Statistics and Data Visualization using R

Medea Skovgaard

The Project

The aim of this project is to demonstrate statistical data analysis and visualization methods using R and RStudio. It can also serve as a tutorial, as all steps are clearly demonstrated and explained in detail.

The Data

For the analyses, the built-in R example dataset chickwts is used. The dataset contains data on chickens fed with six different feed types. In the experiment, the chickens' weights (in grams) were measured six weeks after the start of the feeding period.

The Question

Does the type of feed influence chicken weight?

Exploratory Data Analysis

First, an overview of the data structure is obtained using the str() function to get an initial impression of the data.

```
str(chickwts)
```

```
## 'data.frame':    71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
```

The dataset consists of 71 observations and two variables. One variable records the chickens' weights as numeric values, and the other indicates the type of feed as a categorical factor with six different levels (casein, horsebean, and four other feed types).

Next, summary statistics of the weights are calculated separately for each feeding group to examine whether chicken weight differs between feed types. This step provides an initial overview of differences in central tendency and variability across diets. The aggregate() function computes the minimum, quartiles, median, mean, and maximum separately for each feed type.

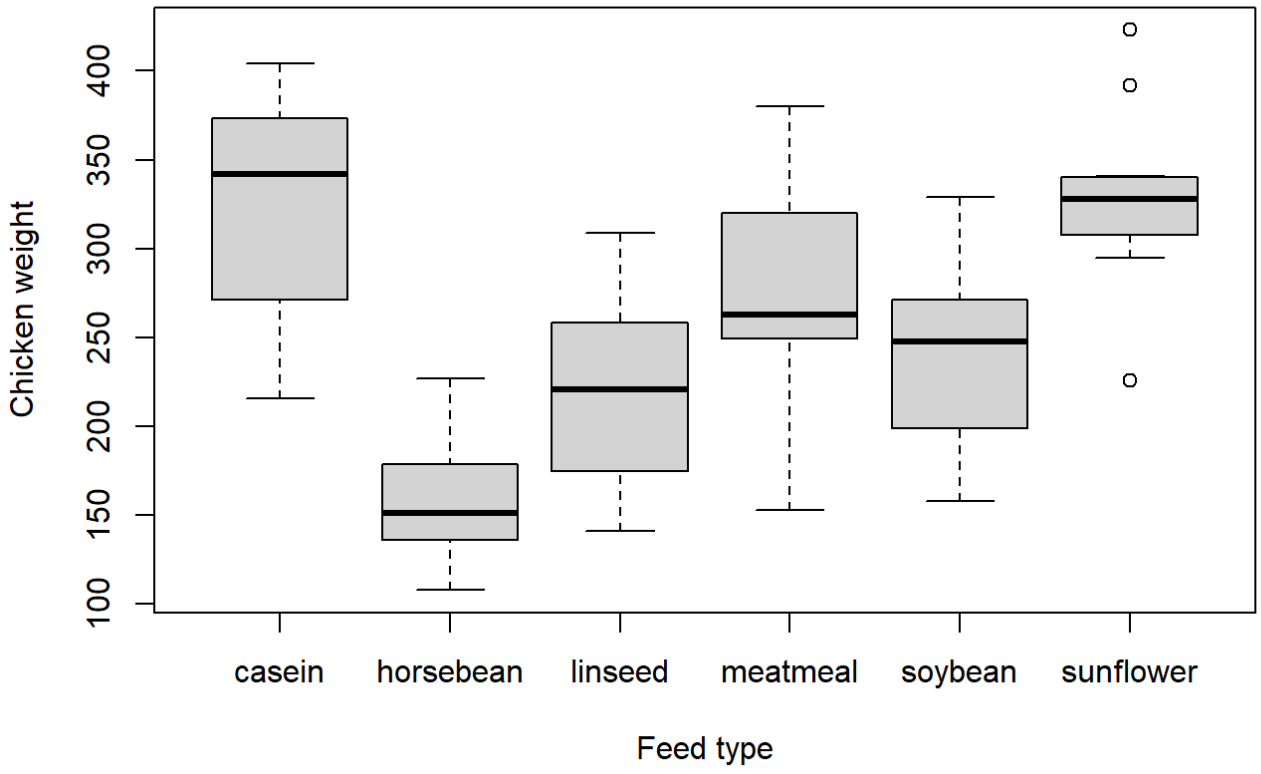
```
aggregate(weight ~ feed, data = chickwts, summary)
```

```
##        feed weight.Min. weight.1st Qu. weight.Median weight.Mean weight.3rd Qu.
## 1    casein    216.0000       277.2500      342.0000    323.5833       370.7500
## 2 horsebean    108.0000       137.0000      151.5000    160.2000       176.2500
## 3   linseed    141.0000       178.0000      221.0000    218.7500       257.7500
## 4  meatmeal    153.0000       249.5000      263.0000    276.9091       320.0000
## 5   soybean    158.0000       206.7500      248.0000    246.4286       270.0000
## 6 sunflower    226.0000       312.7500      328.0000    328.9167       340.2500
##   weight.Max.
## 1    404.0000
## 2    227.0000
## 3    309.0000
## 4    380.0000
## 5    329.0000
## 6    423.0000
```

