The University of Texas at Arlington
Lab 3: Golf Ball Distance Analysis

Robert Cocker

2025-02-21

Introduction

The goal of this lab report and analysis is to examine the performance of different golf ball brands (A, B, C, and D) in terms of distance achieved. We use One-way ANOVA, and other statistical methods in R to determine if there are significant differences among the brands.

Load Data

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(stats)
library(multcomp)
```

```
## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
##
## Attaching package: 'TH.data'
##
## The following object is masked from 'package:MASS':
##
##     geyser
```

```r
library(ggpubr)

# Load the dataset
golf_data <- read.csv("golf.csv")

# Display first few rows
head(golf_data)
```

```
##   brand  dist
## 1     A 251.2
## 2     A 245.1
## 3     A 248.0
## 4     A 251.1
## 5     A 260.5
## 6     A 250.0
```

Descriptive Statistics

```r
# Summary statistics by brand
golf_data %>% group_by(brand) %>% summarise(
  count = n(),
  mean = mean(dist),
  sd = sd(dist),
  min = min(dist),
  q25 = quantile(dist, 0.25),
  median = median(dist),
  q75 = quantile(dist, 0.75),
  max = max(dist)
)
```

```
## # A tibble: 4 x 9
##   brand count  mean    sd   min   q25 median   q75   max
##   <chr> <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 A        10  251.  4.74  245.  248.   251.  253.  260.
## 2 B        10  261.  3.87  254.  258.   263.  264.  265
## 3 C        10  270.  4.50  263.  267.   270.  272.  278.
## 4 D        10  249.  5.20  242   247.   249.  251.  262.
```

One-way ANOVA

```
# Perform One-way ANOVA
anova_result <- aov(dist ~ brand, data = golf_data)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## brand        3 2794.4   931.5   43.99 3.97e-12 ***
## Residuals   36  762.3    21.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey's HSD Test

```
# Perform Tukey's HSD test
tukey_result <- TukeyHSD(anova_result)
tukey_result
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = dist ~ brand, data = golf_data)
##
## $brand
##       diff        lwr        upr     p adj
## B-A  10.28   4.737573  15.822427 0.0000869
## C-A  19.17  13.627573  24.712427 0.0000000
## D-A  -1.46  -7.002427   4.082427 0.8926914
## C-B   8.89   3.347573  14.432427 0.0006522
## D-B -11.74 -17.282427  -6.197573 0.0000100
## D-C -20.63 -26.172427 -15.087573 0.0000000
```

