Score Project

By: CDT Bridget Ge & Claire Tsay

Introduction

- Soccer
- Foot dominance vs. free kick accuracy
- Using the Two-sample t-test to compare foot dominance and free-kick accuracy

Data

- Data Source obtained from Kaggle
- 25,000 match data from 2008 to 2016 in 11 European countries with more than 10,000 participating players
- Seven tables: Country, League, Match, Player, Player_Attributes, Team, and Team_Attributes

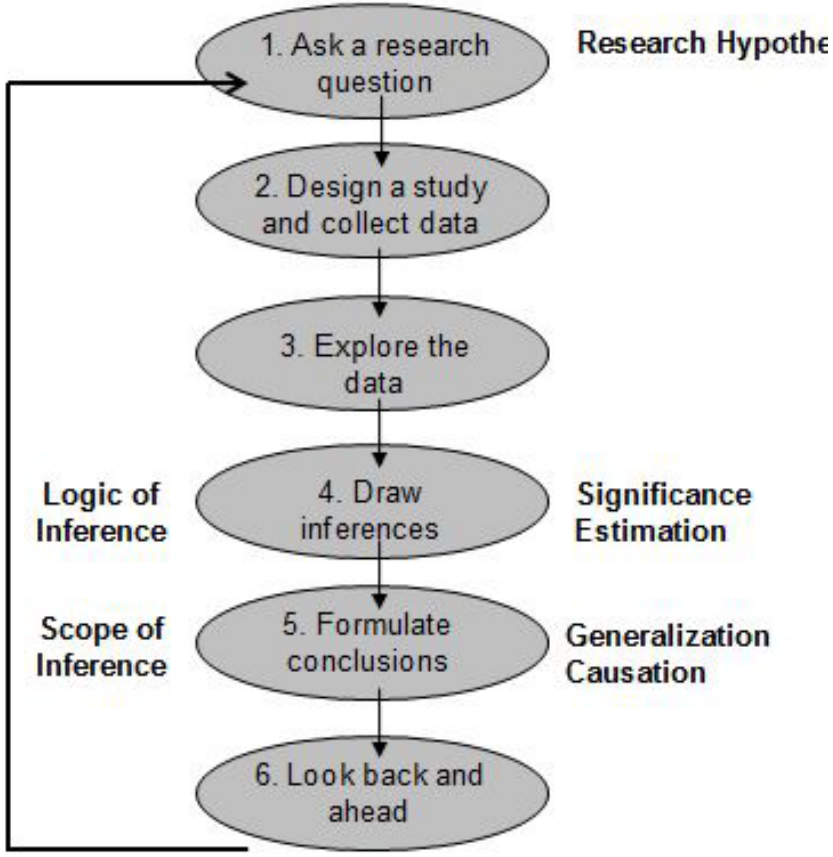
Learning Goals

- Apply the six steps of the statistical investigation method to compare two groups on a quantitative response.
- Calculate the five-number summary (quartiles) and create histograms and boxplots to explore the data from two groups with a quantitative response variable.
- Assess the statistical significance of the observed difference between the two groups.
- Use the 2SD method to estimate a confidence interval for the difference between two means.
- Determine the strength of evidence using the theory-based approach two-sample t-test for comparing two means.

Methods/Instructional Content

- Introduction to Statistical Investigations (chapter 6: comparing two means)
  - Six steps of statistical investigation
  - $t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{Statistic}{SE(standard\ error)}$

- Intermediate Statistical Investigations
  - General layout of the module
  - Types of questions to include in the module



Step 1: Ask a Research Question

- Let’s identify the research question first!
  - Research question: Is there an association between dominant foot and free kick accuracy for players in European football leagues?

