STAT 3333 - Final Project

Tuan Pham - STAT Infererence

2023-11-30

```
##         date home_team away_team home_score away_score tournament    city
## 1 1872-11-30  Scotland   England          0          0   Friendly Glasgow
## 2 1873-03-08   England  Scotland          4          2   Friendly  London
## 3 1874-03-07  Scotland   England          2          1   Friendly Glasgow
## 4 1875-03-06   England  Scotland          2          2   Friendly  London
## 5 1876-03-04  Scotland   England          3          0   Friendly Glasgow
## 6 1876-03-25  Scotland     Wales          4          0   Friendly Glasgow
##    country neutral
## 1 Scotland   FALSE
## 2  England   FALSE
## 3 Scotland   FALSE
## 4  England   FALSE
## 5 Scotland   FALSE
## 6 Scotland   FALSE
```

Part 3: Permutation Test

Claim: Perform the permutation test to check whether the Argentina national team has scored more in the friendly games than in the World Cup.

Hypothesis

> Ho: mean goals scored at Friendly games - mean goals scored at World Cup <= 0
>
> Ha: mean goals scored at Friendly games - mean goals scored at World Cup > 0

Argentina Games at World Cup

```r
Arg.WC <- international.games %>%
  filter(tournament == 'FIFA World Cup') %>%
  filter(home_team == 'Argentina' | away_team == 'Argentina')
```

Argentina Goals Scored at World Cup

```r
Arg.goal.scored.as.Home.Team.inWC <- Arg.WC %>%
  filter(home_team == 'Argentina') %>%
  select('Goal scored' = home_score)
Arg.goal.scored.as.Away.Team.inWC <- Arg.WC %>%
  filter(away_team == 'Argentina') %>%
  select('Goal scored' = away_score)
Arg.goal.scored.WC <- rbind(Arg.goal.scored.as.Home.Team.inWC, Arg.goal.scored.as.Away.Team.inWC)
Arg.goal.scored.WC <- subset(Arg.goal.scored.WC, drop = TRUE)
mean(Arg.goal.scored.WC)
```

```
## [1] 1.727273
```

Argentina Games at Friendlies

```
Arg.friendly <- international.games %>%
  filter(tournament == 'Friendly') %>%
  filter(home_team == 'Argentina' | away_team == 'Argentina')
```

Argentina Goal Scored at Friendly Games

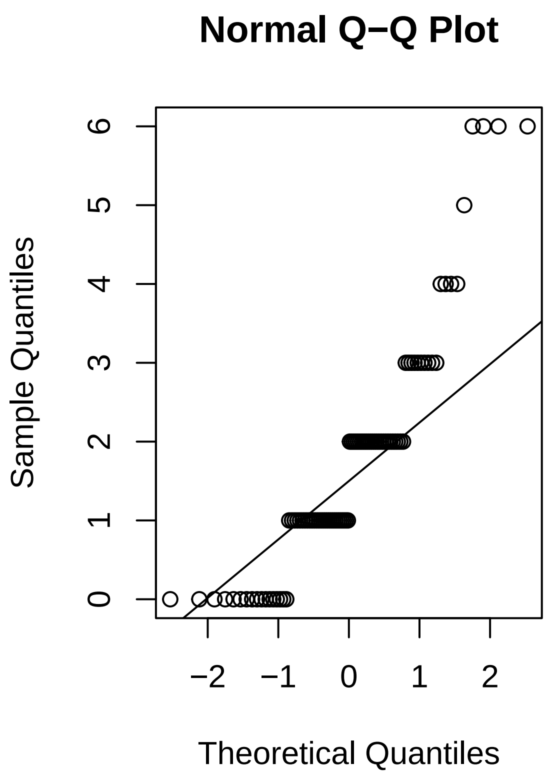
```
Arg.goal.scored.as.Home.Team.inFL <- Arg.friendly %>%
  filter(home_team == 'Argentina') %>%
  select('Goal scored' = home_score)
Arg.goal.scored.as.Away.Team.inFL <- Arg.friendly %>%
  filter(away_team == 'Argentina') %>%
  select('Goal scored' = away_score)
Arg.goal.scored.FL <- rbind(Arg.goal.scored.as.Home.Team.inFL, Arg.goal.scored.as.Away.Team.inFL)
Arg.goal.scored.FL <- subset(Arg.goal.scored.FL, drop = TRUE)
mean(Arg.goal.scored.FL)
```

