# BB503/BB602 - R Training - Week IX

Ege Ulgen

## One-sample t-Test

The dataset we'll use contains information on 78 people using one of three diets.

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

head(diet_df)
```

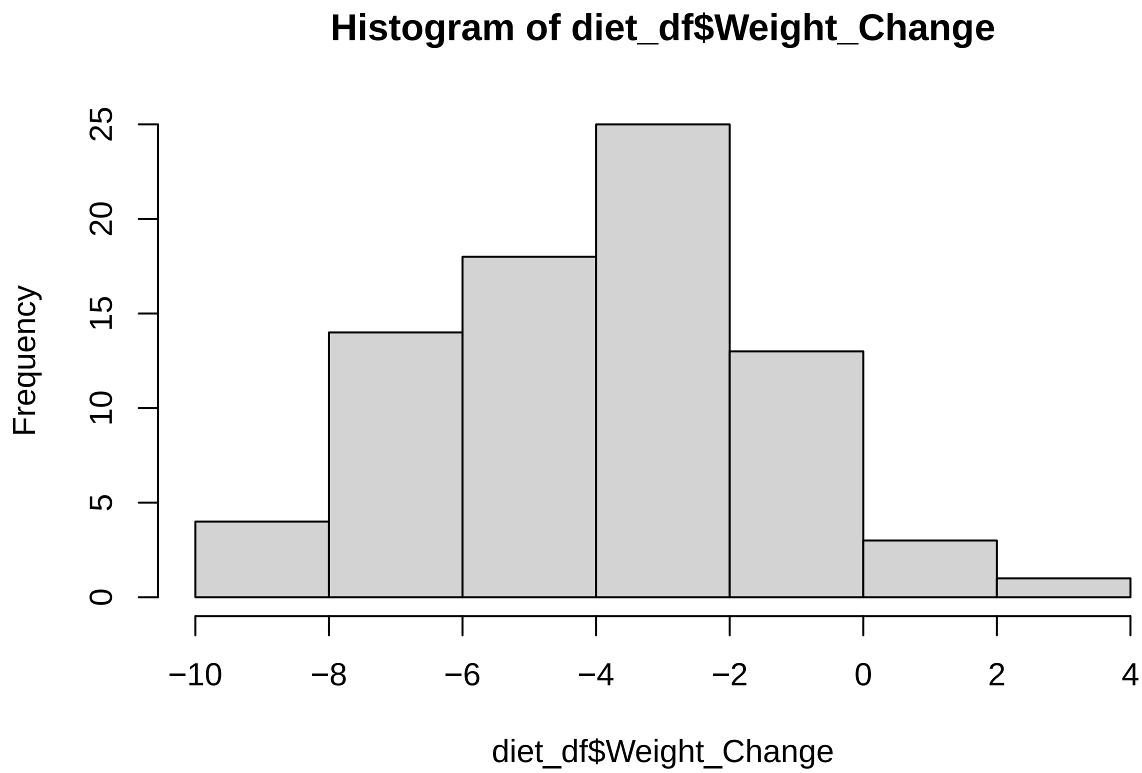
```
##   Person gender Age Height pre.weight Diet weight6weeks
## 1     25     NA  41    171         60    2         60.0
## 2     26     NA  32    174        103    2        103.0
## 3      1      0  22    159         58    1         54.2
## 4      2      0  46    192         60    1         54.0
## 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)
```

