HW1_problem_2

Huibin Chang

2025-08-30

Problem 2: Customer churn

```
# ---- Setup ----
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
library(tidyverse)
library(scales)
```

2.1 Data cleaning

I kept all the columns although the analysis in this problem is on Churn and Senior status. For both SeniorCitizen and Churn, I created an 0-1 integer variable for ease of analysis.

```
# ---- 2.1 Data preprocessing ----
telcom_churn <- read.csv("./data/Telco_Customer_Churn.csv", header=TRUE, na.strings=c("","NA"))
glimpse(telcom_churn)
```

```
## Rows: 7,043
## Columns: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW~
## $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",~
## $ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes~
## $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"~
## $ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2~
## $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
## $ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "~
## $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt~
## $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "~
## $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N~
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y~
## $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes~
## $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye~
## $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes~
## $ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One ~
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", ~
## $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "~
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7~
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949~
## $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y~
```

```
#telcom_churn %>% summarize(across(everything(), ~sum(is.na(.)), .names = "na_{.col}"))

telcom_churn <- telcom_churn %>%
```

```
  mutate(
    Churn = factor(Churn, levels = c("No","Yes")),
    SeniorCitizen = if (is.numeric(SeniorCitizen)) {
      factor(SeniorCitizen, levels = c(0,1), labels = c("No","Yes"))
    } else {
      factor(SeniorCitizen, levels = c("No","Yes"))
    },
    senior_int = case_when(
      SeniorCitizen %in% c(1, "1", "Yes", "yes", TRUE) ~ 1L,
      SeniorCitizen %in% c(0, "0", "No",  "no",  FALSE) ~ 0L,
      TRUE ~ NA_integer_
    ),
    churn_int = case_when(
      Churn %in% c("Yes","yes", 1, "1", TRUE) ~ 1L,
      Churn %in% c("No", "no",  0, "0", FALSE) ~ 0L,
      TRUE ~ NA_integer_
    )
  )

telcom_churn_clean <- telcom_churn %>%
  filter(!is.na(senior_int), !is.na(churn_int))

# Sanity checks
summary(telcom_churn_clean$Churn)
```

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

```
summary(telcom_churn_clean$SeniorCitizen)
```

```
##   No  Yes
## 5901 1142
```

```
telcom_churn_clean %>%
  summarise(
    n = n(),
    seniors = sum(senior_int == 1),
    non_seniors = sum(senior_int == 0),
    churn_rate_overall = mean(churn_int)
  )
```

```
##      n seniors non_seniors churn_rate_overall
## 1 7043    1142        5901          0.2653699
```

2.2 Probability of custumer churn

```
# ---- 2.2 Probability of customer churn ----
n <- nrow(telcom_churn_clean)
n_churn <- sum(telcom_churn_clean$churn_int)
pi_hat <- n_churn / n
alpha <-  0.05
q <- qnorm((1 - alpha/2))
var_hat <- (pi_hat * (1 - pi_hat)) / n
se <- sqrt(var_hat)
ci_lower <- pi_hat - q * se
ci_upper <- pi_hat + q * se
# Also using other methods to verify my own calculation.
binom.test(n_churn, n)$conf.int
```

```
## [1] 0.2550860 0.2758483
## attr(,"conf.level")
## [1] 0.95
```

```
prop.test(n_churn, n, correct = TRUE)$conf.int
```

```
## [1] 0.2551180 0.2758793
## attr(,"conf.level")
## [1] 0.95
```

