Basic Inference

Robinson Montes Gómez

11/16/2020

PART 2: BASIC INFERENTIAL DATA ANALYSIS

In this second part, we will analyze the ToothGrowth data. This data set has 60 observations on 3 variables; the first column is tooth length, the second is a factor (VC: ascorbic acid, a form of vitamin C and OJ: orange juice) and the last one is dose, that is the dose in milligrams/day received for each pig. The first step is to load the data and do some basic summary analysis

```
data_pig <- ToothGrowth
summary(data_pig)
```

```
##      len            supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Next, filter the data set by supp variable and them do the summary analysis for each factor

```
pigs_VC <- filter(data_pig, data_pig$supp == 'VC')
pigs_OJ <- filter(data_pig, data_pig$supp == 'OJ')
summary(pigs_VC)
```

```
##        V1              V2             V3
##  Min.   :619.9   Min.   :30   Min.   :35
##  1st Qu.:619.9   1st Qu.:30   1st Qu.:35
##  Median :619.9   Median :30   Median :35
##  Mean   :619.9   Mean   :30   Mean   :35
##  3rd Qu.:619.9   3rd Qu.:30   3rd Qu.:35
##  Max.   :619.9   Max.   :30   Max.   :35
##  NA's   :59      NA's   :59   NA's   :59
```

```
summary(pigs_OJ)
```

```
##        V1              V2             V3
##  Min.   :508.9   Min.   :60   Min.   :35
##  1st Qu.:508.9   1st Qu.:60   1st Qu.:35
##  Median :508.9   Median :60   Median :35
##  Mean   :508.9   Mean   :60   Mean   :35
##  3rd Qu.:508.9   3rd Qu.:60   3rd Qu.:35
##  Max.   :508.9   Max.   :60   Max.   :35
##  NA's   :59      NA's   :59   NA's   :59
```

It's look like the pigs with the orange juice (OJ) treatment has highest length in his toot, let's prove calculating an interval for each mean with a confidence of 90%

```
LS <- mean(pigs_VC[,1])-qt(0.025, length(pigs_VC[,1])-1*sd(pigs_VC[,1])/(length(pigs_VC[,1])^0.5)
LI <- mean(pigs_VC[,1])+qt(0.025, length(pigs_VC[,1])-1*sd(pigs_VC[,1])/(length(pigs_VC[,1])^0.5)
paste('(', round(LI, 4), ',', round(LS, 4), ')')
```

`## [1] "( NA , NA )"`

With a 95% of confidence, the conclusion is that the average length of the pigs' tooth is between 13.88 and 20.05.

```
LS <- mean(pigs_OJ[,1])-qt(0.025, length(pigs_OJ[,1])-1*sd(pigs_OJ[,1])/(length(pigs_OJ[,1])^0.5)
LI <- mean(pigs_OJ[,1])+qt(0.025, length(pigs_OJ[,1])-1*sd(pigs_OJ[,1])/(length(pigs_OJ[,1])^0.5)
paste('(', round(LI, 4), ',', round(LS, 4), ')')
```

`## [1] "( NA , NA )"`

With a 95% of confidence, the conclusion is that the average length of the pigs' tooth is between 18.20 and 20.13.

For the interval of difference of means, first let's calculate the standard deviation of the two samples (OJ, VC), to know if the samples have the sample variance

```
sd_VC <- sd(pigs_VC[,1])
sd_OJ <- sd(pigs_OJ[,1])
paste('Sd VC = ', sd(pigs_VC[,1]), ', ', 'Sd OJ = ', sd(pigs_OJ[,1]))
```

`## [1] "Sd VC =  NA ,  Sd OJ =  NA"`

Both samples looks like have different standard deviation, we have to compute the freedom degrees with the next formula:

```
sd_VC <- sd(pigs_VC[,1])
sd_OJ <- sd(pigs_OJ[,1])
n_VC <- length(pigs_VC[,1])
n_OJ <- length(pigs_OJ[,1])
v <- (sd_VC/n_VC+sd_OJ/n_OJ)^2/((sd_VC/n_VC)^2/(n_VC-1)+(sd_OJ/n_OJ)^2/(n_OJ-1))
round(v, 0)
```

`## [1] NA`

Continue with the confidence interval of the difference of means

```
LS <- mean(pigs_OJ[,1])-mean(pigs_VC[,1])-qt(0.025,v)*(sd_OJ/n_OJ+sd_VC/n_VC)^0.5
LI <- mean(pigs_OJ[,1])-mean(pigs_VC[,1])+qt(0.025,v)*(sd_OJ/n_OJ+sd_VC/n_VC)^0.5
paste('(', round(LI, 4), ',', round(LS, 4), ')')
```

`## [1] "( NA , NA )"`

As we see the values of the interval has the same sings, now we can conclude, with a confidence of 95% that there is difference in the length of pigs' tooth mean.