EDDA_assign1_group33

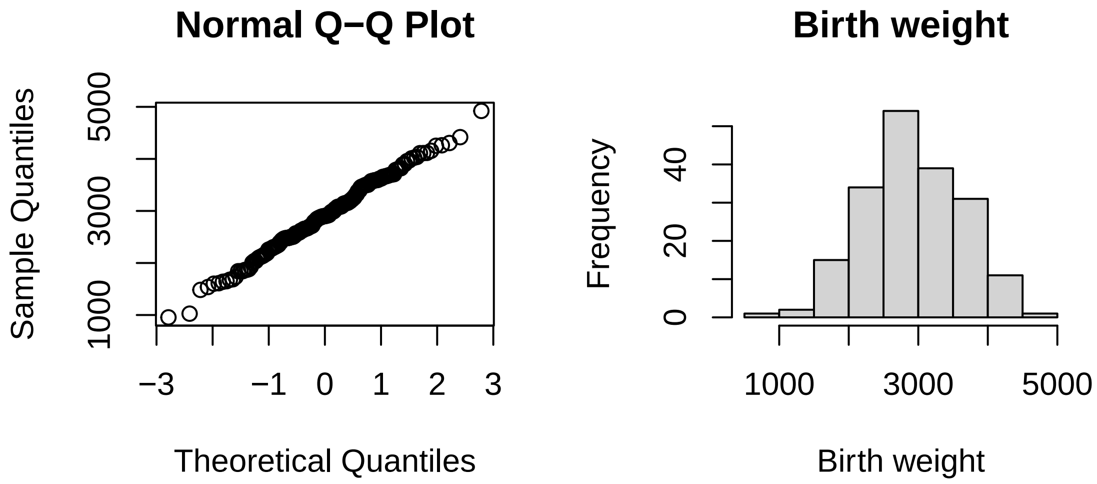
Wenbo Sun, Yuhao Qian, Meifang Li

2/19/2021

Exercise 1. Birthweights

a) We firstly draw a QQ-norm plot and a histogram for birthweight. The QQ-norm plot shows that the data lies almost on a straight line and the histogram is bell-shaped, so the birthweight should follow the normal distribution. Then we implement Shapiro-Wilk test to confirm this inference, where p-value is 0.8995 (>0.05). We use sample mean, which is 2913.293, as a point estimate for μ. Since we checked the normality of birthweight, a one-sample t-test gives us a 90% confidence interval for μ, which is $[2829.202, 2997.384]$.

```
B=read.csv('birthweight.txt');bw=B$birthweight;par(mfrow=c(1,2))
qqnorm(bw); hist(bw,xlab = 'Birth weight', main = 'Birth weight');shapiro.test(bw)
```



```
##
##  Shapiro-Wilk normality test
##
## data:  bw
## W = 1, p-value = 0.9
```

```
point_esti=mean(bw); t.test(bw,mu=point_esti,conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  bw
## t = 0, df = 187, p-value = 1
## alternative hypothesis: true mean is not equal to 2913
## 90 percent confidence interval:
##  2829 2997
## sample estimates:
## mean of x
##      2913
```

b) We perform a one-sample t-test three times for different α, namely 0.05, 0.1 and 0.01, to verify the assumption that mean birthweight is bigger than 2800. The p-value are all 0.01357 so H_1 (birthweight>2800) holds for $\alpha = 0.05, 0.1$ but rejected for $\alpha = 0.01$. In other words, the mean birthweight is bigger than 2800 at 95% and 90% confidence level but we can not say this at 99% confidence level.

```
t.test(bw,mu=2800,alternative = 'greater',conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  bw
## t = 2, df = 187, p-value = 0.01
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829  Inf
## sample estimates:
## mean of x
##      2913
```

```
t.test(bw,mu=2800,alternative = 'greater',conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  bw
## t = 2, df = 187, p-value = 0.01
## alternative hypothesis: true mean is greater than 2800
## 90 percent confidence interval:
##  2848  Inf
## sample estimates:
## mean of x
##      2913
```

```
t.test(bw,mu=2800,alternative = 'greater',conf.level = 0.99)
```

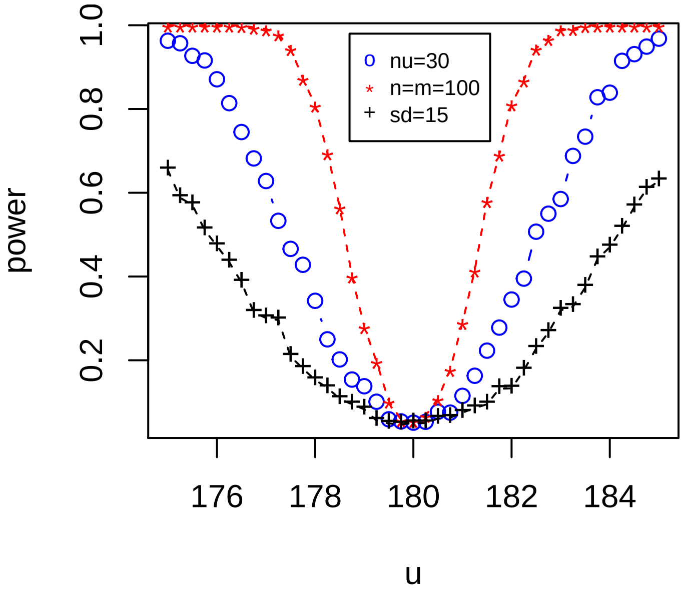
```
##
##  One Sample t-test
##
## data:  bw
## t = 2, df = 187, p-value = 0.01
## alternative hypothesis: true mean is greater than 2800
## 99 percent confidence interval:
##  2794  Inf
## sample estimates:
## mean of x
##      2913
```

