MATH3424 HW1

Chow, Hau Cheung Jasper (hcjchow / 20589533)

September 20, 2021

```r
# Q2
x <- c(240, 243, 250, 254, 264, 279, 284, 285, 290, 298, 302, 310, 312,
       315, 322, 337, 348, 384, 386, 520)
t.test(x, alternative="two.sided", mu=200, paired=FALSE, conf.level=0.98)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 7.7194, df = 19, p-value = 2.836e-07
## alternative hypothesis: true mean is not equal to 200
## 98 percent confidence interval:
##  274.5847 347.7153
## sample estimates:
## mean of x
##    311.15
```

```r
# ROUGH WORK
# Q4a-d, Q8e, Q10f, Q10g p-value hypothesis tests:
t_stat = 1.008
pt(-t_stat, df=12, lower.tail=TRUE)+pt(t_stat, df=12, lower.tail=FALSE)
```

```
## [1] 0.3333482
```

```r
t_stat = 1.008 # one sided upper
pt(t_stat, df=12, lower.tail=FALSE)
```

```
## [1] 0.1666741
```

```r
t_stat = 1.2405
pt(-t_stat, df=12, lower.tail=TRUE)+pt(t_stat, df=12, lower.tail=FALSE)
```

```
## [1] 0.2384974
```

```r
t_stat = 0.2498
pt(-t_stat, df=12, lower.tail=TRUE)+pt(t_stat, df=12, lower.tail=FALSE)
```

```
## [1] 0.806967
```

```r
t_stat = -1.754 # one sided upper
pt(t_stat, df=17, lower.tail=FALSE)
```

```
## [1] 0.9512807
```

```r
t_stat = 11.45816
pt(-t_stat, df=94, lower.tail=TRUE)+pt(t_stat, df=94, lower.tail=FALSE)
```

```
## [1] 1.536333e-19
```

```r
t_stat = 3.932812
pt(-t_stat, df=94, lower.tail=TRUE)+pt(t_stat, df=94, lower.tail=FALSE)
```

```
## [1] 0.0001605822
```

```r
# ROUGH WORK
# Q4
qt(0.025,df=12,lower.tail=FALSE)
```

```
## [1] 2.178813
```

```r
qt(0.05,df=12,lower.tail=FALSE)
```

```
## [1] 1.782288
```

```r
# Q5
qt(0.01,df=12,lower.tail=FALSE)
```

```
## [1] 2.680998
```

```r
# ROUGH WORK
# Q7
x_7 <- c(1,2,3,4,4,5,6,6,7,8,9,9,10,10)
y_7 <- c(23,29,49,64,74,87,96,97,109,119,149,145,154,166)
fit <- lm(y_7 ~ x_7)
y_hat_7 <- predict(fit)

sum((y_7-mean(y_7))^2) # 7b
```

```
## [1] 27768.36
```

```r
sse <- sum((y_7-y_hat_7)^2)
sse # 7c
```

```
## [1] 348.8484
```

```r
cor(x_7,y_7)
```

```
## [1] 0.9936987
```

```r
cor(y_7,y_hat_7)
```

```
## [1] 0.9936987
```

```r
# ROUGH WORK
# Q8
qt(0.02,df=17,lower.tail=FALSE)
```

```
## [1] 2.223845
```

```r
# Q10
setwd("/Users/jchow/Downloads/MATH3424 R") # need to set the starting directory as this
data <- read.table(file="Heights.txt",header=TRUE,col.names=c("Husband","Wife"))
head(data)
```

```
##   Husband Wife
## 1     186  175
## 2     180  168
## 3     160  154
## 4     186  166
```

```
## 5     163  162
## 6     172  152
```

```r
# 10a
cov(data$Husband, data$Wife)
```

```
## [1] 69.41294
```

```r
# 10b
# 2.54cm = 1 inch
data_inch <- data/2.54
cov(data_inch$Husband, data_inch$Wife)
```

```
## [1] 10.75903
```

```r
# 10c
cor(data$Husband, data$Wife)
```

```
## [1] 0.7633864
```

```r
# 10d
husband_fake <- data$Husband
wife_5cm <- data$Husband-5
cor(husband_fake, wife_5cm)
```

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

The above makes intuitive sense. For ANY two arbitrary data points generated in this manner, $(x_0, x_0 - 5), (x_1, x_1 - 5)$,, the slope of the line connecting them is $(x_1 - 5 - (x_0 - 5))/(x_1 - x_0) = (x_1 - x_0)/(x_1 - x_0) = 1$ Since the slope of the line is 1 for ALL points generated in such a manner, the predictor and response are perfectly (positively) linear correlated. Hence we expect the correlation to be 1.

```r
# 10e
heights_model = lm(data$Wife ~ data$Husband, data = data)
out <- summary(heights_model)
out
```

```
##
## Call:
## lm(formula = data$Wife ~ data$Husband, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4685  -3.9208   0.8301   3.9538  11.1287
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.93015   10.66162   3.933 0.000161 ***
## data$Husband  0.69965    0.06106  11.458  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.928 on 94 degrees of freedom
## Multiple R-squared:  0.5828, Adjusted R-squared:  0.5783
## F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

The fitted linear model (as seen in the above summary) is thus $Y = 41.93015 + 0.69965X$.

```r
# 10f
# Assuming two sided?
```

```r
beta_1 <- 0
beta_1_std_err <- coef(out)[, "Std. Error"][2]
beta_1_hat <- coef(out)[, "Estimate"][2]
t_1 <- (beta_1_hat - beta_1)/(beta_1_std_err)
t_1_alpha <- qt(0.05/2, df=length(data$Husband)-2, lower.tail=FALSE)
t_1
```

```
## data$Husband
##     11.45816
```

```r
t_1_alpha
```

```
## [1] 1.985523
```

n =length of Husband column in the dataframe (aka number of datapoints)

Since $|t_1| > t_{n-2,\alpha/2} = 11.45816 > 1.985523$, we reject the null hypothesis H_0 at significance level $\alpha = 0.05$.

Conduct hypothesis test for Q10g.

```r
# 10g
# Assuming two sided?
beta_0 <- 0
beta_0_std_err <- coef(out)[, "Std. Error"][1]
beta_0_hat <- coef(out)[, "Estimate"][1]
t_0 <- (beta_0_hat - beta_0)/(beta_0_std_err)
t_0_alpha <- qt(0.05/2, df=length(data$Husband)-2, lower.tail=FALSE)
t_0
```

```
## (Intercept)
##    3.932812
```

```r
t_0_alpha
```

```
## [1] 1.985523
```

Since $|t_0| > t_{n-2,\alpha/2} = 3.932812 > 1.985523$, we reject the null hypothesis H_0 at significance level $\alpha = 0.05$.