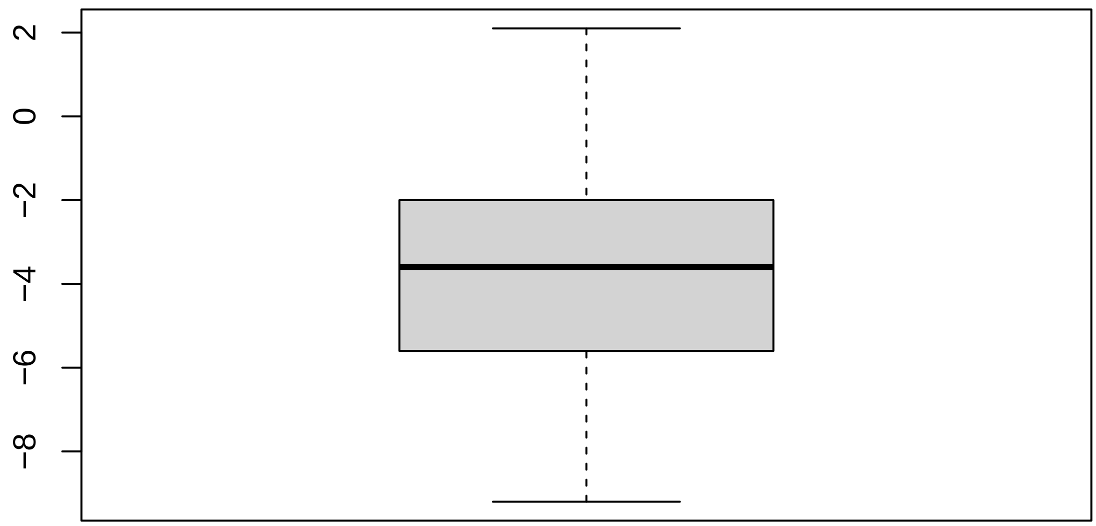
```
##      Person      gender       Age           Height      pre.weight    Diet
##  Min.   : 1.0   0   :43   Min.   :16.0   Min.   :141   Min.   : 58.0   1:24
##  1st Qu.:20.2   1   :33   1st Qu.:32.2   1st Qu.:164   1st Qu.: 66.0   2:27
##  Median :39.5   NA's: 2   Median :39.0   Median :170   Median : 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
```

Let's look at the overall distribution of weight change and the distributions by Diet type:

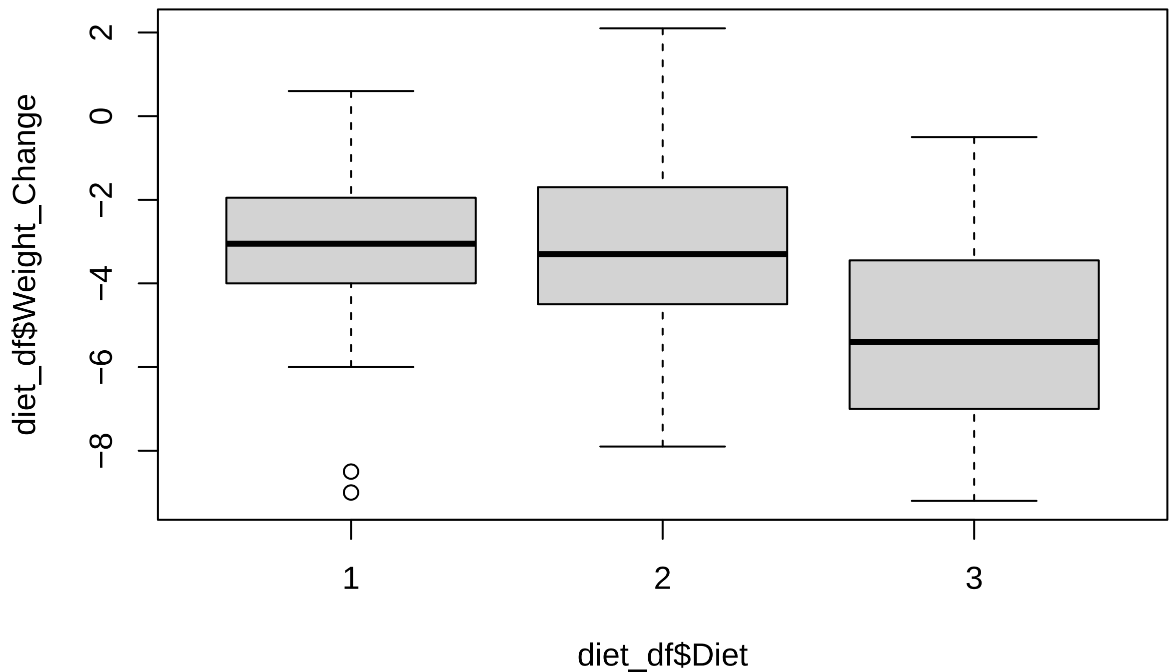
```
# overall
hist(diet_df$Weight_Change)
```




