Assignment 1

Alexia Salomons, Nathan Maxwell Jones, Yauheniya Makarevich, group 71

27 February 2023

Exercise 1.

The data set birthweight.txt contains the birthweights (in grams) of 188 newborn babies. Denote the underlying mean birthweight by μ.

```
birthweight <- readLines("data/birthweight.txt")
birthweight <- as.double(birthweight[2:length(birthweight)])
birthweight
```

```
##   [1] 1538 2617 2691 2401 3596 3153 2500 3186 3971 4011 3699 1483 2478 2101 3314
##  [16] 3098 3157 2657 2115 3680 2514 2007 2140 3289 2947 2656 2799 3229 3593 3182
##  [31] 2263 1639 3215 3654 3085 2828 2992 2667 3138 2254 2865 2304 2485 1602 1842
##  [46] 2811 2878 2449 3580 2128 1026 3476 3957 2434 2902 2758 2351 3575 3787 3653
##  [61] 2673 2705 2899 2478 3446 4252 2504 2901 2579 2301 3466 4922 2391 1611 2162
##  [76] 2567 2585 2894 2710 2788 2470 1733 2502 2851 2628 3814 3571 2483 3616 3152
##  [91] 2663 1866 3038 3500 2173 1687 3502 3395 2327 2464 1650 2034 2712 2498 3008
## [106] 2892 3589 2686 2529 2569 3494 3258 2593 2913 2274 2911 3085 3463 2890 3485
## [121] 1870 3821 1678 2634 2316 3253 2652 4262 3386 3056 2191 2846 3087 3079 2037
## [136] 3627 4024 1841 3039 3148  955 4418 3666 3165 3534 4116 4110 4111 2907 3703
## [151] 3075 3208 2989 3323 3083 2620 3707 3113 2479 2988 3375 3141 2076 2338 2568
## [166] 3618 2980 1926 2923 2863 4305 3893 2721 4156 2279 4038 3588 2920 1843 2965
## [181] 3543 3670 3902 3688 3504 3798 1883 2876
```

```
length(birthweight)
```

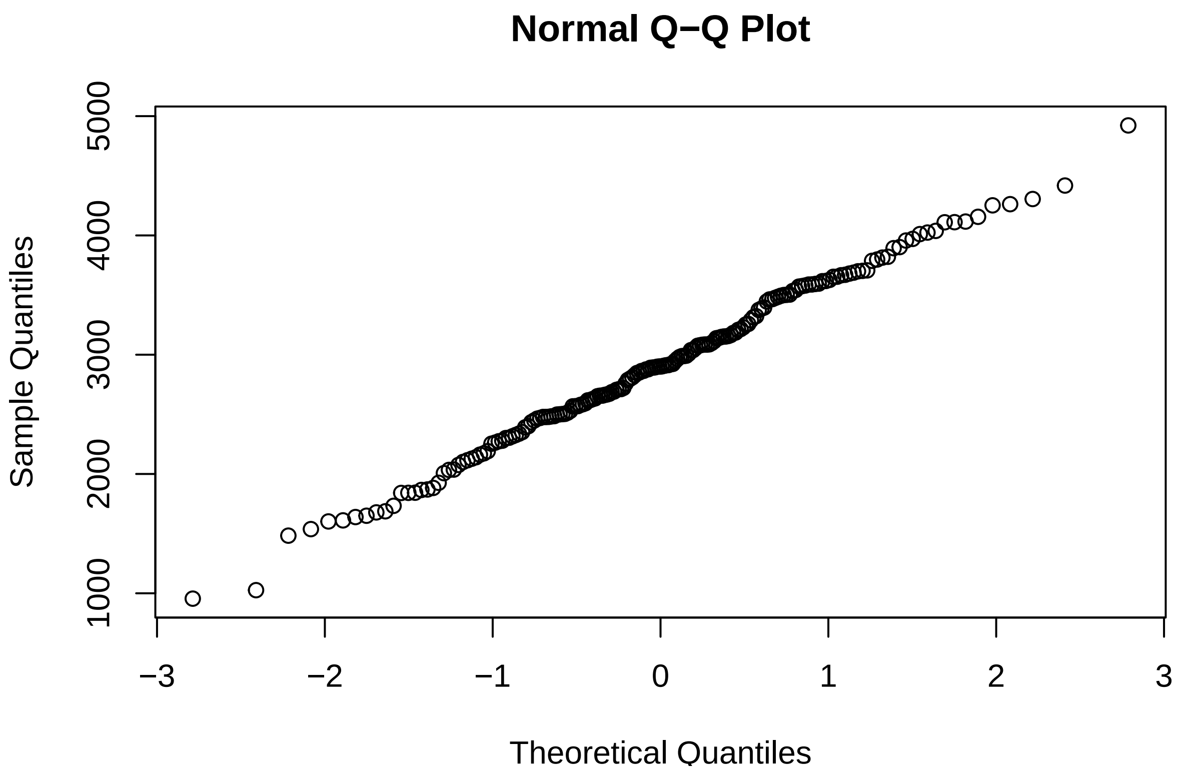
```
## [1] 188
```

```
birthweight_mean <- mean(birthweight)
birthweight_mean
```