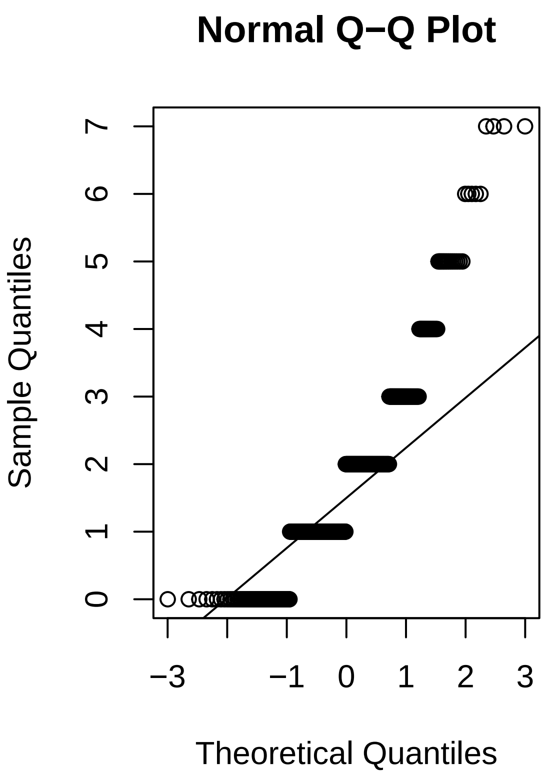
```
## [1] 1.779891
```

```
Arg.goal.scored <- c(Arg.goal.scored.FL, Arg.goal.scored.WC) # Combine 2 vectors in preparing for resam
```

Data Normality Check

```
par(mfrow=c(1,2))
qqnorm(Arg.goal.scored.WC)
qqline(Arg.goal.scored.WC)

qqnorm(Arg.goal.scored.FL)
qqline(Arg.goal.scored.FL)
```





Data are not normal for both World Cup Games and frienly ones.

Mean difference between Argentina goals scored in World Cup vs Friendly games

```
Arg.mean.goal.dff.observed <- mean(Arg.goal.scored.FL) - mean(Arg.goal.scored.WC)
Arg.mean.goal.dff.observed
```