The summary statistics show clear differences in chicken weights between feed types. Chickens fed with sunflower and casein diets have the highest average and median weights, while those fed with horsebean have the lowest. The meatmeal, soybean, and linseed feeds result in intermediate weight levels, suggesting that feed type is associated with substantial variation in chicken weight.

Finally, a boxplot of the data is created. Visualizing the data is important because it can reveal patterns and potential issues that are often difficult to detect from numerical summaries alone.

```
boxplot(weight ~ feed, data = chickwts,
        ylab = "Chicken weight",
        xlab = "Feed type")
```



The same patterns observed in the summary statistics can also be seen when the data are plotted.

One-way ANOVA

While the summary statistics and the data visualization suggest differences between feed types, an Analysis of Variance (ANOVA) is used to formally test whether these differences in group means are statistically significant or could be explained by random variation.

The hypotheses are as follows: H_0: All feed types have the same mean weight. H_1: At least one feed type has a different mean weight.

```
anova_results <- aov(weight ~ feed, data = chickwts)
summary(anova_results)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows the results of comparing chicken weights across the six feed types.

Df (Degrees of Freedom): The feed factor has 5 degrees of freedom (6 groups minus 1), and the residuals (within-group variation) have 65 degrees of freedom.

Sum Sq (Sum of Squares): 231,129 represents the variation between the feed groups, while 195,556 represents the variation within the groups.

Mean Sq (Mean Square): This is the sum of squares divided by the corresponding degrees of freedom. For feed, it's 46,226; for residuals, it's 3,009.

F value: The F statistic is 15.37, which measures how much the group means differ relative to the variation within groups.