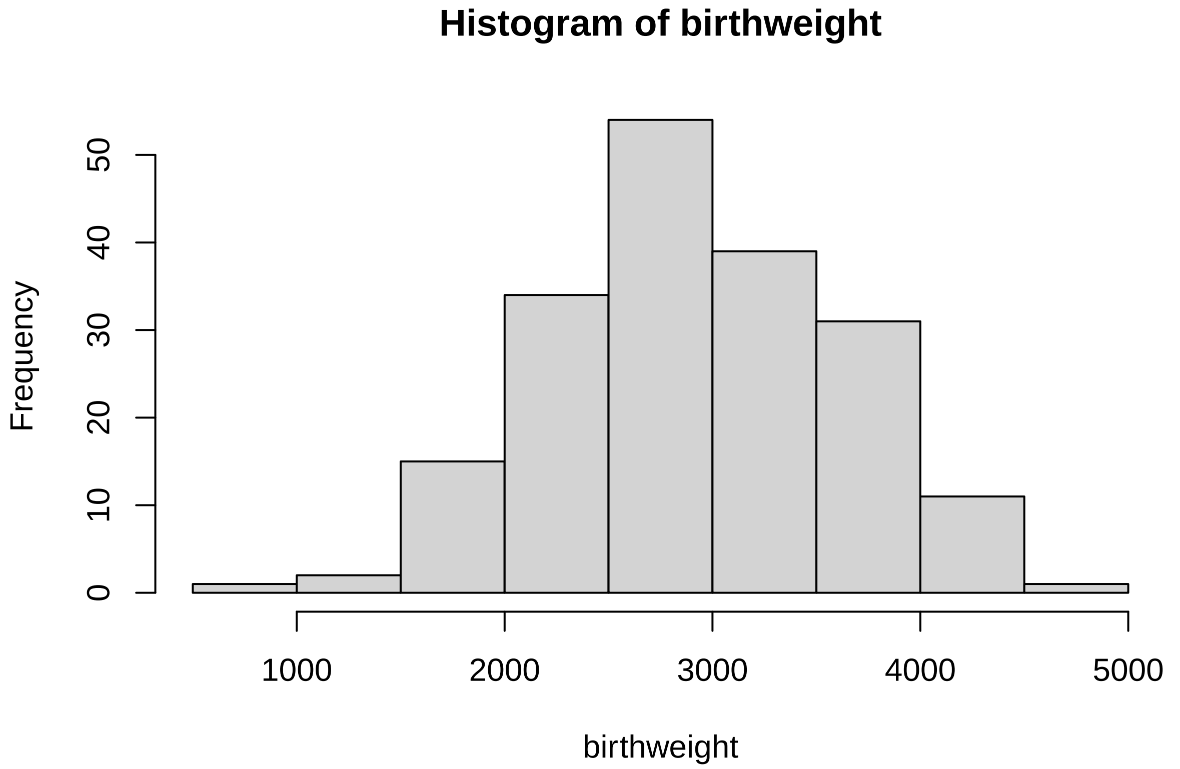
```
## [1] 2913
```

a) Check normality of the data. Assuming normality (irrespective of your conclusion about normality), construct a bounded 96%-CI for μ. Evaluate the sample size needed to provide that the length of the 96%-CI is at most 100. Compute a bootstrap 96%-CI for μ and compare it to the above CI.

```
qqnorm(y = birthweight)
```



```
hist(birthweight)
```



```
# add density line
# lines(density(birthweight), col="blue",lwd=2)
```

```r
shapiro.test(birthweight)
```