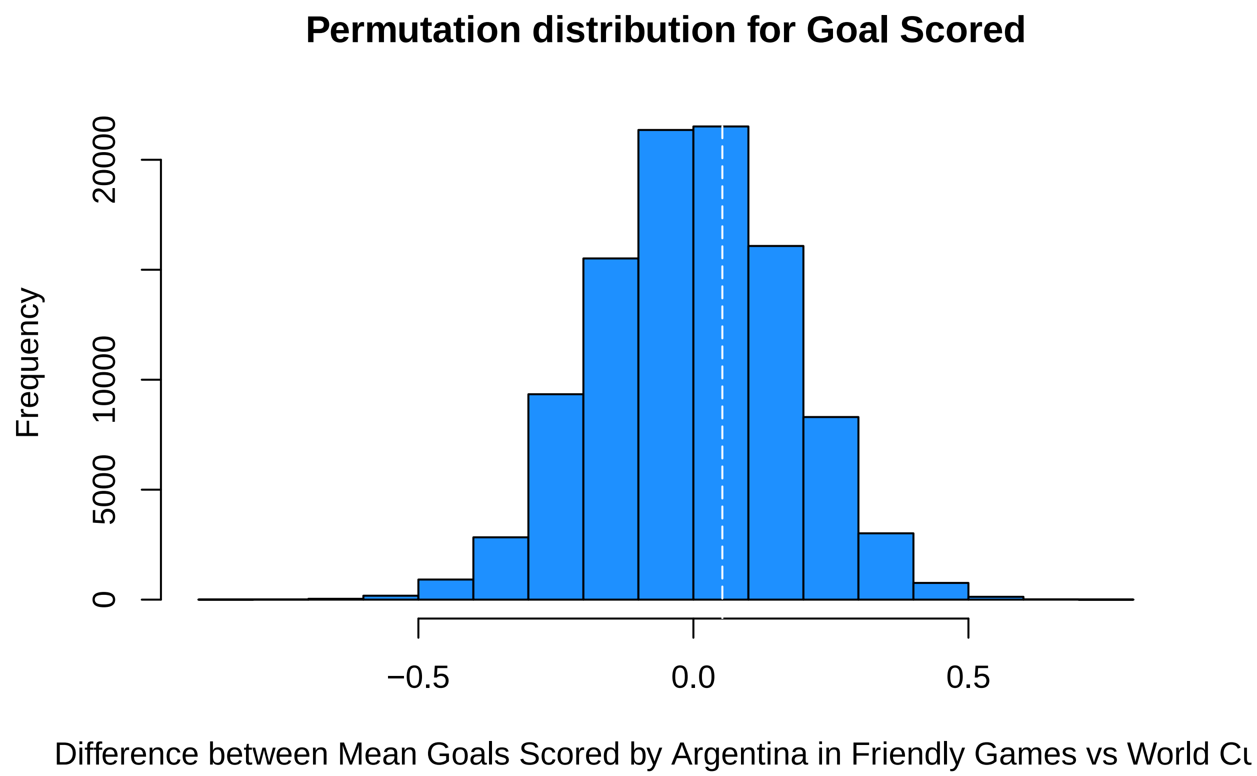
```
## [1] 0.05261858
```

Resampling by Permutation

```
 N <- 10^5 - 1  #set number of times to repeat this process
 result <- numeric(N) # space to save the random differences

 for(i in 1:N) {
  index <- sample(456, size = 368, replace = FALSE) # sample of numbers from 1:30
  result[i] <- mean(Arg.goal.scored[index]) - mean(Arg.goal.scored[-index])
}
```

Resampling Plot

```
hist(result, xlab = "Difference between Mean Goals Scored by Argentina in Friendly Games vs World Cup",
abline(v = Arg.mean.goal.dff.observed, col = "White", lty = 5)
```



Obtain the p-value

```
(sum(result >= Arg.mean.goal.dff.observed) + 1)/(N + 1)  #P-value
```

```
## [1] 0.40248
```

Since P-value is greater than the 0.05 significance level, we failed to reject the null. Consequently, we don't have sufficient evidence to claim that there's a statistically significant difference between the mean goals scored at friendly games versus those scored at the World Cup based on the data analyzed.

Typically, it's expected that strong national teams such as Argentina would score more goals in friendly games compared to crucial tournaments like the World Cup. However, surprisingly, the statistical data doesn't provide evidence to back up this belief.

Percent Bias

```
Arg.bias <- 100 * ((mean(Arg.goal.scored.FL) - mean(Arg.goal.scored.WC))/mean(Arg.goal.scored.FL))
Arg.bias
```

```
## [1] 2.95628
```

The bias between the two subjects' datasets is 2.96%, which indicates the difference in sample sizes contributes to a small bias in comparison of mean between two groups.

Brazil Games at World Cup

```r
Bra.WC <- international.games %>%
  filter(tournament == 'FIFA World Cup') %>%
  filter(home_team == 'Brazil' | away_team == 'Brazil')
```

Brazil Goals Scored at World Cup

```r
Bra.goal.scored.as.Home.Team.inWC <- Bra.WC %>%
  filter(home_team == 'Brazil') %>%
  select('Goal scored' = home_score)

Bra.goal.scored.as.Away.Team.inWC <- Bra.WC %>%
  filter(away_team == 'Brazil') %>%
  select('Goal scored' = away_score)

Bra.goal.scored.WC <- rbind(Bra.goal.scored.as.Home.Team.inWC, Bra.goal.scored.as.Away.Team.inWC)
Bra.goal.scored.WC <- subset(Bra.goal.scored.WC, drop = TRUE)
mean(Bra.goal.scored.WC)
```

```
## [1] 2.078947
```

