

# BB503/BB602 - R Training - Week X

Ege Ulgen

## Analysis of Variance (ANOVA)

We'll work on the diet data from last week. Remember there are 3 different diets:

```
diet_df <- read.csv("../data/Diet_R.csv")
```

```
head(diet_df)
```

```
##   Person gender Age Height pre.weight Diet weight6weeks
## 1     25    NA  41   171        60     2         60.0
## 2     26    NA  32   174       103     2        103.0
## 3      1     0  22   159        58     1         54.2
## 4      2     0  46   192        60     1         54.0
## 5      3     0  55   170        64     1         63.3
## 6      4     0  33   171        64     1         61.1
```

```
# turn categorical variables into factor
```

```
diet_df$Diet <- as.factor(diet_df$Diet)
```

```
diet_df$gender <- as.factor(diet_df$gender)
```

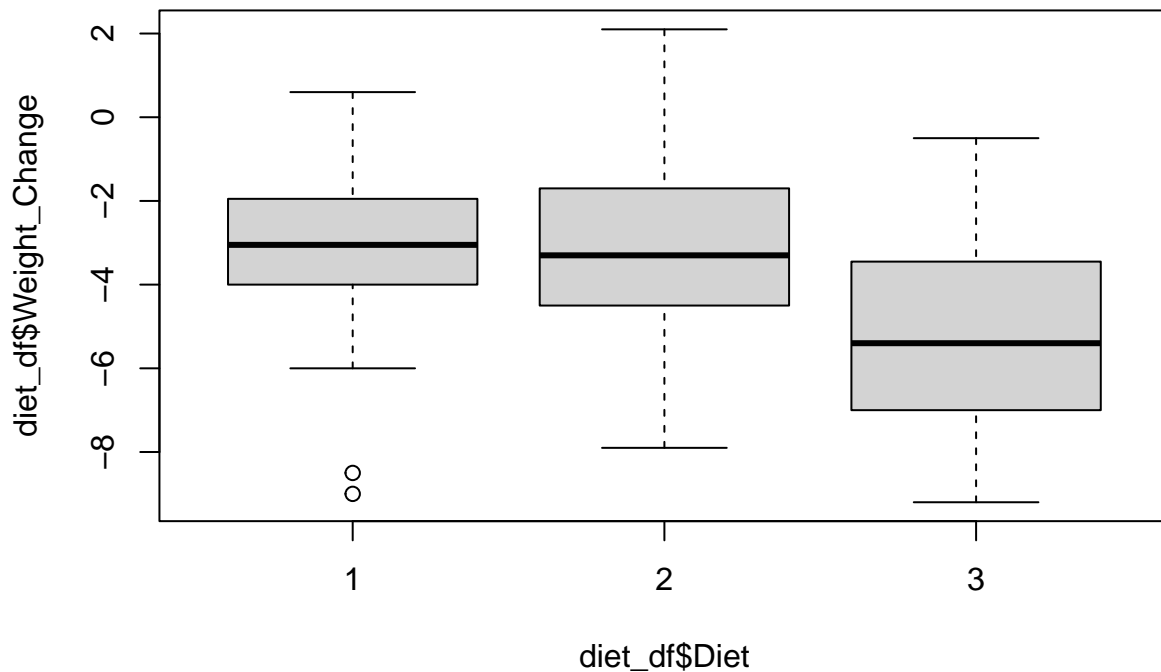
```
# create new variable
```

```
diet_df$Weight_Change <- diet_df$weight6weeks - diet_df$pre.weight
```

```
summary(diet_df)
```

```
##      Person      gender      Age      Height      pre.weight      Diet
## Min.   : 1.0    0   :43  Min.   :16.0  Min.   :141  Min.   : 58.0  1:24
## 1st Qu.:20.2    1   :33  1st Qu.:32.2  1st Qu.:164  1st Qu.: 66.0  2:27
## Median :39.5   NA's: 2  Median :39.0  Median :170  Median : 72.0  3:27
## Mean   :39.5                Mean   :39.2  Mean   :171  Mean   : 72.5
## 3rd Qu.:58.8                3rd Qu.:46.8  3rd Qu.:175  3rd Qu.: 78.0
## Max.   :78.0                Max.   :60.0  Max.   :201  Max.   :103.0
## weight6weeks  Weight_Change
## Min.   : 53.0  Min.   : -9.20
## 1st Qu.: 61.9  1st Qu.: -5.55
## Median : 69.0  Median : -3.60
## Mean   : 68.7  Mean   : -3.84
## 3rd Qu.: 73.8  3rd Qu.: -2.00
## Max.   :103.0  Max.   :  2.10
```

```
boxplot(diet_df$Weight_Change~diet_df$Diet)
```