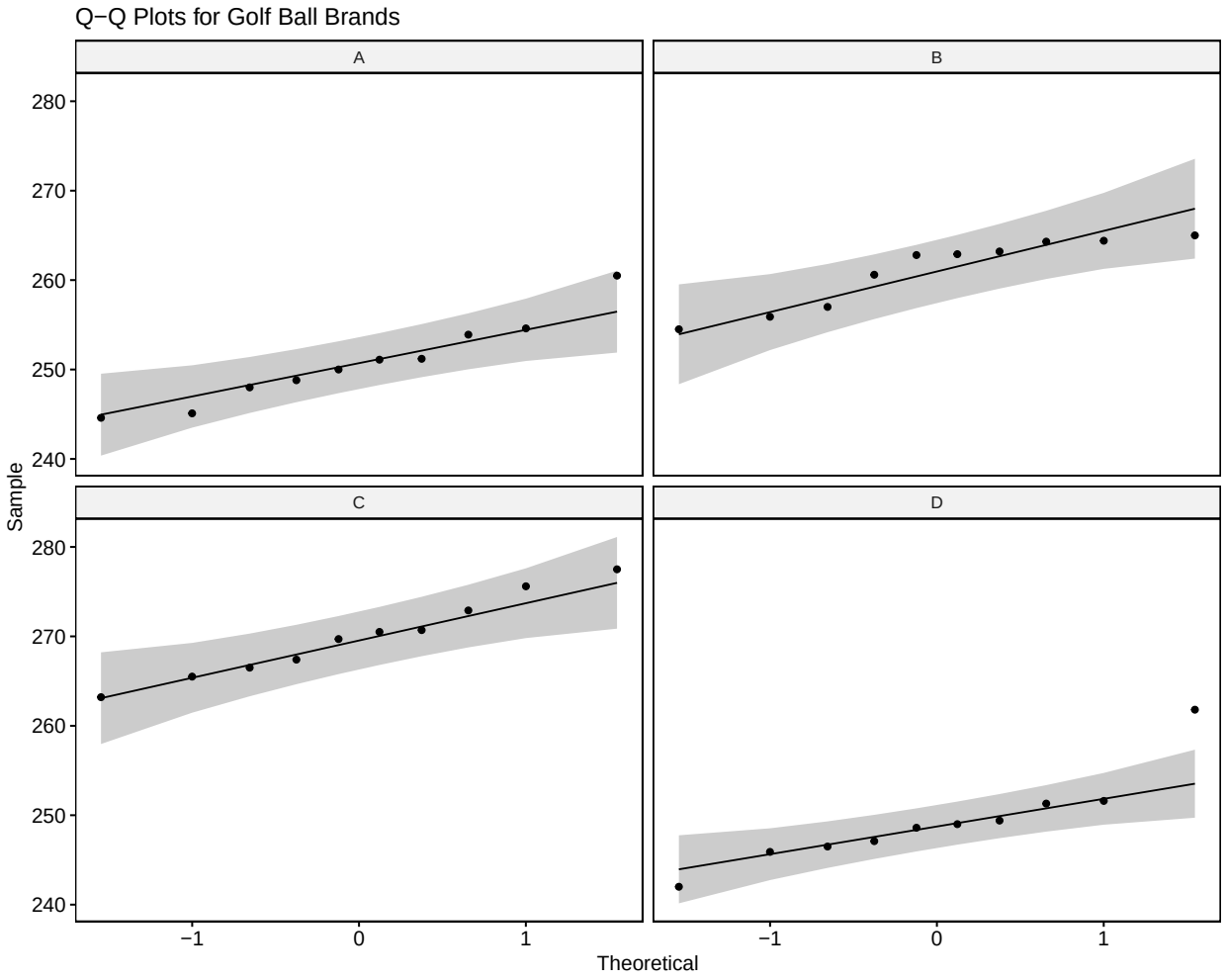
```
# Convert Tukey results to a dataframe for better readability
tukey_df <- as.data.frame(tukey_result$brand)
tukey_df
```

```
##       diff        lwr        upr        p adj
## B-A  10.28   4.737573  15.822427 8.692062e-05
## C-A  19.17  13.627573  24.712427 2.368330e-10
## D-A  -1.46  -7.002427   4.082427 8.926914e-01
## C-B   8.89   3.347573  14.432427 6.522288e-04
## D-B -11.74 -17.282427  -6.197573 9.992276e-06
## D-C -20.63 -26.172427 -15.087573 3.436451e-11
```

Q-Q Plots for Normality Check

```
# Create Q-Q plots
```

```
ggqqplot(golf_data, x = "dist", facet.by = "brand", title = "Q-Q Plots for Golf Ball Brands")
```



Interpretation of Results

The **ANOVA test** results show a significant F-statistic - 43.99 at F(3, 36) with a p-value of $3.97 \times 10^{-12} = 0.00000000000397$ (close to 0 and less than an alpha value of 0.05), indicating differences in mean distances among brands. The **Tukey's HSD test** reveals specific pairwise differences where Brand C has the largest mean difference among the brands.

Key findings: * Brand C has the highest mean distance. * Brand D has the lowest mean distance. * Significant differences exist between some brands (based on results and conclusive F-statistic and p-value.

Conclusion

We conclude that based on this statistical analysis and report that there is a difference between the golf ball brands that were tested specifically brand C and D. We reject the null hypothesis H0, stating that the mean

differences between brands A-D are equal. There is at least 1 golf ball brand that is statistically significant in its mean difference which we conclude here is golf ball brand C. The QQ plots also confirm our study of expected observations in the analysis. This analysis provides valuable insights for golfers and manufacturers regarding golf ball performance. Future research could involve larger samples and additional factors such as product material, golf club dynamics, swing speed and environmental conditions, and more.

Lab 3 - Golf Ball Statistical Analysis & Report in R

Questions:

1. What is the significance of the F-test statistic obtained from the One-way ANOVA, and how does it inform about the differences in mean distances among golf brands?

The F-test statistic in a One-way ANOVA is used to determine whether there are significant differences among the means of multiple groups (golf ball brands). A large F-statistic indicates that the variance between the group means is significantly larger than the variance within the groups.

