Statiscal Inference project part 2

Nguyen Son Linh

5/29/2020

Overview

In this part, we try to infer statistically with the ToothGrowth dataset

Setup

Packages needed

```
library(ggplot2)
```

Part 2: Basic Inferential Statistics instructions

2.1 Load the data

```
data("ToothGrowth")
```

As seen, the sample mean is close to the theoretical mean

2.2 Basic summary and exploratory data analysis.

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
summary(ToothGrowth)
```

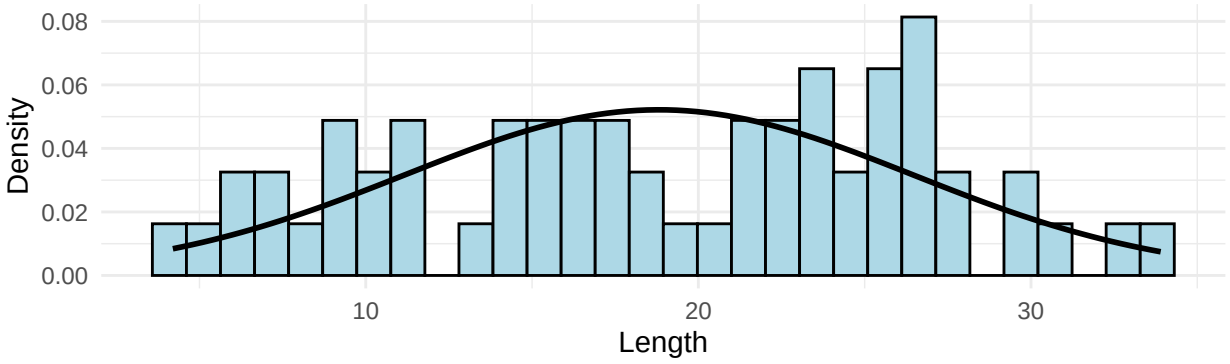
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

With regard to the R documentation, the data consists of length of odontoblasts in 60 guinea pigs. 10 guinea pigs were assigned to each group, which receive one of three dose levels of vitamin C, delivered through either ascorbic acid of orange juice

We make the histogram of the len variable

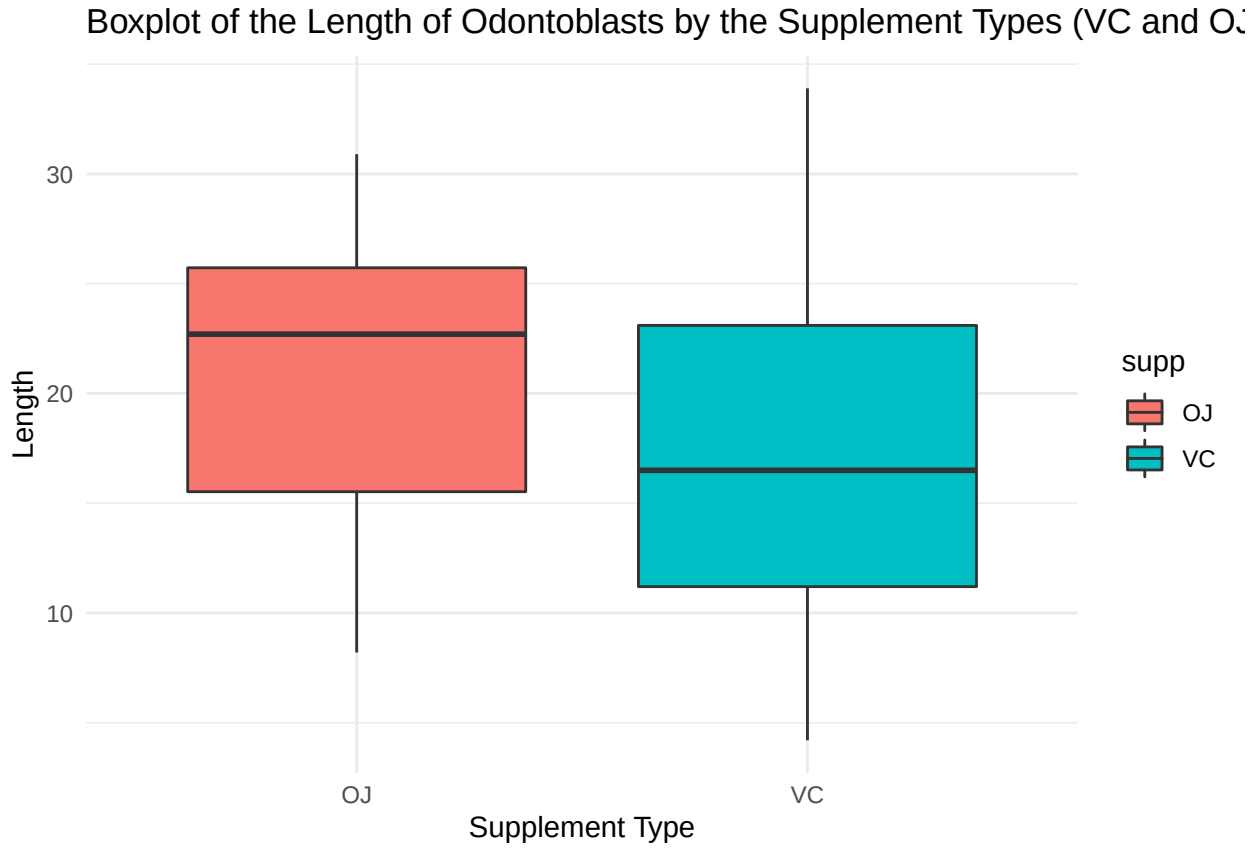
```
g <- ggplot(ToothGrowth,aes(len))
g + geom_histogram(aes(y = ..density..),colour="black",fill="lightblue") +
    stat_function(fun=dnorm,args=list( mean=mean(ToothGrowth$len), sd=sqrt(var(ToothGrowth$len))),geom=
    scale_x_continuous("Length")+
    ylab("Density") +
    theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

