Assignment 1

Paul Jungnickel, Wessel Beumer, Kushnava Singha, group 10

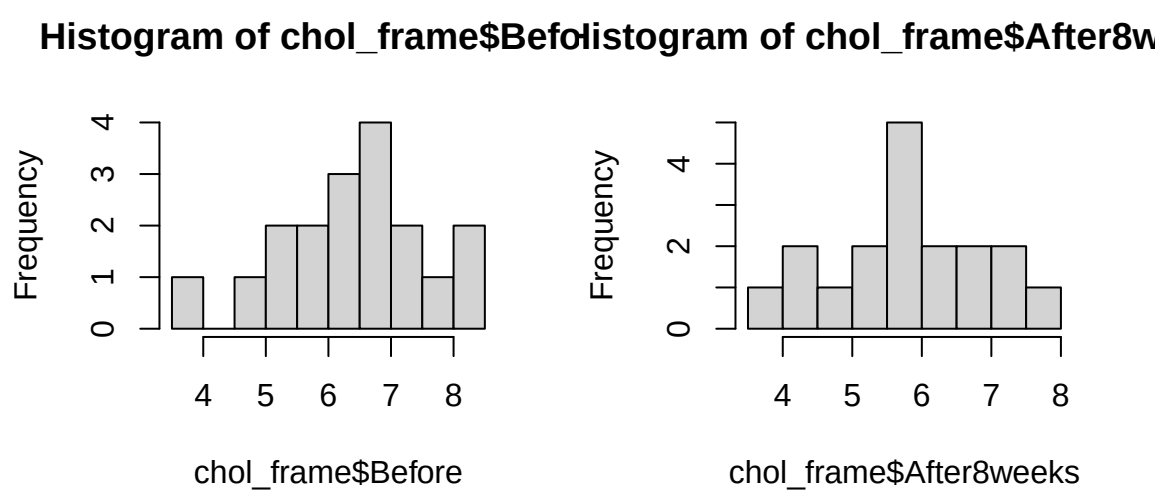
23 02 2025

Exercise 1

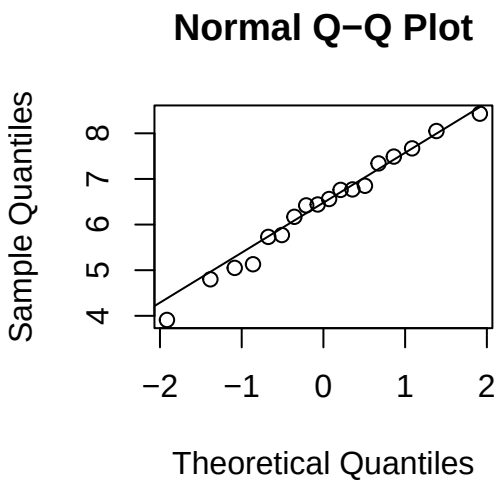
a)

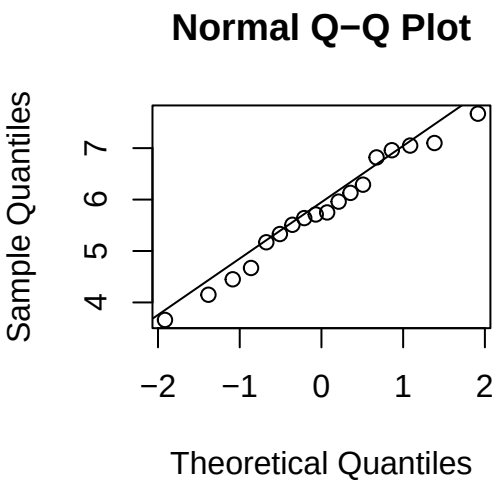
```
cholesterol <- read.table("../data/cholesterol.txt", header = TRUE);
chol_frame <- data.frame(cholesterol);
```

```
par(mfrow=c(1,2));
p1 <- hist(chol_frame$Before, 8);
p2 <- hist(chol_frame$After8weeks, 6)
```




```
qqnorm(chol_frame$Before);
qqline(chol_frame$Before);
qqnorm(chol_frame$After8weeks);
qqline(chol_frame$After8weeks); # two QQ-plots next to each other
```





The histograms of the two samples both appear normal, even though the sample size is very small. The Q-Q plots confirm this.

`cor(chol_frame$Before, chol_frame$After8weeks)` 0.991

The correlation between the two samples is almost one i.e. a high cholesterol at the start also implies a high cholesterol level after 8 weeks.

b) The experiment outcomes are paired , since the same person is studied at two time units

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

```
##
##  Paired t-test
##
## data:  chol_frame$Before and chol_frame$After8weeks
## t = 15, df = 17, p-value = 3e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.540 0.718
## sample estimates:
## mean difference
##           0.629
```

The null hypothesis H0 (no difference in mean) is rejected with a mean decrease of 0.63 after 8 weeks.

