# EDDA Assignment1 - Group 22

Adwitiya Mandal, Oromia Sero, Priyakshi Goswami

2023-02-27

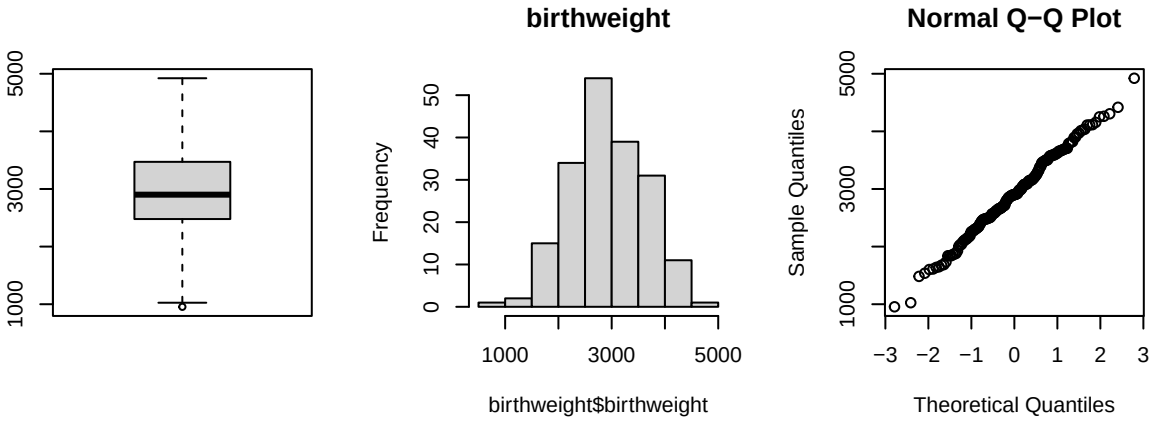
## Exercise 1. Birthweights

```
birthweight=read.table(file='birthweight.txt',header=TRUE)
```

$\mu$ = underlying mean birthweight.

### a)

**Normality check:**




Looking at the histogram and QQ plot of the data, the data seems to appear to satisfy normality. Next, we also do a Shapiro-Wilk test to confirm our graphical checks.

```
shapiro.test(birthweight$birthweight)[[2]] #p-value = 0.8995
```

```
## [1] 0.8995395
```

Since the p value is larger than 0.05, the null hypothesis (i.e. the population is normally distributed) can not rejected. So, there is no evidence that the data is not normally distributed.

#### 96%-CI for mean

We need to construct a 96%-CI for $\mu$. We don't know $\sigma$ here, so we will estimate it by $s$. Then the $(1-\alpha)$ - confidence interval of $\mu$ is given by $\overline{X} \mp t_{\alpha/2}\frac{s}{\sqrt{n}}$. We already know $n = 188$. We can estimate $\overline{X}$ and $s$ from the data. Our $\alpha = 0.04$

```
X_bar = mean(birthweight$birthweight)
s = sd(birthweight$birthweight)
X_bar
```

```
## [1] 2913.293
```

```
s
```

```
## [1] 697.5002
```

Now,we can calculate our margin of error.$t_{\alpha/2} = t_{0.04/2} = t_{0.02}$

```
n=188
qt(0.98,df=n-1)
```

```
## [1] 2.068173
```

```
me = qt(0.98,df=n-1)*s/sqrt(n)
X_bar - me
```

```
## [1] 2808.084
```

```
X_bar + me
```

```
## [1] 3018.501
```

So, the 96%-CI for $\mu$ is [2808.08,3018.50].

**Sample size needed to provide that the length of the 96%-CI is at most 100.**

Length of 96%-CI is $2 * t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$. Therefore

$$
\begin{aligned}
2 * t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &\leq 100 \\
\implies t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &\leq 50 \\
\implies n &\geq (\frac{t_{\frac{\alpha}{2}} * s}{50})^2
\end{aligned}
$$

```
(qt(0.98,df=n-1)*s/50)^2
```

```
## [1] 832.3819
```

Sample size needed is atleast 833.

**Bootstrap 96%-CI for mean**

```
B=1000
Tstar=numeric(B)
for(i in 1:B){
  Xstar=sample(birthweight$birthweight,replace=TRUE) # generate samples
  Tstar[i]=mean(Xstar)} #calculate bootstrap estimates

Tstar20=quantile(Tstar,0.02) # deterimine T*(alpha/2)
Tstar98=quantile(Tstar,0.98) # deterimine T*(1-alpha/2)
#sum(Tstar<Tstar20)
T1=mean(birthweight$birthweight)
c(2*T1-Tstar98,2*T1-Tstar20)
```

```
##      98%       2%
## 2810.999 3018.641
```

Bootstrap 96%-CI is [2811.66,3019.23].

When comparing the two CI-s, we observe that the Bootstrap-CI is narrower.In other words, this shows a reduced margin of error for the mean in the bootstrap sample data.

## b)

Our $H_1$ (alternative hypothesis) is that the mean of birthweights is greater than 2800gram ($H_1 : \mu > 2800$). And $H_0$ is ($H_0 : \mu \leq 2800$). Since, the data is normal, we can use a one sample t-test to check the above hypotheses.

One-sided t-test to check the claim mean birthweight is bigger than 2800 gram.

```
t.test(birthweight$birthweight,mu=2800,alt="g")
```

```
##
##  One Sample t-test
##
## data:  birthweight$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
```

With a small p-value = 0.01357 ($< 0.05$), $H_0$ is rejected. Hence, the expert's claim that $\mu > 2800$ holds for the data.

The 95%-CI is $[2829.202, \infty]$. Since, we have set the argument of alternative='greater' for a one-sided test, the CI constructed is one-sided and the natural border is extended to $\infty$(Inf) in the other side. As we conducted a one-sided test, the INF on the right side shows that we are only concerned with the lower bound of the confidence interval.

Sign-test: We can then use the binomial sign test to test whether the proportion of birthweights above 2800 grams is significantly higher than 0.5

```
sum(birthweight$birthweight>2800)
```

```
## [1] 107
```

```
binom.test(107,n,p=0.5,alt="g") #exact binomial test
```

```
##
##  Exact binomial test
##
## data:  107 and n
## 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
```

Checking the p-value, we can reject $H_0$. So, the expert's claim holds.

