Experimental Design and Data Analysis

Assignment 1

Tommaso Castiglione Ferrari 2673807

Daniyal Selani 2692551

Simone Korteling 2671463

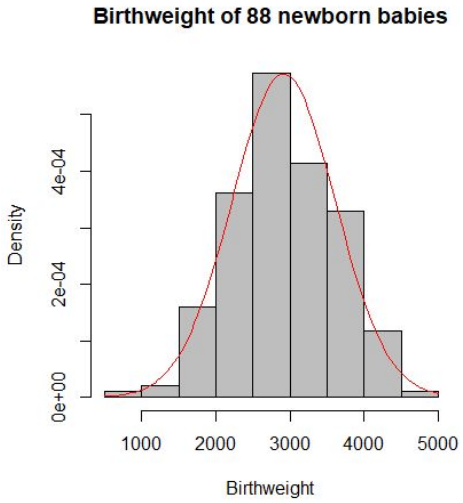
Group 71

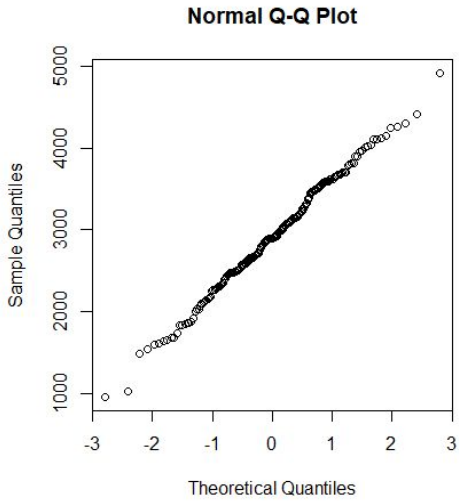
Exercise 1. Birthweight

a) Check normality of the data. Compute a point estimate for μ. Derive, assuming normality (irrespective of your conclusion about normality of the data), a bounded 90% confidence interval for μ.

Normality

First, to check for normality we made an histogram and a QQplot of the data.





Graph 1.1. Histogram of 88 newborn babies; *Graph 1.2 QQplot of 88 newborn babies*

In both the histogram and the QQplot all the points fall approximately along the reference line, therefore we assume normality. Based on the observed data of 188 newborn babies, we computed a point estimate of $\hat{\mu} = 2913{,}293$.

Based only on the observed data, we computed a bounced 90% confidence interval(CI) that contains the true value of the parameter with probability at least $1 - 0.10$. Because the standard deviation is unknown we estimated it by s and the CI is based on a t-distribution.

To compute the boundaries of the CI, we used 1) the estimated μ , 2) the estimated standard deviation and 3) the error. We then computed the lower_bound and upper_bound by subtracting and adding the error from the center of the distribution.

Code:

```
stddev= sd(data, na.rm=TRUE)
n = 188
qt(.95, n-1)

me <- qt(.95, 187)* stddev/sqrt(n-1)

lower_bound = mean(data) - me
upper_bound = mean(data) + me
```

Resulting in a 90% confidence interval of: 2829> $\hat{\mu}$ < 2998

b) An expert claims that the mean birthweight is bigger than 2800, verify this claim by using a t-test. What is the outcome of the test if you take $\alpha = 0.1$? And other values of α?

To test whether the mean birthweight is bigger than 2800 we are testing the following hypothesis: H_0: $\hat{\mu}$ equal/smaller than 2800 vs. H_1: $\hat{\mu}$ bigger than 2800.

Output:

```
One Sample t-test

data:  data
t = 2.2271, df = 187, p-value = 0.01357
alternative hypothesis: true mean is greater than 2800
90 percent confidence interval:
 2847.868      Inf
sample estimates:
mean of x
 2913.293
```

Looking at the results of our t-test with alpha level of $\alpha = 0.1$, our p-value= 0.014 is significant. So we reject our H_0 hypothese, and conclude that the mean birthweights of the population is greater than 2800. Performing a t-test with both an alpha level of $\alpha = 0.05$ and $\alpha = 0.01$ results again in p-value = 0.014. So we reject our H_0 hypothesis, and conclude again that the mean μ of birthweights is greater than 2800.

c) In the R-output of the test from b), also a confidence interval is given, but why is it different from the confidence interval found in a) and why is it one-sided?

