# Assignment 1

Group 47: Luca Cavellini, Paul Hosek, Robert Satzger

2/20/2023

## Preparation: Defining Re-Usable Assumption Checks

To standardize assumption checks throughout the assignment, we use the following functions. `check_normality` plots the histogram, qq-plot, and boxplots for all variables of interest and returns the shapiro-wilk test. `check_vars` extends assumption checks for ANOVAs by showing the qq-plot of the residuals and plotting the fitted values against the residuals.

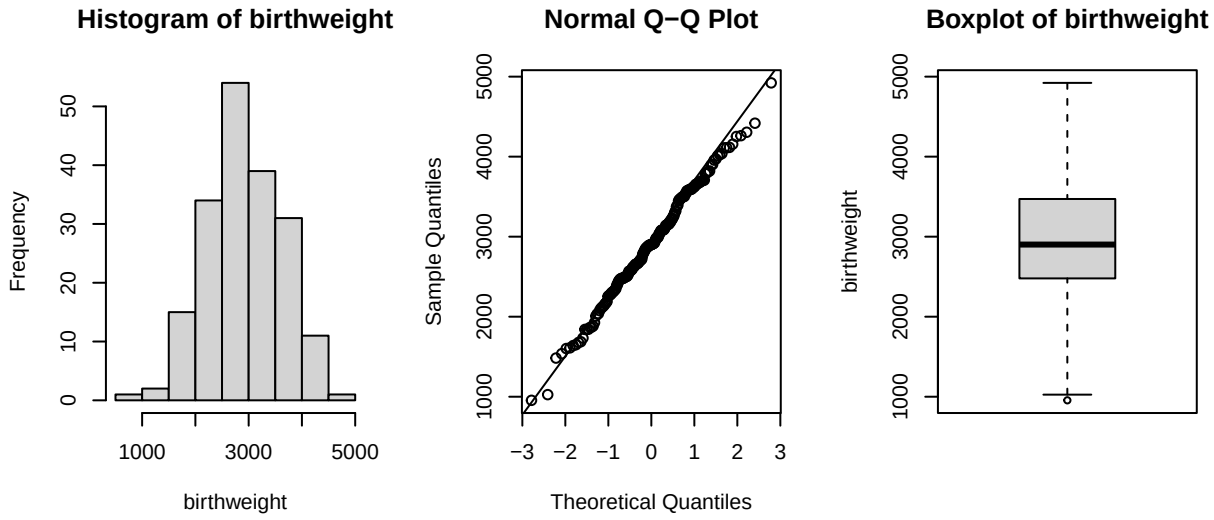
```
check_normality <- function(column_names, data, groups = NULL){
  shapiros = vector("list",length(column_names))
  par(mfrow=c(length(column_names),3+length(groups)/2))
  for (i in 1:length(column_names)){
    col_name <- column_names[[i]]
    col <- data[[col_name]]
    hist(col, xlab = col_name, main = paste("Histogram of", col_name))
    qqnorm(col)
    qqline(col)
    if (missing(groups)){
      boxplot(col, ylab=col_name, main = paste("Boxplot of", col_name))
    }else{
      for (g in groups){
        boxplot(col ~ data[[g]], ylab=col_name, xlab = g,
                main = paste("Boxplot of", col_name, "by", g))
      }
    }
    shapiros[[i]] <- shapiro.test(col)
  }
  return(shapiros)
}

check_vars <- function(anova_model){
  par(mfrow=c(1,2))
  qqnorm(residuals(anova_model), main = "Normal Q-Q Plot of Residuals")
  plot(fitted(anova_model), residuals(anova_model),
       main = "Residuals vs Fitted Values")
}
```

## Exercise 1. Birthweights

### a)

To test for a normal distribution of the sample, the output below show a histogram, a Q-Q plot, and a Shapiro-Wilk test.

```
library(ggplot2)
birthweight <- read.table('data/birthweight.txt', header = TRUE)
shapiro <- check_normality("birthweight", birthweight)
```




The histogram appears to have a bell-shaped curve with no dominant skew to either side, suggesting a normal distribution of the sample data. This is also supported by the roughly linear shape displayed in the Q-Q plot. Moreover, the Shapiro-Wilk test for normality is not significant with $p = 0.90$. Thus, we cannot reject the null hypothesis stating that the sample differs from a normal distribution. Based on these three indications, we therefore assume that the birth weights follow a normal distribution.

```
birthweight <- birthweight$birthweight
alpha = .04
t = qt(1-alpha/2, df = length(birthweight)-1) # t of alpha/2
margin = t * sd(birthweight) / sqrt(length(birthweight))
lower_bound = mean(birthweight) - margin
upper_bound = mean(birthweight) + margin
```

Based on our sample, the 96% confidence interval for $\mu$ is $CI = [2808.08, 3018.50]$.

```
alpha = .04
ci_length = 100
margin = ci_length / 2
q = qnorm(1-alpha/2) # approximating with z-scores instead of t-values because t-distribution
n = q^2 * sd(birthweight)^2 / margin^2
```

Ensuring a confidence interval with a maximum length of 100 requires a sample size $n \geq 821$.

```r
alpha = .04
B = 1000
Tstar = numeric(B)
for (i in 1:B){
  Tstar[i] = mean(sample(birthweight, replace = TRUE))
}
TstarLower = quantile(Tstar, alpha/2)
TstarUpper = quantile(Tstar, 1-alpha/2)
upper_bound = 2*mean(birthweight)-TstarUpper
lower_bound = 2*mean(birthweight)-TstarLower
```

The 96% confidence interval resulting from bootstrapping is $CI = [2808.75, 3010.12]$. As expected, the boundaries of the bootstrapped confidence interval are very similar to the calculated confidence interval reported prior because the sample follows an approximately normal distribution.

**b)**

Based on the assumption of normality established in part **a**, we perform a right-sided one-sample t-test.

```r
mu = 2800
t.test(birthweight, mu = mu, alternative = 'g')
```

```
##
##  One Sample t-test
##
## data:  birthweight
## t = 2.2271, df = 187, p-value = 0.01357
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829.202      Inf
## sample estimates:
## mean of x
##  2913.293
```

```r
binom.test(sum(birthweight>mu),length(birthweight),p=.5,alt='g')
```

```
##
##  Exact binomial test
##
## data:  sum(birthweight > mu) and length(birthweight)
## number of successes = 107, number of trials = 188, p-value = 0.03399
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5065781 1.0000000
## sample estimates:
## probability of success
##              0.5691489
```

**c)** REFINE! The power of the two tests can be compared by simulation. In this case, the

alternative hypothesis is true, stating that the true mean is above 2800. Therefore, one approach would be to randomly sample a mean from a normal distribution with mu in the interval $I = (2800, max(birthweight)]$. With this mean, one could produce a sample $X^*$ and perform each of the two statistical tests, testing its sample mean $\overline{X}^*$ against the hypothesis that the true mean is above 2800.
Consequently, one can calculate the proportion of times each test correctly rejected the null hypothesis.