## c) Comparing the powers of t-test and sign test

We tested $H_0 : \mu \leq 2800$ using the t-test and sign test.

```
n=188
B=1000
psign=numeric(B) ## will contain p-values of sign test
```

```
pttest=numeric(B) ## will contain p-values of t-test
for(i in 1:B) {
  x=rnorm(n,mean=2800.04,sd=1) ## generate data under H1 with mu=2810
  pttest[i]=t.test(x,mu=2800,alt='g')[[3]]
  psign[i]=binom.test(sum(x>2800),n,p=0.5)[[3]] }

sum(psign<0.05)/B # fraction of rejecting H0, the power of the sign test
```

```
## [1] 0.083
```

```
sum(pttest<0.05)/B # fraction of rejecting H0, the power of the t-test
```

```
## [1] 0.157
```

The power of the t-test(0.123) is more than the power of the sign-test(0.061) at $\mu = 2800.04$. This is because t-test has higher performance for normal data than sign-test. Since the power is the probability of avoiding a type II error, meaning probability of rejecting the null hypothesis correctly (when the null hypothesis isn't true), It is seen that the t-test performs better than the sign test and this can be as a result of our data's distribution which is sampled from a normal distribution therefore the sign test won't perform as the t-test.

## d)

Our sample statistic here is $p = P(X < 2600)$. We can estimate this from the sample data $\hat{p}$.

```
p_hat = sum(birthweight$birthweight<2600)/n;p_hat
```

```
## [1] 0.3297872
```

Now, the confidence interval is given by (sample statistic $\mp$ margin of error), $(\hat{p}_l, \hat{p}_r) = (p - me, p + me)$. $\hat{p}_l = 0.25$

$$\begin{aligned} \hat{p}_l &= \hat{p} - me \\ \implies me &= \hat{p} - 0.25 \\ \hat{p}_r = \hat{p} + me &= 2\hat{p} - 0.25 \end{aligned}$$

```
pr = 2*p_hat - 0.25;pr
```

```
## [1] 0.4095745
```

```
me = p_hat - 0.25
```

So, the whole CI is [0.25,0.41].

Then, calculate z alpha/2 quantile

```
x = sqrt((p_hat *(1-p_hat))/n)
z_alpha = me / x
z_alpha
```

```
## [1] 2.326961
```

Now, we get the $\alpha$ and $1 - \alpha$. Our CI-level comes as 0.98.

```
alpha = (1- pnorm(z_alpha))*2
1-alpha
```

```
## [1] 0.9800326
```

## e)

The expert reports that there were 34 male and 28 female babies among 62 who weighted less than 2600 gram, and 61 male and 65 female babies among the remaining 126 babies.The expert claims that the mean weight is different for male and female babies. To check the above claim we have to check if the proportion of male(/female) babies with less than 2600g weight $p_A$ is significantly different from the proportion of male(/female) babies from remaining babies$p_B$. This can be done with a approximate-proportion test.

Our null hypothesis is $H_0 : p_A = p_B$.

```
prop.test(c(34,61),c(62,126))
```

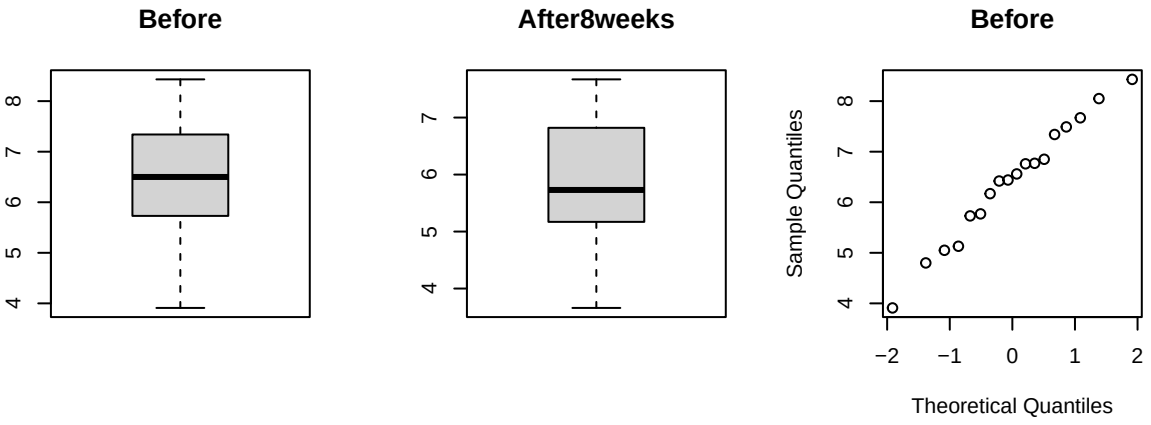
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(34, 61) out of c(62, 126)
## X-squared = 0.45343, df = 1, p-value = 0.5007
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.09929476  0.22781499
## sample estimates:
##    prop 1    prop 2
## 0.5483871 0.4841270
```

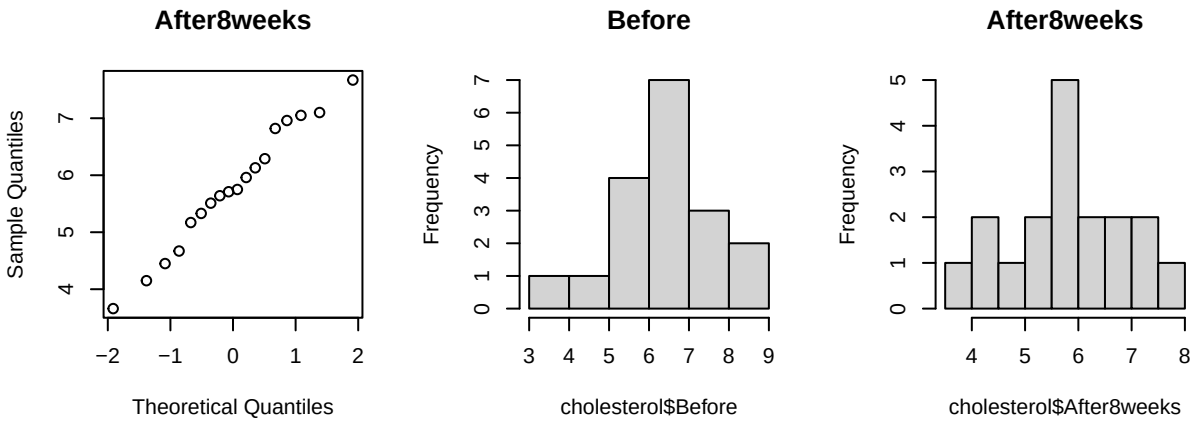
With a p-value of 0.5007, we cannot reject $H_0$. So, the expert's claim did not pass the proportion test.

## Exercise 2. Cholesterol

```
cholesterol=read.table(file="cholesterol.txt",header=TRUE)
```

## a)






Checking the histograms and QQ-plots for the Before and After8weeks columns, we find that the data seems to satisfy normality. We do a Shapiro-Wilk test to confirm. We also view the box-plots for the data to check for any outliers or extremes. We also find no missing data. We don't find any inconsistencies in the data.