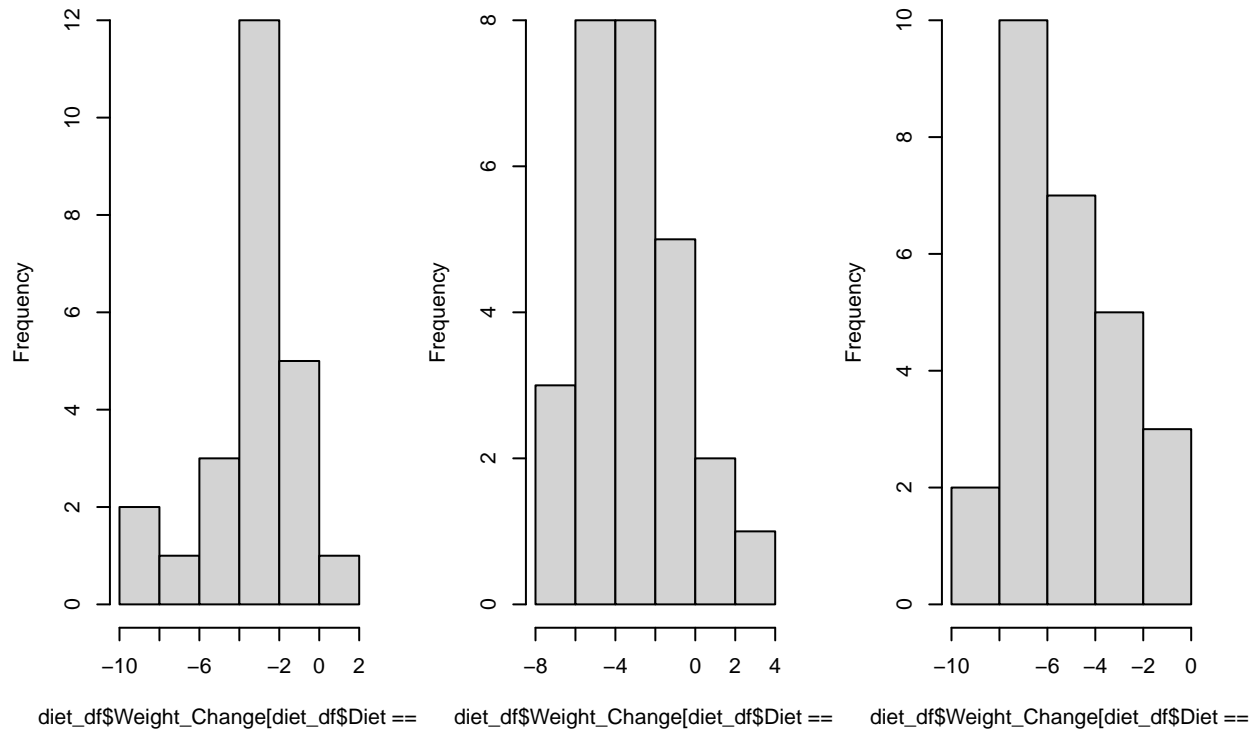
Let's test the hypothesis that at least one of the mean weight changes is different from the others between the three diets.

### 1. Check assumptions, determine $H_0$ and $H_a$ , choose $\alpha$

Inspecting the histograms of weight change by diet, we conclude that they are normally-distributed:

```
par(mfrow = c(1, 3))  
hist(diet_df$Weight_Change[diet_df$Diet == 1])  
hist(diet_df$Weight_Change[diet_df$Diet == 2])  
hist(diet_df$Weight_Change[diet_df$Diet == 3])
```

n of diet\_df\$Weight\_Change[diet\_n of diet\_df\$Weight\_Change[diet\_n of diet\_df\$Weight\_Change[diet



```
par(mfrow = c(1, 1))
```

$H_0 : \mu_1 = \mu_2 = \mu_3$  and  $H_a$  : at least one mean is different

Let's choose  $\alpha = 0.05$

## 2. Calculate the appropriate test statistic

### 2.1. Calculate the grand mean and means per group

```
grand_mean <- mean(diet_df$Weight_Change)
mean1 <- mean(diet_df$Weight_Change[diet_df$Diet == 1])
mean2 <- mean(diet_df$Weight_Change[diet_df$Diet == 2])
mean3 <- mean(diet_df$Weight_Change[diet_df$Diet == 3])
```

### 2.2. Calculate the (total, between group and within group) sum of squared error

Between group sum of squared error =  $\sum n_i(\bar{X}_i - \bar{X})^2$

```
SS_bw <- sum(diet_df$Diet == 1) * (mean1 - grand_mean)^2 +
  sum(diet_df$Diet == 2) * (mean2 - grand_mean)^2 +
  sum(diet_df$Diet == 3) * (mean3 - grand_mean)^2

SS_tot <- sum((diet_df$Weight_Change - grand_mean)^2)

SS_wi <- SS_tot - SS_bw
```

