SI_PA1_template_part2.Rmd

abhibhaku

April 27, 2018

TITLE - Basic Inferential Data Analysis Instructions

OVERVIEW - Goal is to analyze the ToothGrowth data in the R datasets package

```
knitr::opts_chunk$set(echo = TRUE)

# Q1 - Load the ToothGrowth data and perform some basic exploratory data analysis

library(datasets)
library(ggplot2)
data("ToothGrowth")

str(ToothGrowth) # checking the dataset and class of variables
```

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

```
unique(ToothGrowth$len) # checking for unique values in variable len
```

```
##  [1]  4.2 11.5  7.3  5.8  6.4 10.0 11.2  5.2  7.0 16.5 15.2 17.3 22.5 13.6
## [15] 14.5 18.8 15.5 23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5 17.6
## [29]  9.7  8.2  9.4 19.7 20.0 25.2 25.8 21.2 27.3 22.4 24.5 24.8 30.9 29.4
## [43] 23.0
```

```
unique(ToothGrowth$supp) # checking for unique values in variable supp
```

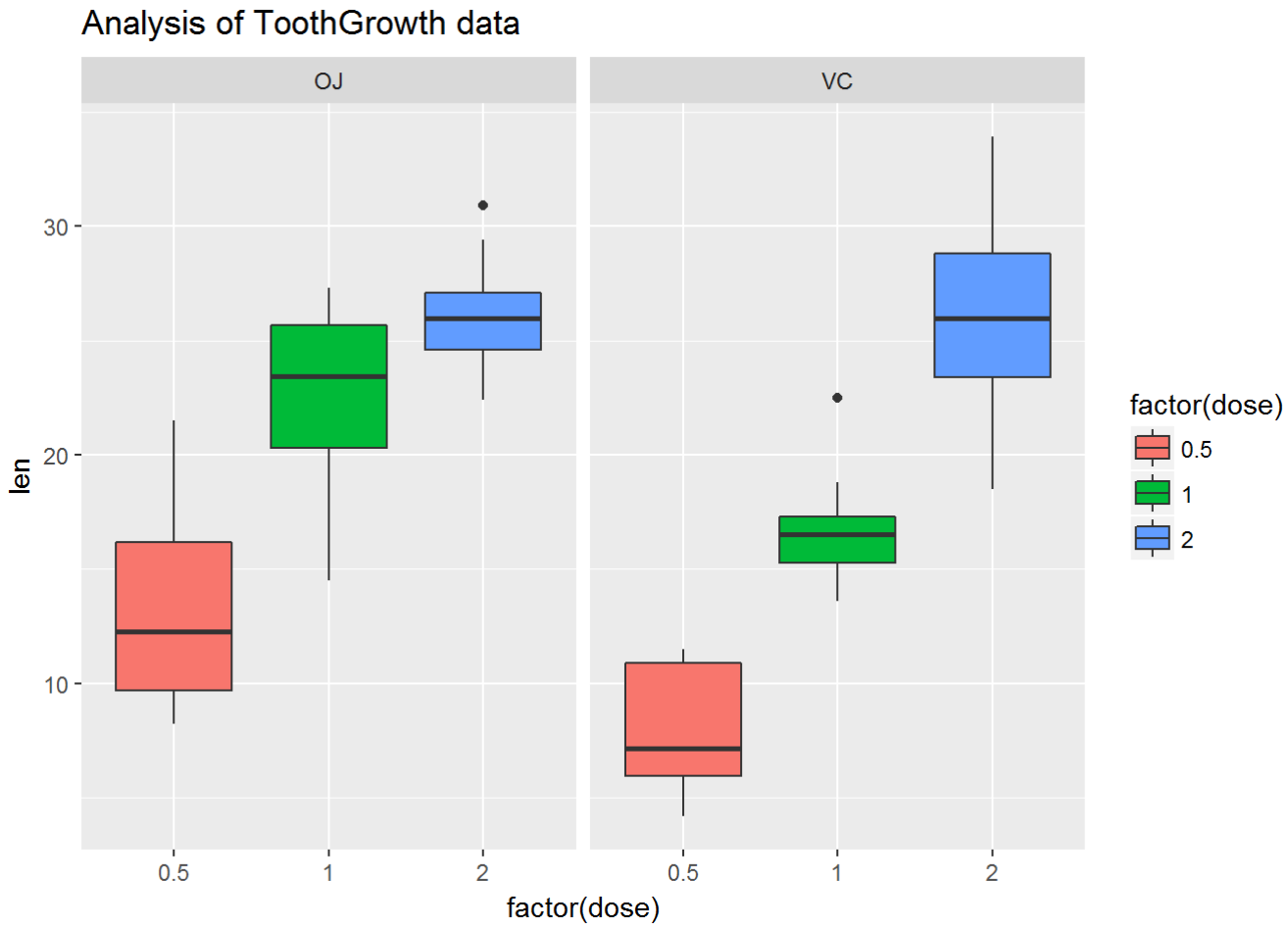
```
## [1] VC OJ
## Levels: OJ VC
```

```
unique(ToothGrowth$dose) # checking for unique values in variable dose
```

```
## [1] 0.5 1.0 2.0
```

```
# plotting boxplot for the dataset

ggplot(ToothGrowth, aes(x=factor(dose), y=len, fill=factor(dose))) + geom_boxplot()+facet_gri
d(.~supp) + ggtitle("Analysis of ToothGrowth data")
```




```
# Result for Q1 - The plot indicates that high length is associated with high dosage for both
 OJ & VC

# Q2 - Provide a basic summary of the data

summary(ToothGrowth) # summarizing each variable of the dataset
```

```
##       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
```

```
# Q3 - Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and d
ose (Only use the techniques from class, even if there's other approaches worth considering)

# Comparing toothgrowth by supplement using a t-test

# Null Hypothesis - Supplement type has no effect on tooth growth

t.test(len~supp,data = ToothGrowth)
```

```
##
##  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
```

```
# p-value comes out to be 0.06, which is greater than 0.05 & the confidence interval contains
 0 - thus, we can't reject the null hypthesis

# Comparing toothgrowth by dosage using a t-test

# Null Hypothesis - higher dosage level is associated with increased tooth growth

# First subsetiing the data into 3 groups as per the dosage levels

Tg1 <- subset(ToothGrowth,dose %in% c(0.5,1.0))
Tg2 <- subset(ToothGrowth,dose %in% c(0.5,2.0))
Tg3 <- subset(ToothGrowth,dose %in% c(1.0,2.0))

# running t-test for dosage 0.5 & 1.0, 0.5 & 2.0 and 1.0 & 2.0

t.test(len~dose,data = Tg1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
t.test(len~dose,data = Tg2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

```
t.test(len~dose,data = Tg3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

```
# For all 3 cases, p-value < 0.05 and zero doesn't lie in their confidence interval range - t
hus, we can reject the null hypthesis

# Conclusions -

# 1) Supplement type has no effect on tooth growth
# 2) High dosage levels lead to increased tooth growth

# Assumptions -

# 1) Variables are independent and identically distributed (i.i.d.)
# 2) Variances of tooth growth are different when using different supplement and dosage
# 3) Sample dataset is represntative of the population and follows a normal distribution
```