Statistical Inference Course Project - Basic inferential data analysis

Yuan Liao

8/26/2020

Overview

This is an R Markdown document for Statistical Inference Course Project - part 2. Here, we're going to analyze the ToothGrowth data in the R datasets package.

```
# Load necessary libs
library(ggplot2)
```

Load data for exploratory data analysis

Load data and take a look.

```
# Load necessary libs
data('ToothGrowth')
summary(ToothGrowth)
```

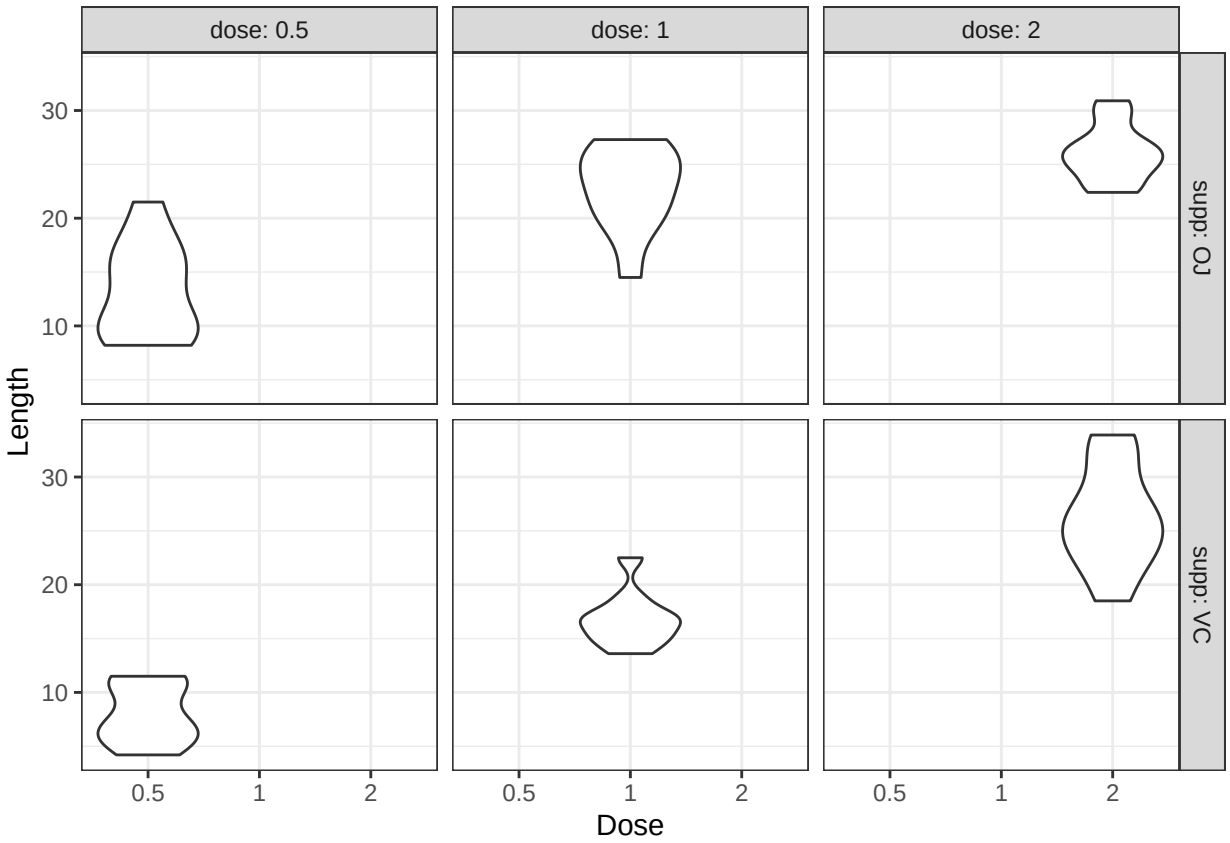
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Show the unique values of dose which looks like a categorical variable.

```
print(unique(ToothGrowth$dose))
```

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

```
# Take a look at the data
ggplot(ToothGrowth, aes(x=factor(dose, levels = c(0.5, 1.0, 2.0)),
                        y=len,
                        group=factor(dose, levels = c(0.5, 1.0, 2.0)))) +
  geom_violin(aes(fill=len)) +
  facet_grid(supp ~ dose, labeller=label_both) +
  labs(x='Dose', y='Length') +
  theme_bw()
```



Compare tooth growth by supp and dose

Two-sample t test is applied to compare the impact of supp and dose on len with confidence interval = 95%.

Hypothesis 1: two supplement groups have different distributions of tooth length

```
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
```

It turns out that p-value is 0.06 which is greater than 0.05. Therefore, the hypothesis is rejected.

Hypothesis 2: dose = 0.5 and does = 1 have different distributions of tooth length

```
t.test(len~dose,data=ToothGrowth[ToothGrowth$dose %in% c(0.5, 1),])
```

```
##
##  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
```

It turns out that p-value < 0.001. Therefore, the hypothesis is accepted.

Hypothesis 3: dose = 1 and does = 2 have different distributions of tooth length

```
t.test(len~dose,data=ToothGrowth[ToothGrowth$dose %in% c(1, 2),])
```

```
##
##  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
```

It turns out that p-value < 0.001. Therefore, the hypothesis is accepted.

Conclusions

With the below assumptions:

1. The analyzed data respresent the true population.
2. Multi-level comparison is approximated by a series of two-sample t tests.
3. Ignore the interaction between supplement and dose on tooth length.
4. Sample/population data follow normal distribution.

There is no effect of supplement on tooth length. The higer the dose, the greater the tooth length.