# Hypothesis Testing with Men's and Women's Soccer Matches

Lunjing Yuan

2023-08-29

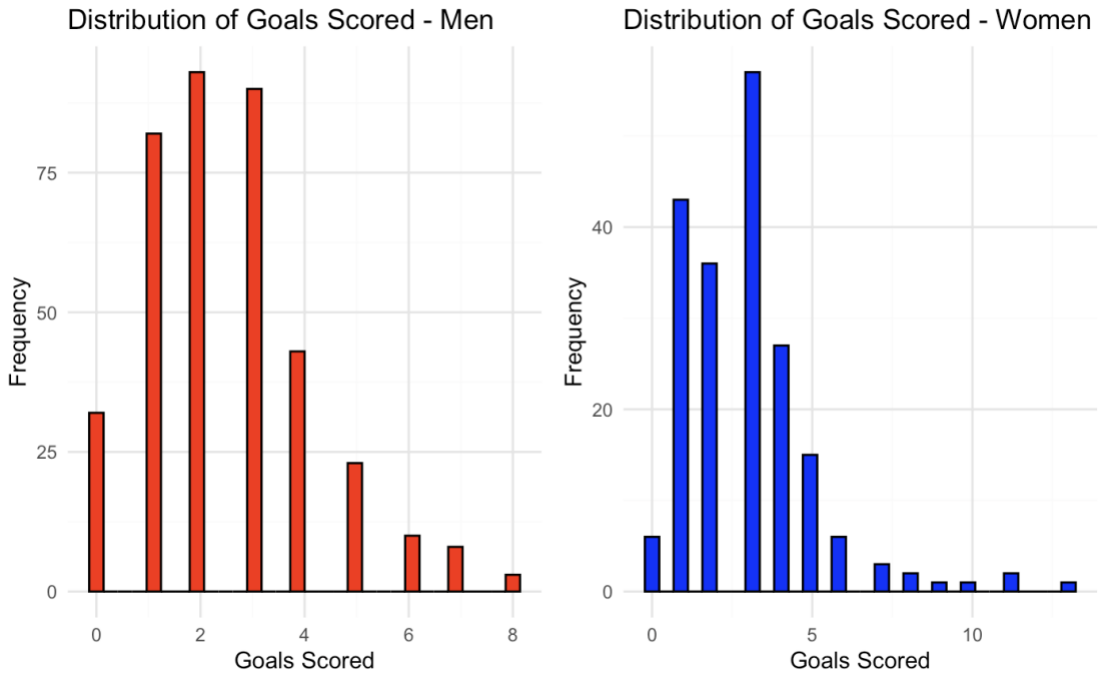
```
#setwd(" ")
setwd("~/Desktop/personal web/r projecrt/Hypothesis Testing with Men's and
Women's Soccer Matches")

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.2.1     ✓ dplyr   1.1.2
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.1.4     ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

# Load men's and women's datasets
women <- read_csv("women_results.csv")

## New names:
## Rows: 4884 Columns: 7
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (3): home_team, away_team, tournament dbl (3): ...1, home_score,
away_score
## date (1): date
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...1`
```

```r
men <- read_csv("men_results.csv")
```

```
## New names:
## Rows: 44353 Columns: 7
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (3): home_team, away_team, tournament dbl (3): ...1, home_score,
away_score
## date (1): date
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...1`
```

```r
#view(women)
#view(men)

# Filtering the matches and creating the test values
men <- men %>%
          filter(tournament == "FIFA World Cup", date > "2002-01-01") %>%
          mutate(goals_scored = home_score + away_score)

women <- women %>%
          filter(tournament == "FIFA World Cup", date > "2002-01-01") %>%
          mutate(goals_scored = home_score + away_score)

# Determine normality using histograms
men_plot <- ggplot(men, aes(x = goals_scored)) +
  geom_histogram(fill = "red", color = "black", bins = 30) +
  ggtitle("Distribution of Goals Scored - Men") +
  xlab("Goals Scored") +
  ylab("Frequency") +
  theme_minimal()

women_plot <- ggplot(women, aes(x = goals_scored)) +
  geom_histogram(fill = "blue", color = "black", bins = 30) +
  ggtitle("Distribution of Goals Scored - Women") +
  xlab("Goals Scored") +
  ylab("Frequency") +
  theme_minimal()

# Add gridlines
men_plot <- men_plot + theme(panel.grid.major = element_line(color =
"gray90"),
                             panel.grid.minor = element_line(color = "gray98"))

women_plot <- women_plot + theme(panel.grid.major = element_line(color =
"gray90"),
```

```
                                panel.grid.minor = element_line(color =
"gray98"))

# Goals scored is not normally distributed, so use Wilcoxon-Mann-Whitney test
of two groups
grid.arrange(men_plot, women_plot, nrow = 1)
```




```
# Run a Wilcoxon-Mann-Whitney test on goals_scored vs. group
test_results <- wilcox.test(
  x = women$goals_scored,
  y = men$goals_scored,
  alternative = "greater"
)

# Determine hypothesis test result using sig. level
p_val <- round(test_results$p.value, 4)
result <- ifelse(p_val <= 0.01, "reject", "fail to reject")

# Create the result data frame
result_df <- data.frame(p_val, result)
result_df

##    p_val result
## 1 0.0051 reject
```