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

```r
# H0 - normal distribution, H1 - not normal
```

NORMAL DISTRIBUTION!

Let's go for CI-96%:

```r
t.test(birthweight, conf.level = 0.96)
```

```
##
##  One Sample t-test
##
## data:  birthweight
## t = 57, df = 187, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 96 percent confidence interval:
##  2808 3019
## sample estimates:
## mean of x
##      2913
```

```r
B <- 1000
alpha <- 0.04
T_star <- numeric(B)

for(i in 1:B) {
  X_star <- sample(birthweight, replace = TRUE)
  T_star[i] <- mean(X_star)
}

T_star_q2 <- quantile(T_star, alpha/2)
T_star_q98 <- quantile(T_star, 1 - alpha/2)

c(2*birthweight_mean - T_star_q98, 2*birthweight_mean - T_star_q2)
```

```
##  98%   2%
## 2812 3022
```

```r
sum(T_star<T_star_q2)
```

```
## [1] 20
```

b) An expert claims that the mean birthweight is bigger than 2800 gram. Verify this claim by using a relevant t-test, explain the meaning of the CI in the R-output for this test. Also propose and perform a suitable sign tests for this problem.

```
t.test(birthweight, alternative = "greater", mu=2800)
```

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

We reject H0(p=0.01337), so H1 is true and mean of the sample is bigger than 2800. CI is infinite on right side, since the test is one-sided.

Binom test

H0: mean <= 2800, H1: mean > 2800.

```
greater_weight <- as.integer(birthweight > 2800)
binom.test(sum(greater_weight), length(greater_weight), p=0.5, alt="g")
```

```
##
##  Exact binomial test
##
## data:  sum(greater_weight) and length(greater_weight)
## number of successes = 107, number of trials = 188, p-value = 0.03
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.507 1.000
## sample estimates:
## probability of success
##                  0.569
```

We reject H0(p=0.03868), so H1 is true and mean of the sample is bigger than 2800.

Both test confirmed the hypothesis that mean of the sample is bigger than 2800.

c) Propose a way to compute the powers of the t-test and sing test from b) at some $\mu > 2800$, comment.

d) Let p be the probability that birthweight of a newborn baby is less than 2600 gram. Using asymptotic normality, the expert computed the left end $\hat{p} = 0.25$ of the confidence interval $[\hat{p_l}, \hat{p_r}]$ for p. Recover the whole confidence interval and its confidence level.

e) The expert also 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. Verify this claim by an appropriate test.

success: w > 2600

```
prop.test(c(61, 65), c(95, 93))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(61, 65) out of c(95, 93)
## X-squared = 0.5, df = 1, p-value = 0.5
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2016  0.0879
## sample estimates:
## prop 1 prop 2
##  0.642  0.699
```

We accept H0: p1-p2=0, where p1, p2 - proportions of the success in population.

Exercise 2

A study tested whether cholesterol was reduced after using a certain brand of margarine as part of a low fat low cholesterol diet. The data set cholesterol.txt contains information on 18 people using margarine to reduce cholesterol: columns Before and After8weeks contain the cholesterol level (mmol/L) respectively before the diet and after 8 weeks on the diet.

```
df <- as.data.frame(read.table("data/cholesterol.txt", header=TRUE))
head(df)
```