The confidence interval is different because in question b the estimated mean population was set to 2800, where in question a we used a mean of μ = 2913. The confidence interval from question b is one-sided because we tested whether the population mean was *greater than* 2800. In this case the alternate hypothesis gives the alternate in only one direction (greater than) of the value of the mean specified in the

null hypothesis. The critical region is only right sided (i.e. a right tailed test) and therefore the confidence interval only contains a *lower* boundary.

Exercise 2. Power function of the t-test

Using the follow code for computing the power of the t-test, we can get the results shown below:

```
get_power=function(n,m,mu,sd,B, sequence){
  power = numeric(length(sequence))

  for (index in 1:length(sequence)){
    nu_val <- sequence[index]
    p <- numeric(B)

    for (b in 1:B) {
      x<-rnorm(n,mu,sd);
      y<-rnorm(m,nu_val,sd)
      p[b] <- t.test(x,y,var.equal=TRUE)[[3]]
    }

    power[index] <- mean(p.value(n, m, mu, nu_val, sd, B)<0.05)
  }
  return(power)
}
```

a) Set n=m=30, mu=180 and sd=5. Calculate now the power of the t-test for every value of nu in the grid seq(175,185,by=0.25). Plot the power as a function of nu.

Graph 2.1

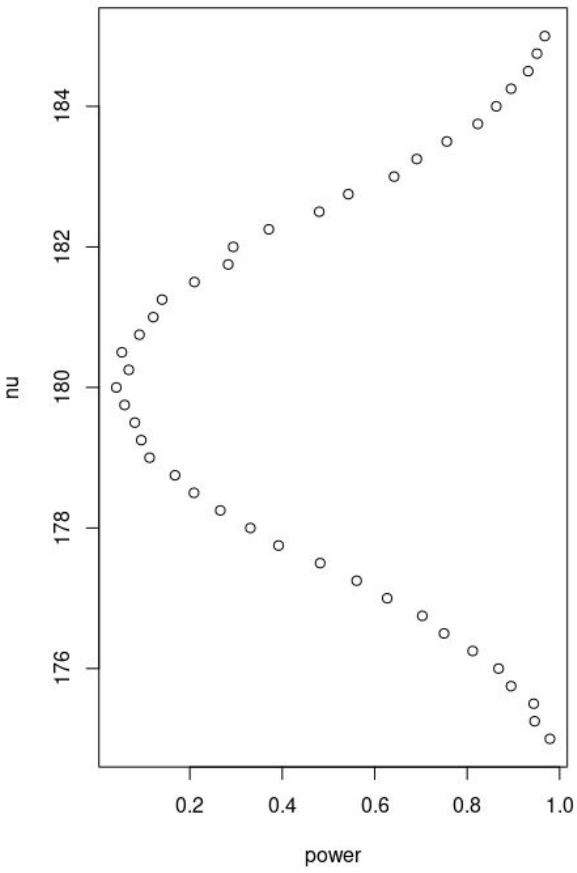
b) Set n=m=100, mu=180 and sd=5. Repeat the preceding exercise. Add the plot to the preceding plot.