### 2.3. Calculate degrees of freedom

```
df_tot <- nrow(diet_df) - 1
df_bw <- 3 - 1
df_wi <- df_tot - df_bw
```

### 2.4. Calculate mean squared errors

```
MSE_bw <- SS_bw / df_bw
MSE_wi <- SS_wi / df_wi
```

### 2.5. Calculate F statistic

```
F_stat <- MSE_bw / MSE_wi
F_stat
```

```
## [1] 6.1974
```

```
anova_table <- data.frame(Df = c(df_bw, df_wi),
                          Sum_Sq = c(SS_bw, SS_wi),
                          Mean_Sq = c(MSE_bw, MSE_wi),
                          F_stat = c(F_stat, NA))

anova_table
```

```
##   Df Sum_Sq Mean_Sq F_stat
## 1  2  71.094  35.5468  6.1974
## 2 75 430.179   5.7357    NA
```

## 3. Calculate critical values/p value

### Critical values

```
F_crit <- qf(1 - 0.05, df1 = df_bw, df2 = df_wi)

F_stat > F_crit
```

```
## [1] TRUE
```

### p value

```
1 - pf(F_stat, df1 = df_bw, df2 = df_wi)

## [1] 0.003229
```

## 4. Decide whether to reject/fail to reject $H_0$

- The calculated test statistic falls within the rejection region
- p value  $< \alpha$

We reject the null hypothesis.

“With 95% confidence, there is enough evidence to say that at least one of the mean weight changes is significantly different than the others.”

“The overall mean weight change was found to be significantly different between diets (ANOVA,  $p = 0.003$ )”

## Using aov()

```
fit <- aov(Weight_Change~Diet, data = diet_df)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Diet           2     71    35.5     6.2 0.0032 **
## Residuals     75    430     5.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can further investigate pairwise differences between mean weight change using the Tukey Honest Significant Differences post hoc test:

```
res <- TukeyHSD(fit)
res
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = Weight_Change ~ Diet, data = diet_df)
##
## $Diet
##      diff      lwr      upr    p adj
## 2-1  0.27407 -1.3325  1.88062 0.91247
## 3-1 -1.84815 -3.4547 -0.24161 0.02014
## 3-2 -2.12222 -3.6808 -0.56365 0.00478
```

## Chi-squared Test

Is there an association between gender and diet group? In other words, do gender frequencies differ between diets?

```
table(diet_df$gender, exclude = FALSE)
```

```
##
##      0      1 <NA>
##  43   33     2
# replace missing gender values with most frequent
diet_df$gender[is.na(diet_df$gender)] <- 0
```

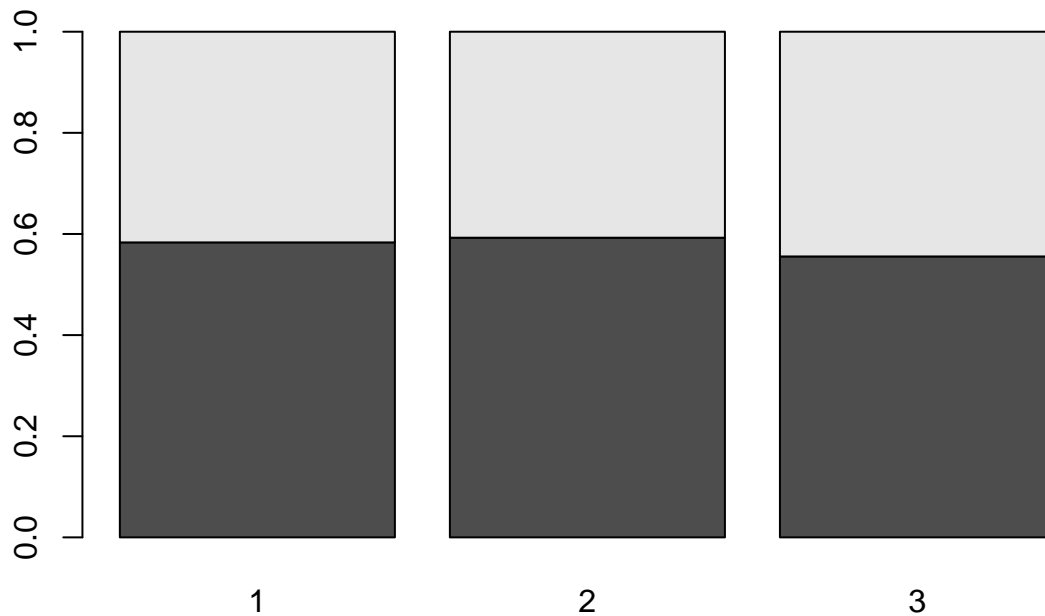
```
# contingency table
tbl <- table(diet_df$Diet, diet_df$gender)
tbl
```