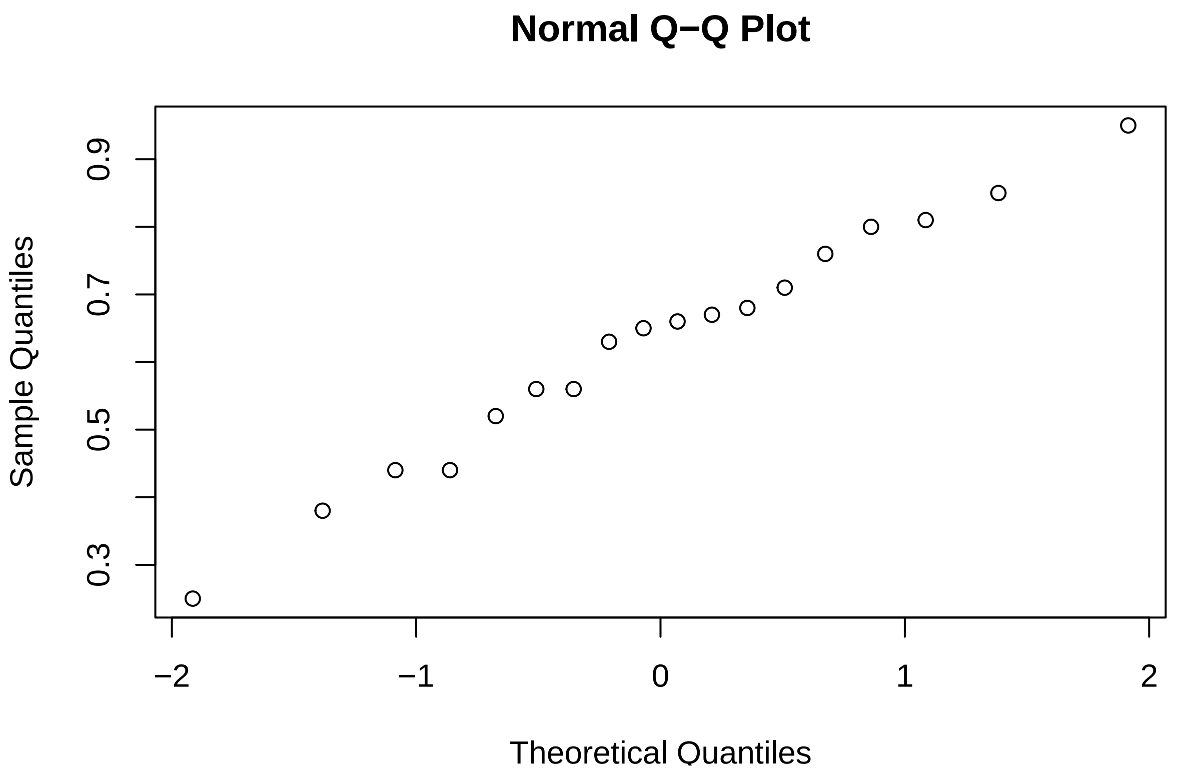
```
##   Before After8weeks
## 1   6.42        5.75
## 2   6.76        6.13
## 3   6.56        5.71
## 4   4.80        4.15
## 5   8.43        7.67
## 6   7.49        7.05
```

a) Make some relevant plots of this data set, comment on normality. Are there any inconsistencies in the data? Investigate whether the columns Before and After8weeks are correlated.