Step 2: Design a Study and Collect Data

- Now identify the observational units and null & alternative hypothesis
  - Observational units: 10,000 players participated in European Football leagues matches from 2008 to 2016
  - Null Hypothesis: There is no association between the preferred foot and free kick accuracy score for European football players
  - Alternative Hypothesis: There is an association between preferred foot and free kick accuracy score for European football players

Step 3: Explore the Data

```
#Data frame with some interesting attributes
select_player_attributes <- Player_Attributes %>%
  drop_na() %>%
  select(preferred_foot, ball_control, free_kick_accuracy, overall_rating)

#Table to show interesting attributes
select_player_attributes %>%
  group_by(preferred_foot) %>%
  summarize(mean_BC = mean(ball_control), mean_FCA = mean(free_kick_accuracy), mean_OR = mean(overall_r
  kable(col.names = c("Preferred Foot", "Mean Ball Control Score", "Mean Free Kick Accuracy", "Mean Rat
```

Preferred Foot	Mean Ball Control Score	Mean Free Kick Accuracy	Mean Rating	Number of Observations
left	65.44723	53.29136	68.65282	44107
right	62.80853	48.13070	68.62965	136247

Step 3: Explore the Data

Five-number summaries of the two populations

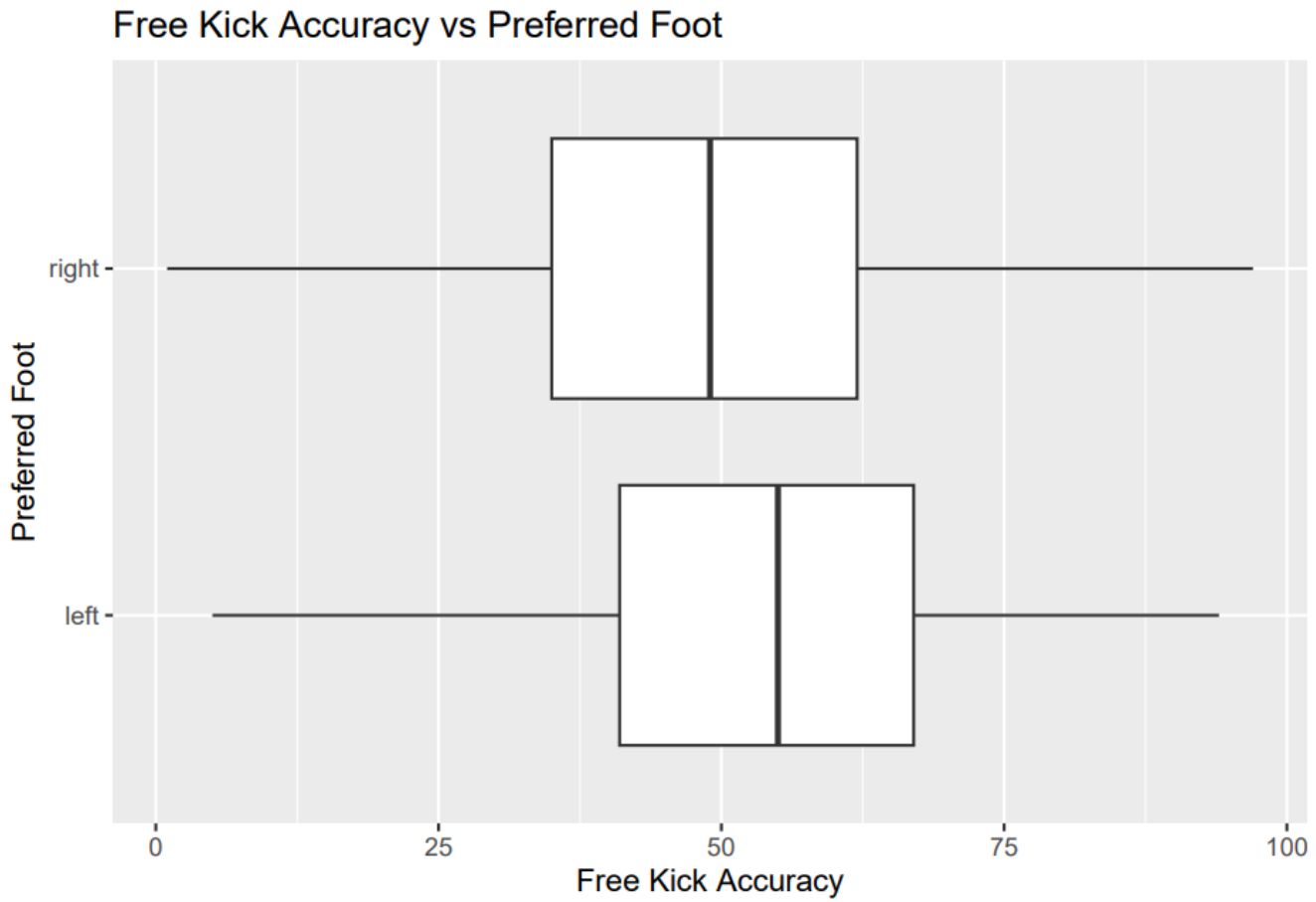
```
#Five number summary of each of the two populations
select_player_attributes %>%
  group_by(preferred_foot) %>%
  summarize(Minimum = min(free_kick_accuracy),
            LowerQuartile = quantile(prob =.25, free_kick_accuracy),
            Median = median(free_kick_accuracy),
            UpperQuartile = quantile(prob=.75, free_kick_accuracy),
            Maximum = max(free_kick_accuracy))
```

```
## # A tibble: 2 x 6
##   preferred_foot Minimum LowerQuartile Median UpperQuartile Maximum
##   <chr>            <int>         <dbl>  <int>         <dbl>   <int>
## 1 left                 5            41     55            67      94
## 2 right                1            35     49            62      97
```

