Statistical Inference Course Project Part 2: Basic Inferential Data Analysis

Daniela Gracia

Synopsis

In this project we will analyze the ToothGrowth data from the R datasets package. We will **1)** perform some basic exploratory data analysis and provide a basic summary of the data **2)** Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose and **3)** state our conclusions. Our main goal is to answer the question **¿Does delivery method and/or dosage affect tooth growth in guinea pigs?**

Data Processing

In this section we will **1)** load the data and perform **2)** exploratory data analysis and a summary of the data.

Loading the data First we load the data:

```
# load the necessary library and data
library(datasets)
data("ToothGrowth")
```

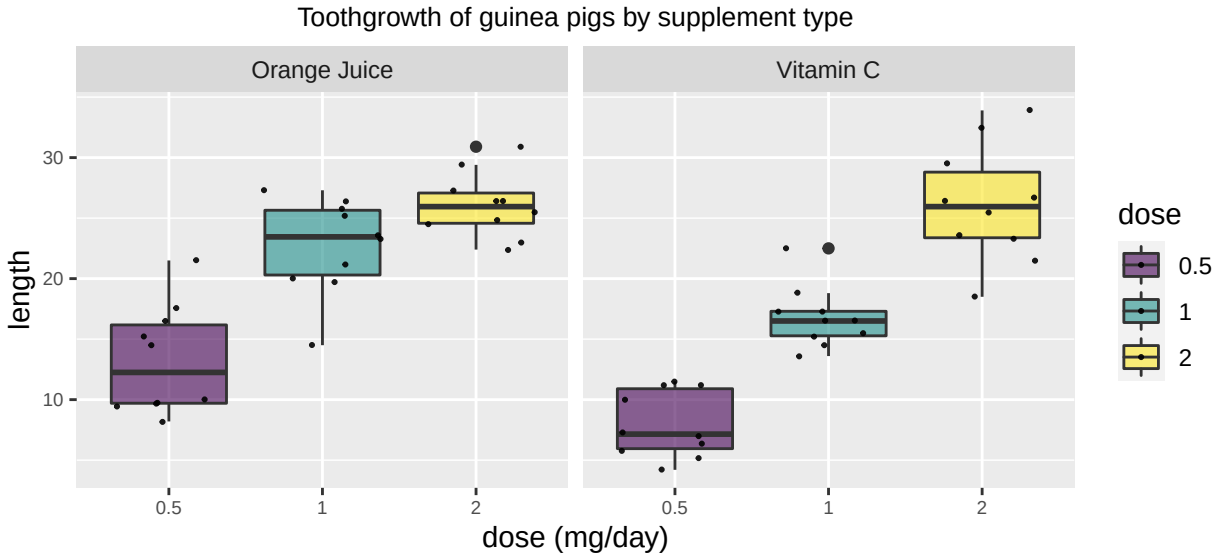
Exploratory data analisys and summary This data set documents the effect of vitamin C on tooth growth in guinea pigs. Lets take a look at its structure:

```
## '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 ...
```

We can see it has 60 observations for 3 variables: *len*, *sup* and *dose*: **len:** numeric, tooth length. **sup:** factor, supplement type, VC (ascorbic acid - vitamin c) or OJ (orange juice). **dose:** numeric, dose of the supplement in milligrams per day. Now we will create a table of sup vs dose to better understand the structure of the data:

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

We can see that for each of the two supplement types, 3 different doses were administered. For a total of ten subjects for each combination and a total of 60 subjects. Now to visualize the distribution of the data we will plot dose vs len for each of the supplement types. We observe two main things: 1) the higher the dose the higher the tooth growth for both supplement types, 2) there seems to be more tooth growth for the orange juice supplement than for the vitamin C supplement. Finally we will create tables to show **A)** Average tooth length for each supplement and **B)** Average tooth length for each dosage.




```
##   supplement average length
## 1         OJ       20.66333
## 2         VC       16.96333
```

```
##   dose (mg/day) average length
## 1           0.5         10.605
## 2             1         19.735
## 3             2         26.100
```

Data Analysis

In this section we will use hypothesis testing to study tooth growth by supp and dose. We will do this by assuming all of our variables are IID random variables, we also assume that the distribution of their averages follows a normal distribution and that the groups have different variance.

Tooth growth by supplement We know that the mean tooth length for orange juice (OJ) is *20.66* and that the mean tooth length for vitamin c (VC) is *16.96*, knowing this we will test the following null and alternative hypothesis for both groups (group 1 = OJ, group 2 = VC):

Null Hypothesis: The type of supplement does not have an effect on tooth growth: $H_0 : \mu_1 = \mu_2$.
Alternative Hypothesis: The type of supplement has an effect on tooth growth: $H_a : \mu_1 \neq \mu_2$.
Now we will use the t.test() function to test this.

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

From this results and confidence interval we see that the p-value (*6.06%*) is higher than the type I error rate $\alpha = 5\%$ and the confidence interval contains zero which means there is a possibility that $\mu_1 - \mu_2 = 0$. Therefore we fail to reject the null hypothesis and conclude that the type of supplement does not have an effect on tooth growth for this data. We will know run the same test using only the information for 0.5 and 1 mg/day dosage options.