c) The H_0 in a) is a two-tailed hypothesis so we construct a confidence interval with a confidence level of $1-\alpha$, while H_0 in b) is a one-tailed hypothesis and it requires a confidence interval with a confidence level of $1-2\alpha$. This is the reason why b) has a different confidence interval with a) and it is one-side.

Exercise 2. Power function of the t-test

a,b,c)

```r
p.value=function(n,m,mu,nu,sd,B=1000){
  p=numeric(B)
  for (b in 1:B) {x=rnorm(n, mu,sd); y=rnorm(m,nu,sd)
  p[b]=t.test(x,y,var.equal = TRUE)[[3]]
  }
  return(p)
}
p_set=function(n,m,mu,sd,B=1000){
  nu_set=seq(175,185,by=0.25);p_set=numeric(length(nu_set))
  for (i in 1:length(nu_set)){
    p=p.value(n,m,mu,nu_set[i],sd)
    p_set[i]= mean(p<0.05)
  }
  return(p_set)
}
nu_set=seq(175,185,by=0.25)
p1=p_set(30,30,180,5,1000) #
p2=p_set(100,100,180,5,1000) #
p3=p_set(100,100,180,15,1000) #
plot(nu_set,p1,xlab = '\nu',ylab = 'power',type = 'b', col="blue") # first plot
points(nu_set, p2, col="red", pch="*") # second plot
lines(nu_set, p2, col="red",lty=2)
points(nu_set, p3, col="black", pch="+") # third plot
lines(nu_set, p3, col="black",lty=2)
legend(178.7,0.98,legend=c("nu=30","n=m=100","sd=15"), col=c("blue","red","black"),
       pch=c("o","*","+"), ncol=1,cex=0.7)
```



d) When comparing a) and b) with different sample sizes, we find that the concave shape of a larger sample size is shaper than that of a smaller size. As for b) and c) with different sd, the concave of larger sd is flatter than small sd. Larger power means it is more likely to reject H_0 (mu=nu) as the p-value is larger. According to the test statistic:

$$T = \frac{\bar{x}_n - \bar{y}_m}{S_{x,y}\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{1}$$

when sample size n,m turns larger, Law of large numbers gives that sample mean is closer to the real mean.

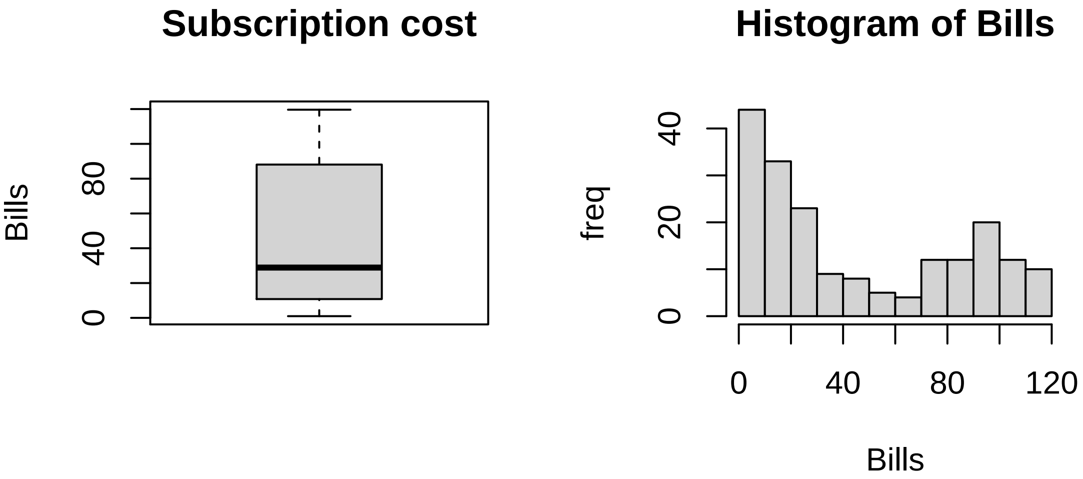
This leads to a larger T so power is stronger. Smaller sd also causes a larger T therefore power is stronger.

Exercise 3. Telecommunication company

a) Firstly, we generate boxplot for raw data and find some values are 0. The data represents the customers' bills. Possible explanations for these zeros might be that customers do not use their sim card, or system failure causes data loss. Considering that the marketing manager hopes to make marketing strategy using the data, zeros do not help in this scenario. Hence, we exclude the zeros in the subsequent analysing.

