Statistical Inference Course Project

luis manuel

10/01/2021

Basic Inferential Data Analysis

now we are going to analyze the data set ToothGrowth, len represents the length of the tooth, supp represents the type of supplement: orange juice(OJ) or ascorbic acid(VC), and dose represents the dose in milligrams per day(0,5, 1 y 2)

```
data(ToothGrowth)
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

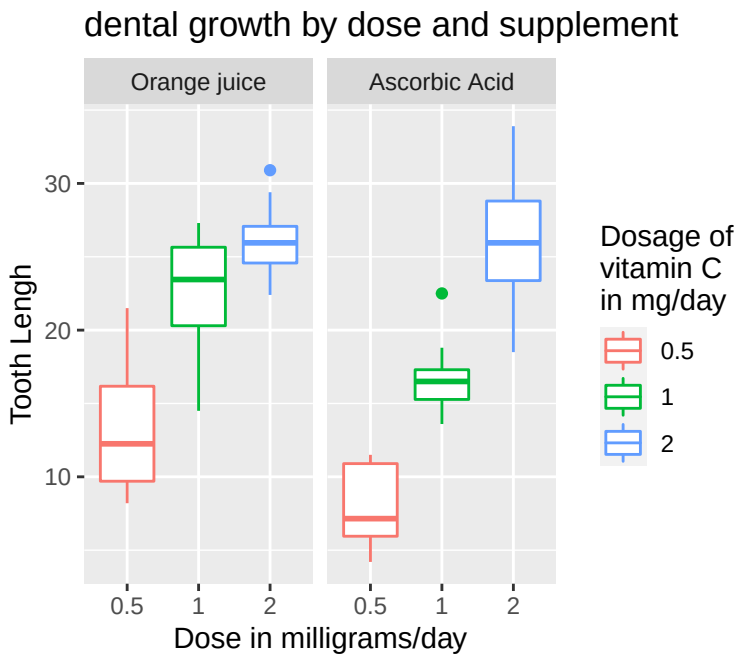
now we are going to analyze the mean and variance of the variable lan for each subgroup of supplement and dose

```
library(dplyr)
ToothGrowth %>% group_by(supp,dose)%>%summarise(mean_len=mean(len),var_len=var(len))
```

```
## # A tibble: 6 x 4
## # Groups:   supp [2]
##   supp   dose mean_len var_len
##   <fct> <dbl>    <dbl>   <dbl>
## 1 OJ      0.5    13.2    19.9
## 2 OJ      1      22.7    15.3
## 3 OJ      2      26.1     7.05
## 4 VC      0.5     7.98    7.54
## 5 VC      1      16.8     6.33
## 6 VC      2      26.1    23.0
```

we can see that the mean is directly proportional to the dose, while the variance for the supplement vc increases as the dose increases, and for the supplement oj it decreases as the dose increases.

```
library(ggplot2)
g<-ggplot(ToothGrowth, aes(x=factor(dose),y= len, color = factor(dose)))
g<-g+geom_boxplot()
      # facet_grid(.~supp)+
g<-g+ facet_grid(.~supp, labeller = as_labeller(
            c("OJ" = "Orange juice",
              "VC" = "Ascorbic Acid")))
g<-g+labs(title = "dental growth by dose and supplement",
           x = "Dose in milligrams/day",
           y = "Tooth Lengh")
g<-g+scale_color_discrete(name = "Dosage of\nvitamin C\nin mg/day")
g
```



so we can confirm what was said above, now we are going to make confidence intervals to compare the means, now we are going to do a hypothesis test for the difference of means between each type of dose per supplement

```
t1<-t.test(len ~ supp, subset(ToothGrowth,dose==0.5), var.equal = FALSE)
t2<-t.test(len ~ supp, subset(ToothGrowth,dose==1), var.equal = FALSE)
t3<-t.test(len ~ supp, subset(ToothGrowth,dose==2), var.equal = FALSE)

HT <- data.frame(p_value = c(t1$p.value, t2$p.value, t3$p.value),
                 Conf.Low = c(t1$conf.int[1],t2$conf.int[1], t3$conf.int[1]),
                 Conf.High = c(t1$conf.int[2],t2$conf.int[2],t3$conf.int[2]),
                 accepted.hypothesis=c(if(t1$p.value<0.05)"Ha"else"H0",
                                       if(t2$p.value<0.05)"Ha"else"H0",
                                       if(t3$p.value<0.05)"Ha"else"H0"),
                 row.names = c("Dosage .05","Dosage 1","Dosage 2"))
HT
```

```
##                p_value  Conf.Low Conf.High accepted.hypothesis
```

```
## Dosage .05 0.006358607  1.719057  8.780943                  Ha
## Dosage 1   0.001038376  2.802148  9.057852                  Ha
## Dosage 2   0.963851589 -3.798070  3.638070                  H0
```

So we can conclude that there is not enough statistical evidence to say that the length means using a dose of $2\frac{mg}{day}$ are different, while for $1\frac{mg}{day}$ and $0.5\frac{mg}{day}$ the means are different, for these last two cases it could be said that the mean of OJ is greater than that of VC