Graph 2.2

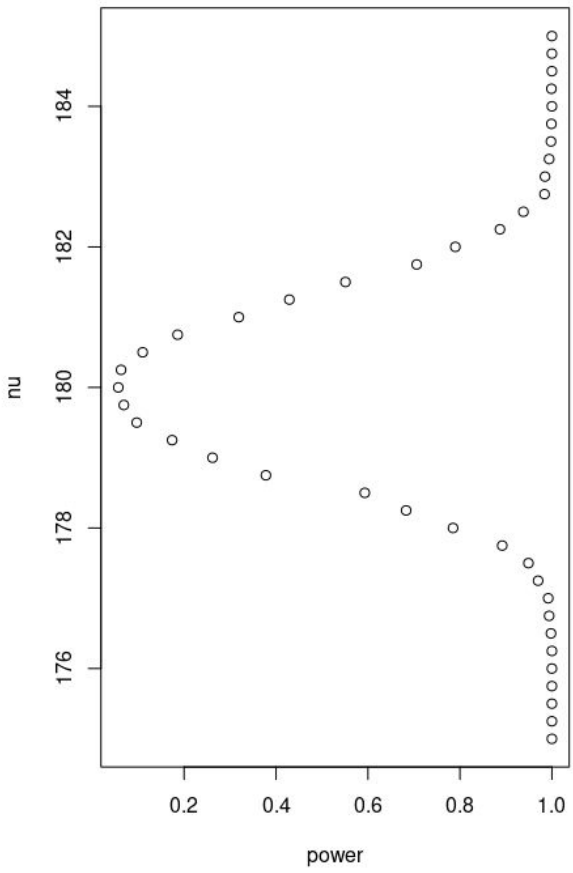
c) Set n=m=30, mu=180 and sd=15. Repeat the preceding exercise.

Graph 2.3

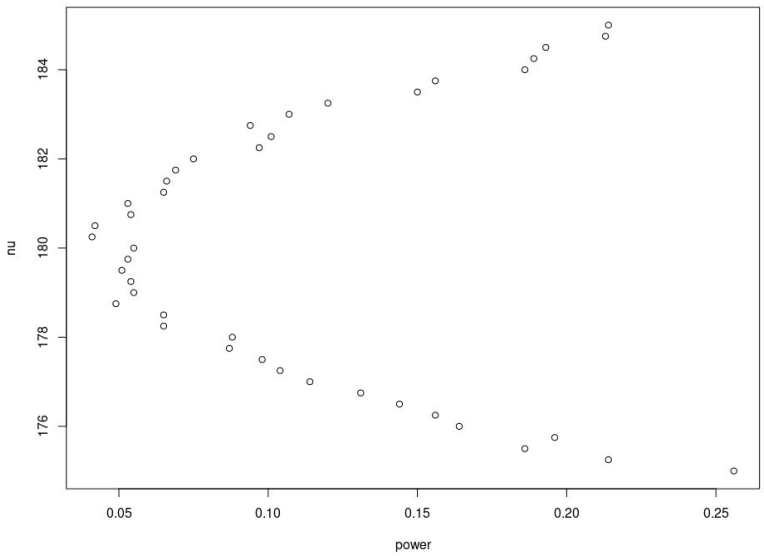
Graph 2.1



Graph 2.2



Graph 2.3



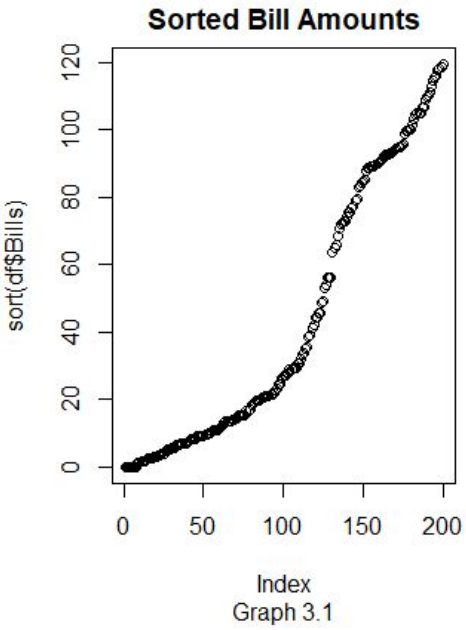
d) Explain your findings.

In the results presented above we can observe two distinct trends in action. The first one correlates the number of samples considered with a more tight and centralized curve - e.g. graph 2.2 -. The second observation considers the different standard deviation considered. In fact, as shown by the graphs, to a smaller standard deviation corresponds a more precise and well defined approximation curve, while when the standard deviation is bigger - e.g. in graph 2.3 - the points are more scattered across when forming the curve.