```
shapiro.test(cholesterol$Before)[[2]]
```

```
## [1] 0.9674667
```

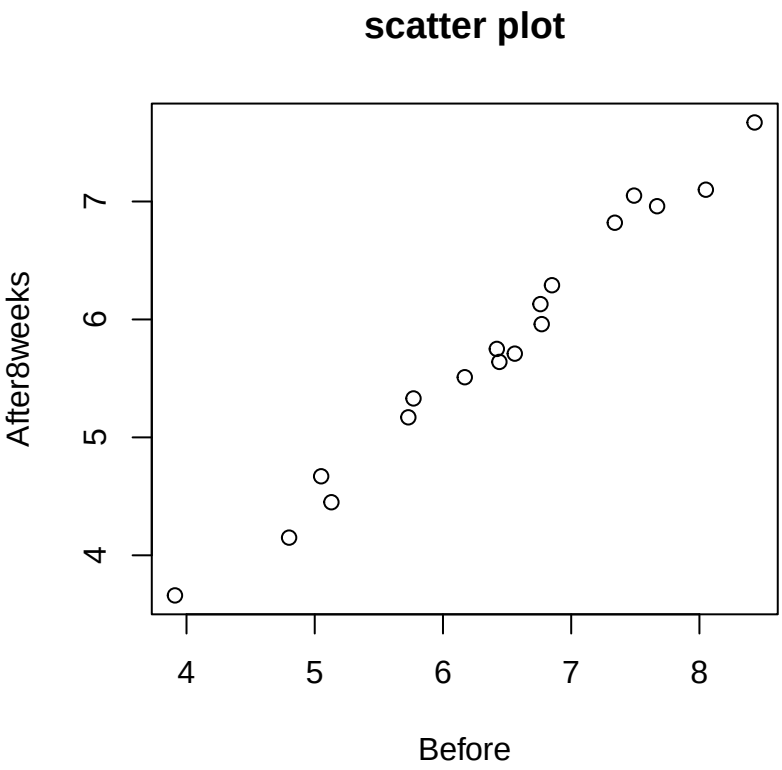
```
shapiro.test(cholesterol$After8weeks)[[2]]
```

```
## [1] 0.9183031
```

```
which(is.na(cholesterol)) #checking for missing data
```

```
## integer(0)
```

**Correlation between Before and After8weeks:**




First,just looking at the scatter plot, the graph suggests a linear relation between the 2 columns. Since both

the columns follow a normal distribution, we perform a Pearson's correlation test on the data to confirm this. The result shows a significant correlation between *Before* and *After8weeks* as shown below.

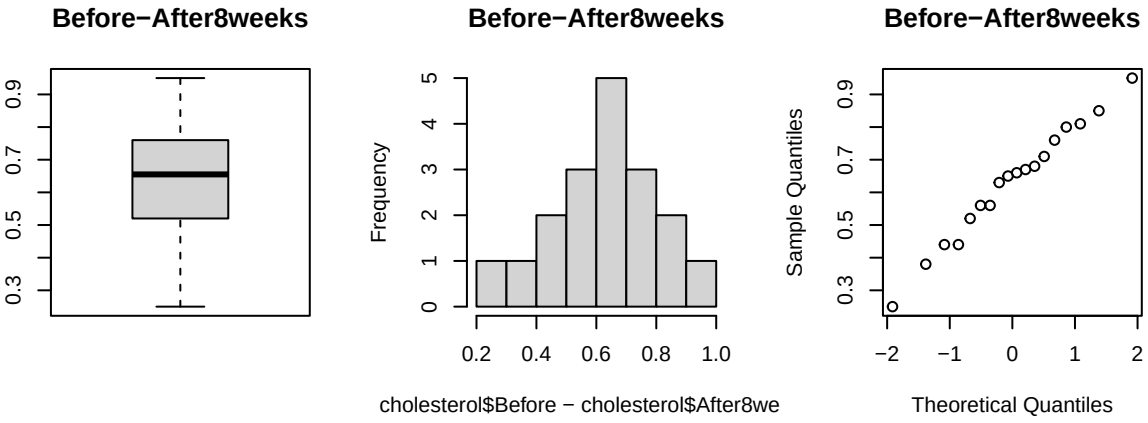
```
#Pearson's Correlation Test
cor.test(cholesterol$Before,cholesterol$After8weeks)
```

```
##
##  Pearson's product-moment correlation
##
## data:  cholesterol$Before and cholesterol$After8weeks
## t = 29.428, df = 16, p-value = 2.321e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9751289 0.9966788
## sample estimates:
##       cor
## 0.9908885
```

## b)

To verify the effect of the diet on the cholesterol, we will check for possible differences in mean between the two columns. The Before and After8weeks are not independent (correlated from (a)) and they are two numerical outcomes of the same experimental unit (cholesterol). So, these two can be treated as two paired samples.

### Check the normality of difference




### Paired t-test

The difference satisfies normality. So, we can go ahead with a paired t-test. Our null hypothesis will be $H_0 : \mu = 0$ where $\mu$ is the mean difference.

```
#paired t-test
t.test(cholesterol$Before,cholesterol$After8weeks,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  cholesterol$Before and cholesterol$After8weeks
## t = 14.946, df = 17, p-value = 3.279e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
```

```
##  0.5401131 0.7176646
## sample estimates:
## mean difference
##       0.6288889
```

From the above analysis,$H_0$ is rejected. So, the difference of mean between the *Before* and *After8weeks* is significant. The mean of the differences of the two columns is different from 0. Hence, t-test suggests the diet does have an effect on the cholesterol level.

**Sign test**

```
x = sum(cholesterol$Before<cholesterol$After8weeks)
binom.test(x,18,p=0.5)
```

```
##
##  Exact binomial test
##
## data:  x and 18
## number of successes = 0, number of trials = 18, p-value = 7.629e-06
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.000000 0.185302
## sample estimates:
## probability of success
##                      0
```

Conclusion: $H_0$ is rejected. The median of the differences of the two columns is different from 0. Hence, t-test suggests the diet does have an effect on the cholesterol level.

**Permutation Test**

Yes, permutation test is applicable since a permutation test is for two paired samples and can be performed with any test statistic that expresses difference between the X and Y within pairs. Here we will use the mean of differences.