```
##
##      0      1
##  1 14 10
##  2 16 11
##  3 15 12
```

```
# relative frequencies
tbl / rowSums(tbl)
```

```
##
##      0      1
##  1 0.58333 0.41667
```

```
## 2 0.59259 0.40741
## 3 0.55556 0.44444
barplot(t(tbl / rowSums(tbl)))
```



```
chisq.test(tbl)
```

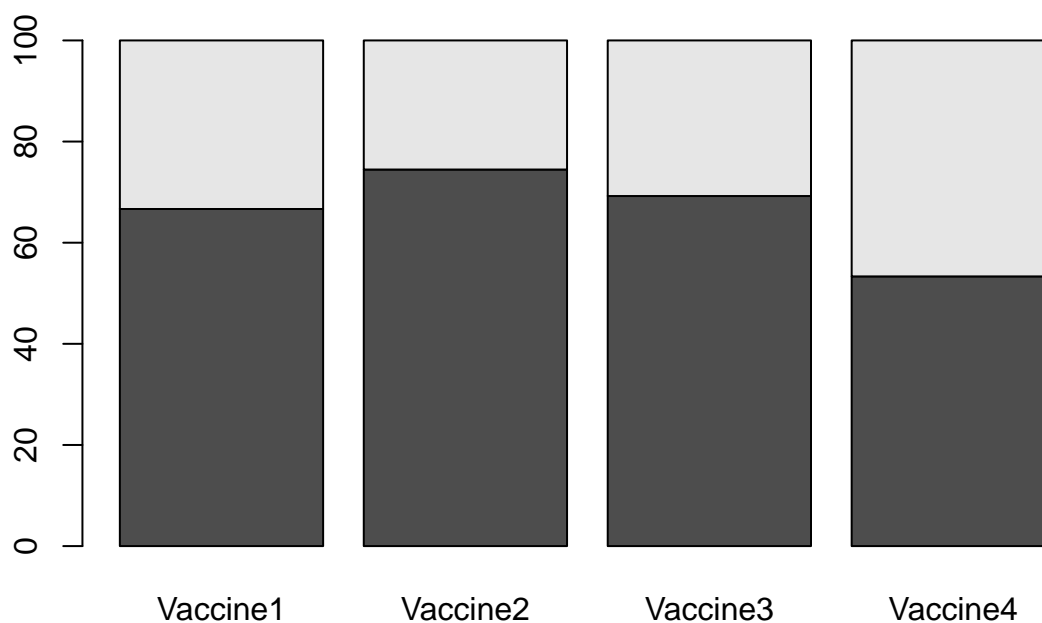
```
##
## Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 0.0817, df = 2, p-value = 0.96
```

Let's also repeat the example from the slides:

```
vac_tbl <- matrix(c(82, 70, 45, 48, 41, 24, 20, 42), ncol = 2)
colnames(vac_tbl) <- c("Protected", "Not")
rownames(vac_tbl) <- paste0("Vaccine", 1:4)
vac_tbl
```

```
##           Protected Not
## Vaccine1         82  41
## Vaccine2         70  24
## Vaccine3         45  20
## Vaccine4         48  42
```

```
# percentages
perc_tbl <- vac_tbl / rowSums(vac_tbl) * 100
barplot(t(perc_tbl))
```



```
chisq.test(vac_tbl)
```

```
##
## Pearson's Chi-squared test
##
## data: vac_tbl
## X-squared = 9.74, df = 3, p-value = 0.021
### Post hoc analysis
post_hoc <- c()
for (i in 1:3) {
  for (j in (i + 1):4) {
    v1 <- rownames(vac_tbl)[i]
    v2 <- rownames(vac_tbl)[j]

    res <- chisq.test(vac_tbl[c(v1, v2), ])
    post_hoc <- rbind(post_hoc,
                      data.frame(v1 = v1, v2 = v2, p = res$p.value))
  }
}
post_hoc$adj_p <- p.adjust(post_hoc$p, method = "fdr")
post_hoc
```

```
##          v1          v2          p    adj_p
## 1 Vaccine1 Vaccine2 0.2741047 0.411157
## 2 Vaccine1 Vaccine3 0.8466564 0.846656
## 3 Vaccine1 Vaccine4 0.0674289 0.135265
```

```
## 4 Vaccine2 Vaccine3 0.5854762 0.702571
## 5 Vaccine2 Vaccine4 0.0045935 0.027561
## 6 Vaccine3 Vaccine4 0.0676325 0.135265
```