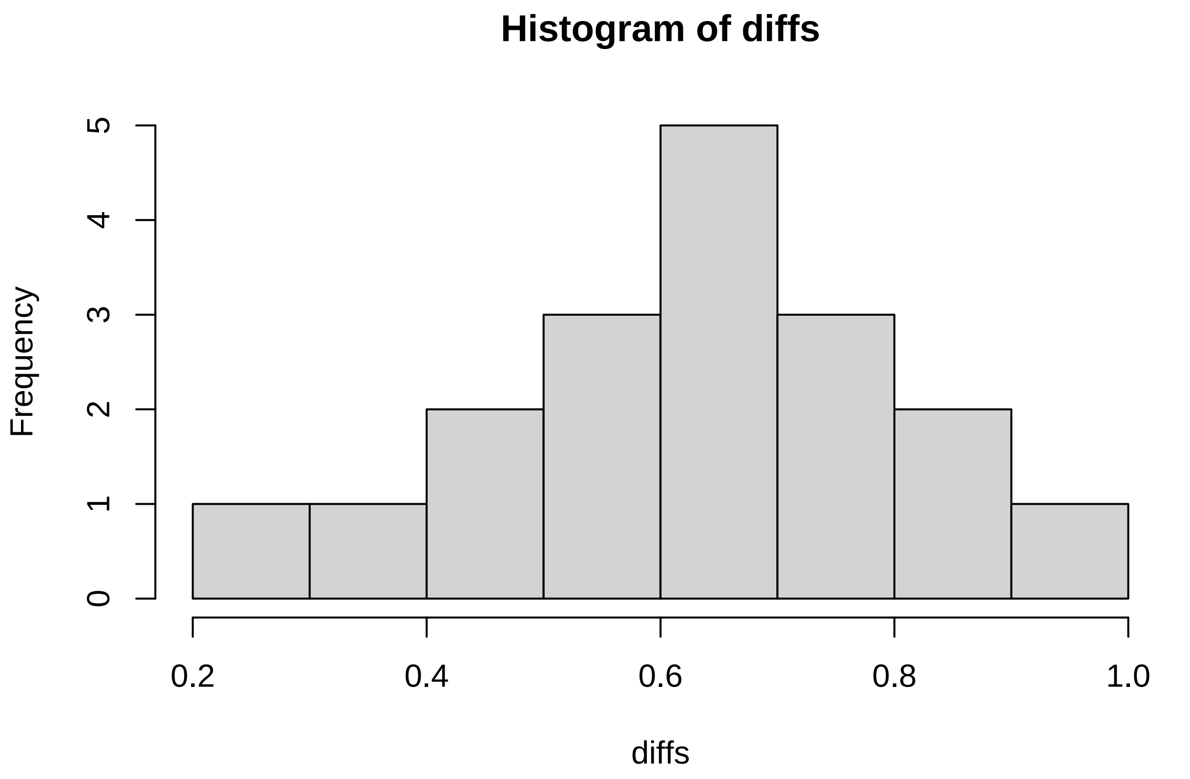
```
diffs <- df[, 1] - df[, 2]
diffs
```

```
##  [1] 0.67 0.63 0.85 0.65 0.76 0.44 0.95 0.38 0.44 0.25 0.81 0.80 0.66 0.71 0.52
## [16] 0.56 0.68 0.56
```

```
qqnorm(diffs)
```



```
hist(diffs)
```



```
shapiro.test(diffs)
```

```
##
```

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

Differences are normally distributed.