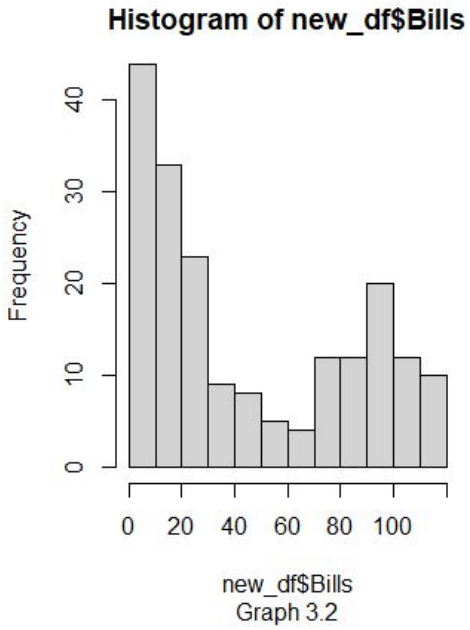
Exercise 3. Telephone Bills

a) Make an appropriate plot of this data set. What marketing advice(s) would you give to the marketing manager? Are there any inconsistencies in the data? If so, try to fix these.

Graph 3.1 shows that there are a few data points in the data that show the monthly fee as 0. As this data represents the first month bills of the new subscribers, we can assume that all of the new subscribers have used their subscription and their bill should not be 0, hence these data points can be removed.



Graph 3.1



Graph 3.2

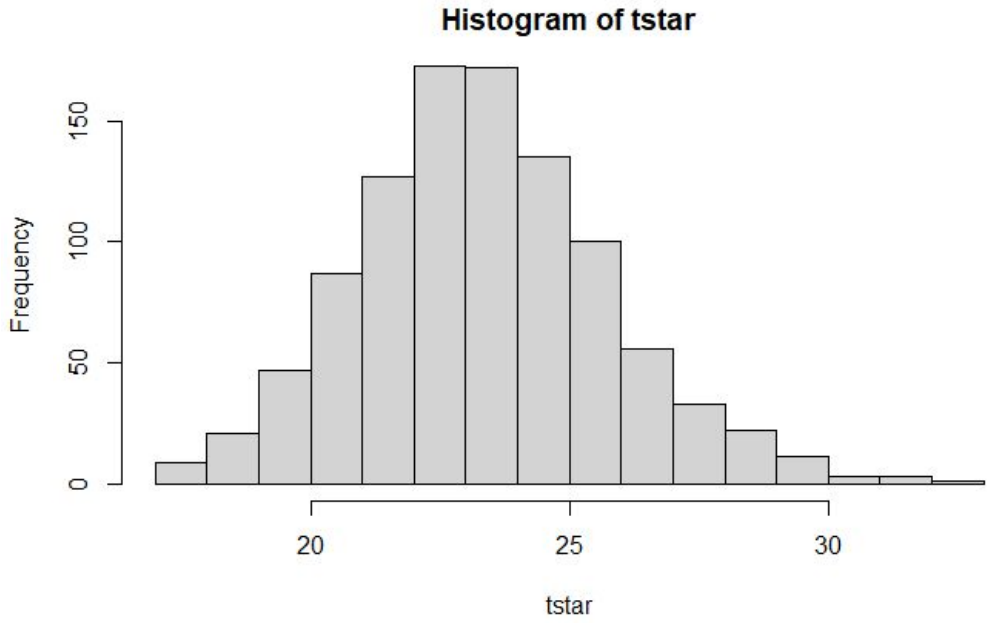
Graph 3.2 shows that most customers rarely use their telephone connections, and some customers use their telephones a lot. However, there are even fewer customers that are somewhere in the middle of the 2 extremes.

b) By using a bootstrap test with the test statistic T = median(X1, . . . , X200), test whether the data telephone.txt stems from the exponential distribution Exp(λ) with some λ from [0.01, 0.1].

We define the null hypothesis H_0 as: the data from telephone.txt stems from the exponential distribution Exp(λ) with some λ.

For the bootstrap test we take the λ parameter for the exponential distribution as 0.03. We set B = 1000 to run our test for the t test statistic t-median.
The median of the original data (t) is 28.905

The resulting distribution from out bootstrap test as outlined can be seen in *graph 3.3*:



We calculate the p-value (using the formula in lecture 2). This results in a p-value of: 0.044. Hence we reject H_0.