Step 3: Explore the Data

Boxplot visualization to compare the five-number summary of each populationion group

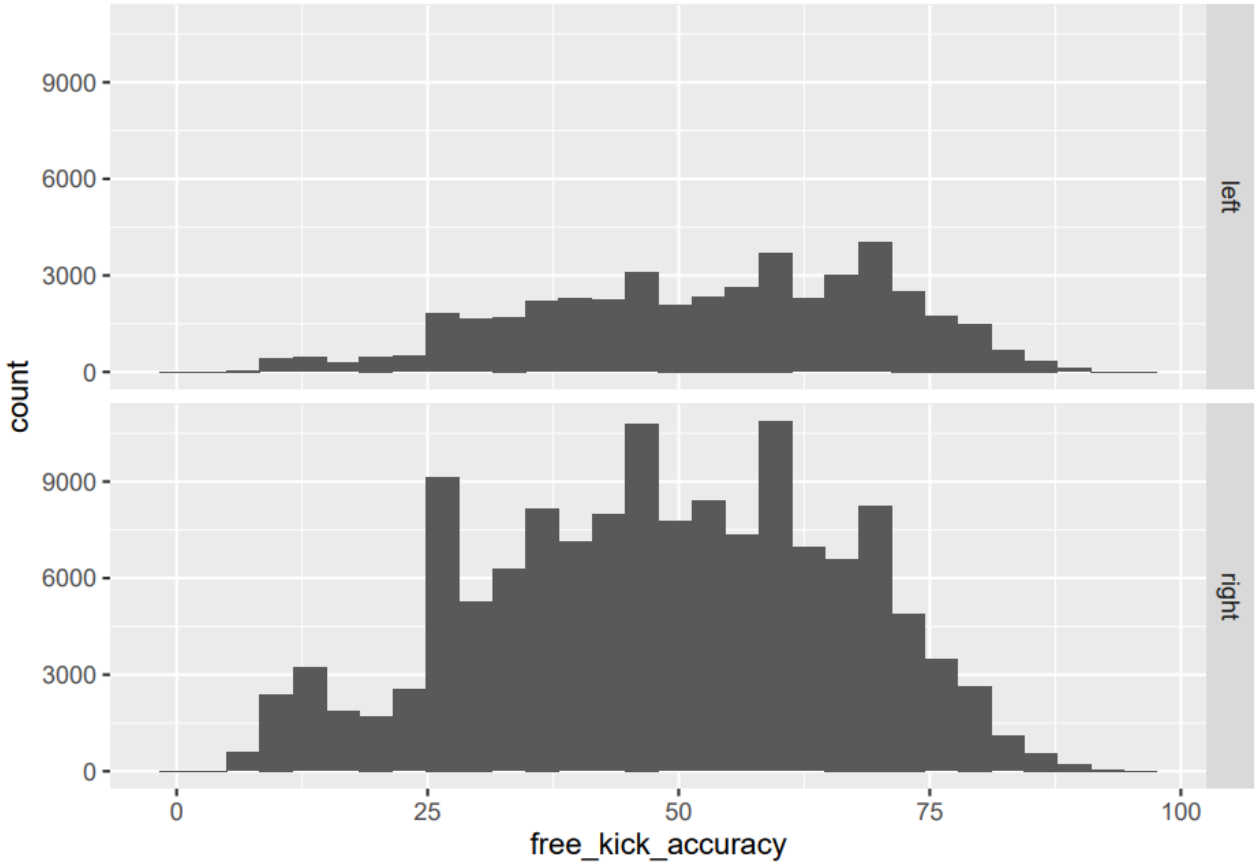
```
#Boxplot to illustrate the five-number summary
select_player_attributes %>%
  ggplot(aes(x = free_kick_accuracy,
             y = preferred_foot)) +
  geom_boxplot()+
  labs(y = "Preferred Foot",
       x = "Free Kick Accuracy",
       title = "Free Kick Accuracy vs Preferred Foot")
```