2.2 Comment

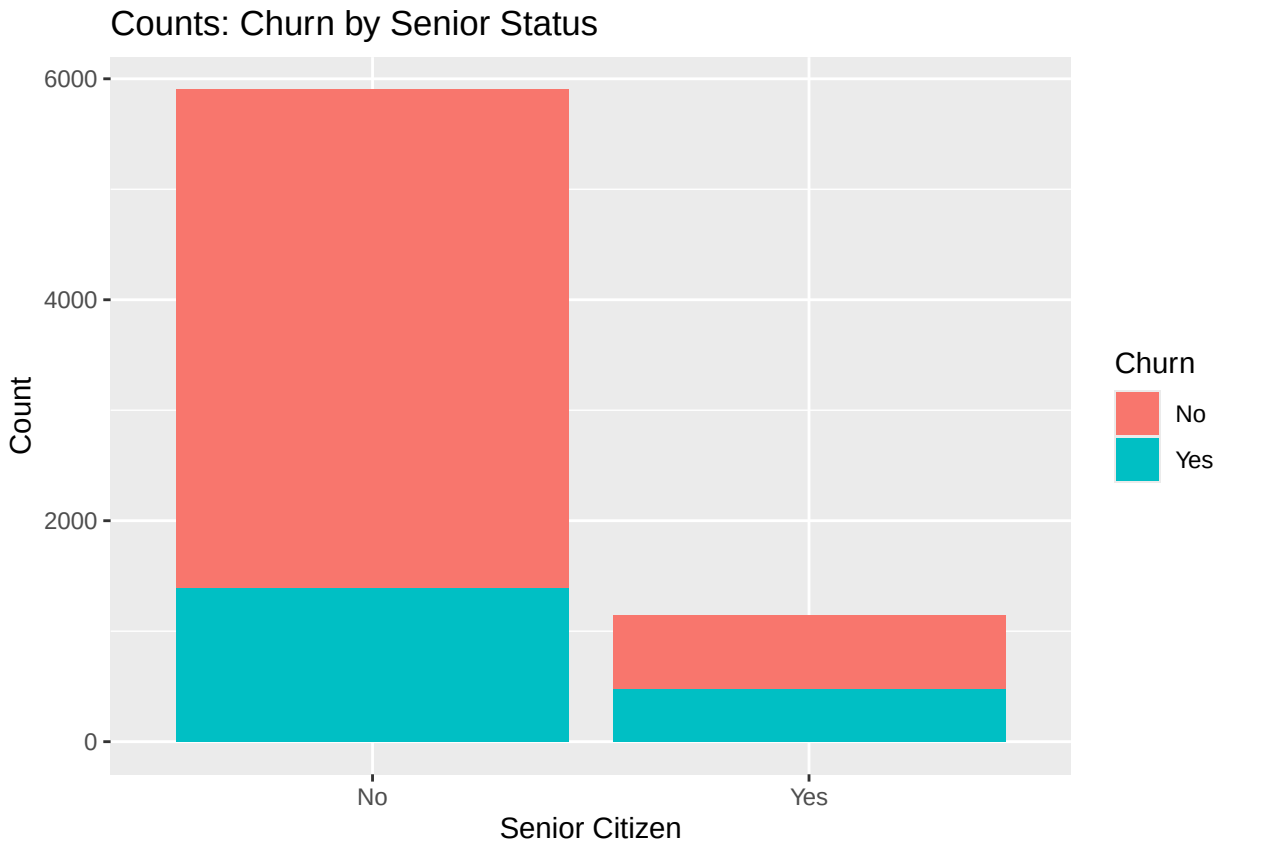
The following code chunk shows that the probability (MLE estimator) $\hat{\pi}$ is 0.265 with a 95% CI between 0.255 and 0.276. Because the CI does not cover 0, $\hat{\pi}$ is therefore **statistically** different from 0. We can of course carry out a 2-sided hypothesis testing and the conclusion is the same.

2.3 Bar plots

I created 2 bar plots, one showing counts and the other percentage of churn for each group. Based on these plots, there is a visually obvious difference between the senior and non-senior groups in terms of churn.

```
# ---- 2.3 Plots (two-bar summary) ----
library(ggplot2)

ggplot(telcom_churn_clean, aes(x = SeniorCitizen, fill = Churn)) +
  geom_bar() +
  labs(title = "Counts: Churn by Senior Status", x = "Senior Citizen", y = "Count")
```