Brazil Games at Friendlies

```r
Bra.friendly <- international.games %>%
  filter(tournament == 'Friendly') %>%
  filter(home_team == 'Brazil' | away_team == 'Brazil')
```

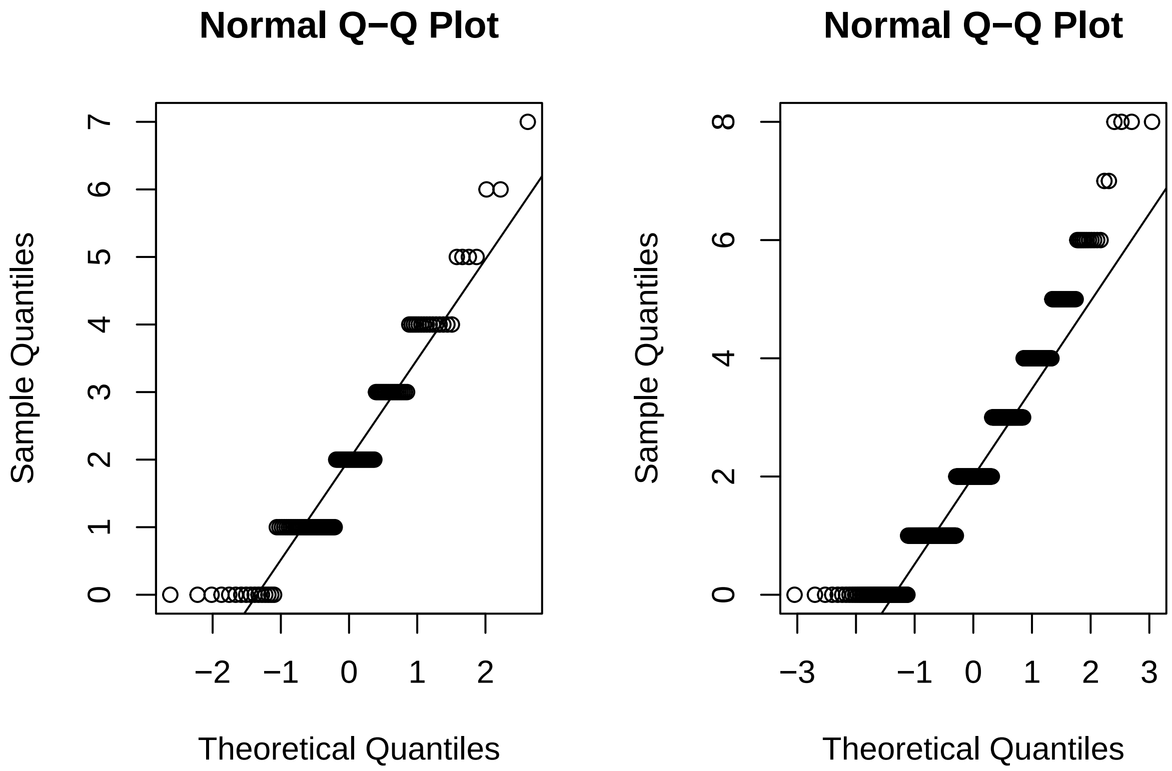
Brazil Goal Scored at Friendly Games

```r
Bra.goal.scored.as.Home.Team.inFL <- Bra.friendly %>%
  filter(home_team == 'Brazil') %>%
  select('Goal scored' = home_score)

Bra.goal.scored.as.Away.Team.inFL <- Bra.friendly %>%
  filter(away_team == 'Brazil') %>%
  select('Goal scored' = away_score)

Bra.goal.scored.FL <- rbind(Bra.goal.scored.as.Home.Team.inFL, Bra.goal.scored.as.Away.Team.inFL)
Bra.goal.scored.FL <- subset(Bra.goal.scored.FL, drop = TRUE)
mean(Bra.goal.scored.FL)
```

```
## [1] 2.208333
```