```
boxplot(diet_df$Weight_Change)
```




```r
# by diet type
boxplot(diet_df$Weight_Change~diet_df$Diet)
```




We can observe that `Weight_Change` seems to follow a normal distribution (we'll later learn how to test normality of a variable). Moreover, diet 3 seems to result in a higher decrease compared to diets 1 & 2. We'll test the difference between mean weight changes between diet 1 and 3 in the next section. For now, let's test whether the overall mean weight change is significantly different than -3.

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

Inspecting the histogram of overall weight change, we concluded that it is normally-distributed.

$H_0 : \mu = -3$ and $H_a : \mu \neq -3$

Let's choose $\alpha = 0.05$

### 2. Calculate the appropriate test statistic

$$t_H = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

```r
t_stat <- (mean(diet_df$Weight_Change) - (-3)) / (sd(diet_df$Weight_Change) / sqrt(nrow(diet_df)))
t_stat
```

```
## [1] -2.9245
```

```r
df <- nrow(diet_df) - 1
```

## 3. Calculate critical values/p value

### Critical values

```
C1 <- qt(0.05/2, df = df)
C2 <- qt(1 - 0.05/2, df = df)
C1; C2
```

```
## [1] -1.9913
```

```
## [1] 1.9913
```

```
# t_stat in rejection zone?
t_stat < C1 | t_stat > C2
```

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

### p value

```
p_val <- 2 * (1 - pt(abs(t_stat), df = df))
p_val
```

```
## [1] 0.0045304
```

### Confidence Interval

The 95% Confidence Interval for $\mu$:

$$95\%\ CI = [\bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}]$$

```
SE <- sd(diet_df$Weight_Change) / sqrt(nrow(diet_df))

mean(diet_df$Weight_Change) + C1 * SE; mean(diet_df$Weight_Change) + C2 * SE
```

```
## [1] -4.4201
```

```
## [1] -3.2696
```

## 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 mean weight change is significantly different than 0."

"The overall mean weight change was found to be significantly different than -3 (t-test p $<$ 0.001, 95% CI =

$$-4.42, -3.27$$

)"

### Using `t.test()`

```
?t.test
t.test(diet_df$Weight_Change, mu = -3)
```

```
##
##  One Sample t-test
##
## data:  diet_df$Weight_Change
## t = -2.92, df = 77, p-value = 0.0045
## alternative hypothesis: true mean is not equal to -3
## 95 percent confidence interval:
##  -4.4201 -3.2696
## sample estimates:
## mean of x
##   -3.8449
```

```r
# change mu
t.test(diet_df$Weight_Change, mu = -4)
```

```
##
##  One Sample t-test
##
## data:  diet_df$Weight_Change
## t = 0.537, df = 77, p-value = 0.59
## alternative hypothesis: true mean is not equal to -4
## 95 percent confidence interval:
##  -4.4201 -3.2696
## sample estimates:
## mean of x
##   -3.8449
```

```r
# change alternative
t.test(diet_df$Weight_Change, mu = -3, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  diet_df$Weight_Change
## t = -2.92, df = 77, p-value = 0.0023
## alternative hypothesis: true mean is less than -3
## 95 percent confidence interval:
##     -Inf -3.3639
## sample estimates:
## mean of x
##   -3.8449
```

```r
# change conf. level
t.test(diet_df$Weight_Change, mu = -3, conf.level = .99)
```

```
##
##  One Sample t-test
##
## data:  diet_df$Weight_Change
## t = -2.92, df = 77, p-value = 0.0045
## alternative hypothesis: true mean is not equal to -3
## 99 percent confidence interval:
##  -4.6079 -3.0818
## sample estimates:
## mean of x
##   -3.8449
```

## Two-sample t-Test

Let's test whether the mean weight changes of diet 1 and 3 are different.

