One-way Analysis of Variance in R

James Quinlan

05/01/2023

In this article we will examine the affect of diet on weight loss (case study) using analysis of variance. First, we cover some background information including the model assumptions, and Fisher's Test statistic.

Background

ANalysis Of Variance (ANOVA or AOV) generalizes the t-test to 3 or more groups. That is, ANOVA is used to determine if the means of two or more groups are the same. Technically, ANOVA can be used for 2 groups, but in practice, if two groups, use Student t-test.

This article specifically addresses one-way ANOVA. There many 'ANOVAs' (e.g., two-way ANOVA, Welch ANOVA, repeated sample ANOVA, etc.). *One-way layout* is an experimental design in which independent measurements are made under each of the several treatments. Let $y_{ij} =$ the jth observation of the ith treatment. These observations are corrupted by random errors. The statistical model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where μ is the overall mean, $\tau_i = \mu_i - \mu$ is the differential effect of the ith treatment, and $\epsilon_{ij} = y_{ij} - \mu_i$ is the random error of the jth observation in the ith treatment. These errors are assumed to be independent, normally distributed with mean 0 and variance σ^2 (see p. 412 in [3]). As always, population means (and variances) are unknown and the sample statistics are used as estimates.

Assumptions of ANOVA

As usual with statistical tests, several assumptions must be satisfied before the findings can be interpreted. Although it is theoretically feasible to execute these tests when one or more assumptions are not satisfied, be careful interpreting the findings and stating conclusions.

The following are the ANOVA assumptions, how to test them, and what alternate tests are available if an assumption is not met:

- **Variable Types**: One-way ANOVA involves a continuous quantitative dependent variable and a qualitative independent variable (with at least 2 levels or treatments that will serve as the comparison groups).
- **Independence**: The data should be independent across groups and within each group. Observations should not influence each other (e.g., a husband and wife probably will affect the outcome of the other). Another example of dependence is when multiple observations are of the

same person, as with medical studies, in which repeated measures ANOVA should be conducted instead.

- **Normality**: Sample data should be normal distributed. *Visually* check the normality assumption with a histogram or a QQ-plot. A normality test such as the *Shapiro-Wilk* or *Kolmogorov-Smirnov* can be used as an analytic test. If the residuals still do not follow a normal distribution even after a transformation such as a logarithmic transformation, use a non-parametric test resistant to non-normal distributions. In particular, the Kruskal-Wallis test: `Kruskal.test(group, data = Data)` has the same purpose as the ANOVA: to compare three or more groups. However, it compares groups using sample medians rather than sample means.
- **Homoscedasticity** Homogeneity in populations, the variances of distinct groups should be equal. Homogeneity may be verified graphically using a boxplot or analytically with a statistical test such as Levene's or Bartlett's. If the null hypothesis of these tests is rejected, then Welch's ANOVA is an alternative to the when the assumption of equal variances is violated. In R, `oneway.test(variable ~ group, var.equal = FALSE)`. Note: Welch ANOVA does not need homogeneity among the variances, but normality is still assumed. While, Kruskal-Wallis test requires neither normalcy nor homoscedasticity of variances.
- **No Outliers**: An outlier is an observation that differs significantly from the others. If there are any significant outliers in the different groups, your ANOVA results may be skewed. There are various approaches for detecting outliers in your data, but you must choose one of the following to deal with them:
    - use a non-parametric variant (i.e., the Kruskal-Wallis test, which used median instead of mean. Why?)
    - transform your data (e.g., logarithmic or Box-Cox transformations)
    - delete them

Fisher's F Test Statistic

The hypotheses to test are:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{(the means are equal for all } k \text{ groups)}$$

The alternative hypothesis states that at least one is different.

$$H_1 : \mu_s \neq \mu_t, \quad \exists s \neq t.$$

Notation

$y_{ij} =$ the jth observation in the ith group.

$n_i =$ the number of sample observations in group i.

$n = n_1 + n_2 + \cdots + n_k$, the total sample size

$\bar{y}_i$ = the sample mean of the ith group.

$\bar{y}$ = the average of all observations.

Using this notation, we identity three sums of squares.

1. The **total sum of squares** of the difference between the measurements and overall mean is:

$$\mathrm{SS}_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}).$$

2. The **within-sample sum of squares** is computed by:

$$\mathrm{SS}_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2.$$

   where s_i^2 is the sample variance of the ith group.

3. The **between-sample sum of squares** is evaluated with:

$$\mathrm{SS}_B = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y}).$$

We have, $\mathrm{SS}_T = \mathrm{SS}_B + \mathrm{SS}_W$. See [4] for proof. Note: several sources express this formula with slightly different notation, specifically, $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$.