It can be expected that the t-test will have a higher power than the sign test because it correctly assumes a normal distribution of the sample. In contrast, the sign test discards a lot of information. This makes it more robust than the t-test but reduces power in cases where assumptions aren't violated.

**d)**

```
p_hat = sum(birthweight<2600)/length(birthweight)
p_left = .25
me = p_hat - p_left
p_right = p_hat + me
z = me/sqrt(p_hat*(1-p_hat)/length(birthweight)) # z of alpha/2
conf_level = 1-pnorm(z, lower.tail = FALSE)*2
```

Taking our sample proportion as $\hat{p}$, we can use the margin of error to compute the right side of the confidence interval. This results in the confidence interval $CI = [0.25, 0.41]$. The interval has a confidence level of 98%.

**e)** The problem can be seen as a comparison of proportions between two independent samples.

```
k_male = 34
k_female = 28
n_male = k_male + 61
n_female = k_female + 65
p_female = k_female/n_female

binom_test = binom.test(k_male, n_male, p=p_female)
```

The binomial test returns $p = 0.26$. Since it is larger than the standard significance level $\alpha = 0.05$, we reject the null hypothesis stating that the two proportions are significantly different. Thus, the expert's claim is not supported by the data.

## Exercise 2. Cholesterol

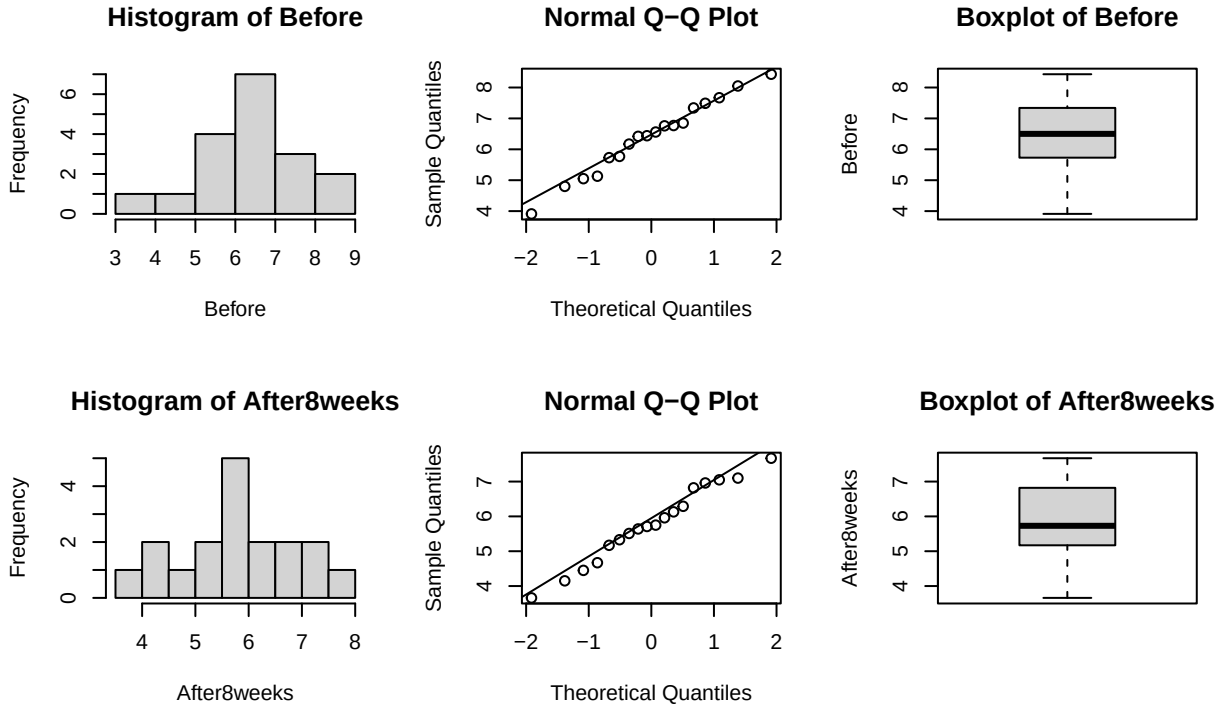
```
library(ggplot2)
library(dplyr)
library(ggpubr)
data <- read.table("data/cholesterol.txt", header = TRUE, sep = " ")
```

**a)**

First plotting the discrete distribution of values, we see that values appear approximately normally distributed. Further, QQ-plots indicate that samples from both measurements are similarly

distributed by inspection. However, we need to confirm this with thorough statistical testing.

```
shapiros <- check_normality(list("Before", "After8weeks"), data)
```

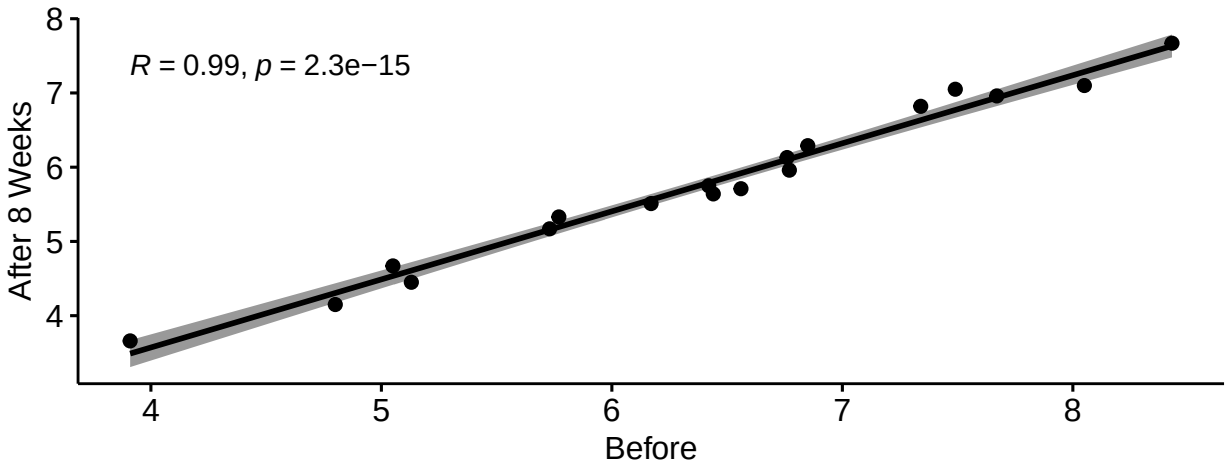



```
# data_long <- data.frame(
#   Time = rep(c("Before", "After 8 Weeks"), each = nrow(data)),
#   Value = c(data$Before, data$After8weeks)
# );
# ggplot(data_long, aes(x = Value, fill = Time)) +
#   geom_histogram(alpha = 0.5, position = "identity", bins = 10) +
#   labs(title = "Histogram of Before and After 8 Weeks", x = "Values", y = "Frequency") +
#   scale_fill_manual(values = c("#E69F00", "#56B4E9"), labels = c("Before", "After 8 Weeks"))
```

