Homework 06 - STAT440

Joseph Sepich (jps6444)

10/11/2020

```
set.seed(42)
```

```
diet_df <- read.csv('./data/diet.csv')
diet_df <- diet_df[complete.cases(diet_df),]
head(diet_df)
```

```
##   X  id        doe        dox        dob          y fail        job month
## 1 1 102 1976-01-17 1986-12-02 1939-03-02 10.8747433    0     Driver     1
## 2 2  59 1973-07-16 1982-07-05 1912-07-05  8.9691992    0     Driver     7
## 3 3 126 1970-03-17 1984-03-20 1919-12-24 14.0095825   13  Conductor     3
## 4 4  16 1969-05-16 1969-12-31 1906-09-17  0.6269678    3     Driver     5
## 5 5 247 1968-03-16 1979-06-25 1918-07-10 11.2744695   13 Bank worker     3
## 6 6 272 1969-03-16 1973-12-13 1920-03-06  4.7446954    3 Bank worker     3
##    energy  height   weight    fat     fibre   energy.grp chd
## 1 22.8601 181.610 88.17984  9.168 1.4000000 <=2750 KCals   0
## 2 23.8841 165.989 58.74120  9.651 0.9350001 <=2750 KCals   0
## 3 24.9537 152.400 49.89600 11.249 1.2480000 <=2750 KCals   1
## 4 22.2383 171.196 89.40456  7.578 1.5570000 <=2750 KCals   1
## 5 18.5402 177.800 97.07040  9.147 0.9910000 <=2750 KCals   1
## 6 20.3073 175.260 61.00920  8.536 0.7650000 <=2750 KCals   1
```

Problem 1

Set a hypothesis testing scenario where you want to test if fat and fibre consumption are correlated. Explicitly say what the null and alternative hypotheses are, as well as the test statistic.

Null Hypothesis:

$$H_0 : \rho = 0$$

Alternative Hypothesis:

$$H_A : \rho \neq 0$$

We can use the following test statistic:

$$T = |r - \rho_0| = |r|$$

This test statistic we are using here translates to the absolute value of the sample correlation coefficient between fat and fibre. Our null hypothesis states that ρ_0 is zero, which is why we only need $|r|$. R has the built in cor command that will calculate r for us: `cor(diet_df$fat, diet_df$fibre)`. The default is pearson's correlation coefficient.

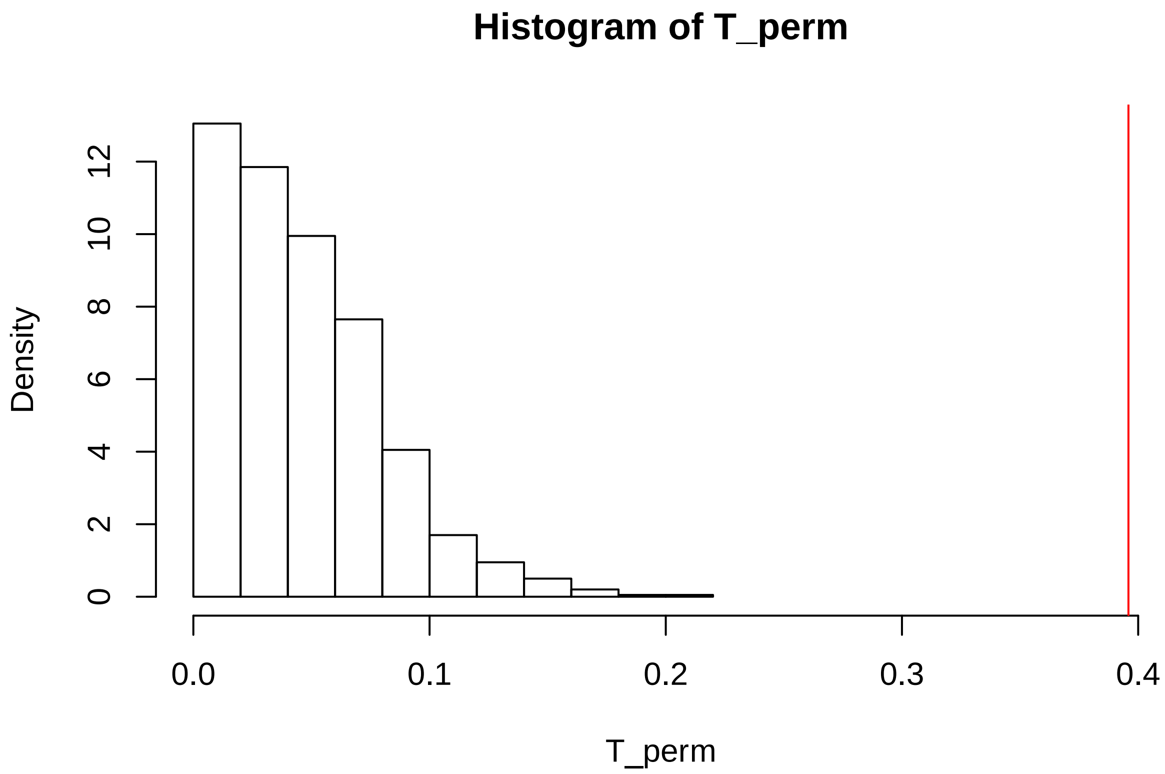
Problem 2

```
K <- 1000

T_obs <- abs(cor(diet_df$fat, diet_df$fibre))

T_perm <- vector('numeric', K)
temp_df <- diet_df
for (i in 1:K) {
    temp_df$fibre <- sample(diet_df$fibre)
    T_perm[i] <- abs(cor(temp_df$fat, temp_df$fibre))
}

x_end <- max(c(T_obs, max(T_perm)))
hist(T_perm, freq=FALSE, xlim = c(0,x_end))
abline(v=T_obs, col='red')
```



Problem 3

The approximate p-value is zero:

```r
print(mean(T_perm >= T_obs))
```

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