Fisher's F test statistic is,

$$F = \frac{\mathrm{SS}_B/(k-1)}{\mathrm{SS}_W/(n-k)} = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

where the numerator is the mean square error in the regression (MSR) and the denominator is the mean square error in the residuals (MSE).

Case Study

Compare three diets to determine if there is a difference in weight loss.

Load Libraries

Load the `tidyverse` library, although the two main libraries we use from `tidyvere` are `dpylr` and `ggplot2` . For more information about tidyverse see: https://dplyr.tidyverse.org/ (https://dplyr.tidyverse.org/) . The `outliers` package will be used to test for outliers in the data which violate model assumptions.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

The `outliers` package will be used to test for outliers in the data which violate model assumptions.

```
library(outliers)
```

Data: Read, Wrangle, Slice

The data set contains information on 78 people using one of three diets to determine which diet is best for losing weight. We test to see if there is a difference in average weight loss between the three different diets. This data *was* obtained from the Mathematics and Statistics Help (MASH) site (https://www.sheffield.ac.uk/mash/statistics/datasets) at the University of Sheffield. However, recently the site has changed and data does not seem to be currently available. Download from this repository (https://raw.githubusercontent.com/jamesquinlan/Intro-Stats-MAT150/main/data/diet.csv) and save to the current working directory. Obtain the path to the working directory using the command `getwd()`. Set the working directory with the function `setwd()`. Note: your path will be different.

```
# Load the data
# raw = read_csv(paste(getwd(),folder,'diet.csv',sep=''))
raw = read_csv('https://raw.githubusercontent.com/jamesquinlan/Intro-Stats-MAT150/mai
n/data/diet.csv')
```

```
## Rows: 78 Columns: 7
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (7): Person, gender, Age, Height, pre.weight, Diet, weight6weeks
##
## 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.
```

Extract the data of interest. In this case, we will need to create a new variable called `Loss` which is the difference between the subjects weight before beginning the diet and then six weeks after.

```
Data = raw %>%
  mutate(Loss = pre.weight - weight6weeks) %>%
  select(Diet,Loss)
```

Convert the `Diet` variable to a factor.

```
Data$Diet = as.factor(Data$Diet)
```

Split weigth loss by group for easier use.

```
Group1 = Data %>% select(Diet,Loss) %>% filter(Diet==1)
Group2 = Data %>% select(Diet,Loss) %>% filter(Diet==2)
Group3 = Data %>% select(Diet,Loss) %>% filter(Diet==3)
```

Attach the data to work with non fully qualified names, e.g., `Loss` vs. `Data$Loss` .

```
attach(Data)    # Don't forget to detach("Data") to avoid conflicts
```

Exploratory Data Analysis (EDA)

Developed by John Tukey in the 1970s, EDA is used to investigate data and summarize their main characteristics, often employing data visualization methods [5]. Before proceeding with statistical tests and analysis, let's look at summary statistics and visualizations.

```
# Summary of Data by variable (Diet and Loss)
summary(Data)
```

```
##  Diet        Loss
##  1:24   Min.   :-2.100
##  2:27   1st Qu.: 2.000
##  3:27   Median : 3.600
##         Mean   : 3.845
##         3rd Qu.: 5.550
##         Max.   : 9.200
```