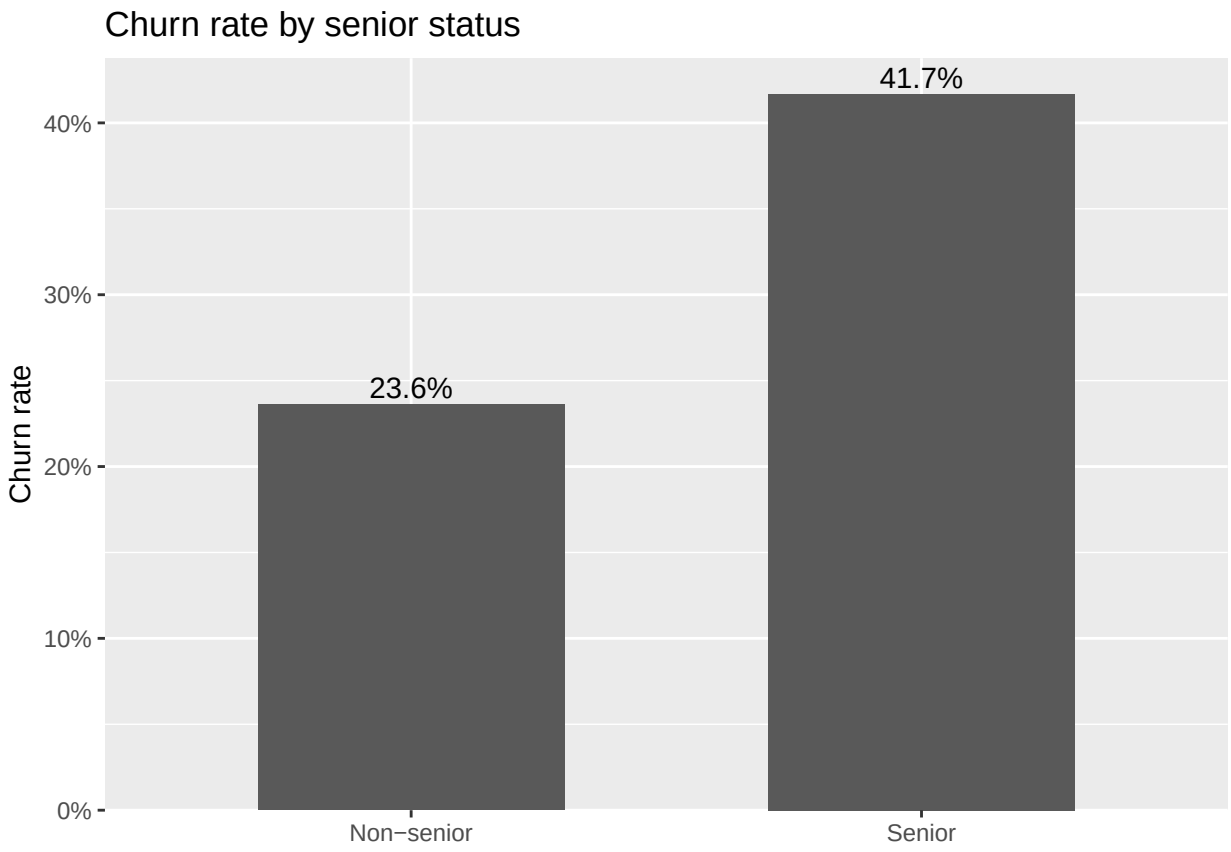



```
churn_by_senior <- telcom_churn_clean %>%
  group_by(senior_int) %>%
  summarise(
    n = n(),
    churn_rate = mean(churn_int),
    .groups = "drop"
  ) %>%
  mutate(group = if_else(senior_int == 1L, "Senior", "Non-senior"))

churn_by_senior
```

```
## # A tibble: 2 x 4
##   senior_int     n churn_rate group
##        <int> <int>      <dbl> <chr>
## 1          0  5901      0.236 Non-senior
## 2          1  1142      0.417 Senior
```

```r
ggplot(churn_by_senior, aes(x = group, y = churn_rate)) +
  geom_col(width = 0.6) +
  geom_text(aes(label = percent(churn_rate, accuracy = 0.1)),
            vjust = -0.3, size = 3.8) +
  scale_y_continuous(labels = percent_format(),
                     expand = expansion(mult = c(0, 0.05))) +
  labs(title = "Churn rate by senior status", x = NULL, y = "Churn rate")
```