After removing the zeros, we make the histogram and boxplot for the sample.

```
tc=read.csv("telephone.txt")
data=tc$Bills[which(tc$Bills != 0.00)]
par(mfrow=c(1,2));boxplot(data,main="Subscription cost",ylab="Bills");
hist(data,breaks=c(10*0:12), xlab="Bills",ylab="freq",main=paste("Histogram of Bills"))
```



Both plots illustrate apparent skewness. Although the max bill achieves around 120 euros, 56% of bills are under 40 euros. Considering that the telecommunication company hopes to start a business in a new country, we suggest that promoting low-cost subscription to expand market share. **b)** To test whether the sample fits a particular exponential distribution, we generate 200 samples using rexp by 1000 times, then compute p-values by comparing artificial sample median and original sample median. The null hypothesis is the sample is generated by an exponential distribution, if a p-value is larger than 0.05, the corresponding λ is a reasonable estimation.

For finding λ value, we search in a linear space, range 0.01 to 0.1 stepped by 0.01. The search process is simply bootstrapping iteratively along the linear space.

```
grids=seq(0.01,0.1,0.01);t=median(data);B=1000;tstar=numeric(B)
n=length(data);m=length(grids);pvs=numeric(m)
for(i in 1:m){
  for(j in 1:B){
      xstar=rexp(n,grids[i]);tstar[j]=median(xstar)
      pl=sum(tstar<t)/B;pr=sum(tstar>t)/B; pvs[i]=2*min(pl,pr)
      }
  }
```

The results are listed in the following table.

λ	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
p-value	0.0	0.07	0.034	0.0	0.0	0.0	0.0	0.0	0.0	0.0

When λ=0.02, the p-value implies that the null hypothesis can not be rejected at 95% confidence level. Besides, except 0.02 and 0.03, other λ value leads to 0 p-value, it is reasonable to imply that true λ value lies between 0.02 and 0.03.

c) A bootstrap confidence interval of a test statistics can be simulated by repeated sampling with replacement. The formula of bootstrap CI is given by:

$$[2T - {T^*}_{(1-\alpha/2)}, 2T - {T^*}_{(\alpha/2)}]$$

where T^* is generated by repeated sampling, and alpha is the confidence level. Here we hope to construct a 95% confidence interval. The margin can be determined by 97.5% and 2.5% quantile.

```
B=1000;Tstar=numeric(B)
for(i in 1:B){
Xstar=sample(data,replace=TRUE);Tstar[i]=median(Xstar)
}
Ts25=quantile(Tstar,0.025);Ts975=quantile(Tstar,0.975)
c(2*median(data)-Ts975,2*median(data)-Ts25)
```

```
## 97.5%  2.5%
##  15.6  36.6
```

d) The central limit theorem indicates if we sample large random samples from a population with replacement, the sample means approximately follow a normal distribution, ignoring the population's distribution. In b), we assume the sample follows exp distribution and roughly estimate λ by the bootstrap median. For utilizing the central limit theorem, we can repeatedly sample from original data and compute mean values. Since sample means follow normal distribution and we know that the mean of an exponential distribution is $1/\lambda$ we can conduct T-test to examine whether the mean value equality is rejected. Rejection means the original sample distribution is different from the exponential distribution with particular λ. By this process, we can estimate λ of the exponential distribution.

In b), we estimate that the lambda value is between 0.02 and 0.03. We narrow down our search space as `seq(0.02,0.03,0.001)` Then we implement the search algorithm as follow:

```
interval=seq(0.02,0.03,0.001);B=1000;pvs=numeric(length(interval));means_a=numeric(B)
for(i in 1:length(interval)){
    for(j in 1:B){
        means_a[j]=mean(sample(data,replace=TRUE))
    }
    pvs[i]=t.test(means_a,mu=1/interval[i])[3]
}
unlist(pvs)
```

```
##  [1] 6.53e-279 6.19e-114  4.71e-02  7.87e-88 1.25e-234 1.98e-323  0.00e+00
##  [8]  0.00e+00  0.00e+00  0.00e+00  0.00e+00
```

The third p-value, λ=0.022, is larger than 0.05, indicates that T test can not reject sample means' equality. It is plausible that the original sample follows the exponential distribution with λ= 0.022.

Since we have an estimation of sample distribution, we can compute the confidence interval of mean. We sample 1000 times and generate a distribution of the exponential distribution. Because the means are normally distributed, we estimate the confidence interval of mean by T-test.

```
means=numeric(1000)
for(i in 1:1000){
means[i]=mean(rexp(length(data),0.022))
}
result=t.test(means,mu=1/0.022)
result
```

```
##
##  One Sample t-test
##
## data:  means
## t = 0.09, df = 999, p-value = 0.9
## alternative hypothesis: true mean is not equal to 45.5
## 95 percent confidence interval:
##  45.3 45.7
## sample estimates:
## mean of x
##      45.5
```

Regarding the confidence interval of the median, we already know the mean of an exponential distribution is $1/\lambda$, and the median is $ln2/\lambda$. Using this relationship, we can transfer the confidence interval of mean to median by multiplying $ln2$. Then we get the confidence interval of the median is:

```
CI=result[4]$conf.int;CI_median=c(CI[1],CI[2])*log(2)
CI_median
```

```
## [1] 31.4 31.6
```