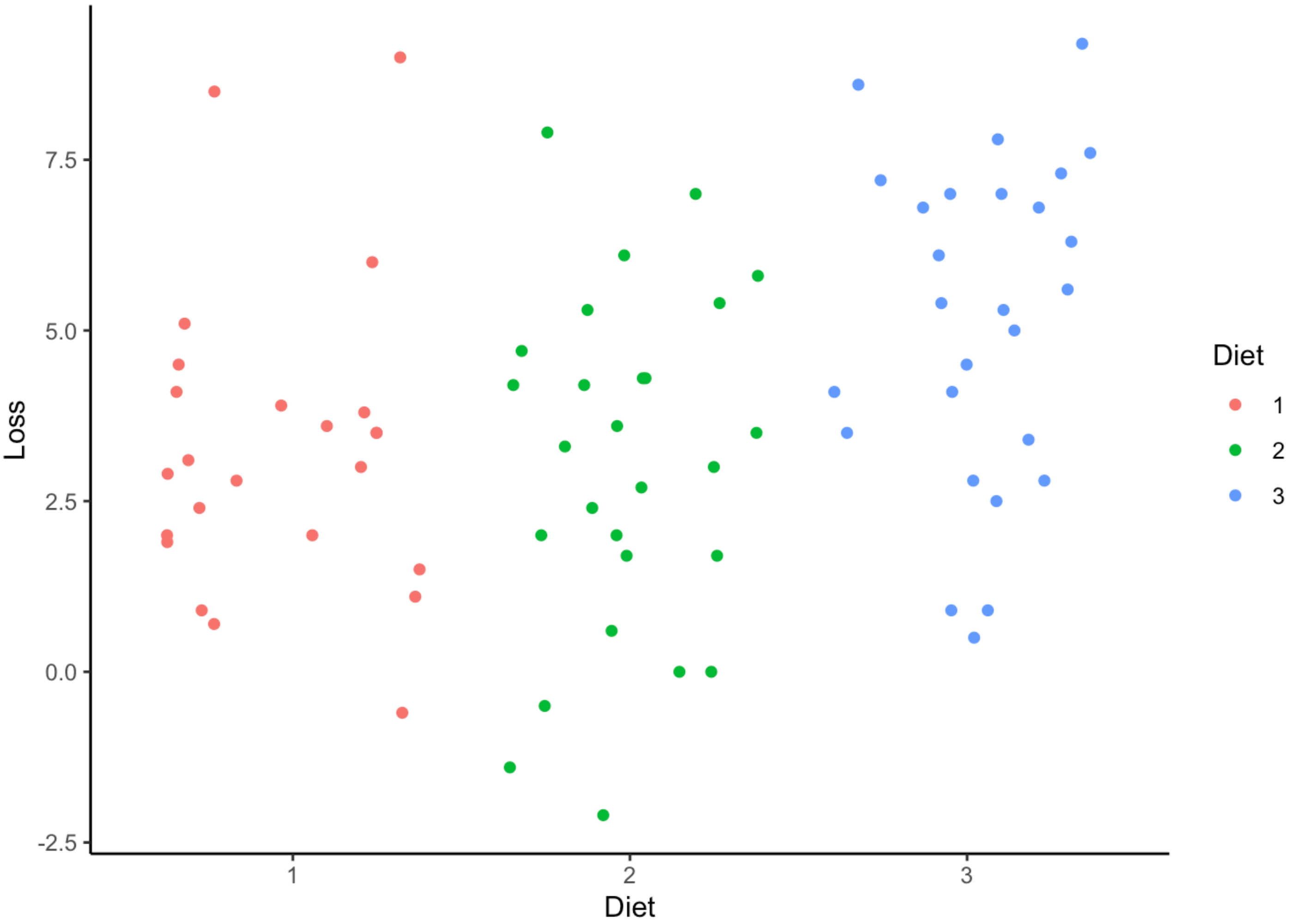
Summarize by diet,

```
Data %>%
  group_by(Diet) %>%
  summarize(
    min = min(Loss),
    q1 = quantile(Loss, 0.25),
    median = median(Loss),
    mean = mean(Loss),
    q3 = quantile(Loss, 0.75),
    max = max(Loss)
  )
```

```
## # A tibble: 3 × 7
##   Diet     min    q1 median  mean    q3   max
##   <fct>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 1     -0.600  1.98   3.05  3.3   3.95  9
## 2 2     -2.10   1.70   3.30  3.03  4.5   7.90
## 3 3      0.5    3.45   5.4   5.15  7     9.2
```