2.4 Contingency table

```
# ---- 2.4 Tables + group rates  ----
ct_table <- with(telcom_churn_clean, table(
  `Senior status` = ifelse(senior_int == 1L, "Senior", "Non-senior"),
  `Churn status`  = ifelse(churn_int  == 1L, "Yes",    "No")
))
#ct_table
#prop_by_row <- prop.table(ct_table, margin = 1)
#prop_by_row
ct_table_sum <- addmargins(ct_table)

group_rates <- telcom_churn_clean %>%
  group_by(senior_int) %>%
  summarise(
    n          = n(),
    churn_yes  = sum(churn_int == 1L),
    churn_rate = churn_yes / n,
    .groups    = "drop"
  ) %>%
  mutate(group = if_else(senior_int == 1L, "Senior", "Non-senior")) %>%
  select(group, n, churn_yes, churn_rate)
#group_rates

rate_senior <- group_rates$churn_rate[group_rates$group == "Senior"]
rate_non    <- group_rates$churn_rate[group_rates$group == "Non-senior"]
diff_rate   <- rate_senior - rate_non
sprintf("Senior churn rate: %.2f%% | Non-senior: %.2f%% | Difference: %.2f percentage points",
        100*rate_senior, 100*rate_non, 100*diff_rate)
```

```
## [1] "Senior churn rate: 41.68% | Non-senior: 23.61% | Difference: 18.08 percentage points"
```

```
ct_table_sum
```

```
##              Churn status
## Senior status   No  Yes  Sum
##    Non-senior 4508 1393 5901
##    Senior      666  476 1142
##    Sum        5174 1869 7043
```

2.4 Comment

Based on the contingency table and the probabilities, there is a pratical difference bewtween the 2 groups in churn.

2.5 two confidence intervals

```
# ---- 2.5 Wald & Agresti-Caffo ----
ct_table
```

```
##              Churn status
## Senior status   No  Yes
##    Non-senior 4508 1393
##    Senior      666  476
```

```
count_senior_churn <- ct_table["Senior", "Yes"]
count_senior_nchurn <- ct_table["Senior", "No"]
count_nsenior_churn <- ct_table["Non-senior", "Yes"]
count_nsenior_nchurn <- ct_table["Non-senior", "No"]

n_seniors <- count_senior_churn + count_senior_nchurn
n_nseniors <- count_nsenior_churn + count_nsenior_nchurn
n_churns <- count_senior_churn + count_nsenior_churn
n_nchurns <- count_senior_nchurn + count_nsenior_nchurn

stopifnot(n_seniors + n_nseniors == sum(ct_table))
stopifnot(n_churns  + n_nchurns  == sum(ct_table))

pi_hat_senior <- count_senior_churn / n_seniors
pi_hat_nsenior <- count_nsenior_churn / n_nseniors
pi_hat_senior; pi_hat_nsenior
```

```
## [1] 0.4168126
```

```
## [1] 0.2360617
```

```
pi_diff_hat <- pi_hat_senior - pi_hat_nsenior
pi_diff_hat
```

```
## [1] 0.1807509
```

```
pi_diff_hat_se <- sqrt((pi_hat_senior * (1 - pi_hat_senior)) / n_seniors +
                         (pi_hat_nsenior * (1 - pi_hat_nsenior)) / n_nseniors)
pi_diff_hat_se
```

```
## [1] 0.01560176
```

```
wald_pi_diff_ci_lower <- pi_diff_hat - q * pi_diff_hat_se
wald_pi_diff_ci_upper <- pi_diff_hat + q * pi_diff_hat_se

pi_hat_senior_ac <- (count_senior_churn + 1) / (n_seniors + 2)
pi_hat_nsenior_ac <- (count_nsenior_churn + 1) / (n_nseniors + 2)
pi_diff_hat_ac <- pi_hat_senior_ac - pi_hat_nsenior_ac
pi_diff_hat_ac
```

```
## [1] 0.1808069
```

```
pi_diff_hat_ac_se <- sqrt((pi_hat_senior_ac * (1 - pi_hat_senior_ac)) / (n_seniors + 2) +
                         (pi_hat_nsenior_ac * (1 - pi_hat_nsenior_ac)) / (n_nseniors + 2))
ac_pi_diff_ci_lower <- pi_diff_hat_ac - q * pi_diff_hat_ac_se
ac_pi_diff_ci_upper <- pi_diff_hat_ac + q * pi_diff_hat_ac_se

ac_pi_diff_ci_upper - ac_pi_diff_ci_lower
```

```
## [1] 0.06111336
```

```
wald_pi_diff_ci_upper - wald_pi_diff_ci_lower
```

```
## [1] 0.06115777
```

```
# Also using existing method to verify my own calculation
prop.test(x = c(count_senior_churn, count_nsenior_churn),
          n = c(n_seniors, n_nseniors),
          correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(count_senior_churn, count_nsenior_churn) out of c(n_seniors, n_nseniors)
## X-squared = 159.43, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1496495 0.2118524
## sample estimates:
##    prop 1    prop 2
## 0.4168126 0.2360617
```

```
#tibble::tibble(
#  metric = "Diff (Senior - Non-senior)",
#  point  = pi_diff_hat,
#  Wald_L = wald_pi_diff_ci_lower, Wald_U = wald_pi_diff_ci_upper,
#  AC_L   = ac_pi_diff_ci_lower,   AC_U   = ac_pi_diff_ci_upper
#)
```

