 95% confidence, that senior citizens within the *population* churn between 63% - 92% more than non-senior citizens, which is consistent with our previous findings.

2.8 Odds ratios (10 points)

Calculate the odds of a senior customer churning compared to a non-senior customer. Then, compute the confidence interval of the odds ratio and comment on the results.

2.8.1 Odds MLE

Since we don't know the population's probabilities we need to use MLEs.

```r
odds1 <- pi.hat1/(1 - pi.hat1)
odds2 <- pi.hat2/(1 - pi.hat2)
```

So we have $\widehat{odds_1} = 0.7147147$, meaning that for senior citizens the probability of churning is estimated to be 0.71 times as large as the probability of not churning.

And $\widehat{odds_2} = 0.3090062$, meaning that for non-senior citizens the probability of churning is estimated to be 0.31 times as large as the probability of not churning.

2.8.2 Odds Ratio MLE

With those results we can also compute the Odds Ratio (using MLEs) as follows:

```r
OR.hat <- odds1/odds2
OR.hat.inv <- odds2/odds1
```

And so we have $\widehat{OR} = \frac{\widehat{odds_1}}{\widehat{odds_2}} = 2.3129461$, meaning that the estimated odds of churning are about 2.31 as large in seniors than in non-seniors.

Or its equivalent: $\frac{1}{\widehat{OR}} = \frac{\widehat{odds_2}}{\widehat{odds_1}} = 0.432349$, meaning that the estimated odds of churning are about 0.43 as large in non-seniors than in seniors.

2.8.3 Odds Ratio CI

And we can finally compute the Odds Ratio CI:

```r
var.log.or <- 1/w1 + 1/(n1-w1) + 1/w2 + 1/(n2-w2)
OR.CI <- exp(log(OR.hat) + qnorm(p = c(alpha/2, 1-alpha/2)) * sqrt(var.log.or))
OR.CI
```

```
## [1] 2.026745 2.639563
```

So we have an interval of $2.03 < \widehat{OR} < 2.64$. And there's sufficient evidence to say that being a senior citizen increases the true odds of churning.