Homework 4

Matthew Bentz

4/19/2022

1.

The confidence interval for a population proportion is $p \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$. Since p on the left side does not affect the length of the interval, we do not have to worry about it, and since the $Z_{\frac{\alpha}{2}}$ value is a constant for any α, we can disregard it as well. So, to maximize the length of our confidence interval, we set the derivative of $\sqrt{\frac{p(1-p)}{n}}$ with respect to p equal to zero. We find that the confidence interval is maximized at $p = \frac{1}{2}$.

2.

```r
control <- c(0.11, 0.11, 0.11, 0.19, 0.21, 0.22, 0.24, 0.25, 0.31)
gallstone <- c(0.18, 0.27, 0.36, 0.37, 0.39, 0.47, 0.37, 0.57)
ulcer <- c(0.29, 0.30, 0.40, 0.45, 0.47, 0.52, 0.57, 1.10)

# --------------------------------------------------
# Here we add NA values for gallstone and ulcer ----
# to allow the creation of a rectangular data frame.

length(ulcer) <- length(control)
length(gallstone) <- length(control)

suppressWarnings(library(tidyr))
CCK = data.frame(control, gallstone, ulcer)
CCK = gather(CCK, key = "diseases", value = "activity", control, gallstone, ulcer)
CCK$diseases = as.factor(CCK$diseases)
head(CCK)
```

```
##   diseases activity
## 1  control     0.11
## 2  control     0.11
## 3  control     0.11
## 4  control     0.19
## 5  control     0.21
## 6  control     0.22
```

```r
cckaov = aov(CCK$activity ~ CCK$diseases)
summary(cckaov)
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## CCK$diseases  2 0.4328 0.21641   7.949 0.00252 **
## Residuals    22 0.5989 0.02722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

3.

```r
n_1 <- 9
n_2 <- 8
n_3 <- 8
S_1 <- sd(control)
S_2 <- sd(gallstone, na.rm = T)
S_3 <- sd(ulcer, na.rm = T)

MS_e <- ((n_1-1)*(S_1^2) + (n_2-1)*(S_2^2) + (n_3-1)*(S_3^2)) /
  ((n_1-1) + (n_2-1) + (n_3-1))
```

The MS_E value when pooling individual treatment variances is 0.02722, which is equal to the ANOVA table's value.

4.

Source	df	Sum of Squares	Mean Square	F
Mixture	5	3,575.065	715.013	51.333
Error	150	2,089.35	13.929	
Total	155	5,664.415		

$$\begin{aligned}
k &= 6 \\
n &= 26 \\
N &= n * k = 156 \\
SST &= 5664.415 \\
MSE &= 13.929 \\
SSE &= MSE * (N - k) = 13.929 * (156 - 6) = 2,089.35 \\
SSTr &= SST - SSE = 5,664.415 - 2,089.35 = 3,575.065 \\
MSTr &= SSTr/(k - 1) = 3,575.065/(6 - 1) = 715.013 \\
F &= MSTr/MSE = 715.013/13.929 = 51.333
\end{aligned}$$

5.

We are given the following:

y = flow rate (m^3/min)
x = pressure drop (in. of water)
Linear regression line of $y = -.12 + .095x$ for $5 < x < 20$

$\beta_1 = 0.095$. We know this from our equation $y = \beta_0 + \beta_1 x + \epsilon$. This means that for every 1 inch of water increase in pressure drop, x, there is a $0.095 m^3$/min increase in the flow rate, y.

The change in flow rate associated when there is no pressure is calculated when $x = 0$. Our flow rate, y, can be calculated from plugging x into the regression line. $y = -.12 + (.095)(0)$. The flow rate, y, is equal to -$0.12 m^3$/min which does not make sense because the flow rate cannot be negative. This behavior is due to the extrapolation of data, since the linear regression line was given for x-values between 5 and 20.

The expected flow rate for a pressure drop of 10 in. and 15 in. can be calculated in the same way. Since these values are interpolated, they can be an accurate estimator for flow rate. We can calculate these values with $y = -.12 + (.095)(10)$ and $y = -.12 + (.095)(15)$. The flow rate for a pressure drop of 10 in. is $0.83 m^3$/min and $1.305 m^3$/min for a pressure drop of 15 in.