Data Normality Check

```r
par(mfrow=c(1,2))
qqnorm(Bra.goal.scored.WC)
qqline(Bra.goal.scored.WC)

qqnorm(Bra.goal.scored.FL)
qqline(Bra.goal.scored.FL)
```



Data are not normal for both World Cup Games and frienly ones.

```
Bra.goal.scored <- c(Bra.goal.scored.FL, Bra.goal.scored.WC) # Combine to vectors of goals
```

Mean difference between Brazil goals scored in World Cup vs Friendly

```
Bra.mean.goal.dff.observed <- mean(Bra.goal.scored.FL) - mean(Bra.goal.scored.WC)
Bra.mean.goal.dff.observed
```

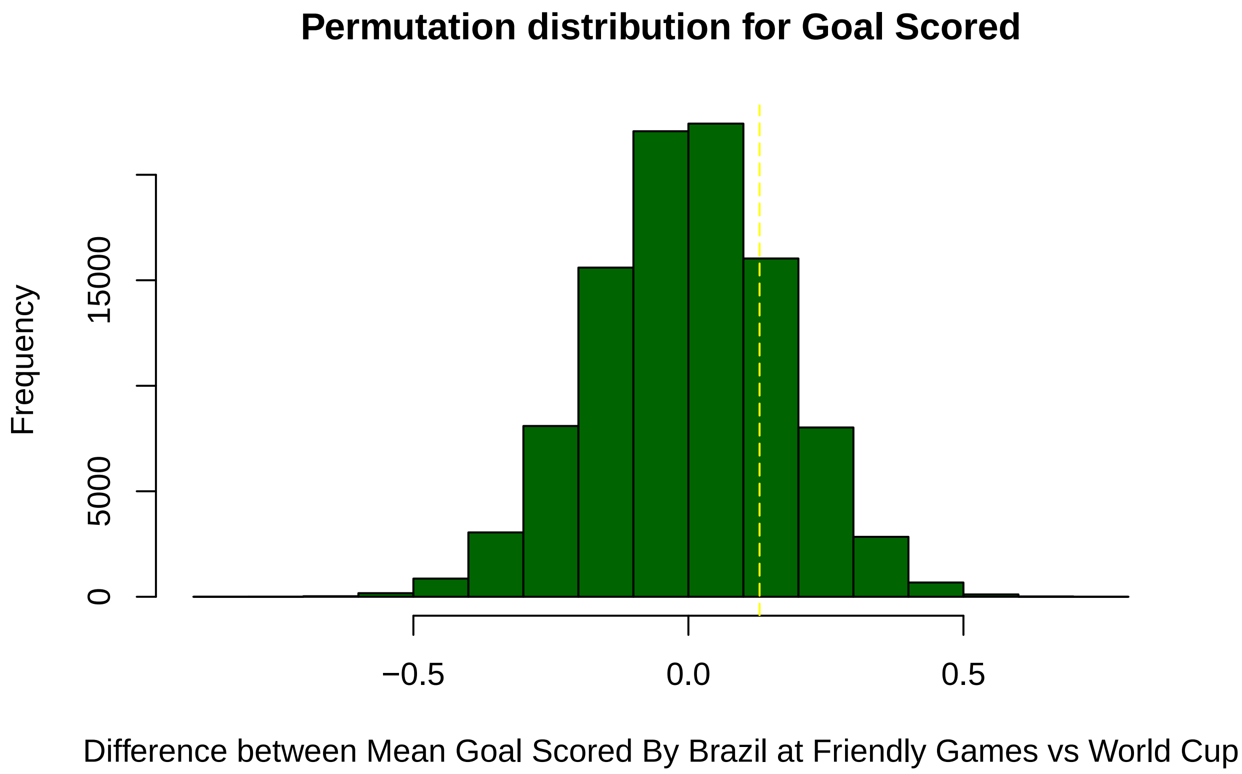
```
## [1] 0.129386
```

Resampling the data

```
 N <- 10^5 - 1  #set number of times to repeat this process
 Bra.result <- numeric(N) # space to save the random differences

 for(i in 1:N) {
  index <- sample(546, size = 432, replace = FALSE)
  Bra.result[i] <- mean(Bra.goal.scored[index]) - mean(Bra.goal.scored[-index])
}
```

