Statistical Inference Course Project

Aiyu Li

01/16/2021

Synopsis

This is a project for Statistical Inference class. There are two parts in this project.

- Part 1: simulation Exercise: investigate the distribution of averages of 40 exponentials and a thousand simulations in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations.
- Part 2: analyze the ToothGrowth data in the R datasets package; provide a basic summary of the data; Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose and provide the compared conclusions.

R environment and reproducibility

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18363)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.6.3  magrittr_1.5    tools_3.6.3     htmltools_0.5.0
##  [5] yaml_2.2.1      stringi_1.4.6   rmarkdown_2.3   knitr_1.30
##  [9] stringr_1.4.0   xfun_0.17       digest_0.6.25   rlang_0.4.7
## [13] evaluate_0.14
```

```r
set.seed(2021)
```

Load required libraries

```r
library(stats)
library(ggplot2)
library(sqldf)
```

```
## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite
```

```r
library(dplyr, warn.conflicts = F)
```

Part 2 (P2) Analyze the tooth growth

P2 Objectives

- Load the ToothGrowth data and perform some basic exploratory data analyses.
- Provide a basic summary of the data.
- Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.
- State your conclusions and the assumptions needed for your conclusions

P2: Load data and basic exploratory analyses

```r
data(ToothGrowth)
str(ToothGrowth)
```

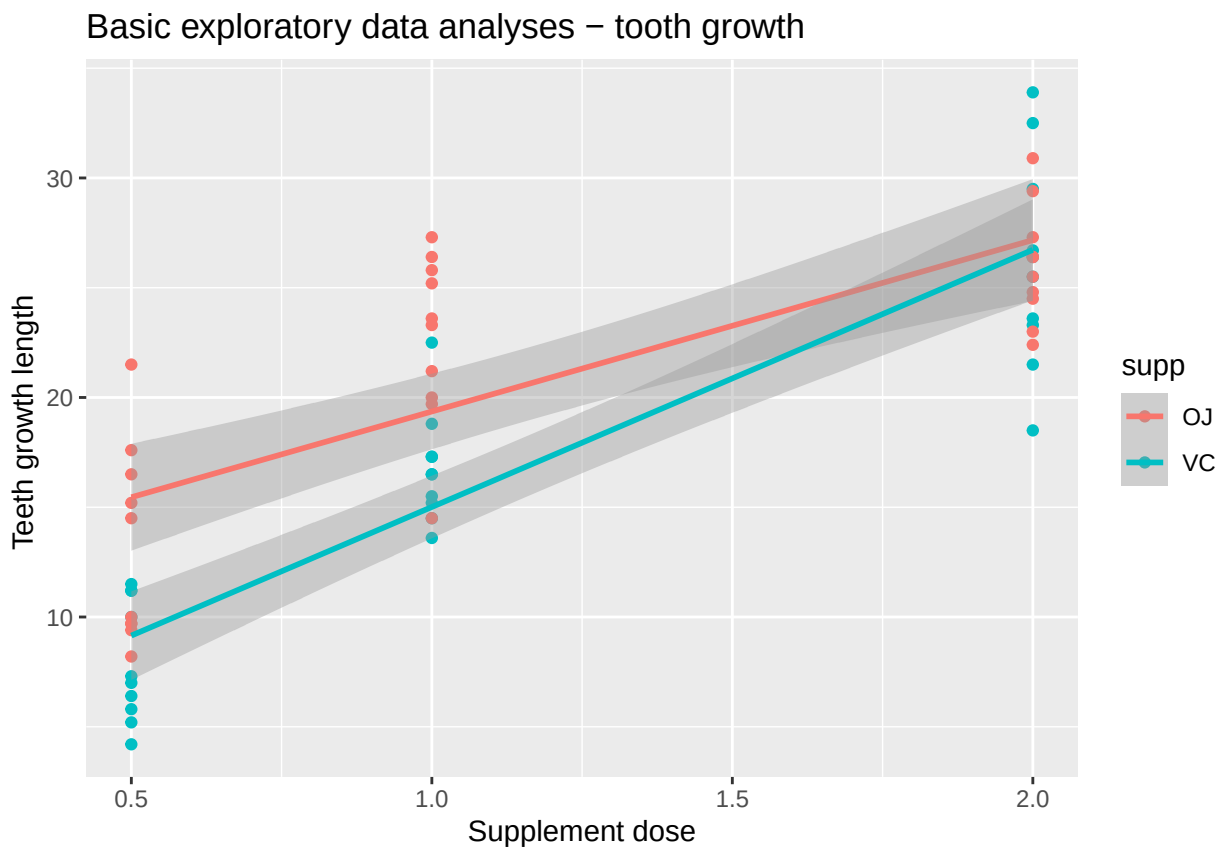
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```r
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
```

```
qplot(dose, len, data = ToothGrowth, color = supp, geom = "point") +
        geom_smooth(method = "lm") +
        labs(title = "Basic exploratory data analyses - tooth growth ") +
        labs(x = "Supplement dose", y = "Teeth growth length ")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Findings from the above plot:

1. When supplement dose increases, the teeth length goes up for both OJ and VC.
2. Supplement OJ has a higher length increase than VC at the lower dose.

P2: Provide a basic summary of the data

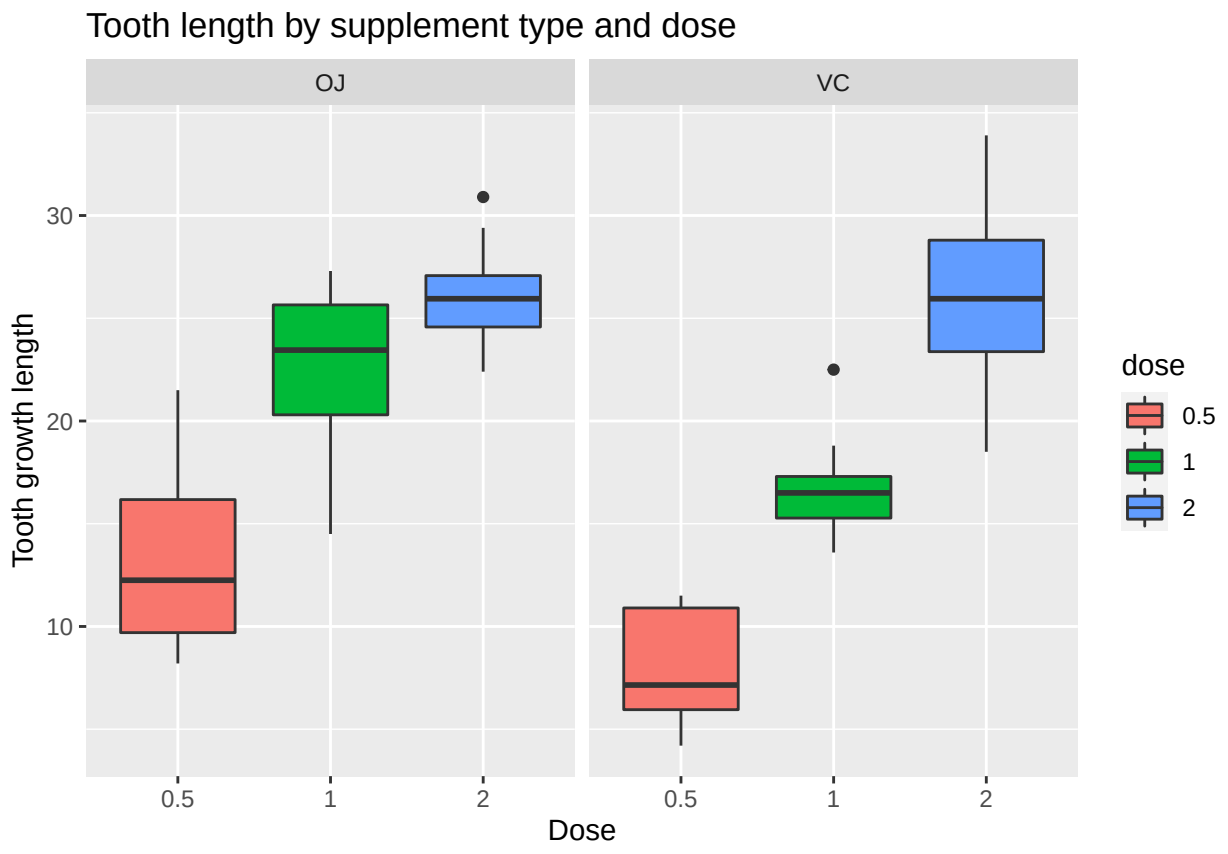
```
GroupSumStats <- sqldf('select supp, dose, max(len) as Max, min(len) as Min,
                        median(len) as Median, avg(len) as Mean,
                        round(stdev(len),2) as stdev, count(*) as n
                        from ToothGrowth group by supp, dose')
GroupSumStats
```