Shapiro-Wilk tests for normality indicate that the assumption of normality is met for both pre- and post-measurement. P-values are $p = 0.97$ for the before measurement and $p = 0.92$ for the post measurement respectively.

Further, we observe a very high correlation of the data.

```
ggscatter(data, x = colnames(data)[1], y = colnames(data)[2],
          add = "reg.line",cor.coef = TRUE, conf.int = TRUE,
          cor.method = "pearson",
          xlab = "Before", ylab = "After 8 Weeks");
```




**b)** First, we use a non-parametric, repeated measures t-test.

```
t.test(data$Before, data$After8weeks, paired = TRUE, alternative = "two.sided");
```

```
##
##  Paired t-test
##
## data:  data$Before and data$After8weeks
## t = 14.946, df = 17, p-value = 3.279e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5401131 0.7176646
## sample estimates:
## mean of the differences
##               0.6288889
```

```
wilcox.test(data$Before, data$After8weeks, paired = TRUE, alternative = "two.sided");
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Before and data$After8weeks
## V = 171, p-value = 7.629e-06
## alternative hypothesis: true location shift is not equal to 0
```

The results indicate, that at $\alpha = 0.05$, there is a significant effect of the diet. However, this must not implyeffect has practical significance and is strong enough for the diet to be useful in practice.

The permutation test is applicable, since it can express any test-statistic including repeated measures and also is non-parameteric.

**c)**

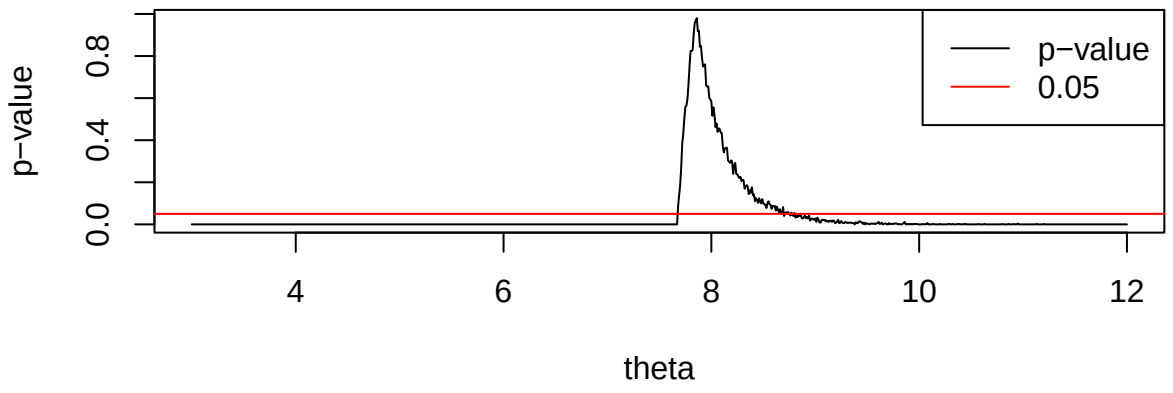
```
nsamples <- 1000;

theta <- replicate(nsamples, max(sample(data$After8weeks,size=18, replace = TRUE)));
cat("The confidence interval of the max is:",quantile(theta, c(0.025, 0.975)),"\n");
```

```
## The confidence interval of the max is: 6.96 7.67
```

This confidence interval could be improved by resampling more often. However, since the sample size is only 18, conclusions about the true population mean are limited. As such, we recommend to collect more samples.

**d)**

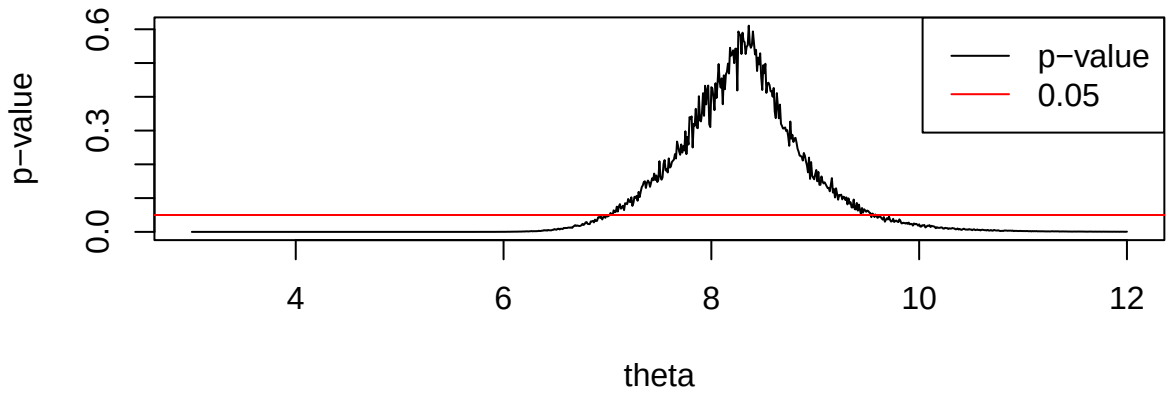
```
nsamples <- 1000;
thetas <- seq(3, 12, by = .01);
t <- max(data$After8weeks)
p_vals <- numeric(length(thetas));
for (i in 1:length(thetas)) {
  res <- replicate(nsamples,max(runif(18,3, thetas[i])))
  pl=sum(res<t)/nsamples
  pr=sum(res>t)/nsamples
  p_vals[i]=2*min(pl,pr)
}

plot(thetas, p_vals, type = "l", xlab = "theta", ylab = "p-value");
abline(h = 0.05, col = "red");
legend("topright", legend = c("p-value", "0.05"),
       col = c("black", "red"), lty = c(1, 1));
```