```
mystat=function(x,y) {mean(x-y)}
B=1000; Tstar=numeric(B)
for (i in 1:B) {
  Xstar=t(apply(cbind(cholesterol$Before,cholesterol$After8weeks),1,sample))
  Tstar[i]=mystat(Xstar[,1],Xstar[,2])} # Calculate the difference in means for the permuted data
myt=mystat(cholesterol$Before,cholesterol$After8weeks) #observed difference in means
pl=sum(Tstar<myt)/B
pr=sum(Tstar>myt)/B
p=2*min(pl,pr);p
```

```
## [1] 0
```

P-value = 0. Conclusion: there is indeed significant difference between the two diet groups. The low fat margarine diet have significant effect.

## c)

Given the *After8weeks* column $(X_1, ...., X_{18}) \sim Unif[3, \theta]$, $\theta > 3$. From this, we know that the mean $\mu = \frac{3+\theta}{2}$. We can also find the estimated mean $\hat{\mu}$ from the sample data.

```
muhat=mean(cholesterol$After8weeks);muhat
```

```
## [1] 5.778889
```

Now, we can estimate $\theta$ as

$$\hat{\mu} \approx \frac{3+\hat{\theta}}{2}$$
$$\hat{\theta} = 2 * \hat{\mu} - 3$$

```
theta=2*muhat-3;theta
```

```
## [1] 8.557778
```

$\hat{\theta} = 8.56$

Using the central limit theorem, we can write

$Z = \frac{\sqrt{(n)}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$

The upper quantile $z_{\alpha/2}$ of $N(0,1)$ distribution is such that $P(Z \geq z_\alpha) = \alpha$ for $Z \sim N(0,1)$.

$$\begin{aligned} 1-\alpha &= P(|Z| \leq z_{\alpha/2}) \\ &= P(\frac{\sqrt{(n)}(\overline{X}-\mu)}{\sigma} \leq z_{\alpha/2}) \\ &= P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\ &= P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \frac{\theta+3}{2} \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \end{aligned}$$

But, since here $\sigma$ is unknown, we will use the estimated standard deviation $s$ from our sample data and the CI is based on a t-distribution and upper t-quantile $t_\alpha$. The confidence interval for $\theta$ is given by: $[2(\overline{X} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}) - 3, 2(\overline{X} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}) - 3]$

Now, we can calculate the 95%- CI for $\theta$.

```
s=sd(cholesterol$After8weeks)
n=18
alpha=0.05
t = qt(1-alpha/2,df=n-1) #upper t-quantile
c_l = 2*(muhat - t*s/sqrt(n)) - 3
c_r = 2*(muhat + t*s/sqrt(n)) - 3
c_l;c_r
```

```
## [1] 7.461842
```

```
## [1] 9.653714
```

We get 95%- confidence interval for $\theta$ - [7.46,9.65].

The CI interval could be improved by using larger confidence level (e.g., 99%), which will result in a wider confidence interval, but it will also increase the probability of capturing the true value. Additionally, increasing the sample size could decrease the standard error in our estimates.

## d)

```
after = cholesterol$After8weeks
# Define the test statistic function
T=function(x) max(x)
theta_vals=seq(3, 12, by = 0.1) # Set up the range of theta values in [3,12] to test
```

```
# Create a vector to store the p-values for each theta value
p_vals=rep(NA, length(theta_vals))
# Perform the bootstrap test for each theta value
for (i in seq_along(theta_vals)) {
  # Generate bootstrap samples
  boot_samples=replicate(10000, runif(18, min = 3, max = theta_vals[i]))
  # Calculate test statistics for bootstrap samples
  boot_t=apply(boot_samples, 2, T)
  # Calculate p-value
  p_vals[i]=mean(boot_t >= max(after))
}
theta_vals[p_vals<0.05]
```

```
##  [1] 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8
## [20] 4.9 5.0 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9 6.0 6.1 6.2 6.3 6.4 6.5 6.6 6.7
## [39] 6.8 6.9 7.0 7.1 7.2 7.3 7.4 7.5 7.6
```

The output above are the $\theta$ values in the range of [3,12] with step size of 0.1 that rejects $H_0$ with p-value<0.05. Hence, approximately, for $\theta > 7.6$ values, $H_0 : X_1, ...., X_{18} \sim Unif[3, \theta]$ is not rejected.

The Kolmogrov-Smirnov test can also be done in this case as the KS test is used to determine whether a sample comes from a specific distribution..

```
theta_val=seq(3, 12, by = 0.1)
p_val=rep(NA, length(theta_val))

for (i in seq_along(theta_val)) {
  unif_sample=runif(18,min=3,max=theta_val[i])
  # Calculate p-value
  p_val[i]=ks.test(after,unif_sample)[[2]]#ks test

}
theta_vals[p_vals>0.05]
```

```
##  [1]  7.7  7.8  7.9  8.0  8.1  8.2  8.3  8.4  8.5  8.6  8.7  8.8  8.9  9.0  9.1
## [16]  9.2  9.3  9.4  9.5  9.6  9.7  9.8  9.9 10.0 10.1 10.2 10.3 10.4 10.5 10.6
## [31] 10.7 10.8 10.9 11.0 11.1 11.2 11.3 11.4 11.5 11.6 11.7 11.8 11.9 12.0
```

For the above values, $H_0 : F_x = F_y$(the distributions are same), is not rejected by the Kolmogrov Smirnov test.

## e)

**Sign test to verify whether median cholesterol level after 8 weeks of low fat diet is less than 6**

Our null hypothesis is $H_0 : m \geq 6$. Now we perform the binomial test.

```
t=sum(cholesterol$After8weeks<6)
binom.test(t,n=18,p=0.5,alt='l')
```

```
##
##  Exact binomial test
##
## data:  t and 18
## number of successes = 11, number of trials = 18, p-value = 0.8811
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
```

```
##  0.0000000 0.8010467
## sample estimates:
## probability of success
##              0.6111111
```

The p-value is greater than 0.05, therefore, we fail to reject the null hypothesis. We do not have sufficient evidence to say that median cholesterol level after 8 weeks of low fat diet is less than 6.

We will also perform a test to check whether the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%.