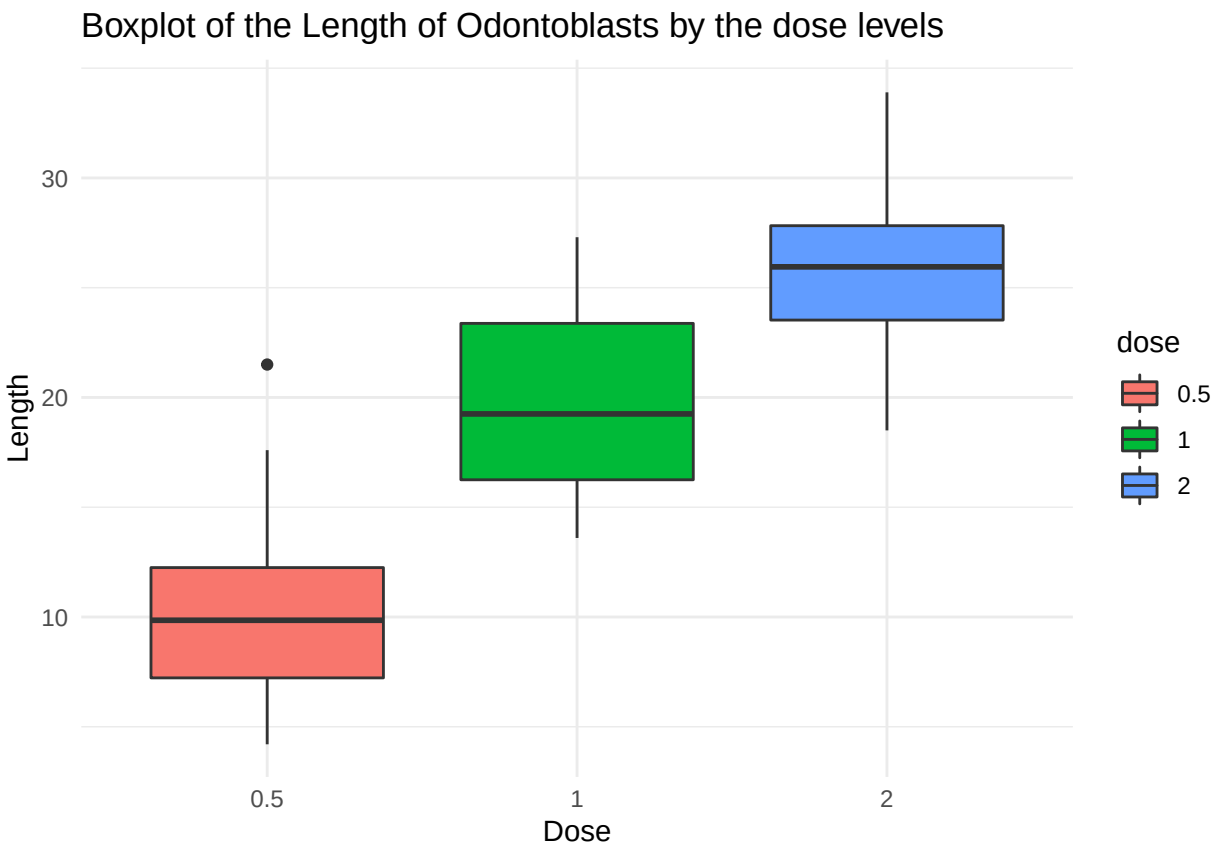


The shape difference between the histogram and the calculated normal distribution may be explained by the dose levels difference and the supplement. Hence we explore the relationship of len and other variables.

```
ToothGrowth$supp <- as.character(ToothGrowth$supp)
g <- ggplot(ToothGrowth,aes(supp,len))
g+geom_boxplot(aes(fill=supp))+
    ggtitle("Boxplot of the Length of Odontoblasts by the Supplement Types (VC and OJ)") +
    xlab("Supplement Type") +
    ylab("Length") +
    theme_minimal()
```




```
ToothGrowth$dose <- as.character(ToothGrowth$dose)
g <- ggplot(ToothGrowth,aes(dose,len))
g + geom_boxplot(aes(fill=dose)) +
    ggtitle("Boxplot of the Length of Odontoblasts by the dose levels") +
    xlab("Dose") +
    ylab("Length")  +
    theme_minimal()
```



The two plot depicts that orange juice has a higher median, while three doses have the same variability, with median increasing as dose levels increase.

2.3 Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering).

Consider these hypotheses:
The null hypothesis $H_0 : \mu = 0$ is that population mean difference (μ) is not different of the length by the supp, given a dose level. The alternative hypothesis $H_1 : \mu \neq 0$ is that there is a population mean difference of length by supp, given a dose level.

Here we would use the t-distribution as the population standard deviation is unknown and there are 60 observations from 60 different guinea pigs

```
dose_levels <- levels(factor(ToothGrowth$dose))
reject <- " "
notreject <- " "
for (level in dose_levels){
    result<-t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == level, ])
    print(paste("For dose",as.character(level)," the t.test result is: "))
    print(result)
    ifelse(result$p.value < 0.05,