```
sub_df <- subset(diet_df, Diet %in% c(1, 3))

### compare variances (F-test)
var.test(sub_df$Weight_Change[sub_df$Diet == 1], sub_df$Weight_Change[sub_df$Diet == 3])
```

```
##
##  F test to compare two variances
##
## data:  sub_df$Weight_Change[sub_df$Diet == 1] and sub_df$Weight_Change[sub_df$Diet == 3]
## F = 0.874, num df = 23, denom df = 26, p-value = 0.75
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.39207 1.99036
## sample estimates:
## ratio of variances
##            0.87445
```

```
# more compactly
var.test(Weight_Change~Diet, data = sub_df)
```

```
##
##  F test to compare two variances
##
## data:  Weight_Change by Diet
## F = 0.874, num df = 23, denom df = 26, p-value = 0.75
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.39207 1.99036
## sample estimates:
## ratio of variances
##            0.87445
```

```
res <- t.test(Weight_Change~Diet, data = sub_df, var.equal = TRUE)
res
```

```
##
##  Two Sample t-test
##
## data:  Weight_Change by Diet
## t = 2.83, df = 49, p-value = 0.0066
## alternative hypothesis: true difference in means between group 1 and group 3 is not equal to 0
## 95 percent confidence interval:
##  0.5380 3.1583
## sample estimates:
## mean in group 1 mean in group 3
##         -3.3000         -5.1481
```

```
res$p.value
```

```
## [1] 0.0066444
```

```
res$conf.int
```

```
## [1] 0.5380 3.1583
```

```
## attr(,"conf.level")
## [1] 0.95
```

```
res$estimate
```

```
## mean in group 1 mean in group 3
##         -3.3000         -5.1481
```

## Paired t-Test

A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat, low cholesterol diet. The subjects consumed on average 2.31g of the active ingredient, stanol easter, a day. This data set contains information on 18 people using margarine to reduce cholesterol over three time points.

```
chol_df <- read.csv("../data/Cholesterol_R.csv")
head(chol_df, 3)
```

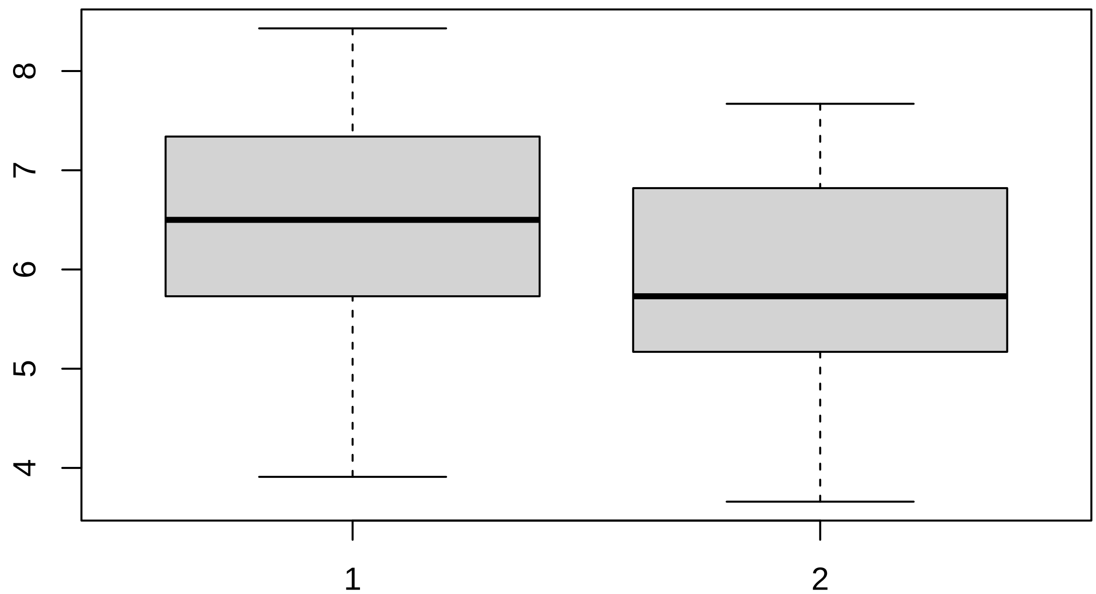
```
##   ID Before After4weeks After8weeks Margarine
## 1  1   6.42        5.83        5.75         B
## 2  2   6.76        6.20        6.13         A
## 3  3   6.56        5.83        5.71         B
```