```
t=sum(cholesterol$After8weeks<4.5);t
```

```
## [1] 3
```

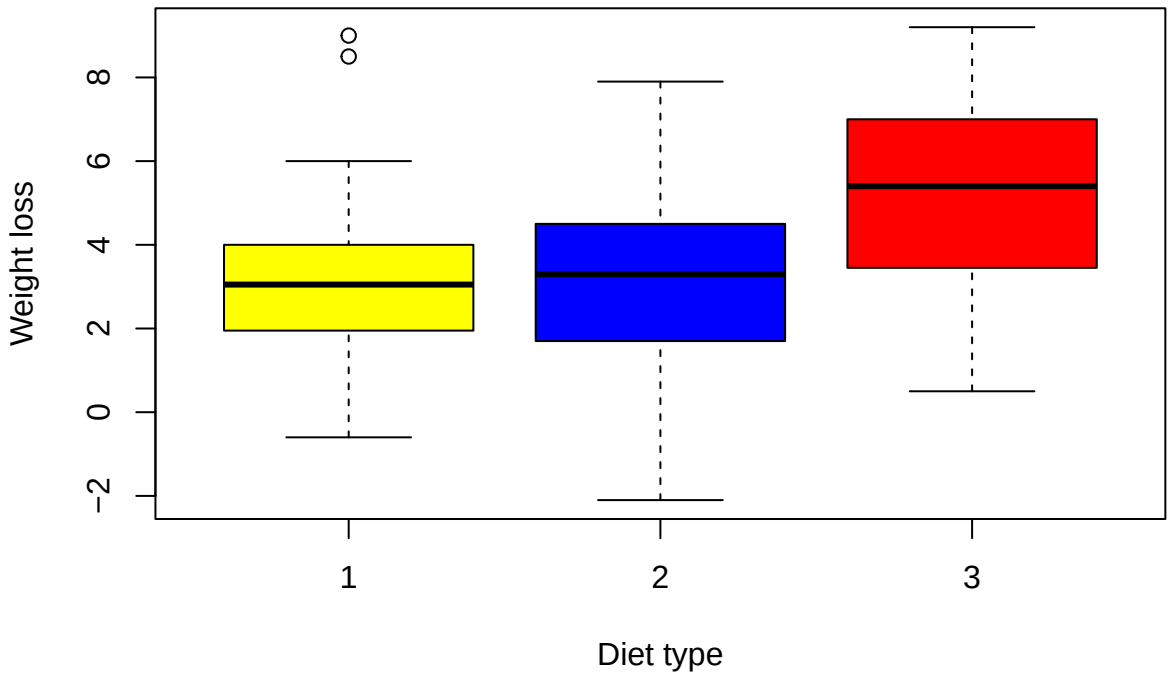
```
binom.test(t,18,p=0.25,alt='l')
```

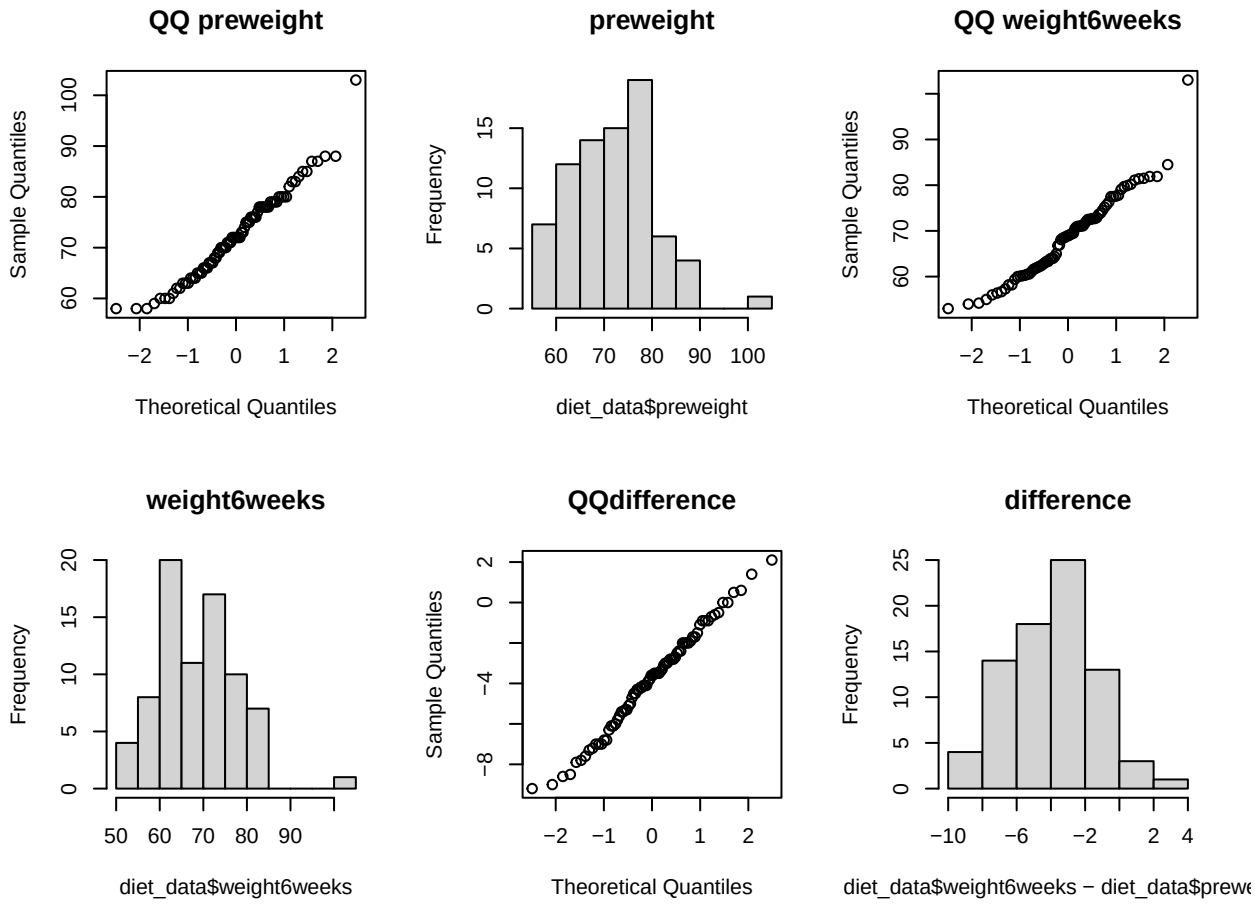
```
##
##  Exact binomial test
##
## data:  t and 18
## number of successes = 3, number of trials = 18, p-value = 0.3057
## alternative hypothesis: true probability of success is less than 0.25
## 95 percent confidence interval:
##  0.0000000 0.3766792
## sample estimates:
## probability of success
##              0.1666667
```

Again, we do not have enough evidence to say whether the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%.

### Exercise 3 Diet

**a)**