Comparing the confidence interval with the result in c), estimating confidence interval by exponential distribution leads to a smaller range. However, the sample median is 28.9, which lies outside the confidence interval, indicating the estimation is not accurate. According to the sample data histogram, the tail is much thicker than the exponential distribution, although the t-test can not reject the hypothesis. Hence, when we seek the actual data distribution, we can not simply accept the hypothesis test's numerical result. Data observation also plays a crucial role in data exploration.

e) The claim that the median bill is smaller than 40 euros can be tested by sign test because of the skewness of bills. The null hypothesis is meidan>=40 and alternative hypothesis is median<40.

```
binom.test(sum(data>=40),length(data),p=0.5,alternative='less')
```

```
##
##  Exact binomial test
##
## data:  sum(data >= 40) and length(data)
## number of successes = 83, number of trials = 192, p-value = 0.04
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.000 0.494
## sample estimates:
## probability of success
##                  0.432
```

The p-value<0.05, null hypothesis meidan>=40 is rejected. The claim that median bill is smaller than 40 euros is accepted.

Then we consider the claim that the fraction of the bills less than 10 euro is less than 25%. When we test the median, we actually construct the alternative hypothesis as the fraction is less than 50%. Hence, set the null hypothesis as p>=0.25 and alternative hypothesis is p<0.25.

```
binom.test(sum(data<10),length(data),p=0.25,alternative='less')
```

```
##
##  Exact binomial test
##
## data:  sum(data < 10) and length(data)
## number of successes = 44, number of trials = 192, p-value = 0.3
## alternative hypothesis: true probability of success is less than 0.25
## 95 percent confidence interval:
##  0.000 0.285
## sample estimates:
## probability of success
##                  0.229
```

The p-value>0.05, which means that in a 95% confidence level, we can not reject the null hypothesis, p>=0.25. The claim that the fraction of bills of less than 10 euros is possibly larger than 25%.

Exercise 4. Energy drink

a) This experiment has two numerical outcomes per experimental unit and Spearman's rank correlation test does not assume normality, so we use Spearman's test to check whether the run times before drink and after are correlated. As can be seen in the following result, the p-value is 0.001056<0.05, so H_0 is rejected and run times before drink and after are correlated.

```
run=read.table("run.txt",header = TRUE)
cor.test(run[,1],run[,2],method = "spearman")
```

```
## Warning in cor.test.default(run[, 1], run[, 2], method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  run[, 1] and run[, 2]
## S = 1904, p-value = 0.4
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##   rho
## 0.172
```

b) The permutation test for two paired samples does not assume normality, so we use it to check whether there is a difference in the two running tasks with softdrink and energy drink respectively. The p-value of softdrink and energy drink are>0.05, so H_0 cannot be rejected and there is no significant difference between before and after with softdrink and energy drink respectively.

```
mystat=function(x,y){mean(x-y)};B=1000;tstar=numeric(B)
lemo=subset(run, drink=="lemo",select = c("before.","after","drink"))#lemo
for (i in 1:B) {
  star=t(apply(cbind(lemo[,1],lemo[,2]),1,sample))
  tstar[i]=mystat(star[,1],star[,2])
}
myt=mystat(lemo[,1],lemo[,2]);pl=sum(tstar<myt)/B;pr=sum(tstar>myt)/B
p=min(pl,pr);p
```

```
## [1] 0.195
```

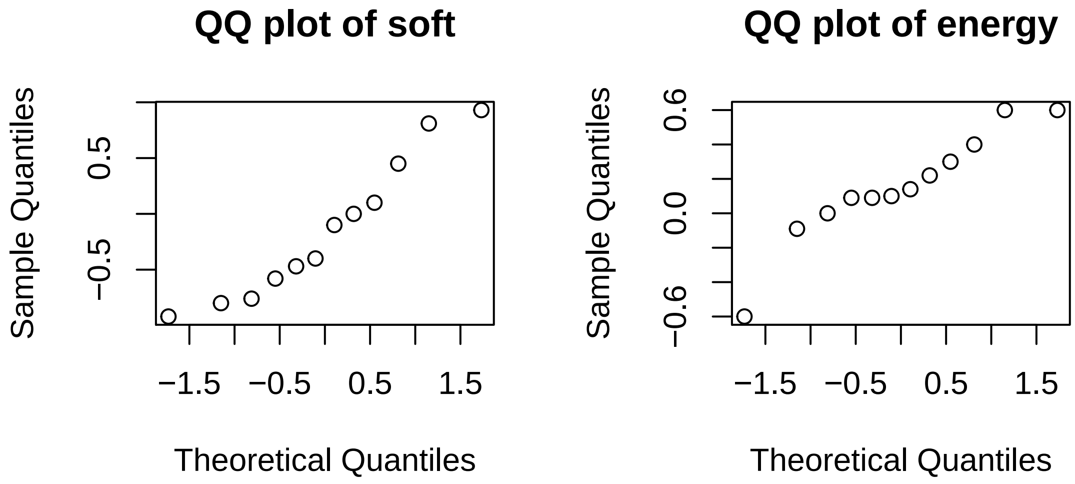
```
energy=subset(run, drink=="energy",select = c("before.","after","drink"))#energy
for (i in 1:B) {
  star=t(apply(cbind(lemo[,1],lemo[,2]),1,sample))
  tstar[i]=mystat(star[,1],star[,2])
}
myt=mystat(lemo[,1],lemo[,2]);pl=sum(tstar<myt)/B;pr=sum(tstar>myt)/B
p=min(pl,pr);p
```

```
## [1] 0.196
```