Plot the samples

```r
hist(Bra.result, xlab = "Difference between Mean Goal Scored By Brazil at Friendly Games vs World Cup",
abline(v = Bra.mean.goal.dff.observed, col = "yellow", lty = 5)
```



```
P-value
```

```r
(sum(result >= Bra.mean.goal.dff.observed) + 1)/(N + 1)  #P-value
```

```
## [1] 0.22954
```

Since P-value is greater than the 0.05 significance level, we failed to reject the null. Consequently, we don't have sufficient evidence to claim that there's a statistically significant difference between the mean goals scored at friendly games versus those scored at the World Cup based on the data analyzed.

```
Bias percentage
```

```r
Bra.bias <- 100 * ((mean(Bra.goal.scored.FL) - mean(Bra.goal.scored.WC))/mean(Bra.goal.scored.FL))
Bra.bias
```

```
## [1] 5.858987
```

The bias between the two subjects' datasets is 5.85%, which indicates the difference in sample sizes contributes to a small bias in comparison of mean between two groups.

After analyzing the data, we haven't found enough evidence to assert a statistically significant difference between the average goals scored in friendly games compared to those scored at the World Cup. Our examination of the two most prominent South American teams in World Cup history yielded notably similar results. Now, let's shift our focus to one of Europe's most renowned teams in World Cup history, the Germany National Team.

Germany Games at World Cup

```
Ger.WC <- international.games %>%
  filter(tournament == 'FIFA World Cup') %>%
  filter(home_team == 'Germany' | away_team == 'Germany')
```

Germany Goals Scored at World Cup

```
Ger.goal.scored.as.Home.Team.inWC <- Ger.WC %>%
  filter(home_team == 'Germany') %>%
  select('Goal scored' = home_score)

Ger.goal.scored.as.Away.Team.inWC <- Ger.WC %>%
  filter(away_team == 'Germany') %>%
  select('Goal scored' = away_score)

Ger.goal.scored.WC <- rbind(Ger.goal.scored.as.Home.Team.inWC, Ger.goal.scored.as.Away.Team.inWC)
Ger.goal.scored.WC <- subset(Ger.goal.scored.WC, drop = TRUE)
mean(Ger.goal.scored.WC)
```

```
## [1] 2.071429
```

Germany Games at Friendly Games

```
Ger.friendly <- international.games %>%
  filter(tournament == 'Friendly') %>%
  filter(home_team == 'Germany' | away_team == 'Germany')
```