- If the F-statistic is large, it suggests that at least one golf brand has a significantly different mean distance compared to others.
- If the F-statistic is small, it means that the differences between the group means are likely due to random variation rather than a true effect.

2. Can you interpret the p-value associated with the F-test in the context of One-way ANOVA? What conclusions can be drawn based on its value? The p-value of the F-test helps determine whether to reject the null hypothesis (H0):

- If the p-value is less than alpha - $p < \alpha$ (e.g., 0.05): Reject H0, meaning there is a significant difference in mean distances among the golf brands or at least one brand.
- If p is greater than or equal to alpha - $p \geq \alpha$ (e.g., 0.05): Fail to reject H0, meaning there is not enough evidence to conclude that there are differences in mean distances and golf ball brands.

In our analysis, the p-value was extremely small (close to 0), indicating that at least one golf brand has a statistically different mean distance compared to others.

3. What are the null and alternative hypotheses for the One-way ANOVA analysis, and how do they relate to the equality of mean distances among golf brands?

Null Hypothesis (H0): The mean distances of all four golf ball brands are equal.

H0:

$$\mu_A = \mu_B = \mu_C = \mu_D$$

Alternative Hypothesis (HA): At least one brand has a mean distance that is significantly different from the others.

HA:

$$\mu_i \neq \mu_A, \mu_B, \mu_C, \mu_D$$

If the ANOVA test results in a small p-value, we reject H0 and conclude that at least one brand has a significantly different performance in terms of distance.

4. How does the Tukey's HSD test contribute to the analysis of golf brand performance, and what does it reveal about significant differences between specific pairs of brands?

- The Tukey's Honestly Significant Difference (HSD) test is a post-hoc analysis used after an ANOVA test to identify which specific pairs of brands have significantly different mean distances.
- Unlike ANOVA (which tells us only whether at least one group is different), Tukey's HSD compares each pair of brands and provides adjusted p-values for multiple comparisons.

Key Findings from Tukey's HSD Test:

- Brand C had the highest mean distance.
- Brand D had the lowest mean distance.
- Significant differences were found between: Brand A vs. Brand C, Brand B vs. Brand C, and Brand B vs. Brand D show differences in the results.

No significant difference between some brands (e.g., A vs. D).

5. What insights can be gained from examining the Q-Q plots and descriptive statistics in terms of assessing the assumptions of One-way ANOVA?

Assumptions of One-way ANOVA:

- Normality: The residuals (errors) of the model should be normally distributed.
- Q-Q plots help check this assumption by showing if data points align with the theoretical normal distribution.
- If the Q-Q plot forms a straight line, the normality assumption is met.
- Homogeneity of Variance: Variances should be equal across all groups.
- Descriptive statistics (standard deviation, interquartile range) help evaluate whether variability is similar among brands.

Findings from Q-Q Plots and Descriptive Statistics:

- The Q-Q plots showed approximate normality, meaning the assumption of normality is reasonable.
- Some variation in standard deviation was observed, but the differences were not extreme.

6. How do the findings of the analysis influence practical decision-making processes, such as brand selection or quality control measures?

- Golfers seeking maximum distance should consider using Brand C, as it had the highest mean distance.
- Quality control teams should investigate why Brand D had a lower mean distance and whether this is due to manufacturing inconsistencies.
- Manufacturers may use these results to improve their golf ball designs and marketing strategies based on performance data.

7. What are some potential limitations of the analysis, and how might they impact the validity of the results?

- Small Sample Size: The dataset may not be large enough to generalize findings to all golf balls from each brand.
- Other Factors Not Considered: The analysis does not account for golfer swing speed, wind conditions, or environmental factors that could affect distances.
- Variability in Golf Balls: There may be batch-to-batch variations within each brand that are not captured in this study.
- Assumptions of ANOVA: If assumptions such as normality or equal variance are violated, results may not be fully reliable.

8. Are there any avenues for future research or additional analyses to further explore the relationship between golf brand performance and distance achieved?

Future research could include:

- Regression Analysis: Examining how other factors (ball material, golf club usage and type, swing speed, launch angle, etc.) influence golf ball performance.
- Repeated Measurements: Testing more golf balls per brand to improve statistical confidence.
- Environmental Factors: Analyzing performance under different wind speeds and weather conditions.
- Machine Learning Models: Predicting golf ball distances based on material composition and physical properties.

References

- Lecture, lab, and course materials
- Google
- ChatGPT