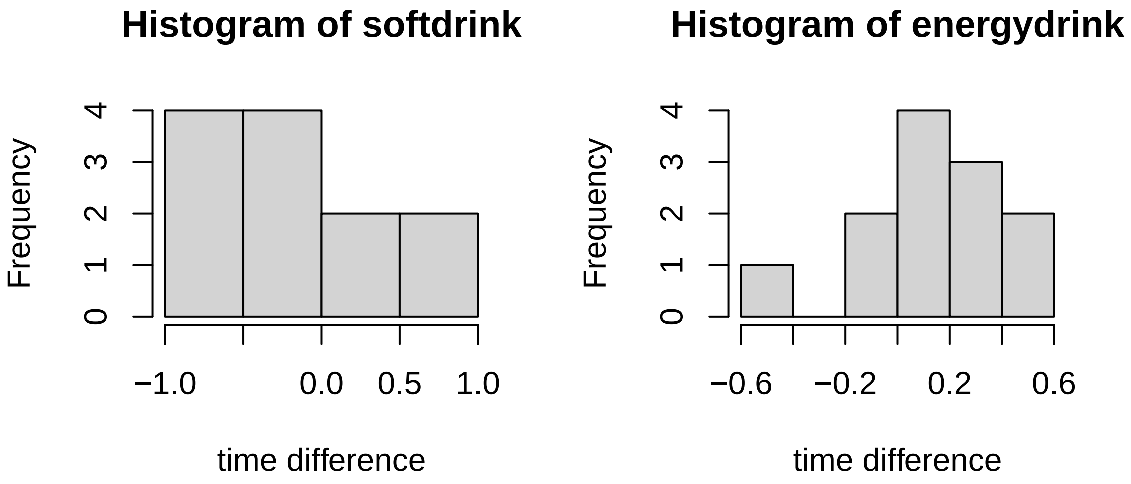
c) When obtaining the time difference between running tasks, we have one numerical outcome per experimental unit and two groups of different drinks. Thus we could use two samples t-test to check if the means of two samples are the same, so we first check whether two samples follow normal distribution. The QQ plots and histograms for softdrink and energy drink cannot tell the normality exactly when the sample number is small so we also use shapiro test to test the normality. As can be seen from the results, p-value for softdrink is 0.3725>0.05 and p-value for energy drink is 0.2788>0.05, both cannot reject H_0, so they both follow normal distribution.

Then we implement the two samples t-test, the result show p-value equals 0.1586, which is larger than 0.05, so H_0 cannot be rejected and the time differences of softdrink and energy drink have equal means. We also use the Mann-Whitney test to test whether the populations are the same. The p-value of this result, which is 0.1332>0.05, indicate that H_0 cannot be rejected and the time differences of softdrink and energy drink have equal medians. All the tests show that time differences are not effected by the type of drink.

```
par(mfrow=c(1,2))
qqnorm(lemo[,1]-lemo[,2],main="QQ plot of soft");
qqnorm(energy[,1]-energy[,2],main="QQ plot of energy")
```



```
hist(lemo[,1]-lemo[,2],main = "Histogram of softdrink",xlab="time difference");
hist(energy[,1]-energy[,2],main = "Histogram of energydrink",xlab = "time difference")
```



```
shapiro.test(lemo[,1]-lemo[,2]);shapiro.test(energy[,1]-energy[,2])
```

```
##
##   Shapiro-Wilk normality test
```

```
##
## data:  lemo[, 1] - lemo[, 2]
## W = 0.9, p-value = 0.4
```

```
##
##  Shapiro-Wilk normality test
##
## data:  energy[, 1] - energy[, 2]
## W = 0.9, p-value = 0.3
```

```
t.test(lemo[,1]-lemo[,2],energy[,1]-energy[,2])
```

```
##
##  Welch Two Sample t-test
##
## data:  lemo[, 1] - lemo[, 2] and energy[, 1] - energy[, 2]
## t = -1, df = 17, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.728  0.129
## sample estimates:
## mean of x mean of y
##    -0.145     0.154
```

```
wilcox.test(lemo[,1]-lemo[,2],energy[,1]-energy[,2]) #Mann-whitney
```

```
## Warning in wilcox.test.default(lemo[, 1] - lemo[, 2], energy[, 1] - energy[, :
## cannot compute exact p-value with ties
```