Alternatively, we can use the Kolmogorov-Smirnov test to test if the distributions are the same. But we must be careful to not accept distributions with a maximum value smaller than the maximum of the sample, because else the maximum lies outside the given range.

```
thetas <- seq(3, 12, by = .01);

# Loop over theta values and perform bootstrap test
ks_res <- numeric();
p_vals <- numeric(length(thetas));
for (i in 1:length(thetas)) {
```

```
  theta <- thetas[i];
  cur_unif <- runif(10000,3,thetas[i]);
  p_vals[i] <- ks.test(data$After8weeks, cur_unif)['p.value'];
}
```




**e)**

```
binom_e <- binom.test(sum(data$After8weeks < 6), nrow(data), alternative = "less")
cat("The p-value does not reject the H0 at alpha=0.05: p-value =",binom_e$p.value)
```

```
## The p-value does not reject the H0 at alpha=0.05: p-value = 0.8810577
```

First, we find true fraction of cholesterol levels less than 4.5.

```
prop <- mean(data$After8weeks < 4.5)
cat("The fraction of cholesterol levels <4.5 is", prop, "..")
```

```
## The fraction of cholesterol levels <4.5 is 0.1666667 ..
```

```
binom_e <- binom.test(sum(data$After8weeks < 4.5), nrow(data), p = 0.25, alternative = "less")
cat("No, the fraction <4.5 is not less than 25%, with p=",binom_e$p.value)
```

```
## No, the fraction <4.5 is not less than 25%, with p= 0.3056892
```

## Exercise 3. Diet

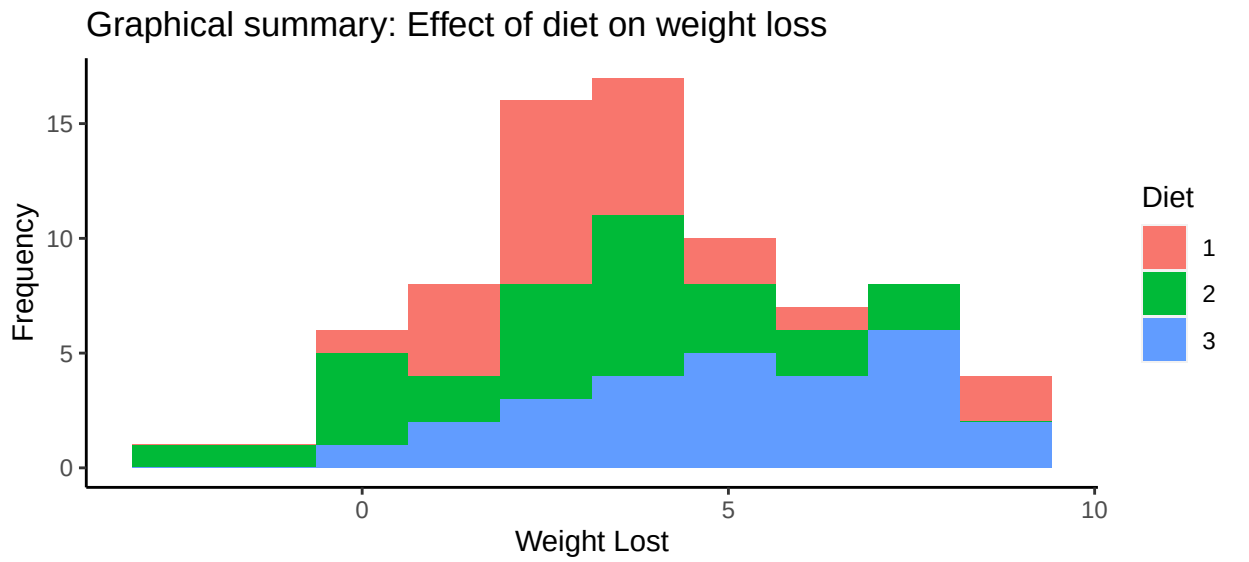
```
library(ggplot2)
library(dplyr)
library(ggpubr)
```

**a)**

```
data <- read.table("data/diet.txt", header=TRUE)
data$diet <- factor(data$diet)
```

```r
data$weight.lost <- data$preweight - data$weight6weeks
```

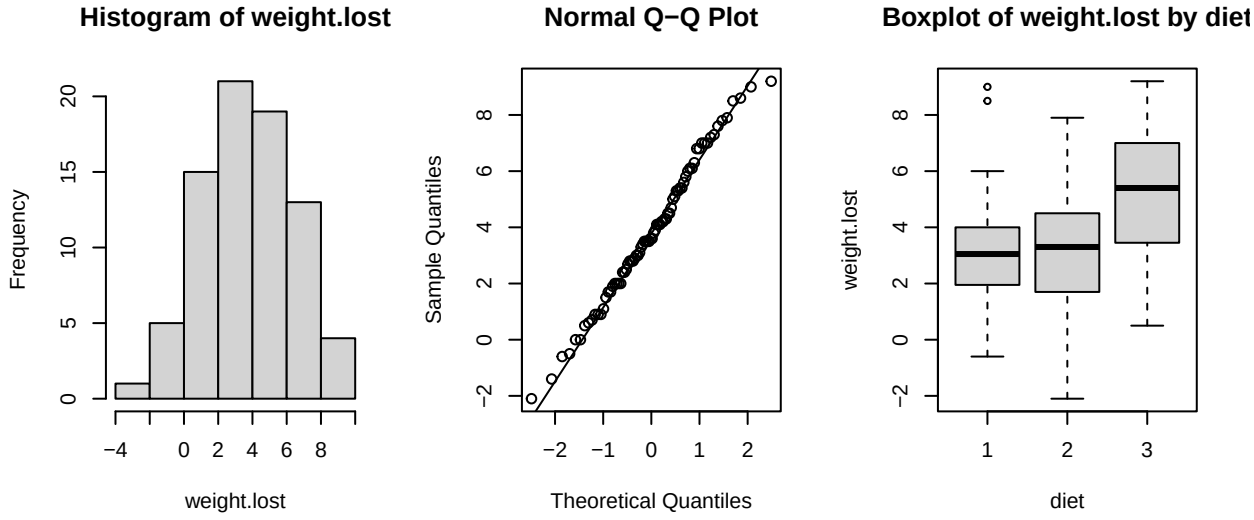
```r
# Assuming your data frame is called `mydata`
# boxplot(weight.lost ~ diet, data = data, xlab = "Diet", ylab = "Weight Lost",
#         main = "Graphical summary: Effect of diet on weight loss")
```




To test if the diet has a general effect on weight loss, we can do a paired t-test. Here, we test if there is a significant weight loss between the measurements.