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

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

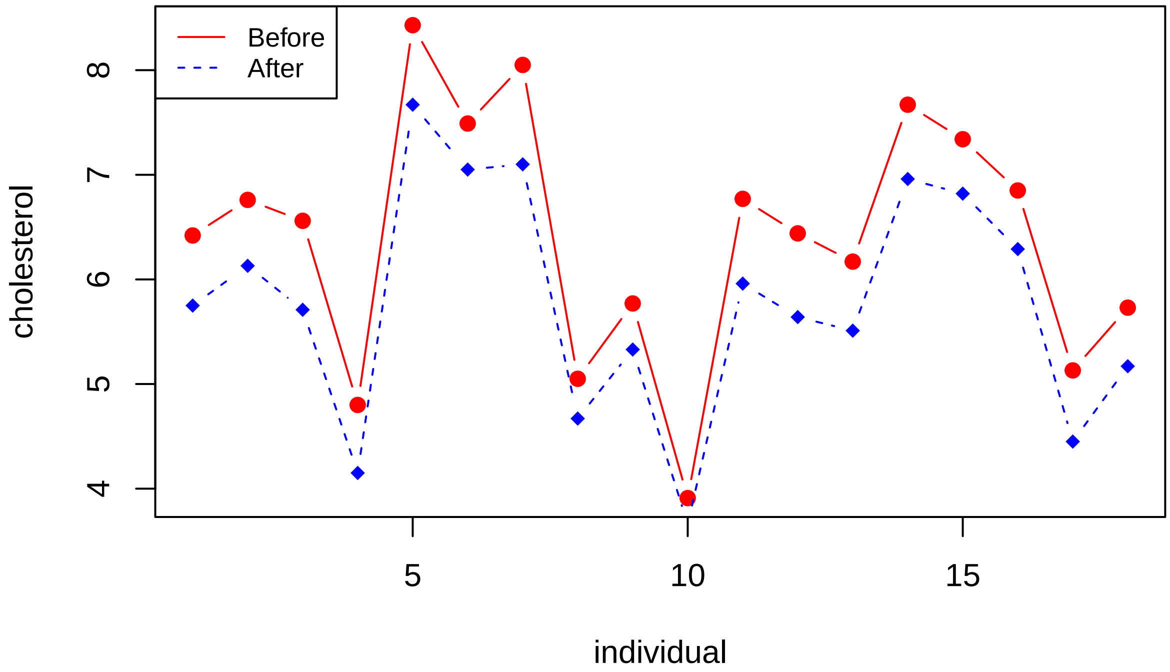
```r
shapiro.test(df[, 2])
```

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

```r
cor.test(df[, 1], df[, 2], method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## t = 29, df = 16, p-value = 2e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.975 0.997
## sample estimates:
##   cor
## 0.991
```

```r
# Create a first line
plot(1:length(df[, 1]), df[, 1], type = "b", pch=19, col = "red", xlab = "individual", ylab = "
lines(1:length(df[, 2]), df[, 2], pch=18, col = "blue", type = "b", lty=2)
legend("topleft", legend=c("Before", "After"), col=c("red", "blue"), lty = 1:2, cex=0.8)
```



b) Apply two relevant tests (cf. Lectures 2, 3) to verify whether the diet with low fat margarine has an effect (argue whether the data are paired or not). Is a permutation test applicable?

Data is paired since it is two different measurements of the same person and two samples are correlated. Relevant test:

1. T-test paired test

```
t.test(df[, 1], df[, 2], paired=2)
```

```
##
##  Paired t-test
##
## data:  df[, 1] and df[, 2]
## 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
```

-> based on p-value we reject null-hypothesis that samples have the same mean. there is a difference between these two samples.

2. Permutation test - it is applicable because we are only testing for a difference between mean, not how they relate to each other.

```
diff_mean <- function(x, y) {
  return(mean(x-y))
}
```

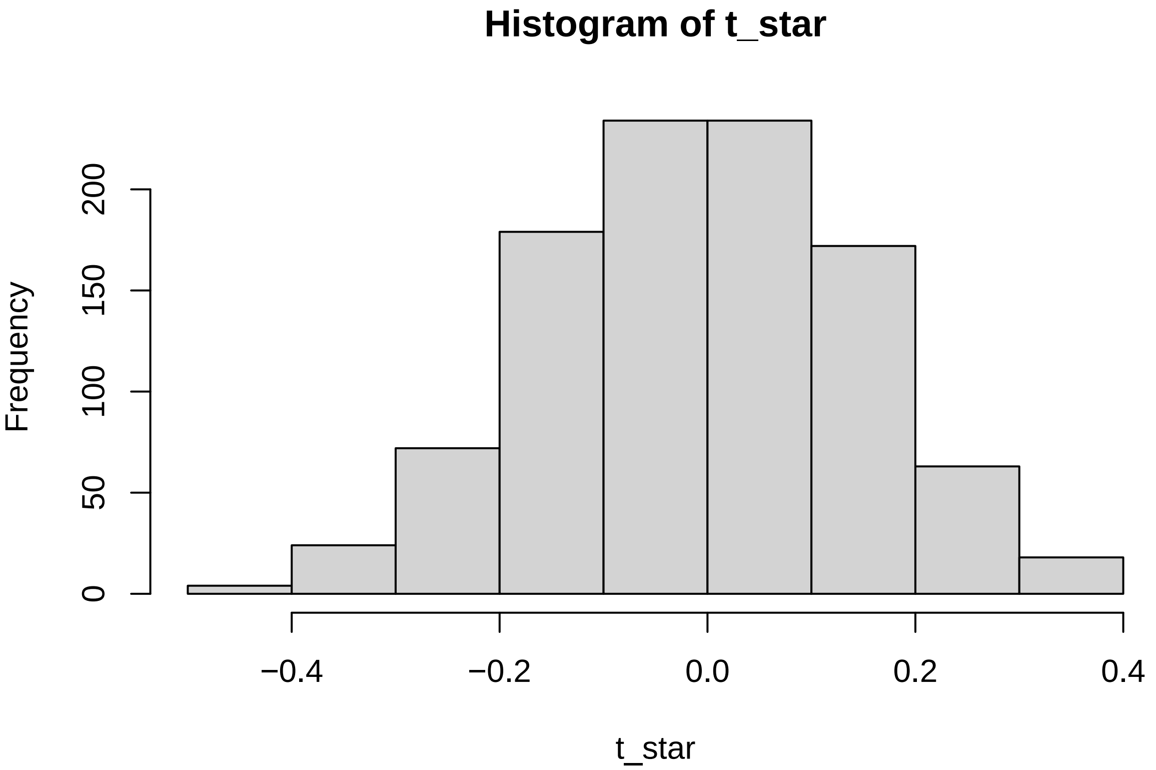
```
# original dataframe
stats <- diff_mean(df[, 1], df[, 2])
stats
```

```
## [1] 0.629
```

```
B <- 1000
t_star <- numeric(B)

for (i in 1:B) {
  diff_star <- t(apply(cbind(df[, 1], df[, 2]), 1, sample))
  t_star[i] <- diff_mean(diff_star[, 1], diff_star[, 2])
}

hist(t_star)
# plot(rep(stats, 2), c(0, 50), col="red", lwd=2)
lines(rep(stats, 2), c(0, 50), col="red", lwd=2)
```



```
# calculating p-value
pl <- sum(t_star < stats) / B
pr <- sum(t_star > stats) / B

p <- 2*min(pl, pr)
p
```