The permutation test can be applied since we have paired samples

```
X = chol_frame$Before
Y = chol_frame$After8weeks
B = 10000
Tstar = numeric(B)
for(i in 1:B){
  xystar = t(apply(cbind(X, Y), 1, sample))
  Tstar[i] = mean(xystar[,1] - xystar[,2])
}
t = mean(X-Y)
pl=sum(Tstar<t)/B;pr=sum(Tstar>t)/B
```

```
p = 2*min(pl, pr)
print(p)
```

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

p is much smaller than 0.95 so there is a significant difference after 8 weeks

Mann-Whitney can be applied but is less strong - designed for samples from different populations

```
wilcox.test(X, Y)
```

```
## Warning in wilcox.test.default(X, Y): cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  X and Y
## W = 210, p-value = 0.1
## alternative hypothesis: true location shift is not equal to 0
```

again the null hypothesis is rejected.

c)

#we have a discrete sample whose true std is unknown -> student dist

```
X <- chol_frame$After8weeks
print("97 percent confidence interval assuming normality:")
```

```
## [1] "97 percent confidence interval assuming normality:"
```

```
mean(X) + qt(c(0.015, 1-0.015), df=length(X)-1) * sd(X) / sqrt(mean(X))
```

```
## [1] 4.69 6.86
```

```
B = 1000
Tstar = numeric(B)
for(i in 1:B){
  Xstar = sample(X, replace=TRUE)
  Tstar[i] = mean(Xstar)
}
Tstar_q = quantile(-Tstar, c(0.015, 1-0.015))
mean(Tstar)
```

```
## [1] 5.79
```

```
mean(X)
```

```
## [1] 5.78
```

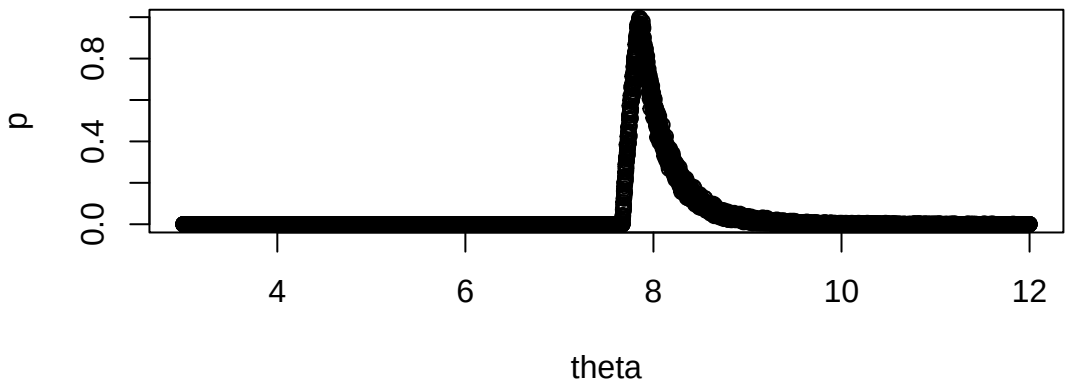
```
2*mean(X) + Tstar_q
```

```
##  1.5% 98.5%
##  5.19  6.33
```

with bootstrapping we get a significantly smaller confidence interval.

d)

```
B = 1000
Tstar = numeric(B)
theta = seq(from=3, to=12, by=0.001)
num_test_pts = length(theta)
p = numeric(num_test_pts)
for(n in 1:num_test_pts){
  for(i in 1:B){
    Xstar = runif(length(X), 3,theta[n])
    Tstar[i] = max(Xstar)
  }
  t = max(X)
  pl=sum(Tstar<t)/B;pr=sum(Tstar>t)/B
  p[n] = 2*min(pl, pr)
}
plot(theta, p)
```




```
print(theta[p>0.95])
```

```
##  [1] 7.83 7.85 7.85 7.85 7.85 7.85 7.85 7.85 7.86 7.86 7.86 7.86 7.86 7.86 7.86
## [16] 7.86 7.86 7.87 7.88 7.88 7.89
```

for theta between 7.83 and 7.88 we have p>0.95 i.e. do not reject H0.

Kolmogorov Smirnov can be applied by collecting independent samples from U(3, theta) Since we are now looking at the entire distribution we always reject the null hypothesis

```
suppressWarnings({
B = 1000
Tstar = numeric(B)
theta = seq(from=3, to=12, by=0.01)
num_test_pts = length(theta)
p = numeric(num_test_pts)
for(n in 1:num_test_pts){

  Y = runif(B, 3, theta)
  k_test = ks.test(X, Y)
  p[n] = k_test$p.value
```

```
}
print(theta[p>0.95])
})
```

```
## numeric(0)
```

e) use bootstrapping to get a confidence interval on the median of after8weeks

```
X = chol_frame$After8weeks
B = 1000
Tstar = numeric(B)
for(i in 1:B){
  Xstar = sample(X, replace=TRUE)
  Tstar[i] = median(Xstar)
}
Tstar_q = quantile(-Tstar, c(0.05, 1-0.05))
median(X)
```

```
## [1] 5.73
```

```
2*median(X) + Tstar_q
```

```
##   5%  95%
## 5.17 6.06
```

The median has a higher-than 5% chance of being larger than 6 - not statistically significant

again use bootstrapping to get a confidence interval on the 25th percentile

```
X = chol_frame$After8weeks
B = 100000
Tstar = numeric(B)
for(i in 1:B){
  Xstar = sample(X, replace=TRUE)
  Tstar[i] = quantile(Xstar, 0.25)
}
Tstar_q = quantile(-Tstar, c(0.05, 1-0.05))
quantile(X, 0.25)
```

```
##  25%
## 5.21
```

```
2*quantile(X, 0.25) + Tstar_q
```

```
##   5%  95%
## 4.75 5.97
```