Code:

```
B<- 1000
t <- median(new_df$Bills)
tstar <- numeric(B)
n = length(new_df$Bills)
l = 0.03
for (i in 1:B){
  xstar <- rexp(n,l)
  tstar[i] <- median(xstar)}
pr <- sum(tstar>t)/B
pl <- sum(tstar<t)/B
p <- 2*min(pl,pr)
```

c) Construct a 95% bootstrap confidence interval for the population median of the sample.

Constructing a 95% bootstrap confidence interval can be done by sampling the data B number of times, with replacement, for the chosen t statistic, in this case median, and calculating the value using the formula provided in lecture 2. For a 95% confidence interval, with alpha = 0.05

The resulting values are:
[16.505 36.575]

Code:
```
B <- 1000
tstar <- numeric(B)
t<-median(new_df$Bills)
for (i in 1:B){
  xstar <- sample(new_df$Bills, replace=TRUE)
  tstar[i] <-median(xstar)}
tstar25<-quantile(tstar,0.025)
tstar975<-quantile(tstar, 0.975)
ci = c(2*t-tstar975,2*t-tstar25)
```

d) Assuming X1, . . . Xn ~ Exp(λ) and using the central limit theorem for the sample mean, estimate λ and construct again a 95% confidence interval for the population median. Comment on your findings.

Assuming the data is from an exponential distribution, lambda = 1/sample mean
```
lbd <- 1/mean(new_df$Bills)
```

Lambda = 0.02202461

Median = 28.905

```
t <- qt(1-((0.025/2)*(sd(new_df$Bills)/sqrt(length(new_df$Bills)))),df=n-1)
med-t; med+t
```

```
[1] 27.08197
[1] 30.72803
```

For a 95% confidence interval, with alpha = 0.05 the population median is between 27.082 and 30.728. The confidence interval is small, which indicates that we can be confident about our findings.

e) Using an appropriate test, test the null hypothesis that the median bill is bigger or equal to 40 euro against the alternative that the median bill is smaller than 40 euro. Next, design and perform a test to check whether the fraction of the bills less than 10 euro is less than 25%.

The distribution in the data cannot be assumed to be symmetric, hence we will use the binomial test to test the null hypothesis (H_0) that the median is equal to 40. We set p=0.5.

```
> binom.test(40, length(new_df$Bills), 0.5)
```

Output:

```
        Exact binomial test

data:  40 and length(new_df$Bills)
number of successes = 40, number of trials = 192, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1532039 0.2726995
sample estimates:
probability of success
          0.2083333
```

We get a p-value < 2.2e-16, and hence we can reject H_0.

To test the null hypothesis (H_0) that the fraction of the bills are less than $10 is 25%, we can perform the same test but set p=0.25. We do this because if the fraction of bills less than $10 is 25%, then the probability of a bill being less than $10 is 0.25.

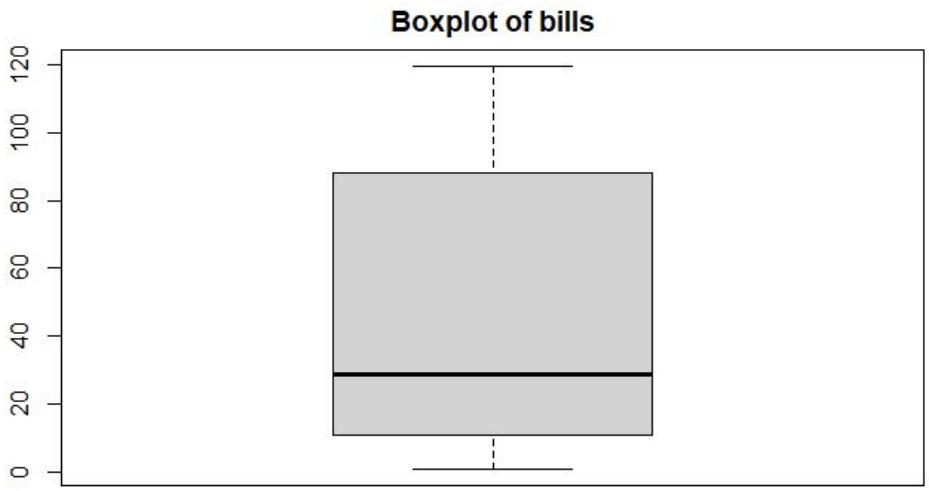
```
>binom.test(10, length(new_df$Bills), 0.25)
```