```
# turn categorical variable into factor
chol_df$Margarine <- as.factor(chol_df$Margarine)

summary(chol_df)
```

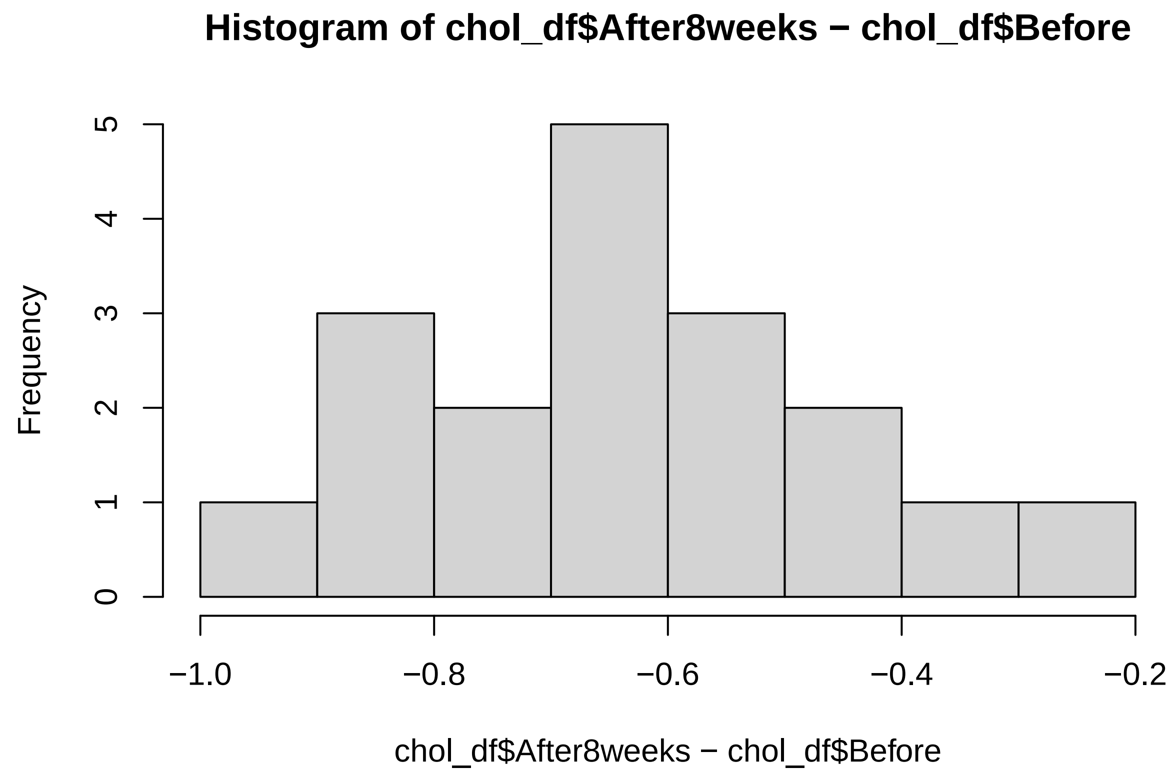
```
##        ID            Before       After4weeks     After8weeks    Margarine
##  Min.   : 1.00   Min.   :3.91   Min.   :3.70   Min.   :3.66   A:9
##  1st Qu.: 5.25   1st Qu.:5.74   1st Qu.:5.17   1st Qu.:5.21   B:9
##  Median : 9.50   Median :6.50   Median :5.83   Median :5.73
##  Mean   : 9.50   Mean   :6.41   Mean   :5.84   Mean   :5.78
##  3rd Qu.:13.75   3rd Qu.:7.22   3rd Qu.:6.73   3rd Qu.:6.69
##  Max.   :18.00   Max.   :8.43   Max.   :7.71   Max.   :7.67
```

For the overall data, let's compare whether there is a significant change between `Before` and `After8weeks`:

```
boxplot(chol_df$Before, chol_df$After8weeks)
```




```r
# check normality
hist(chol_df$After8weeks - chol_df$Before)
```




```r
t.test(chol_df$After8weeks, chol_df$Before, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  chol_df$After8weeks and chol_df$Before
## t = -14.9, df = 17, p-value = 3.3e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.71766 -0.54011
## sample estimates:
## mean difference
##        -0.62889
```

```r
t.test(chol_df$After8weeks - chol_df$Before)
```

```
##
##  One Sample t-test
##
## data:  chol_df$After8weeks - chol_df$Before
## t = -14.9, df = 17, p-value = 3.3e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.71766 -0.54011
## sample estimates:
## mean of x
##  -0.62889
```