Pr(>F): The p-value is extremely small (5.94×10^{-10}), indicating that the differences in mean weights between feed types are highly significant.

The significance codes indicate the level of statistical significance, with *** showing a highly significant result.

Interpretation: Since the p-value is far below the typical significance level of 0.05, we reject the null hypothesis. This means that at least one feed type leads to a significantly different mean chicken weight compared to the others.

TukeyHSD test

After an ANOVA indicates that there are significant differences between group means, the TukeyHSD test helps identify exactly which groups differ from each other. It performs all pairwise comparisons between the group means.

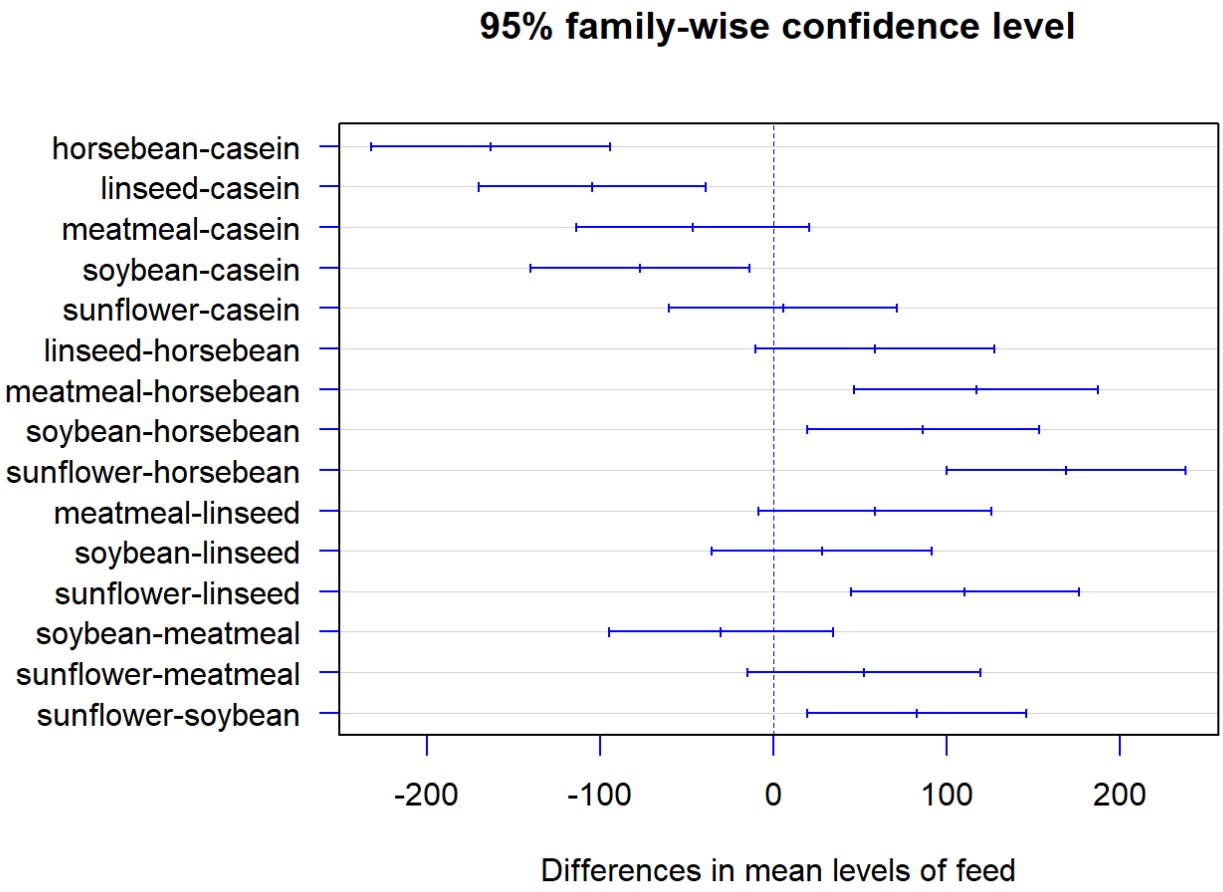
```
tukey_results <- TukeyHSD(anova_results)
```

The TukeyHSD test compares all pairs of feed types to identify which differences in mean chicken weight are statistically significant. The results show the estimated difference in means (diff), the 95% confidence interval (lwr and upr), and the adjusted p-value (p adj) for each pairwise comparison. Chickens fed horsebean, linseed, and soybean are generally lighter than those fed casein, with horsebean showing the largest decrease. Sunflower consistently produces heavier chickens compared to horsebean, linseed, and

soybean. Meatmeal is similar to casein and not significantly different from most other feeds.

Below, we visualize the TukeyHSD test results as a plot of confidence intervals for the differences between group means.

```
par(mar = c(5, 10, 4, 2))
plot(tukey_results, las = 1, col = "blue")
```



Interpretation

The type of feed influences chicken weight. The ANOVA results show a statistically significant difference in mean weight between feed types, indicating that not all feeds lead to the same growth outcomes. The subsequent TukeyHSD test further reveals that some feed types result in significantly higher or lower chicken weights than others, confirming that feed composition has a meaningful effect on chicken growth.