2.5 Comment

The MLE estimates for senior churn probability is 0.417 and 0.236 for non-seniors. The estimate for difference is therefore 0.18075.

The Wald CI ($\alpha = 0.05$) is between 0.15017 and 0.21133.

For Agresti-Caffo CI, the adjusted difference (according to the textbook and the async) is 0.18081, and the Agresti-Caffo CI is between 0.15025 and 0.21136.

2.6 Hypothesis testing

```
# ---- 2.6 Hypothesis testing ----
z_value <- pi_diff_hat / pi_diff_hat_se
z_value
```

```
## [1] 11.58529
```

```
p_value <- 2 * (1 - pnorm(abs(z_value)))
p_value
```

```
## [1] 0
```

```
p_value < 0.001
```

```
## [1] TRUE
```

```
# So even given a significance level of (1 - 0.001 = 99.9%),
# We can reject the null that there is no difference.
```

2.6 Comment

Given the null hypothesis, the test statistic is 11.59 and using the limiting distribution (Gaussian), its corresponding (two-sided) p-value is 0. Because the test statistic is far beyond the critical value, we reject the null in favor of the alternative.

2.7 Relative risk (rr)

```
# ---- 2.7 Relative risk ----
rr_hat <- pi_hat_senior / pi_hat_nsenior
rr_hat
```

```
## [1] 1.765694
```

```
var_log_rr <- (1 / count_senior_churn) - (1 / n_seniors) + (1 / count_nsenior_churn) - (1 / n_nseniors)
se_log_rr <- sqrt(var_log_rr)
log_rr_ci_lower <- log(rr_hat) - q * se_log_rr
log_rr_ci_upper <- log(rr_hat) + q * se_log_rr
rr_ci_lower <- exp(log_rr_ci_lower)
rr_ci_upper <- exp(log_rr_ci_upper)
```

2.7 Comment

The relative risk is 1.766, with a 95% CI (calculated first using log odds then exponentiate) between 1.626 and 1.918.

That is the senior group is 177% as likely as non-senior group to churn, or that the senior group is 77% more likely to churn compared to the non-senior group.

This is consistent with findings from previous parts where the senior churn estimate is 0.417 and the the non-senior churn estimate is 0.236

2.8 Odds ratio

```
# ---- 2.8 Odds ratio  ----
odds_senior <- pi_hat_senior / (1 - pi_hat_senior)
odds_nsenior <- pi_hat_nsenior / (1 - pi_hat_nsenior)
odds_ratio <- odds_senior / odds_nsenior
var_log_or <- (1 / count_senior_churn) + (1 / count_senior_nchurn) + (1 / count_nsenior_churn) + (1 / co
se_log_or <- sqrt(var_log_or)
log_or_ci_upper <- log(odds_ratio) + q * se_log_or
log_or_ci_lower <- log(odds_ratio) - q * se_log_or
or_ci_upper <- exp(log_or_ci_upper)
or_ci_lower <- exp(log_or_ci_lower)
```

2.8 Comment

Odds (MLE) for senior to churn: 0.715.

Odds (MLE) for non-senior to churn: 0.309.

The odds ratio (MLE) is 2.313.

The 95% CI for odds ratio is (calculated first using log then exponentiate) between 2.027 and 2.64.