```
shapiro.test(diet_data$weight6weeks-diet_data$preweight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet_data$weight6weeks - diet_data$preweight
## W = 0.98991, p-value = 0.802
```

The individual distributions of *preweight* and *weight6weeks* does not look normal. But the difference between the two appears to be normal.

### Paired t-test to test the claim that diet effects weight loss

We can check this claim by testing if the sample outcomes of *preweight* and *weight6weeks* columns have a significant difference. We perform a two-paired sample t-test for this. $H_0 : \mu = 0$ where $\mu$ is mean of differences.

No reason to suspect that the differences as not from a normal population.

```
#Paired t-test
t.test(diet_data$preweight,diet_data$weight6weeks,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  diet_data$preweight and diet_data$weight6weeks
## t = 13.309, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
```

```
## 95 percent confidence interval:
##  3.269602 4.420141
## sample estimates:
## mean difference
##        3.844872
```

$H_0$ is rejected. The diet has a significant effect of diet on the weight loss.

```
wilcox.test(diet_data$preweight,diet_data$weight6weeks,paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  diet_data$preweight and diet_data$weight6weeks
## V = 2892.5, p-value = 1.372e-13
## alternative hypothesis: true location shift is not equal to 0
```

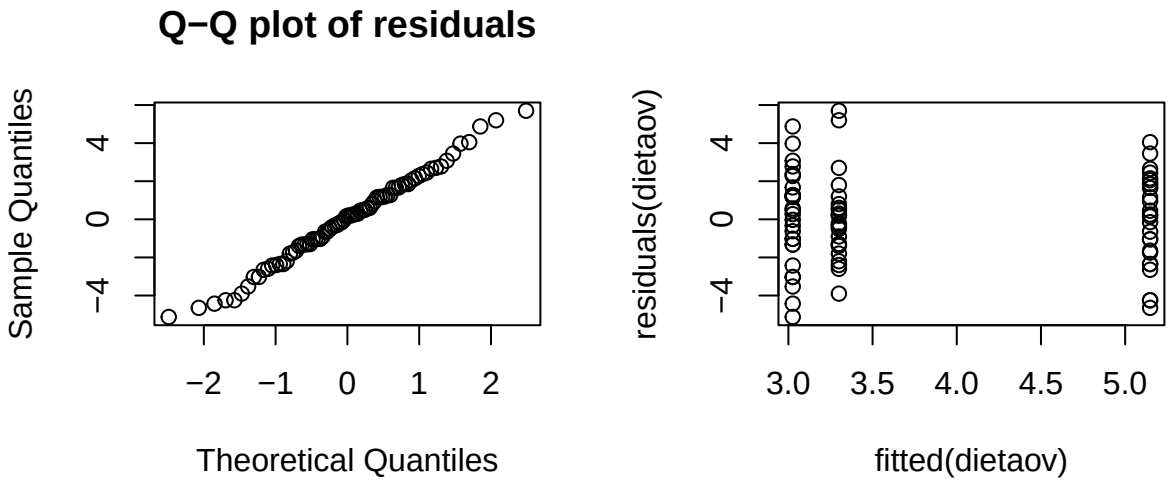
p-value = 1.372e-13. So, the diet works.

### b) one-way ANOVA to check if type of diet has an effect on lost weight

Diet is the is the factor here with 3 different categories. $H_0 : \mu_1 = \mu_2 = \mu_2$ no factor effect.

```
diet_data$diet = factor(diet_data$diet,levels=c("1","2","3"))
#is.factor(diet_data$diet); is.numeric(diet_data$diet)
dietaov=lm(weight.lost~diet,data=diet_data)
```

We have to check the assumptions of normality of errors and the homogeneity variance of residuals before applying ANOVA.




The residuals look normal. Plot of fitted values against residuals show no pattern.

```
anova(dietaov)
```

```
## 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
```

p-value = 0.003229.So $H_0$ can be rejected with significance level $\alpha = 0.01$.This suggests type of data effects weight loss.

```
summary(dietaov)
```

```
##
## Call:
## lm(formula = weight.lost ~ diet, data = diet_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1259 -1.3815  0.1759  1.6519  5.7000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.3000     0.4889   6.750 2.72e-09 ***
## diet2        -0.2741     0.6719  -0.408  0.68449
## diet3         1.8481     0.6719   2.751  0.00745 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.395 on 75 degrees of freedom
## Multiple R-squared:  0.1418, Adjusted R-squared:  0.1189
## F-statistic: 6.197 on 2 and 75 DF,  p-value: 0.003229
```

All three type of diets lead to weight loss with Diet 3 being the best. We can check this by observing the $\mu + \alpha_i > 0$ for every diet type.

Kruskal-Wallis test can also be used in this situation. It is a non-parametric alternative to one way-ANOVA and does not rely on the normality assumption as it's based on ranks.

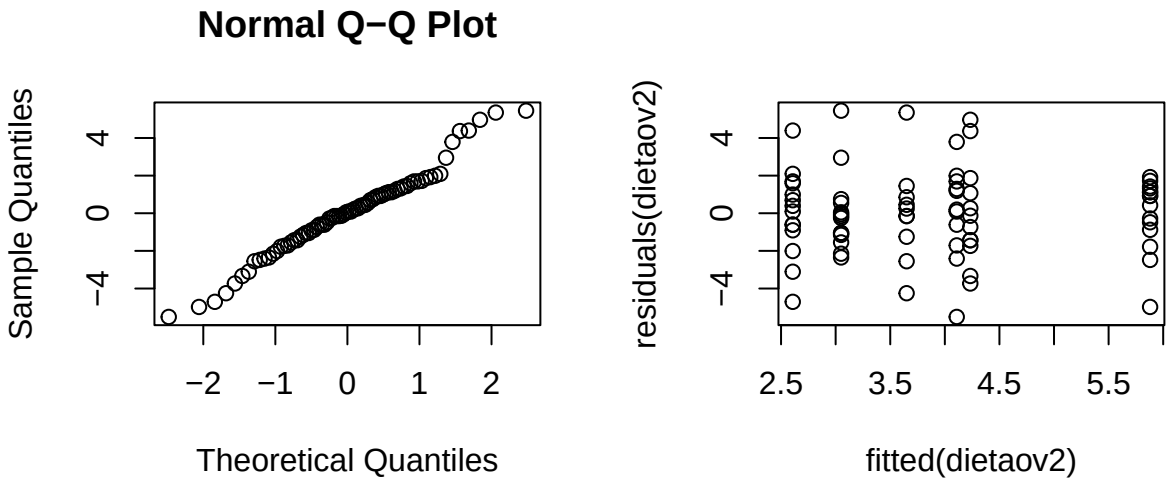
```
kruskal.test(weight.lost~diet, data=diet_data)
```

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

The p-value for testing $H_0 : F_1 = F_2 = F_3$ is 0.005416, hence $H_0$ is rejected. The Kruskal-Wallis test arrive at the same conclusion after all, but still one-way ANOVA is more powerful for this case.

## c)Two-way ANOVA