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

c) $\mu = (\theta + 3)/2, \theta > 3$

```r
mu <- mean(df[, 2])
mu # bar{x}
```

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

```r
s <- sd(df[, 2])
s
```

```
## [1] 1.1
```

```r
theta_hat <- 2*mu - 3
theta_hat
```

```
## [1] 8.56
```

```r
alpha <- 0.05

n <- length(df[, 2])
t_alpha <- qt(1 - alpha/2, df=n)
t_alpha
```

```
## [1] 2.1
```

```r
theta_l <- theta_hat - t_alpha*s/sqrt(n)
theta_r <- theta_hat + t_alpha*s/sqrt(n)
c("[", theta_l,",", theta_r, "]")
```

```
## [1] "["                "8.01211959907622" ","                "9.10343595647934"
## [5] "]"
```

We can improve the CI by having more individuals in the samples.

d)

```r
t <- max(df[, 2])
t
```

```
## [1] 7.67
```

```r
n <- length(df[, 2])
n
```

```
## [1] 18
```

```r
for (theta in 3:12) {
  B <- 1000
  t_star <- numeric(B)

  for (i in 1:B) {
    x_star <- runif(n, min = 3, max = theta)
    t_star[i] <- max(x_star)
  }

  pl <- sum(t_star < t)/B
  pr <- sum(t_star > t)/B

  p <- 2* min(pl, pr)
  print(paste("Theta =", theta, ", ", p))
}
```

```
## [1] "Theta = 3 ,  0"
## [1] "Theta = 4 ,  0"
## [1] "Theta = 5 ,  0"
## [1] "Theta = 6 ,  0"
## [1] "Theta = 7 ,  0"
## [1] "Theta = 8 ,  0.594"
## [1] "Theta = 9 ,  0.018"
## [1] "Theta = 10 ,  0.004"
## [1] "Theta = 11 ,  0"
## [1] "Theta = 12 ,  0"
```

e) Medium and proportion tests

```
less_chol <- as.integer(df[, 2] < 6)
binom.test(sum(less_chol), length(less_chol), p=0.5, alt="g")
```

```
##
##  Exact binomial test
##
## data:  sum(less_chol) and length(less_chol)
## number of successes = 11, number of trials = 18, p-value = 0.2
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.392 1.000
## sample estimates:
## probability of success
##                  0.611
```