           print(paste("Since p-value < 5%, reject null hypothesis for dose level",as.character(level))
           print(paste("Since p-value > 5%, fail to reject null hypothesis for dose level",as.character
    "\n"
    ifelse(result$p.value < 0.05,
```

```
            reject <- paste(reject, " and ", as.character(level)),
            notreject <- paste(notreject, " and ", as.character(level)))

}
```

[1] “For dose 0.5 the t.test result is:”

```
Welch Two Sample t-test
```

data: len by supp t = 3.1697, df = 14.969, p-value = 0.006359 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 1.719057 8.780943 sample estimates: mean in group OJ mean in group VC 13.23 7.98

[1] “Since p-value $< 5\%$, reject null hypothesis for dose level 0.5” [1] “For dose 1 the t.test result is:”

```
Welch Two Sample t-test
```

data: len by supp t = 4.0328, df = 15.358, p-value = 0.001038 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 2.802148 9.057852 sample estimates: mean in group OJ mean in group VC 22.70 16.77

[1] “Since p-value $< 5\%$, reject null hypothesis for dose level 1” [1] “For dose 2 the t.test result is:”

```
Welch Two Sample t-test
```

data: len by supp t = -0.046136, df = 14.04, p-value = 0.9639 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -3.79807 3.63807 sample estimates: mean in group OJ mean in group VC 26.06 26.14

[1] “Since p-value $> 5\%$, fail to reject null hypothesis for dose level 2”

2.4 Conclusion

- In the initial exploratory data analysis, the dose levels were apparently a factor boosting tooth growth, whereas the supplement type seems unlikely.
- After conducting the t-test for supplement types at each dose level, we find a strong evidence against the null hypothesis at dose level 0.5 and 1. At dose level 2 that did not happen.
- Assume independent, normally distributed data