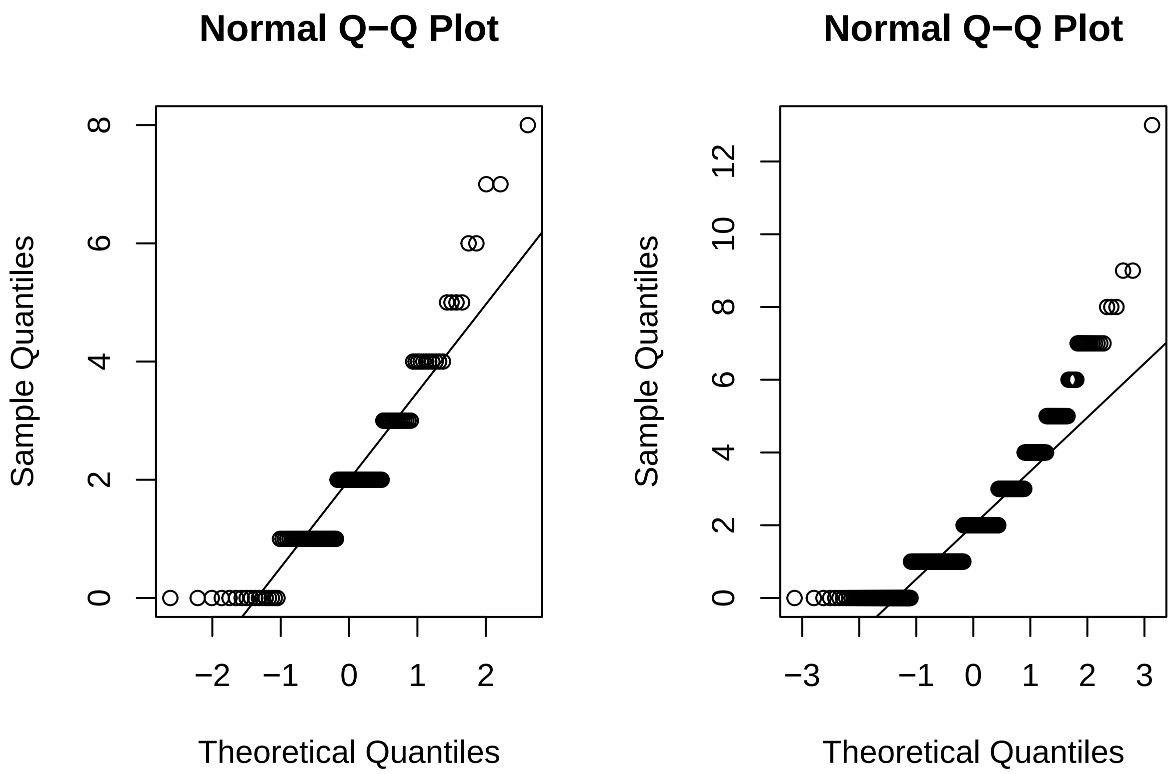
Germany Goals Scored at Friendly Games

```
Ger.goal.scored.as.Home.Team.inFL <- Ger.friendly %>%
  filter(home_team == 'Germany') %>%
  select('Goal scored' = home_score)

Ger.goal.scored.as.Away.Team.inFL <- Ger.friendly %>%
  filter(away_team == 'Germany') %>%
  select('Goal scored' = away_score)

Ger.goal.scored.FL <- rbind(Ger.goal.scored.as.Home.Team.inFL, Ger.goal.scored.as.Away.Team.inFL)

Ger.goal.scored.FL <- subset(Ger.goal.scored.FL, drop = TRUE)
mean(Ger.goal.scored.FL)
```

```
## [1] 2.151986
```

Data Normality Check

```
par(mfrow=c(1,2))
qqnorm(Ger.goal.scored.WC)
qqline(Ger.goal.scored.WC)

qqnorm(Ger.goal.scored.FL)
qqline(Ger.goal.scored.FL)
```



Data are not normal for both World Cup Games and frienly ones.

```
Ger.goal.scored <- c(Ger.goal.scored.FL, Ger.goal.scored.WC) #Combine two vectors, prepare for resampli
```

Obtain the mean difference of the observation

```
Ger.mean.goal.dff.observed <- mean(Ger.goal.scored.FL) - mean(Ger.goal.scored.WC)
Ger.mean.goal.dff.observed
```

```
## [1] 0.08055761
```