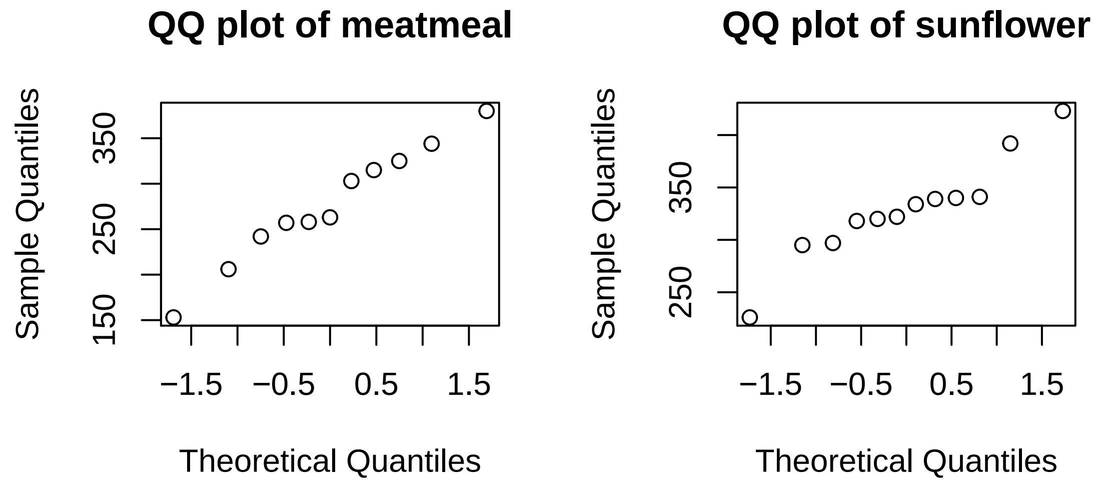
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  lemo[, 1] - lemo[, 2] and energy[, 1] - energy[, 2]
## W = 46, p-value = 0.1
## alternative hypothesis: true location shift is not equal to 0
```

d) The design of experiment in b) does show some defects because the experiment outcomes should not have any type of dependence on each other except for the “treatment”. If the subjects must perform two tasks, like running two times in this experiment, they should be allowed sufficient time between the tasks to recover and forget. However, they run again with only a half-an-hour break so this experiment cannot exclude the influences of the break time and the results cannot be treated as the exclusive effects of the type of drink. This flaw of design can be eliminated in experiment in c) because each subject would compute the time differences between before and after drink. When comparing all outcomes of these subjects, the influences of break time are treated equally to everyone so the results can respond to the effects of the type of drink.

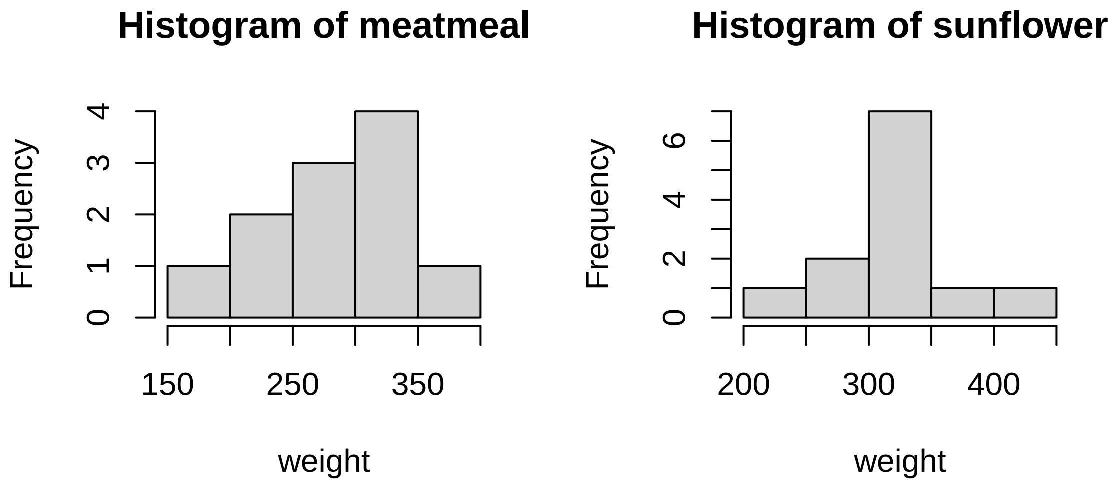
Exercise 5. Chick weights

a) The meatmeal and sunflower are two different groups so we should perform the two sample t-test which is not paired. We first check whether two samples follow normal distribution. The QQ plots and histograms for meatmeal and sunflower look like normal distribution but we also use shapiro test to test the normality. As can be seen from the results, p-value for meatmeal is 0.9612>0.05 and p-value for sunflower is 0.3603>0.05, both cannot reject H_0, so they both follow normal distribution. Then we implement the two samples t-test(not paired), Mann-Whitney test and Kolmogorov-Smirnov test respectively. As shown in the table, the p-value of t-test is 0.04441<0.05, H_0 of equal means is rejected; the p-value of Mann-Whitney test is 0.06882>0.05, H_0 of equal medians cannot be rejected, which means the underlying distribution of meatmeal and sunflower are the same; the p-value of Kolmogorov-Smirnov test is 0.1085>0.05, H_0 of equal means cannot be rejected, which also means the weight of meatmeal and sunflower have the same distribution.