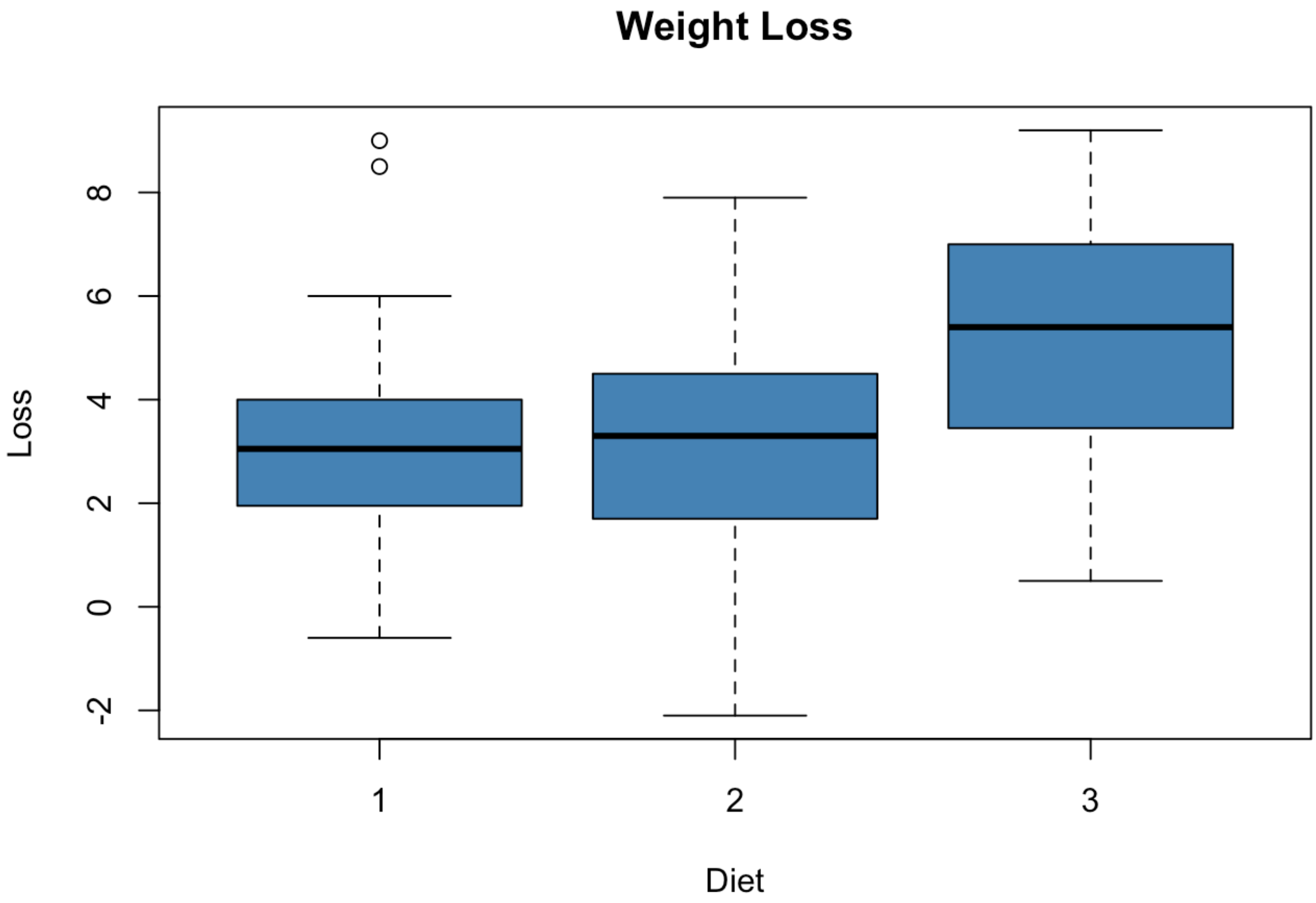
Generate a scatter plot using three colors to identify the diet with `ggplot`.

```
ggplot(Data) +
  aes(x = Diet, y = Loss, color = Diet) +
  geom_jitter() +
  theme_classic()
```



Here is a **boxplot** for each of the three groups.

```
# Boxplot
boxplot(Loss ~ Diet,
  data = Data,
  main = "Weight Loss",
  xlab = "Diet",
  ylab = "Loss",
  col = "steelblue",
  border = "black"
)
```



Results of EDA

We see Diet 3 seems to be the “best” with both the largest average weight lost and maximum weight lost. In addition, Diet 1 and 2 saw some weight gain. From the boxplot, there appear to be a few outliers and Diet 3 also shows higher mean. Variance in the data seems to be mostly equal.

Assumption Verification

Variable Types

Weight loss is a continuous quantitative dependent variable and diet is the independent categorical variable with at least three levels.

Independent Observations

We will assume, without verification, that the design of experiment was such that participants were random selection and randomly assignment to each diet.

Outliers

Let's check for outliers next since if present, we can remove them before checking the other assumptions. The boxplot in EDA suggests Diet 1 has two outliers. Statistical tests for outliers include: Grubb's test, Dixon's test (small samples $n \leq 25$), and Rosner's test (detect multiple outliers).

Grubb's Test

We can use `grubbs.test()` function from the `outliers` package to test if the single highest value is an outlier [1]. The function takes several parameters (see `?grubbs.test`):

```
grubbs.test(x, type = 10, opposite = FALSE, two.sided = FALSE)
```

where:

- `x` : a numeric vector of data values
- `type` : 10 is a test for one outlier (side is detected automatically and can be reversed by opposite parameter). 11 is a test for two outliers on opposite tails, 20 is test for two outliers in one tail.
- `opposite` : a logical indicating whether you want to check not the value with largest difference from the mean, but opposite (lowest, if most suspicious is highest etc.)
- `two-sided` : Logical value indicating if there is a need to treat this test as two-sided

The hypotheses for Grubb's Test are:

- H_0 : There is NO outlier in the data.
- H_1 : There IS an outlier in the data.