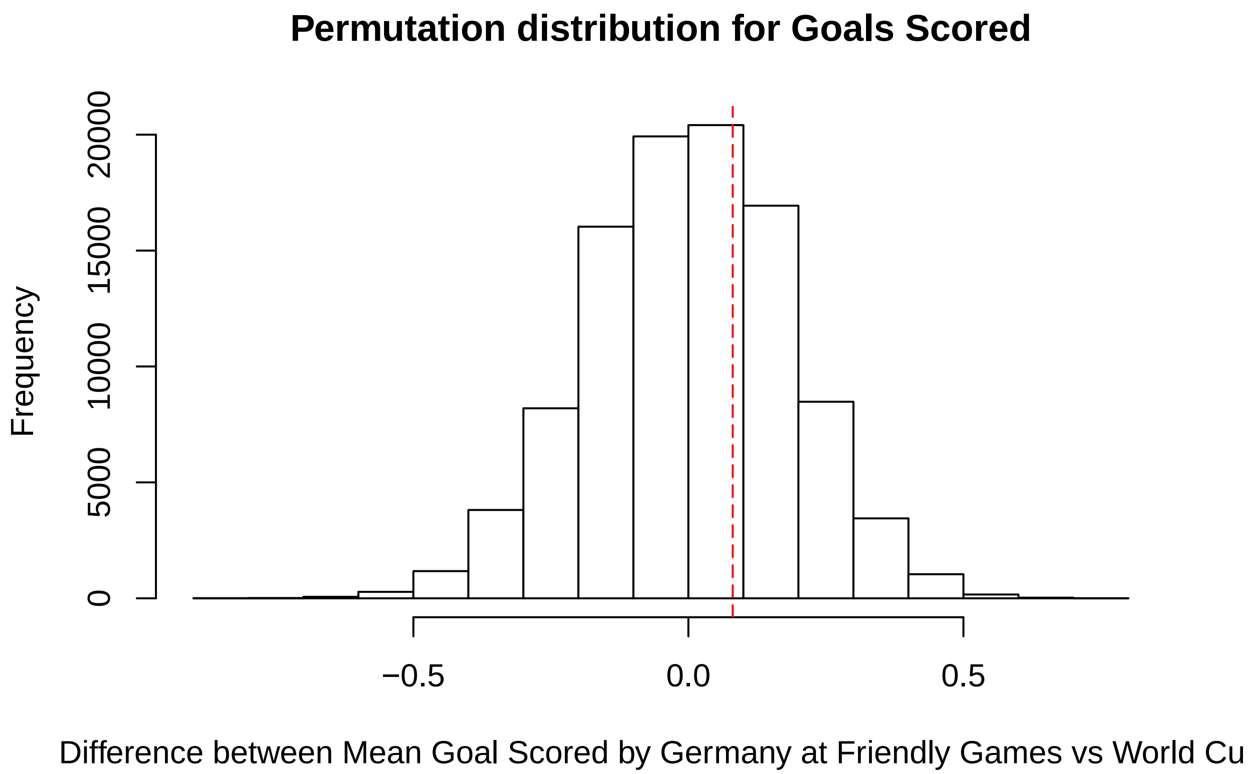
Resampling the data

```
 N <- 10^5 - 1  #set number of times to repeat this process
 Ger.result <- numeric(N) # space to save the random differences
```

```
for(i in 1:N) {
  index <- sample(691, size = 579, replace = FALSE)
  Ger.result[i] <- mean(Ger.goal.scored[index]) - mean(Ger.goal.scored[-index])
}
```

Plot the samples

```
hist(Ger.result, xlab = "Difference between Mean Goal Scored by Germany at Friendly Games vs World Cup"
abline(v = Ger.mean.goal.dff.observed, col = "Red", lty = 5)
```



Obtain p-value

```
(sum(result >= Ger.mean.goal.dff.observed) + 1)/(N + 1)  #P-value
```

```
## [1] 0.34049
```

Since P-value is greater than the 0.05 significance level, we again failed to reject the null. Consequently, we don't have sufficient evidence to claim that there's a statistically significant difference between the mean goals scored at friendly games versus those scored at the World Cup based on the data analyzed.

Bias percentage between two datasets

```
Ger.bias <- 100 * ((mean(Ger.goal.scored.FL) - mean(Ger.goal.scored.WC))/mean(Ger.goal.scored.FL))
Ger.bias
```

```
## [1] 3.743407
```

> The bias between the two subjects' datasets is 3.74%, which indicates the difference in sample sizes contributes to a small bias in comparison of mean between two groups.

Conclusion:

The World Cup holds the highest prestige among tournaments globally, leading us to instinctively assume that scoring during this event is more challenging compared to friendly games. However, based on the data analysis conducted among the three most successful teams in World Cup history, we can infer that there's no statistically significant evidence supporting this assertion. Whether playing at friendly games or the World Cup, the difficulty in scoring goals appears to be increasingly similar and without significant distinction.