First, testing the assumption of normality, we see that it is met. The data follows approximately a normal distribution

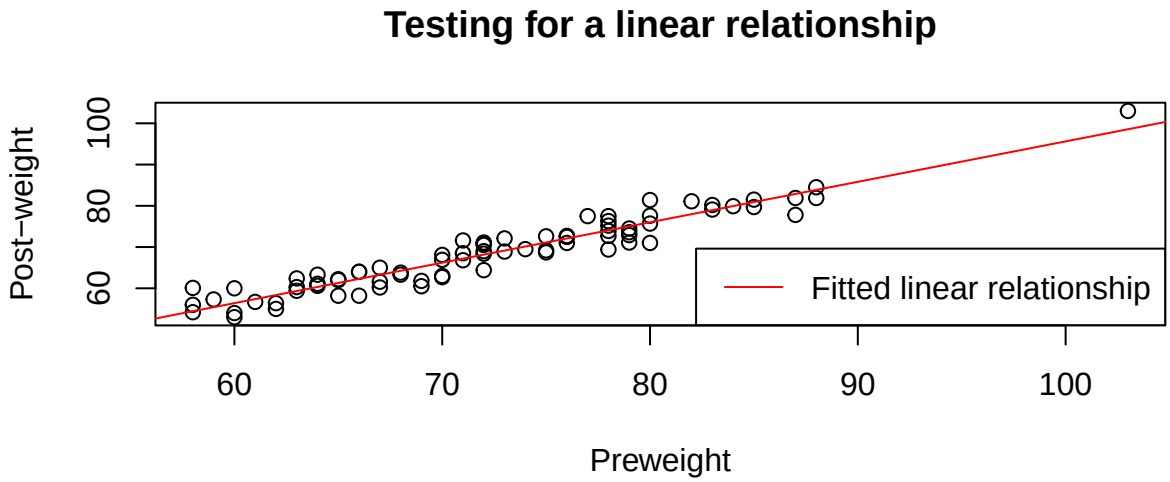
```r
shapiro <- check_normality("weight.lost", data, "diet")
```




Upon visual inspection, the assumption of a normal distribution appears to be met. The Shapiro-Wilk test does not find a violation of the normality assumption, either, with $p = 0.80$.

Further, the assumption of a linear relationship of the independent and dependent variables is met. The data spreads evenly around the the fitted linear line and does not follow a higher order interaction.

```
plot(data$preweight, data$weight6weeks, main = "Testing for a linear relationship", xlab = "Pre
abline(lm(data$weight6weeks ~ data$preweight), col = "red")
legend("bottomright", legend = "Fitted linear relationship", col = "red", lty = 1)
```




The results of the repeated measures t-test show us that at alpha=0.05, there is a significant difference between pre- and post weight. This indicates a statistically significant effect of the diet. However, since there are three different types of diets, it is possible that if combined they have a significant effect on weight loss, but there are differences between the diets. It is possible, that one of the diets leads to no weight loss, but the other two do. As such, we need to conduct a one-way ANOVA to investigate the effect of the type of diet.

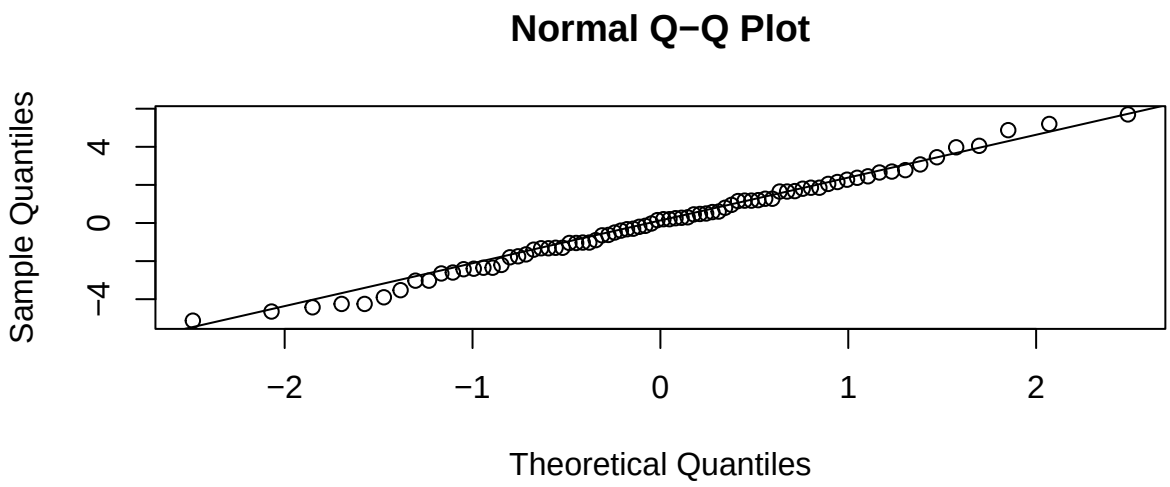
```
t.test(data$weight6weeks, data$preweight,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  data$weight6weeks and data$preweight
## t = -13.309, df = 77, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.420141 -3.269602
## sample estimates:
## mean of the differences
##               -3.844872
```