The boxplots in our EDA indicate only Diet 1 has potential outliers. We will perform Grubb's Test on Diet 1 data. The p-values are calculated using `qgrubbs` function.

```
grubbs.test(Group1$Loss)
```

```
## 
##  Grubbs test for one outlier
## 
## data:  Group1$Loss
## G = 2.54448, U = 0.70627, p-value = 0.0747
## alternative hypothesis: highest value 9 is an outlier
```

The p-value is slightly greater than 0.05 so we will not reject the null hypothesis. If Grubb’s Test suggested there were outliers present, then we could remove them using the commands below and continue our analysis and tests on the modified data `Data2` .

```
Outliers = boxplot.stats(Group1$Loss)$out
todrop = which(Data$Loss >= Outliers[1] & Data$Diet == 1)
Data2 <- Data[-todrop,]
```

Normality

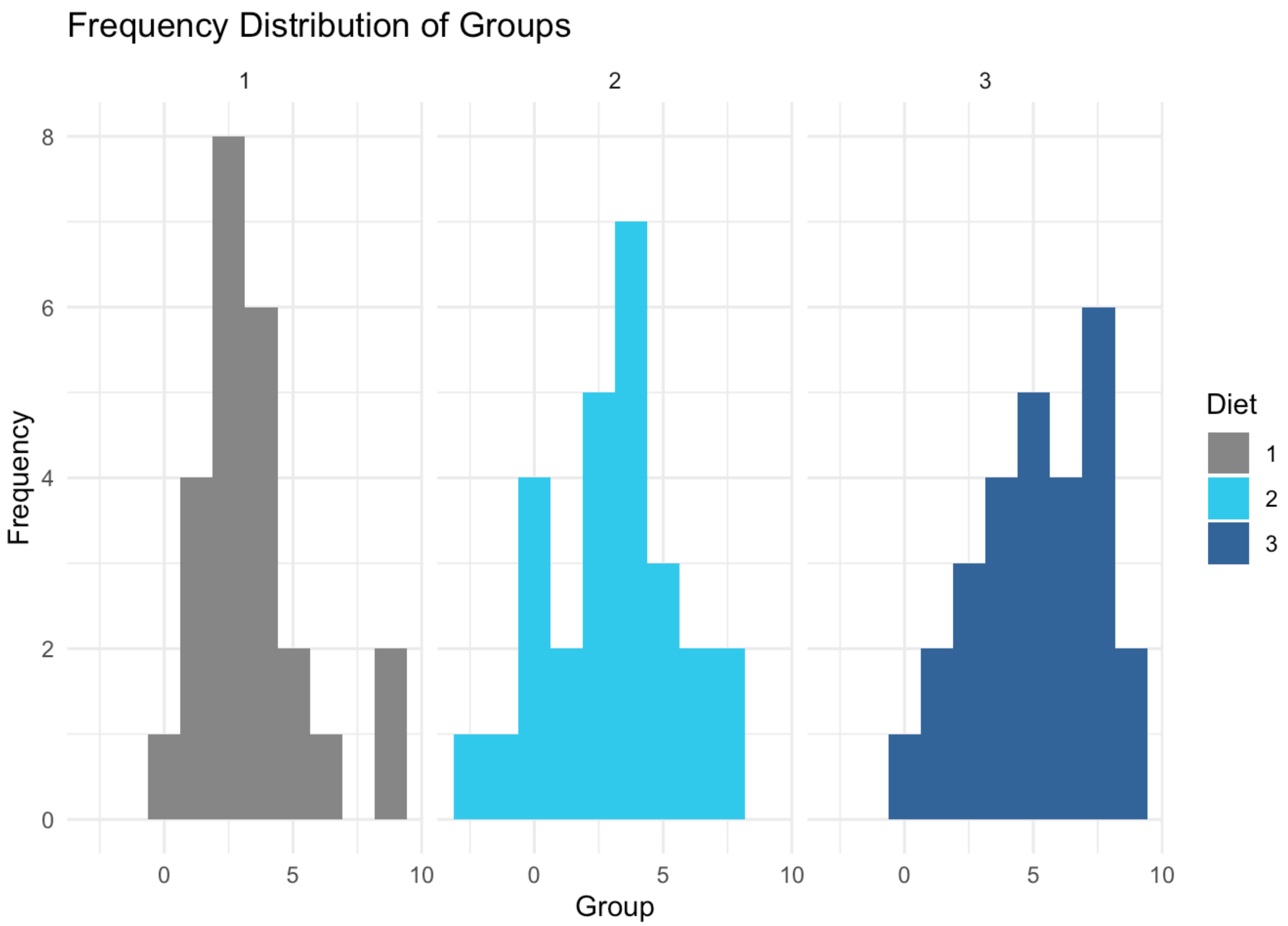
The dependent variable should be approximately normally distributed for each level of the predictor variable.

Visually we can check normality using:

- Histograms
- Density Plots
- Q-Q Plots

First, generate a histogram for each group.