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.0503, df = 36.553, p-value = 0.004239
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  1.875234 9.304766
## sample estimates:
## mean in group OJ mean in group VC
##           17.965           12.375
```

In this case the p-value (*0.42%*) is much smaller, and significantly less than α and the confidence interval does not contain zero. Therefore in this case we reject the null hypothesis and accept the alternative hypothesis $H_a : \mu_1 \neq \mu_2$. We can conclude that the type of supplement does have an effect on tooth growth and particularly the orange juice supplement has a higher impact.

Tooth growth by dosage We know that the mean tooth growth for 0.5 mg dose is *10.61*, for 1 mg dose it is *19.74*, and for the 2 mg dose it is *26.1*. Knowing this we will test the following null and alternative hypothesis for both groups (group 1 = 0.5 mg dose, group 2 = 1 mg dose and group 3 = 2 mg dose):

Null Hypothesis: The dosage does not have an effect on tooth growth: $H_0 : \mu_1 = \mu_2;\ \mu_2 = \mu_3;\ \mu_1 = \mu_3$.
Alternative Hypothesis: The dosage has an effect on tooth growth: $H_a : \mu_3 > \mu_2 > \mu_1$.

First we test $H_a : \mu_2 > \mu_1$

```
##
##  Welch Two Sample t-test
##
## data:  half_dose$len and one_dose$len
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Second we test $H_a : \mu_3 > \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  one_dose$len and two_dose$len
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Finally we test $H_a : \mu_3 > \mu_1$

```
##
##  Welch Two Sample t-test
##
## data:  half_dose$len and two_dose$len
## t = -11.799, df = 36.883, p-value = 2.199e-14
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -13.27926
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

For all the tests above the p-value is near zero, the t-statistic is greater than the .975th quantile (*0.83*) and the confidence interval does not contain zero, therefore we reject the null hypothesis. We can conclude that we accept the alternative hypothesis $H_a : \mu_3 > \mu_2 > \mu_1$ and therefore dosage does have an effect on tooth growth.

Conclusions

Assuming our variables are IID and behave as predicted by the CLT. Our analysis showed the following: **1)** Dosage has an effect on tooth growth, specifically the higher the dosage the higher the tooth length. **2)** Delivery method has an effect on tooth growth, specifically orange juice is more effective than vitamin c in the 0.5 and 1 mg dosage.

Appendix

Below is all the code used to make calculations and create figures in this report.

```
# first we see the structure of the data
str(ToothGrowth)
```

```
# create table
table(ToothGrowth$supp,ToothGrowth$dose)
```

```
# load packages
library(viridis)
library(ggplot2)

# transform dose variable into a factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# new facet label names for supp variable
supp.labs <- c("Orange Juice", "Vitamin C")
names(supp.labs) <- c("OJ", "VC")

# create boxplot divided in facets by supplement type using ggplot2
plot1 <- ggplot(ToothGrowth, aes(x=dose, y=len, fill=dose)) +
        geom_boxplot() +
        facet_grid(.~supp, labeller = labeller(supp = supp.labs)) +
        geom_jitter(color="black", size=0.4, alpha=0.9) +
        scale_fill_viridis(discrete = TRUE, alpha=0.6) +
        labs(x="dose (mg/day)",y="length",title="Toothgrowth of guinea pigs by supplement type") +
        theme(plot.title = element_text(hjust = 0.5, size =13),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10))
plot1
```

```
# load necessary packages
library(dplyr)
library(knitr)

# create table of supplement and average length
supplement <- ToothGrowth %>% group_by(supp) %>%
        summarise(average_length = mean(len))
colnames(supplement) <- c("supplement","average length")
supplement <- as.data.frame(supplement)

# create table of dosage and average length
dosage <- ToothGrowth %>% group_by(dose) %>%
        summarise(average_length=mean(len))
colnames(dosage) <- c("dose (mg/day)","average length")
dosage <- as.data.frame(dosage)

# calculate means
meanoj <- round(supplement[[1,2]], digits = 2)
meanvc <- round(supplement[[2,2]], digits = 2)
half <- round(dosage[[1,2]], digits = 2)
one <- round(dosage[[2,2]], digits = 2)
two <- round(dosage[[3,2]], digits = 2)

# display tables
supplement
dosage
```

```r
# run t.test()
t.test(data=ToothGrowth, alternative = "two.sided",len~supp, paired=FALSE, var.equal=FALSE)
pval1 <- round(t.test(data=ToothGrowth, alternative = "two.sided",len~supp, paired=FALSE, var.equal=FALSE)$p.
```

```r
# run t.test()
exclude_2dose <- filter(ToothGrowth,dose!="2")
t.test(data=exclude_2dose, alternative = "two.sided",len~supp, paired=FALSE, var.equal=FALSE)
pval2 <- round(t.test(data=exclude_2dose, alternative = "two.sided",len~supp, paired=FALSE, var.equal=FALSE)$
```

```r
# create a data frame for each dosage
half_dose <- filter(ToothGrowth,ToothGrowth$dose=="0.5")
one_dose <- filter(ToothGrowth,ToothGrowth$dose=="1")
two_dose <- filter(ToothGrowth,ToothGrowth$dose=="2")

#run t.test()
t.test(half_dose$len, one_dose$len, paired=FALSE, var.equal=FALSE,alternative ="l" )
```

```r
#run t.test()
t.test(one_dose$len, two_dose$len, paired=FALSE, var.equal=FALSE,alternative ="l" )
```

```r
#run t.test()
t.test(half_dose$len, two_dose$len, paired=FALSE, var.equal=FALSE,alternative ="l" )
quantile_95<-round(pt(.975,37), digits=2)
```