**b)**

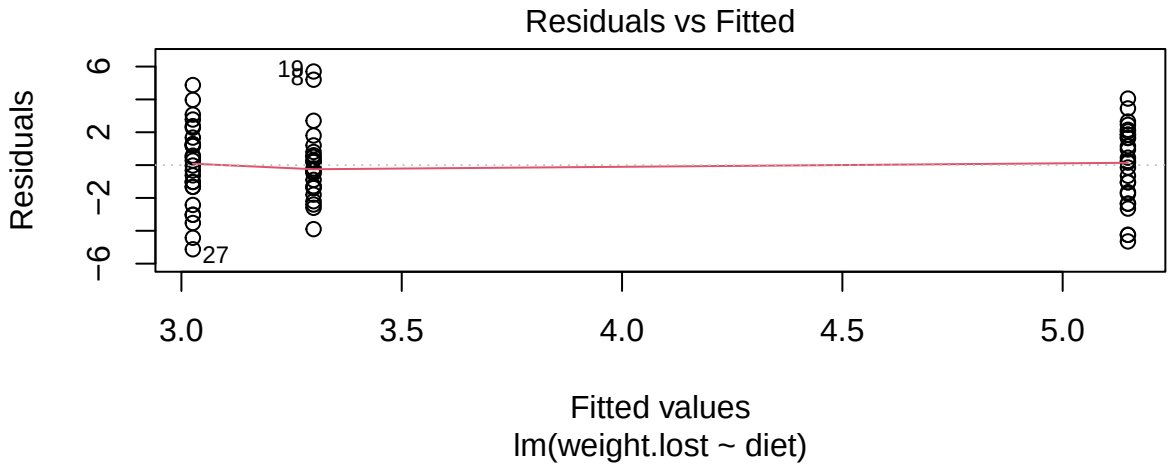
First, testing for normality, we can see that data is approximately normal distributed. However, extreme values on both tails spread some doubt about the normality assumption.

```
one_w_anova <- lm(weight.lost ~ diet, data)
qqnorm(one_w_anova$residuals); qqline(one_w_anova$residuals)
```




Secondly, testing for the independence of residuals. For this, we can plot the order of observation against the residuals. Here, we see data to be randomly distributed within blocks.

```
plot(one_w_anova, which = 1)
```




We now can do a one-way ANOVA to test the effect of type of diet. The result indicate, that there is a significant difference in the type of diet on weight loss (at alpha=0.5). Based on the result of the ANOVA alone, we cannot conclude which type of diet is different.

```
res_ow <-anova(one_w_anova)
res_ow
```

```
## Analysis of Variance Table
```

```
##
## Response: weight.lost
##           Df Sum Sq Mean Sq F value   Pr(>F)
## diet       2  71.09  35.547  6.1974 0.003229 **
## Residuals 75 430.18   5.736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To identify which type of diet is best for loosing weight, we will conduct post-hoc t-tests between all pairs of the ANOVA. Here, we must account for the multiple comparisons and adjust our sigificance level accordingly. As such, we will use Bonferroni correction.

```
post_hoc <- pairwise.t.test(data$weight.lost, data$diet, p.adjust.method = "bonferroni")
post_hoc
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$weight.lost and data$diet
##
##   1      2
## 2 1.0000 -
## 3 0.0224 0.0051
##
## P value adjustment method: bonferroni
```

We can see that there is a significant difference between diets 1 and 3 and diets 2 and 3. However, no significant difference between diet 1 and 2 was found.

```
means <- aggregate(data$weight.lost, list(data$diet),mean)
names(means)[2] <- "Average weight loss"
names(means)[1] <- "Diet Type"

means
```

```
##   Diet Type Average weight loss
## 1         1            3.300000
## 2         2            3.025926
## 3         3            5.148148
```

Comparing the means, we can see the direction of the relationships. Based on the significant differences found before, we can conclude that diet 3 is more effective than diets 1 and 2. It is the best diet for losing weight. But, all three diets lead to weight loss.

It is also possible to alternatively conduct a Kruskal-Wallis test instead of the one-way ANOVA. Again, a significant difference at the same significance level between groups was found.

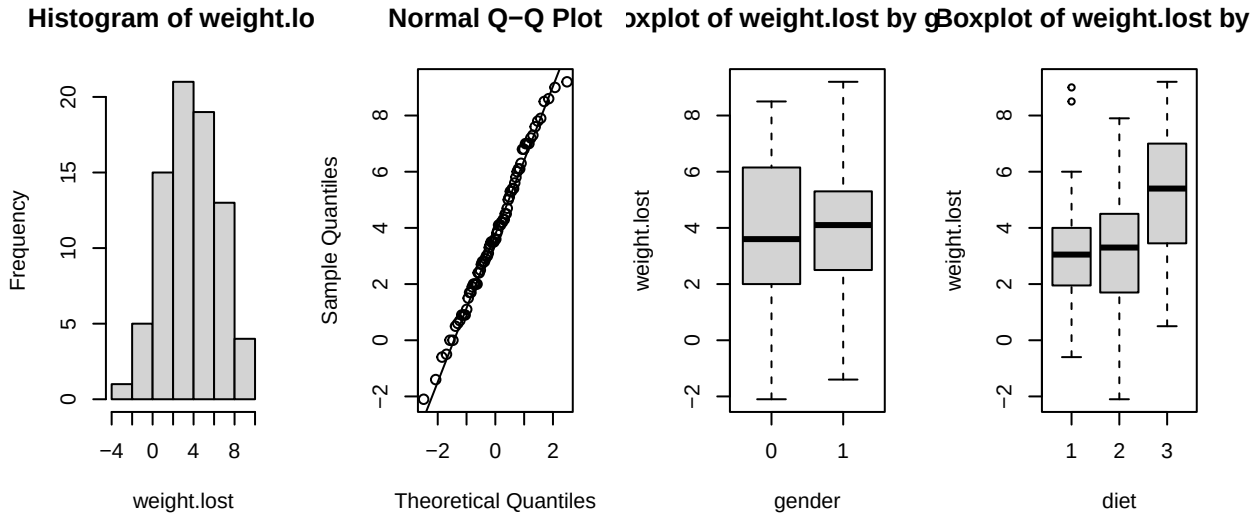
```
kruskal.test(weight.lost ~ diet, data)
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  weight.lost by diet
## Kruskal-Wallis chi-squared = 10.437, df = 2, p-value = 0.005416
```

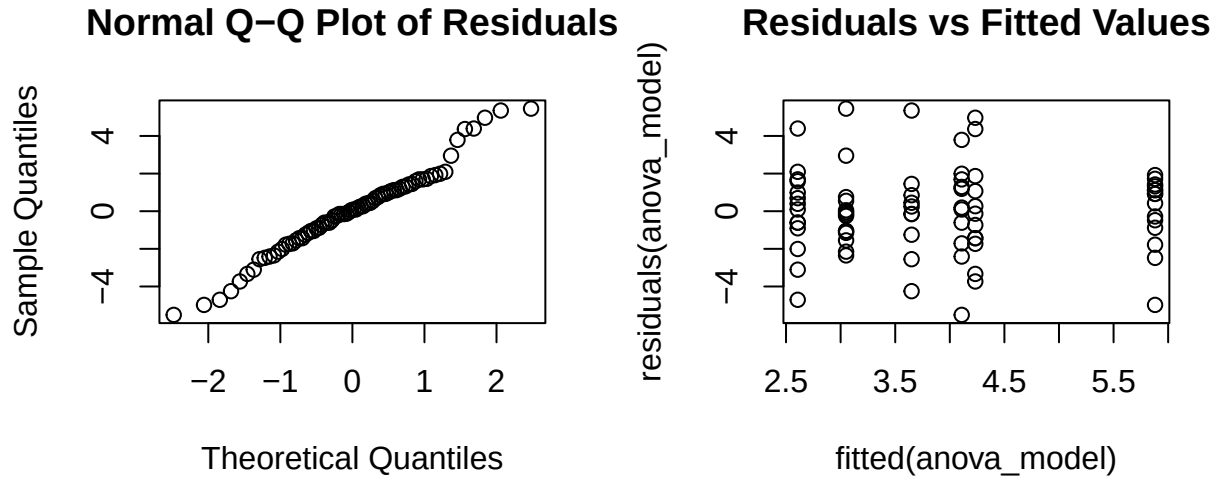
c)