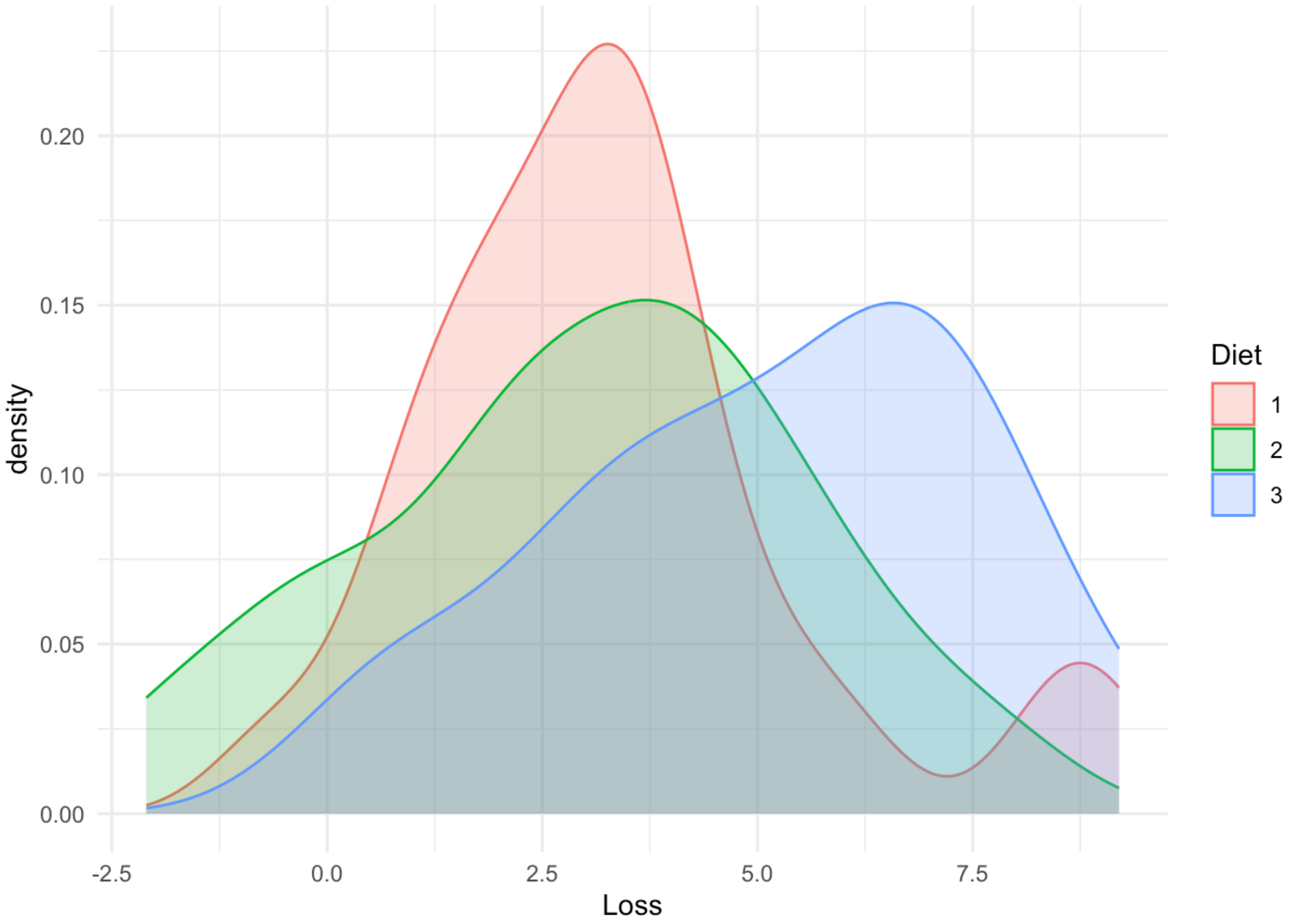
```
ggplot(Data) +
  aes(x = Loss ,fill = Diet) +
  geom_histogram(bins = 10) +
  theme_minimal() +
  facet_wrap(~Diet) +
  labs(x = "Group", y = "Frequency", title = "Frequency Distribution of Groups" ) +
  scale_fill_manual(values=c("#888888",
                             "#35CCEC",
                             "#336699"))
```



Maybe they are normal.