```
diet_data$gender = factor(diet_data$gender,levels=c("0","1"))
dietaov2=lm(weight.lost~diet*gender,data=diet_data)
```




The ANOVA assumptions seems to be satisfied.

```
anova(dietaov2)
```

```
## 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
```

Gender does not have any significant effect on weight.lost. Interaction between gender and diet type has a slight significance. Effect of diet type on weight lost is very significant.

## d) Dropped

## e)

The approach from c - (two-way ANOVA) is preferable where we have both gender and diet because their interaction has significant effect, it is worth to include gender in our prediction model.

The predicted weight.lost from the diet(1,2,3) and the genders(0,1) is shown below.

```
predict(dietaov2, data.frame(diet="1", gender="0"), type="response")[[1]]
```

```
## [1] 3.05
```

```
predict(dietaov2, data.frame(diet="1", gender="1"), type="response")[[1]]
```

```
## [1] 3.65
```

```
predict(dietaov2, data.frame(diet="2", gender="0"), type="response")[[1]]
```

```
## [1] 2.607143
```

```
predict(dietaov2, data.frame(diet="2", gender="1"), type="response")[[1]]
```

```
## [1] 4.109091
```

```
predict(dietaov2, data.frame(diet="3", gender="0"), type="response")[[1]]
```

```
## [1] 5.88
```

```
predict(dietaov2, data.frame(diet="3", gender="1"), type="response")[[1]]
```

```
## [1] 4.233333
```

## Exercise 4

```
library(MASS)
```

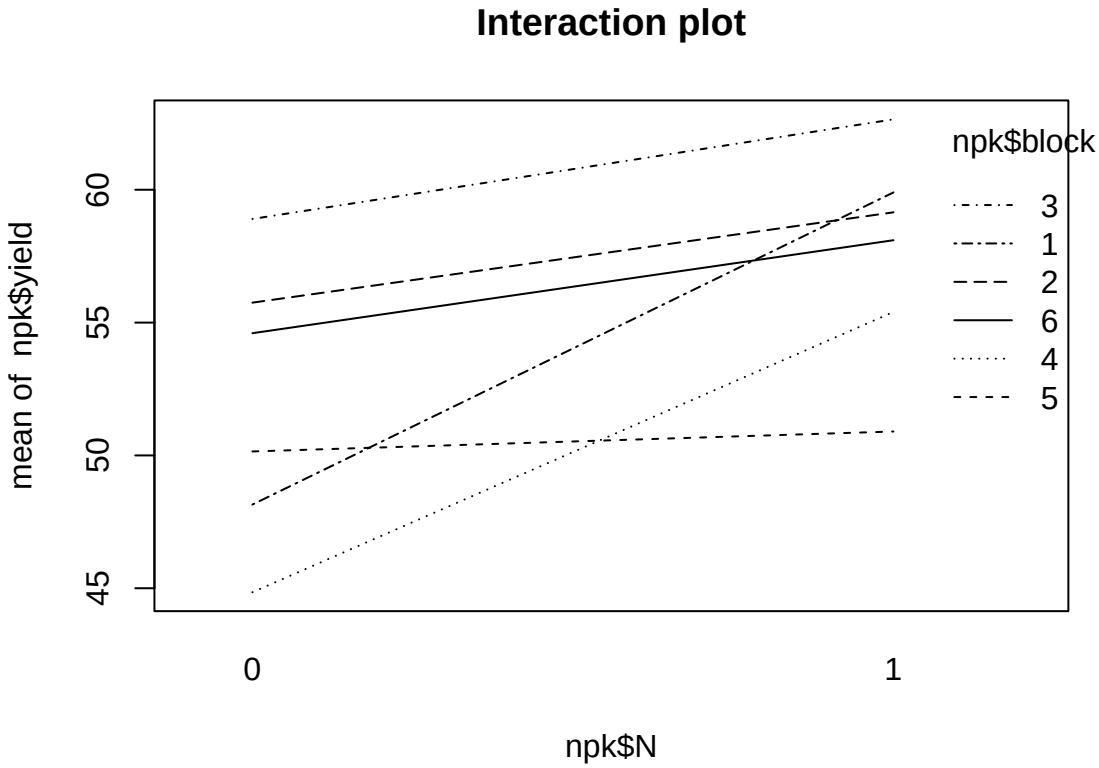
### a)

```
random_plots <- cbind(rep(1:24),rep(1:6, each = 4),
                            replicate(3, c(replicate(6, sample(c(1,1,0,0))))))

npk_rand <- data.frame(random_plots)

header <- c("plot", "block", "N", "P", "K")
colnames(npk_rand) <- header

head(npk_rand)
```

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

### b)

```
interaction.plot(npk$N, npk$block, npk$yield,main='Interaction plot')
```




The purpose of taking the factor block into account is to control for any variability in the soil or other environmental factors that could affect the yield. It allows us to better isolate the effect of the treatment(nitrogen) on the response variable(yield).

## c)