```
meatmeal=subset(chickwts, feed=="meatmeal",select = c("weight","feed"))
sunflower=subset(chickwts, feed=="sunflower",select = c("weight","feed"))
par(mfrow=c(1,2))
qqnorm(meatmeal[,1],main="QQ plot of meatmeal");qqnorm(sunflower[,1],main="QQ plot of sunflower")
```



```
hist(meatmeal[,1],main = "Histogram of meatmeal",xlab="weight");
hist(sunflower[,1],main = "Histogram of sunflower",xlab="weight")
```



```
shapiro.test(meatmeal[,1]);shapiro.test(sunflower[,1])
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  sunflower[, 1]
## W = 0.9, p-value = 0.4
```

```
t.test(meatmeal[,1],sunflower[,1])#not paired
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  meatmeal[, 1] and sunflower[, 1]
## t = -2, df = 19, p-value = 0.04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -102.57   -1.44
## sample estimates:
## mean of x mean of y
##       277       329
```

```
wilcox.test(meatmeal[,1],sunflower[,1])
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  meatmeal[, 1] and sunflower[, 1]
## W = 36, p-value = 0.07
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(meatmeal[,1],sunflower[,1])
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  meatmeal[, 1] and sunflower[, 1]
## D = 0.5, p-value = 0.1
## alternative hypothesis: two-sided
```

Category	t-test	Mann-Whitney test	Kolmogorov-Smirnov test
p-value	0.04441	0.06882	0.1085

b) We perform a one-way ANOVA test and find that the p-value for testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ is $5.936 * 10^{-10}$ <0.05, hence H_0 of equal means is rejected and the effects of six feed supplement on the weight of chicks are significant and at least two distributions are different.

```
dataaov=lm(weight~feed,data = chickwts)
anova(dataaov);summary(dataaov)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value  Pr(>F)
## feed       5 231129   46226    15.4 5.9e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -123.91  -34.41    1.57   38.17  103.09
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      323.58      15.83   20.44  < 2e-16 ***
## feedhorsebean   -163.38      23.49   -6.96  2.1e-09 ***
## feedlinseed     -104.83      22.39   -4.68  1.5e-05 ***
## feedmeatmeal     -46.67      22.90   -2.04  0.04557 *  
## feedsoybean      -77.15      21.58   -3.58  0.00067 ***
## feedsunflower      5.33      22.39    0.24  0.81249    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.9 on 65 degrees of freedom
## Multiple R-squared:  0.542,  Adjusted R-squared:  0.506 
## F-statistic: 15.4 on 5 and 65 DF,  p-value: 5.94e-10
```

The estimated weights for six feed supplement are shown in the following table. The best feed supplement is sunflower, whose estimated weight is 328.916, largest in all the six supplement.

Category	casein	horsebean	linseed	meatmeal	soybean	sunflower
Estimated weight	323.583	160.20	218.75	276.909	246.428	328.916

c) Because the six samples are small, separate QQ plots are not so useful. We already have tested the normality of meatmeal and sunflower in a) so we use shapiro test to test the other four supplement. As can be seen from the results, all p-values are larger than 0.05 so they all follow normal distribution. We also check the normality of residuals, the QQ plot look like normal distribution and the p-value of shapiro test for residuals is 0.6272>0.05, so we can say that the residuals follow normal distribution and the assumptions do not fail in this experiment.

```r
casein=subset(chickwts, feed=="casein",select = c("weight","feed"))
horsebean=subset(chickwts, feed=="horsebean",select = c("weight","feed"))
linseed=subset(chickwts, feed=="linseed",select = c("weight","feed"))
soybean=subset(chickwts, feed=="soybean",select = c("weight","feed"))
shapiro.test(casein[,1]);shapiro.test(horsebean[,1]);shapiro.test(linseed[,1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  casein[, 1]
## W = 0.9, p-value = 0.3
```