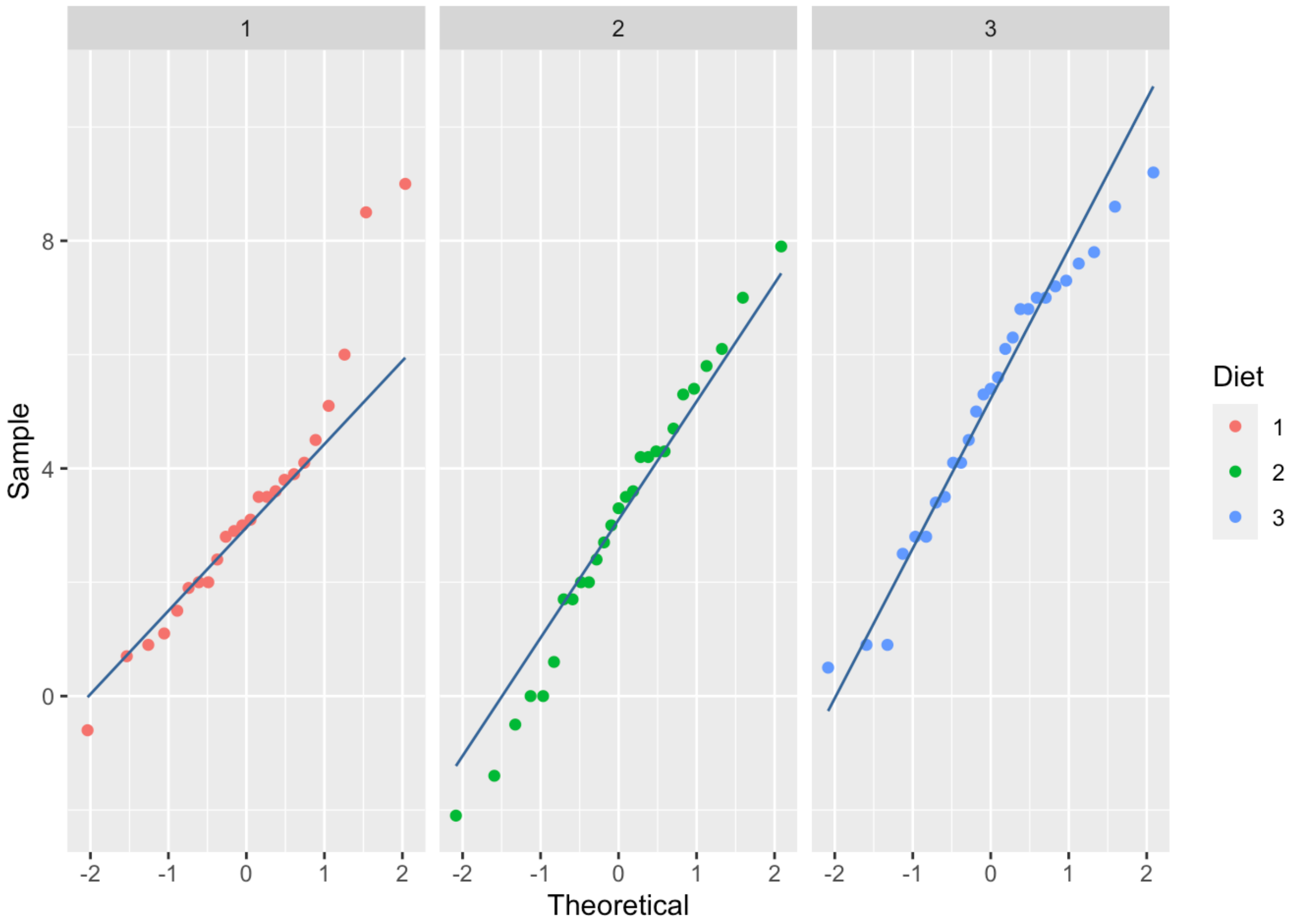
Each distribution is displayed using a density plots colored by diet.

```
ggplot(Data) +
  aes(x = Loss, color = Diet, fill = Diet) +
  geom_density(alpha = 0.25)+
  theme_minimal()
```



Last, the Q-Q Plots are given for each diet.

```
ggplot(data = Data, aes(sample = Loss, col = Diet)) +
  geom_qq() +
  geom_qq_line(color="#336699") +
  xlab("Theoretical") +
  ylab("Sample") +
  facet_wrap(~Diet)
```



By inspection of these plots, it appears that the Diet 2 distribution is approximately normal. Diet 1 has a pronounced hump on the right and Diet 3 is skewed to the left. In the case of Diet 1 in particular, we can see the effect of the two outliers reflected in all three plots. So we may have reason to question whether Diet 1 and Diet 3 are normally distributed. On the other hand, the apparent deviation from normality could have happened by chance - the result of sampling variation. This suggests a hypothesis test, which can tell us that likelihood.