Output:

```
        Exact binomial test

data:  10 and length(new_df$Bills)
number of successes = 10, number of trials = 192, p-value = 5.661e-13
alternative hypothesis: true probability of success is not equal to 0.25
95 percent confidence interval:
 0.02525533 0.09369536
sample estimates:
probability of success
          0.05208333
```

We get a significant p-value = 5.661e-13, hence we reject H_0 that the fraction of the bills are less than $10 is 25%.

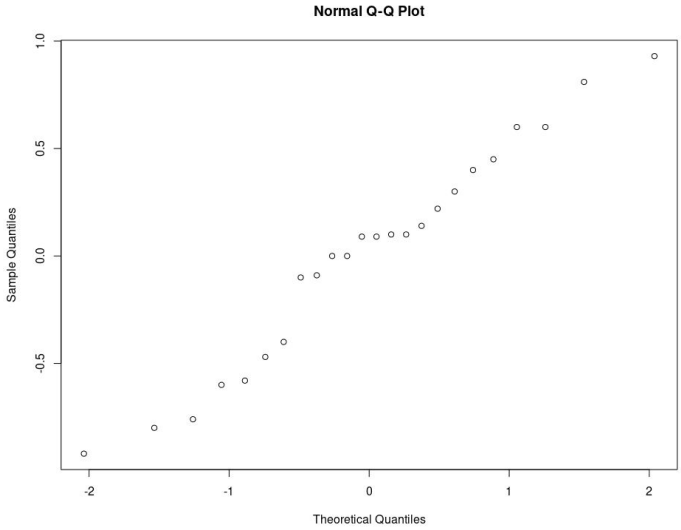
These results can be further corroborated with a box plot:



Exercise 4. Energy drink

a) Disregarding the type of drink, test whether the run times before drink and after are correlated.

Asserting the normality of the difference between the "before" and "after" performances with *qqnorm(data[,1]-data[,2])*



We can use the paired t-test to determine whether the run times before and after the drinks are correlated:

t.test(data[,1],data[,2],paired=TRUE)

This returns a p-value = 0.96, so we can determine that the run times before and after the drink are correlated.

b) Test separately, for both the softdrink and the energy drink conditions, whether there is a difference in speed in the two running tasks.

Using Pearson's product-moment correlation, we can determine if there is a difference in speed in the two running tasks. Considering the two drinks to be

lemo <- data[1:12, 1:2]
energy <- data[13:24, 1:2]

We have for the softdrink

```
cor.test(lemo[, 1], lemo[, 2])
```

That returns a p-value of 0.018, which suggests that there is a difference between before and after the two runs.

Considering the energy drink, we have:

```
cor.test(energy[,1], energy[,2])
```

That gives us a p-value of 0.009. Once again, this p-value allows us to determine that there is a difference between before and after the two runs, considering the energy drink consumption.

c) For each pupil compute the time difference between the two running tasks. Test whether these time differences are effected by the type of drink.

For testing this, we can use a permutation test for two paired samples, by testing the null hypothesis of no difference between the distributions of the two populations - aka the two types of drinks -. So, if the p-value is greater than 0.05, we can assume that there actually is a difference in the running tasks, considering the two types of drinks.

```
lemo_diff = lemo[,1] - lemo[,2]
energy_diff = energy[,1] - energy[,2]
mystat=function(x,y) {mean(x-y)}
tstar=numeric(1000)
for (i in 1:1000)
{
  datastar=t(apply(cbind(lemo_diff,energy_diff),1,sample))
  tstar[i]=mystat(datastar[,1],datastar[,2])
}
myt=mystat(lemo_diff,energy_diff)
pl=sum(tstar<myt)/B
pr=sum(tstar>myt)/B
p=2*min(pl,pr)
```