```
two_w_anova <- lm(weight.lost ~ diet * gender, data)
check_normality("weight.lost", data, list("gender", "diet"))
```




```
## [[1]]
##
##  Shapiro-Wilk normality test
##
## data:  col
## W = 0.98991, p-value = 0.802
```

```
check_vars(two_w_anova)
```



```
print(anova(two_w_anova))
```

```
## Analysis of Variance Table
##
## Response: weight.lost
##             Df Sum Sq Mean Sq F value   Pr(>F)
## diet         2  60.53 30.2635  5.6292 0.005408 **
## gender       1   0.17  0.1687  0.0314 0.859910
## diet:gender  2  33.90 16.9520  3.1532 0.048842 *
## Residuals   70 376.33  5.3761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**e)**

## Exercise 4. Yield of peas

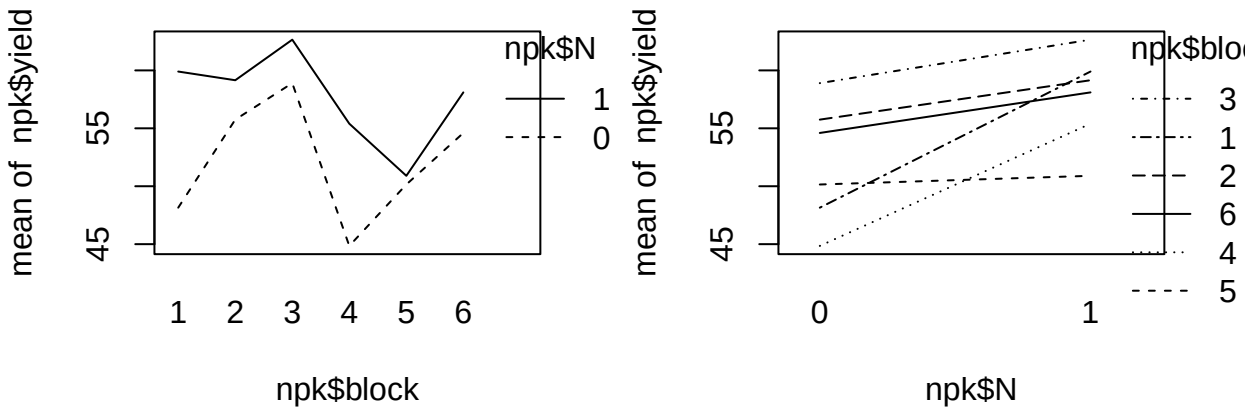
**a)**

```
library(MASS)

random_plots <- cbind(
  rep(1:24),
  rep(1:6, each = 4),
  replicate(3, c(replicate(6, sample(c(1,1,0,0)))))
)
plots_df <- data.frame(random_plots)
header <- c("plot", "block", "N", "P", "K")
colnames(plots_df) <- header
plots_df[1:8,]
```

```
##   plot block N P K
## 1    1     1 1 0 0
## 2    2     1 0 1 1
## 3    3     1 1 1 1
## 4    4     1 0 0 0
## 5    5     2 0 1 0
## 6    6     2 0 0 1
## 7    7     2 1 1 0
## 8    8     2 1 0 1
```

**b)**

The following plots show the average yield per block for plots treated with or without Nitrogen (N).

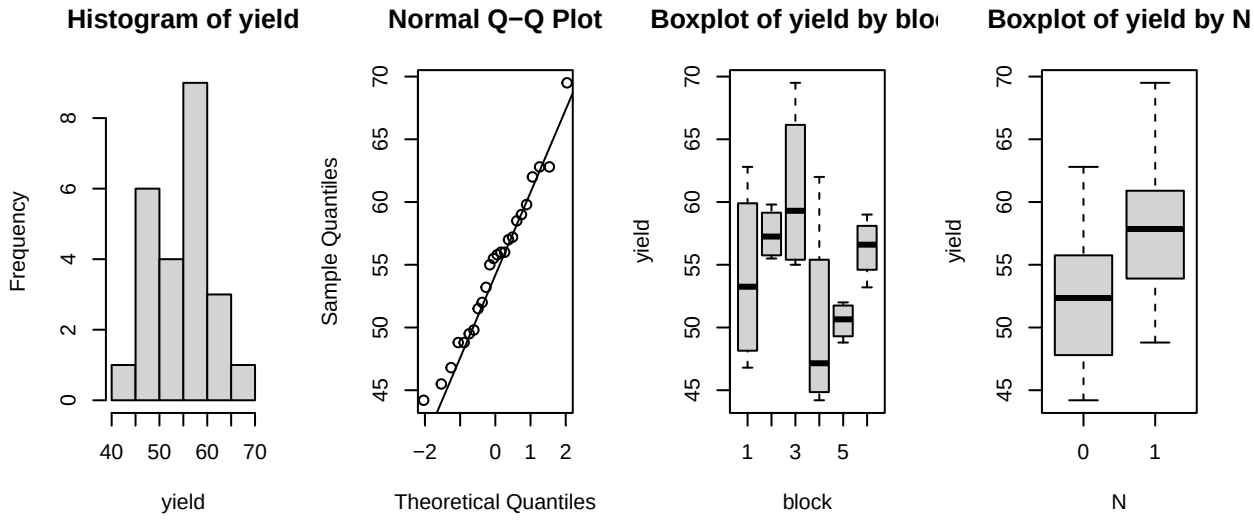



The study generally follows an incomplete block design. This would suggest interactions cannot be evaluated. However, taking only the factors block and N into account, there are two observations per condition. This means, the interactions may be evaluated with caution. The plots suggest that the treatment effect (i.e., the effect of Nitrogen on yield) may differ between blocks. Taking this potential interaction of block and Nitrogen into account can help explain some of the variance in the treatment effect. Therefore, it can help improve the model fit. By performing a full two-way ANOVA, we can test whether the interaction is, in fact, present.