Step 3: Explore the Data

- Distribution of free kick accuracy of the two populations

```
#Boxplot to illustrate the five-number summary
select_player_attributes %>%
  ggplot(aes(x = free_kick_accuracy,
             y = preferred_foot)) +
  geom_boxplot()+
  labs(y = "Preferred Foot",
       x = "Free Kick Accuracy",
       title = "Free Kick Accuracy vs Preferred Foot")
```



Step 4: Draw Inferences Beyond the Data

- Conduct the two-sample t-test

```
#Look at Data
select_player_attributes %>%
  group_by(preferred_foot) %>%
  summarise(xbar = mean(free_kick_accuracy),
            s = sd(free_kick_accuracy),
            n = n())
```

```
## # A tibble: 2 x 4
##   preferred_foot  xbar     s      n
##   <chr>          <dbl> <dbl>  <int>
## 1 left            53.3  17.3  44107
## 2 right           48.1  17.8 136247
```

```
#Calculate Standardized Statistics
xbar_left = 53.291
xbar_right = 48.131
s_left = 17.325
s_right = 17.796
n_left = 44107
n_right = 136247
sd = sqrt(s_left^2/n_left+s_right^2/n_right)
null = 0
statistic = xbar_left-xbar_right
t = (statistic-null)/sd
```

Step 4: Draw Inferences Beyond the Data

- Confidence interval at 99% confidence

```
n = n_left+n_right
pvalue = 2*pt(t,n-2, lower.tail = FALSE)
pvalue
```

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

```
#calculate Confidence interval at 99% confidence
multiplier = qt(.995,n-2)
se = sd
CI = c(statistic - multiplier*se, statistic + multiplier*se)
CI
```

```
## [1] 4.91388 5.40612
```

Step 5: Formulate Conclusions

- Our P-value is zero, what does that imply? What conclusions can we draw?
  - P-value of zero -> it is extremely unlikely that the observed difference between the two groups was due to chance
  - Reject the null hypothesis

Step 6: Look Back and Ahead

- Identify improvements for the study so the results can be generalized to a larger population
  - Potentially include data from other countries
  - Increase sample size
  - Use data collected more recently
- Discuss applications of t-statistics in other sports
  - Baseball

Conclusions

- Free-kick Accuracy vs. Dominant Foot
- Two-sample T-test
- Other applications of the two-sample t-test

Work Cited

- Simulation-based Statistical Inference. https://www.causeweb.org/sbi/?p=617 - . Accessed May 6, 2024.

- European Soccer Database. Kaggle. https://www.kaggle.com/datasets/hugomathien/soccer/data. Accessed March 3, 2024.

- Tintle, N., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T. & Vander Stoep, J. Introduction to Statistical Investigations, (2nd ed).