The obtained p-value is 0.17, so we can safely assume that the type of drink actually affects the time difference between the two running tasks.

d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

For question b, a possible objection is that the second run, even if provided with energy drink or soft drink, is going to be altered by the fact that a physical challenging activity was already performed shortly before. This can alter the second run time obtained, by possibly reducing the speed. This can create a false negative, or at least a reduced effect of the energy drink/softdrink consumption on the performance of the students, due the fact that is the second run, and so having performed an already physically taxating exercise.
This does not apply to question c, because the way the test is being performed, it only considers the difference of the results of the two drinks, normalizing the input to not be strictly dependent on the second run - like point b -.

Exercise 5. Chick weights

a) Test whether the distributions of the chicken weights for meatmeal and sunflower groups are different by performing three tests: the two samples t-test (argue whether the data are paired or not), the Mann-Whitney test and the Kolmogorov-Smirnov test. Comment on your findings.

T-test

There is one numerical outcome per experimental unit (chicken weights) and there are two levels of experimental units (i.e. 'meatmeal' and 'sunflower'). The chickens are randomly assigned to either one of these levels. This means the data is *not paired* and we will use a non-paired t-test to see whether the distributions of chickens weights differ for the two different levels of feed supplements.

The two samples t-test assumes that samples come from a normal population. We test the null hypothesis $H_0 : \mu$ equals v vs. $H_1 : \mu$ not equal v

```
Output:
Welch Two Sample t-test

data:  meatmeal_sunflower$weight by meatmeal_sunflower$feed
t = -2.1564, df = 18.535, p-value = 0.04441
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -102.572435   -1.442716
sample estimates:
mean in group 4 mean in group 6
      276.9091        328.9167
```

The results of the T-test show that the effect of levels on the weight of chickens is significant with $p = 0.04$ ($p < 0.05$), hence H_0 is rejected.

Mann-Whitney test

We test the hypothesis H_0: $F = G$ that the populations are the same. So we test whether the ranks of the distributions are the same between chickens who ate sunflower and meatmeal supplements.

Output:

```
        Wilcoxon rank sum exact test

data:  meatmeal_sunflower$weight by meatmeal_sunflower$feed
W = 36, p-value = 0.06882
alternative hypothesis: true location shift is not equal to 0
```

The results of the Mann-Whitney test show that the levels are not significant with a p-value of 0.069 ($p>0.05$), hence H_0 is not rejected and we conclude that there are no significant differences between the chicken weights for meatmeal and sunflower groups.

Kolmogorov-Smirnov test

We test the null hypothesis H_0: $F = G$ that the populations are the same. We computed the highest vertical difference in summed histograms to check whether the populations come from the same distributions.

Output:

```
Two-sample Kolmogorov-Smirnov test

data:  sunflower and meatmeal
D = 0, p-value = 1
alternative hypothesis: two-sided
```

The result of the Kolmogorov-Smirnov test shows a p-value = 1 ($p>0.05$). Hence, H_0 is not rejected and we conclude that chicken weights for meatmeal and sunflower groups are not significantly different from each other.

b) Conduct a one-way ANOVA to determine whether the type of feed supplement has an effect on the weight of the chicks. Give the estimated chick weights for each of the six feed supplements. What is the best feed supplement?

First we made a linear restriction on the parameters to perform an ANOVA. We made a linear model with 'feed' as a factor with 6 different levels. We then performed an ANOVA test with 'weight' as dependent variable and 'feed' as independent and tested whether the weights of the chickens can be explained by the different food supplements.