**c)**

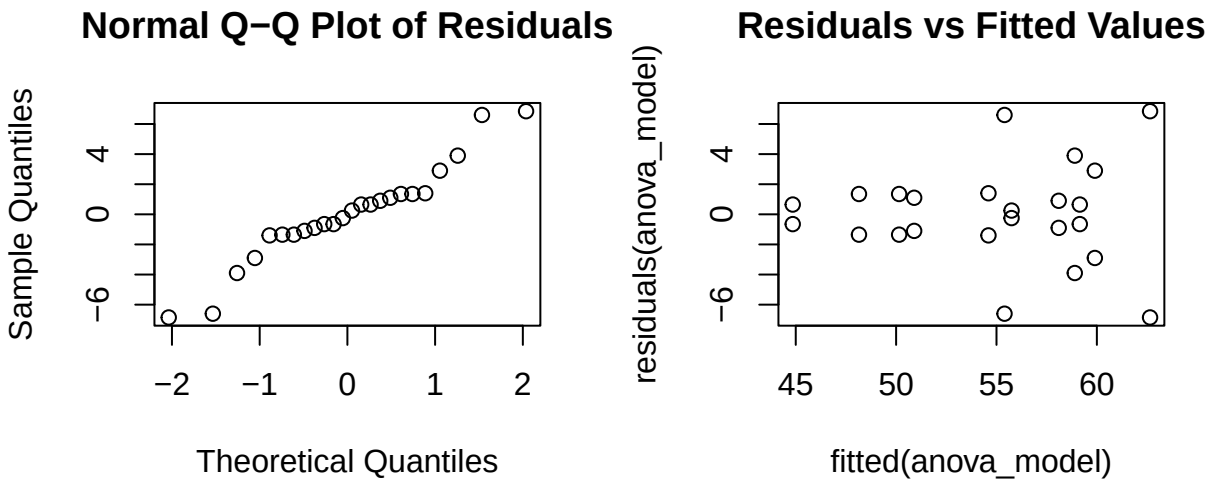
[FIXED EFFECTS VS FULL ANOVA?...] To perform an ANOVA, we first check the normality assumption with the following plots.

```
npk$block <- as.factor(npk$block)
npk$N <- as.factor(npk$N)
shapiro <- check_normality("yield", npk, c("block","N"))
```



The histogram and the QQ-plot indicate a normal distribution. Moreover, the Shapiro-Wilk test returns $p = 0.87$. Therefore, we do not find a violation of the normality assumption. Second, we check for homogeneity of variances.

```
lm_aov1 = lm(yield ~ block * N, data = npk)
check_vars(lm_aov1)
```




[DOES NOT SEEM TO BE MET... Suggestions in slides: Otherwise, you may consider transforming the data (e.g. use log Y ); using a different test; omit some (outlying) data-points (careful!); something else (there is no fix that always works).]

```
aov1 = anova(lm_aov1)
print(aov1)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## block      5 343.29  68.659  3.3592 0.03967 *
## N          1 189.28 189.282  9.2607 0.01021 *
## block:N    5  98.52  19.704  0.9640 0.47690
## Residuals 12 245.27  20.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There does not appear to be an interaction effect of block x Nitrogen on yield. Therefore, block does not affect our research question about the effect of Nitrogen on yield. A more sensible way to take block into account in the analysis would be in an additive model.

The Friedman test is not applicable in this case because there are two observations of the outcome variable (yield) per block x Nitrogen combination.

**d)**

[TEST ASSUMPTIONS FOR ONE MODEL?; TESTS DO NOT MAKE MUCH SENSE...]

```r
models <- list(
  lm(yield ~ block * N + P + K, npk),
  lm(yield ~ block * P + K + N, npk),
  lm(yield ~ block * K + N + P, npk)
)
aovs <- vector("list", 3)

for (i in 1:length(models)){
  aovs[[i]] <- anova(models[[i]])
  print(aovs[[i]])
}
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  4.8465 0.016439 *
## N          1 189.28 189.282 13.3611 0.004423 **
## P          1   8.40   8.402  0.5931 0.459045
## K          1  95.20  95.202  6.7201 0.026843 *
## block:N    5  98.52  19.704  1.3908 0.306583
## Residuals 10 141.67  14.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  4.0678 0.028234 *
## P          1   8.40   8.402  0.4978 0.496588
## K          1  95.20  95.202  5.6404 0.038947 *
## N          1 189.28 189.282 11.2143 0.007381 **
## block:P    5  71.40  14.280  0.8460 0.547341
## Residuals 10 168.79  16.879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  4.0407 0.028799 *
## K          1  95.20  95.202  5.6028 0.039477 *
## N          1 189.28 189.282 11.1397 0.007521 **
## P          1   8.40   8.402  0.4945 0.497989
## block:K    5  70.27  14.054  0.8271 0.558263
## Residuals 10 169.92  16.992
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[WHICH MODEL IS BEST?]

**e)**

Below, we base our normality assumption on the checks done in **c**.

```
library(lme4)
mixed_model <- lmer(yield ~ N + (1|block), REML=FALSE, data = npk)
mixed_model1 <- lmer(yield ~ (1|block), REML=FALSE, data = npk)
anova(mixed_model1, mixed_model)
```

```
## Data: npk
## Models:
## mixed_model1: yield ~ (1 | block)
## mixed_model: yield ~ N + (1 | block)
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## mixed_model1    3 159.38 162.91 -76.690   153.38
## mixed_model     4 153.48 158.20 -72.742   145.48 7.8953  1   0.004956 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[COMPARE TO C]