```
##
##  Shapiro-Wilk normality test
##
## data:  horsebean[, 1]
## W = 0.9, p-value = 0.5
```

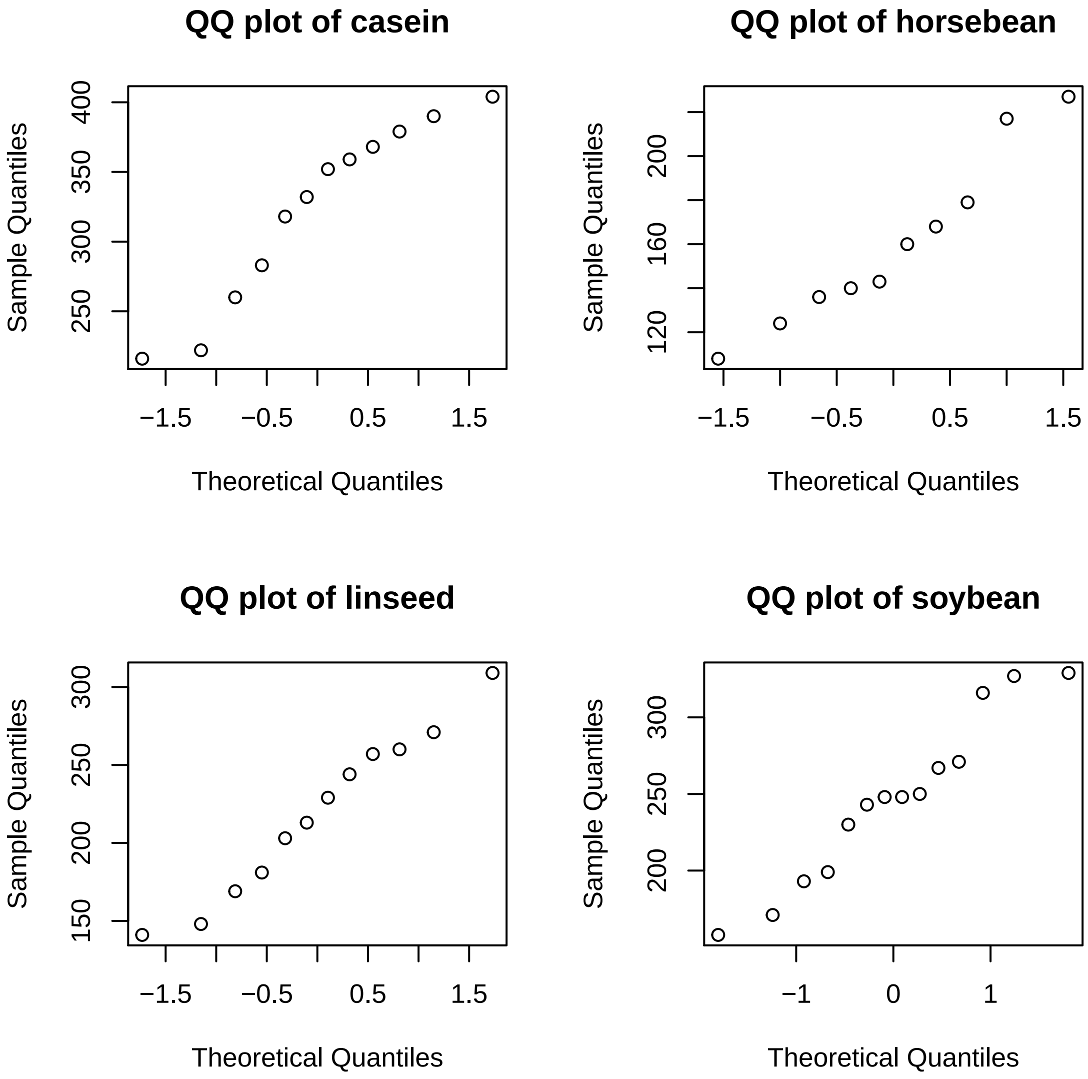
```
##
##  Shapiro-Wilk normality test
##
## data:  linseed[, 1]
## W = 1, p-value = 0.9
```

```r
shapiro.test(soybean[,1])
```

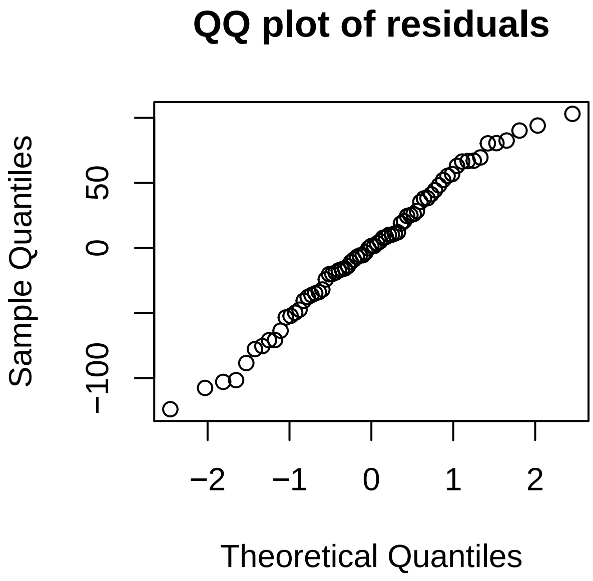
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  soybean[, 1]
## W = 0.9, p-value = 0.5
```

```
par(mfrow=c(2,2));qqnorm(casein[,1],main="QQ plot of casein");
qqnorm(horsebean[,1],main="QQ plot of horsebean");
qqnorm(linseed[,1],main="QQ plot of linseed");
qqnorm(soybean[,1],main="QQ plot of soybean")
```



```
par(mfrow=c(1,1));qqnorm(residuals(dataaov),main="QQ plot of residuals")
```



```
shapiro.test(residuals(dataaov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(dataaov)
## W = 1, p-value = 0.6
```

d) The one-way ANOVA relies on the normality. The Kruskal-Wallis test is a nonparametric alternative to one-way ANOVA which could test whether samples are from same distributions. If the normality assumption fails, we could use the Kruskal-Wallis test which is based on rank.

We perform the Kruskal-Wallis test and the p-value for testing $H_0 : F_1 = F_2 = F_3 = F_4 = F_5 = F_6$is $5.113 * 10^{-7}$<0.05, hence H_0 is rejected and the samples are not from the same distribution. This conclusion is the same as we get in b).

```
attach(chickwts);kruskal.test(weight,feed)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight and feed
## Kruskal-Wallis chi-squared = 37, df = 5, p-value = 5e-07
```