Output:

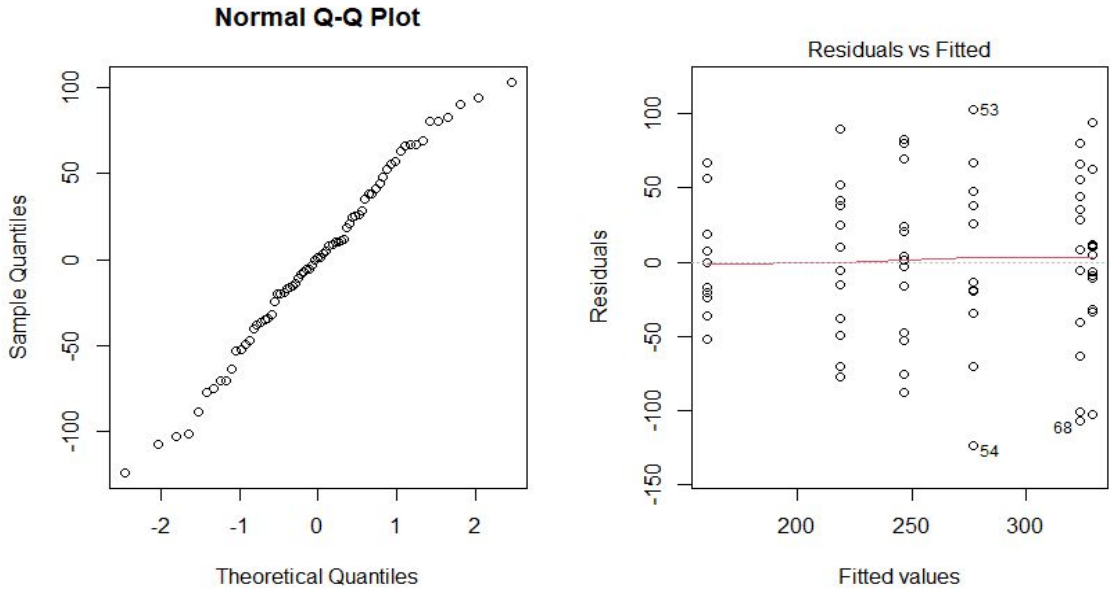
```
Analysis of Variance Table

Response: chick$weight
          Df Sum Sq Mean Sq F value    Pr(>F)
chick$feed  5 231129   46226  15.365 5.936e-10 ***
Residuals  65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ is $p<0.001$, hence H_0 is rejected, i.e. the factor 'feed' is significant.

Furthermore, with the *command fitting(chickmodel)* we obtained the estimated means of the different levels, resulting in the following means:
Casian : $\mu = 323.58$
Horsebean : $\mu = 160.2$
Linseed : $\mu = 218.75$
Meatmeal : $\mu = 276.9$
Soybean : $\mu = 246.43$
Sunflower: $\mu = 328.92$

This would mean that the chickens who ate 'sunflower' have the highest mean weight, and thus we interpret sunflower as the best food supplement.

c) Check the ANOVA model assumptions by using relevant diagnostic tools.
By computing the estimators of the residuals we can check for the assumption of ' normality of the population'. The QQplot for the residuals looks as follows (see left graph). We also checked the assumption 'homogeneity of variances' by the residuals versus fits plot (see right graph).



Graph 5.1. Shows normality QQplot; Graph 5.2. Shows relationship between residuals and the means of different levels.

In the left QQplot above, as all the points fall approximately along the reference line, we can assume normality. This conclusion is supported by the Shapiro-Wilk test with a non significant $p = 0.63$. In the right plot above, there is no evident relationship between residuals and the mean of the different levels of the factor feed, which is good. So, we can also assume the homogeneity of variances.

d) Does the Kruskal-Wallis test arrive at the same conclusion about the effect of feed supplement as the test in b)? Explain possible differences between conclusions of the Kruskal-Wallis and ANOVA tests.

We performed an Kruskal-Wallis test to check whether this test arrived with the same results as the one-way ANOVA-test. The results are as follows:

```
Kruskal-Wallis rank sum test

data:  chick$weight and chick$feed
Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

The p-value for testing $H_0 : F_1 = F_2 = F_3 = F_4 = F_5 = F_6$ is 5.113e-07, hence H_0 is rejected. Even though we arrive at the same conclusion (i.e. rejecting H_0), the p-value is different compared to the ANOVA test. The differences could be due to the fact that an ANOVA is a parametric test while kruskal test is a non-parametric approach. Thus, kruskal test does not need any distributional assumption and doesn't rely on the assumptions of normality. Instead the test is based on differences between levels in ranks.