```
npk$N = factor(npk$N)
peas_mod = lm(yield~N*block,data=npk)
anova(peas_mod)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value  Pr(>F)
## N          1 189.28 189.282  9.2607 0.01021 *
## block      5 343.29  68.659  3.3592 0.03967 *
## N:block    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
```

Since, the interaction doesn't have significance, we run the additive model below.

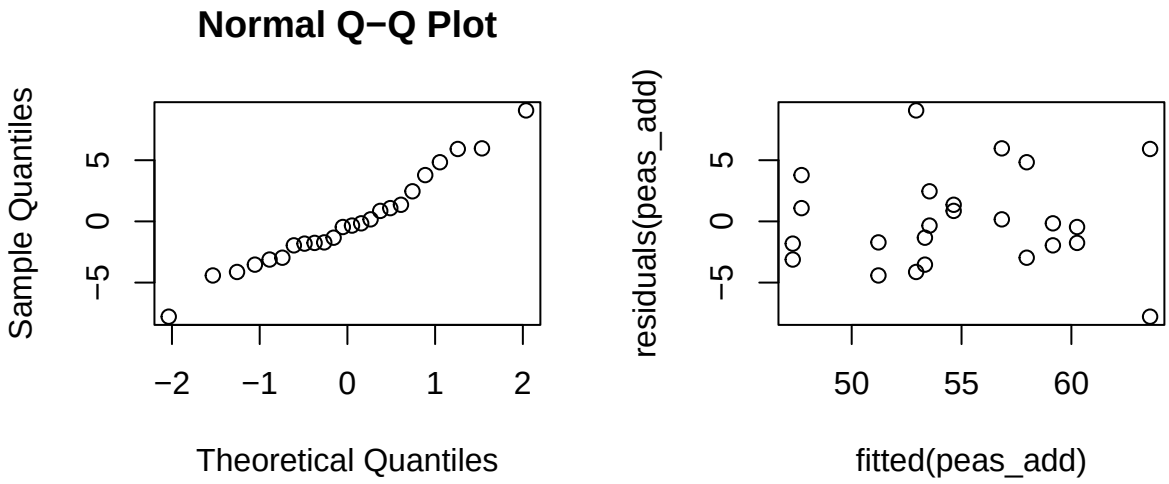
```
peas_add = lm(yield~block+N,data=npk)
anova(peas_add)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## block      5 343.29  68.659  3.3951 0.026173 *
```

```
## N          1 189.28 189.282  9.3598 0.007095 **
## Residuals 17 343.79  20.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It was sensible to include *block* as a factor in the two-way ANOVA model since block factor has significant effect on the model. From the summary of the model we can identify the third block having a significant effect.

```
##checking the model assumptions
par(mfrow=c(1,2))
qqnorm(residuals(peas_add))
plot(fitted(peas_add),residuals(peas_add))
```




The normality of the above qq plot is slightly doubtful. Some data-points seem extreme. It would be a good idea to perform an extra test suitable for non-normal data.

The Friedman test requires a complete block design, where each experimental unit appears only once in a single block. This dataset is a balanced incomplete block design, with each experimental unit (plot) receiving a combination of factors. Therefore, the Friedman test function does not apply to this dataset.

## d)

```
npk$P = as.factor(npk$P)
npk$K = as.factor(npk$K)
```

To investigate models with all factors combined, while restricting to (one-pairwise) interaction we carried out the following 3 models, model_1,model_2,model_3.

model_1 shows no significance in the interaction between block and N. P doesn't show any significant effect on the yield, whereas N and K does.

```
model_1 = lm(yield~P+K+N*block,data=npk)
anova(model_1)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## P          1   8.40   8.402  0.5931 0.459045
## K          1  95.20  95.202  6.7201 0.026843 *
```

```
## N          1 189.28 189.282 13.3611 0.004423 **
## block      5 343.29  68.659  4.8465 0.016439 *
## N:block    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
```

model_2 shows no significance in the interaction between block and P. P doesn't show any significant effect on the yield, whereas N and K does.

```
model_2 = lm(yield~N+K+P*block,data=npk)
anova(model_2)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## N          1 189.28 189.282 11.2143 0.007381 **
## K          1  95.20  95.202  5.6404 0.038947 *
## P          1   8.40   8.402  0.4978 0.496588
## block      5 343.30  68.659  4.0678 0.028234 *
## P:block    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
```

model_3 shows no significance in the interaction between block and K.

```
model_3 = lm(yield~N+P+K*block,data=npk)
anova(model_3)
```

```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## N          1 189.28 189.282 11.1397 0.007521 **
## P          1   8.40   8.402  0.4945 0.497989
## K          1  95.20  95.202  5.6028 0.039477 *
## block      5 343.29  68.659  4.0407 0.028799 *
## K:block    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
```

Since the pairwise interactions are not significant, in order to check the exact p-values for the factors, we run the additive two-way ANOVA model, model_4. The K, N has significant effect, and P doesn't.

```
model_4 = lm(yield~N+P+K+block,data=npk)
summary(model_4)
```

```
##
## Call:
## lm(formula = yield ~ N + P + K + block, data = npk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0000 -1.7083 -0.0833  2.2458  6.4833
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   53.800      2.450  21.955 8.13e-13 ***
## N1             5.617      1.634   3.438  0.00366 **
## P1            -1.183      1.634  -0.724  0.47999
## K1            -3.983      1.634  -2.438  0.02767 *
## block2         3.425      2.830   1.210  0.24483
## block3         6.750      2.830   2.386  0.03068 *
## block4        -3.900      2.830  -1.378  0.18831
## block5        -3.500      2.830  -1.237  0.23512
## block6         2.325      2.830   0.822  0.42412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.002 on 15 degrees of freedom
## Multiple R-squared:  0.7259, Adjusted R-squared:  0.5798
## F-statistic: 4.966 on 8 and 15 DF,  p-value: 0.003761
```

model_4 is preferred since one (pair- wise) interaction term of factors N, P and K with block has no effect for all the other models. It should be noted that treatment P has no significant effect on the yields of peas.

## e)

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
mixed_mod = lmer(yield~N+(1|block), REML=FALSE, data=npk)
#summary(mixed_mod)

mixed_mod_1 = lmer(yield~(1|block), REML=FALSE, data=npk)
#summary(mixed_mod_1)

anova(mixed_mod_1, mixed_mod)
```

```
## Data: npk
## Models:
## mixed_mod_1: yield ~ (1 | block)
## mixed_mod: yield ~ N + (1 | block)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## mixed_mod_1    3 159.38 162.91 -76.690   153.38
## mixed_mod      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
```

To perform mixed effects analysis the code above performed an ANOVA test between the random effect model with and without treatment in it. The p-value for treatment is lower with this model than in c), which is the fixed effect model. The p-values for both cases are less than 0.05, indicating the significance of nitrogen treatment. Both models arrive at the same conclusion, but the second model is preferred since it has the lowest p-value.