Shapiro-Wilk’s Test for Normality

Shapiro-Wilk’s and Kolmogorov-Smirnov are two statistical tests of normality. The null hypothesis is,

$$H_0 \ : \text{the data is normally distributed}$$

Of course, the alterative hypothesis is, H_1 : the data is **not** normally distributed. Here we employ the Shapiro-Wilk test.

```
shapiro.test(Group1$Loss)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Group1$Loss
## W = 0.92553, p-value = 0.07749
```

```
shapiro.test(Group2$Loss)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Group2$Loss
## W = 0.98559, p-value = 0.9612
```

```
shapiro.test(Group3$Loss)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Group3$Loss
## W = 0.96013, p-value = 0.372
```

Since p-values are all greater than 0.05, we fail to reject the null hypothesis.

Homoscedasticity

ANOVA assumes homogeneity in the variances between groups, that is, variances are equal between groups. We can check this assumption in R using:

Variance per Diet

```
cbind(var(Group1$Loss), var(Group2$Loss), var(Group3$Loss))
```

```
##          [,1]     [,2]     [,3]
## [1,] 5.018261 6.367379 5.738746
```

Bartlett’s Test for Homogeneity of Variances

Bartlett’s Test uses the following null and alternative hypotheses:

- H_0 : The variance among each group is equal.
- H_1 : At least one group has a variance that is not equal to the rest.

In R, Bartlett’s test for equal variance is,

```
bartlett.test(Loss ~ Diet, data=Data)
```

```
## 
##  Bartlett test of homogeneity of variances
## 
## data:  Loss by Diet
## Bartlett's K-squared = 0.33745, df = 2, p-value = 0.8447
```

We fail to reject the null hypothesis (equal variance).

One-way ANOVA Test

The quickest route to testing the hypothesis is to use `oneway.test` function in R. If it was determined that the variance were unequal, simply change the parameter, `var.equal = FALSE` , in the function.

```
oneway.test(Loss ~ Diet, data = Data, var.equal = TRUE)
```

```
## 
##  One-way analysis of means
## 
## data:  Loss and Diet
## F = 6.1974, num df = 2, denom df = 75, p-value = 0.003229
```

ANOVA Model in R

To get the fitted model, use `aov` function.

```
# Fit the one-way ANOVA model
model = aov(Loss ~ Diet, data = Data)
```

The `summary` function provides degrees of freedom, sum of squares, mean square error, the value of the F test statistic, and the p-value.

```
# View the model output
summary(model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## Diet         2   71.1   35.55   6.197 0.00323 **
## Residuals   75  430.2    5.74                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **sources of variation** are listed in the left column of the table. These are referred to in some textbooks (i.e., [3]) as **between sample variation** (or variability in the model) and **within sample variation**.

Conclusions

The critical value of F_α at the $\alpha = 0.05$ level is is 3.119 with $\mathrm{df}_w = 2$ and $\mathrm{df}_b = 75$ (in R `qf(0.95,2,75)`). Because the computed value $F = 6.197$ is greater than the critical value, we reject the null hypothesis of equality of the mean scores for the three diets. The p-value is computed using `1 - pf(6,197,2,75)` and is given in the last columns table under `Pr(>F)` as 0.00323. Thus, there is an extremely low probability that the means differ only by chance. From the three sample means, we observed that the mean weight loss of Diet 3 was larger than both the mean weight loss of Diet 1 and Diet 2.

Summary

- One-way ANOVA is used to test the hypothesis that the means of 3 or more groups are equal.
- Dependent variable is quantitative and the independent variable is categorical with at least 3 levels (if only two levels, use Student's t-test).
- Fisher's F Statistic is used the criteria to reject the null hypothesis given several assumptions are satisfied.
    - The F test statistic formula is, F = MSR/MSE.
- The assumptions for the statistical model state that the observations are:
    - Independent
    - Normally distributed
    - Equal variance among groups
    - No outliers
- To test these assumptions:
    1. Verify the observations were randomly selected.
    2. Run a Shapiro-Wilk's Test to test normality
    3. Run Barrlett's test for equality of variance
    4. Look at box and whisker's plot and run Grubb's test for outliers. Remove them if present.
- Reject the null hypothesis if F test statistic is greater than the critical value at the given significance level.

References

[1]. Grubbs, F.E. (1950). *Sample Criteria for testing outlying observations*. Ann. Math. Stat. 21, 1, 27-58.

[2]. Keppel, G., & Wickens, T. D. (1973). *Design and analysis: a researcher's handbook* Prentice-Hall. Englewood Cliffs, NJ.

[3]. Ott, R. L., & Longnecker, M. T. (2015). *An introduction to statistical methods and data analysis*. Cengage Learning.

[4]. Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

[5]. Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160). Reading, Mass.

[6]. Peck, R., & Devore, J. L. (2011). *Statistics: The Exploration & Analysis of Data*. Cengage Learning.