```
##   supp dose  Max  Min Median  Mean stdev  n
## 1   OJ  0.5 21.5  8.2  12.25 13.23  4.46 10
```

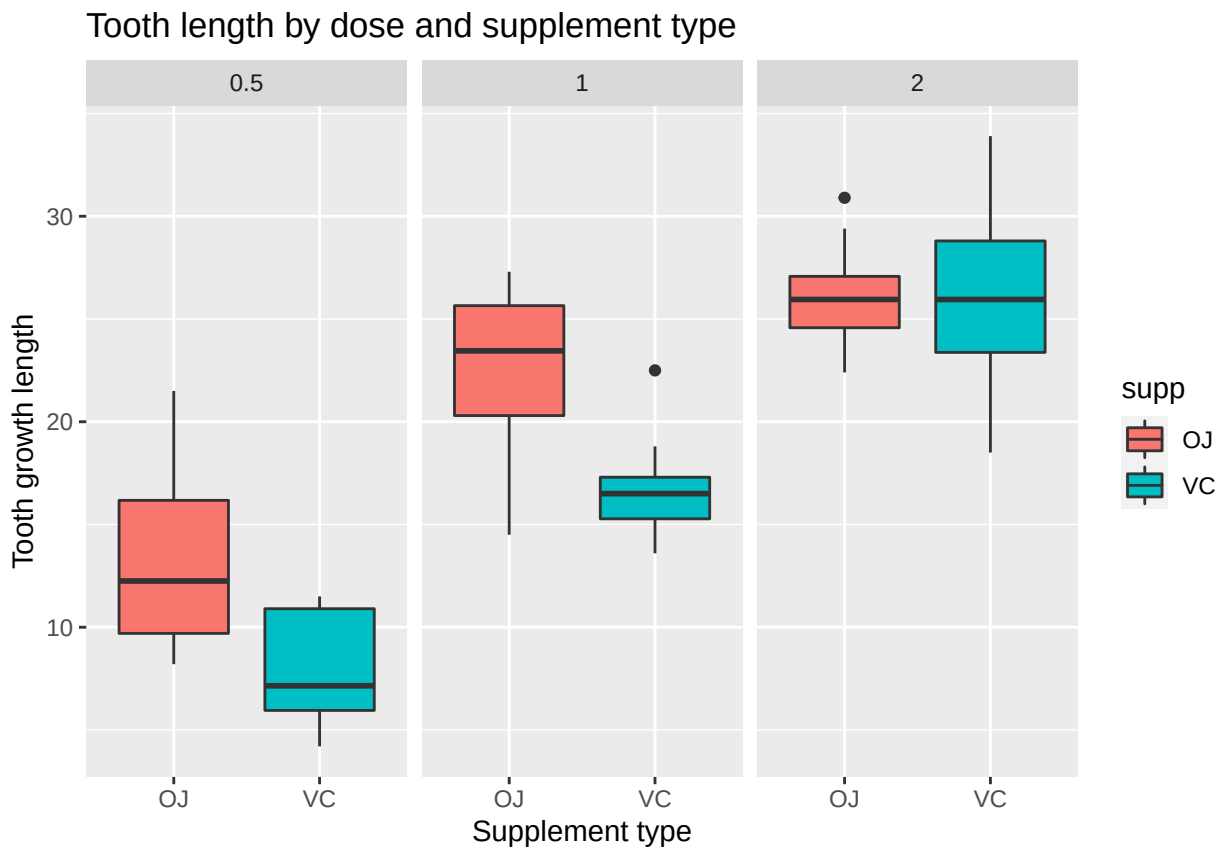
```
## 2   OJ  1.0 27.3 14.5  23.45 22.70  3.91 10
## 3   OJ  2.0 30.9 22.4  25.95 26.06  2.66 10
## 4   VC  0.5 11.5  4.2   7.15  7.98  2.75 10
## 5   VC  1.0 22.5 13.6  16.50 16.77  2.52 10
## 6   VC  2.0 33.9 18.5  25.95 26.14  4.80 10
```

```
#update dose as factor for analysis
ToothGrowth$dose<-as.factor(ToothGrowth$dose)

ggplot(aes(x=dose, y=len), data=ToothGrowth) +
        geom_boxplot(aes(fill=dose)) +
        xlab("Dose") +
        ylab("Tooth growth length") +
        facet_grid(~ supp) +
        ggtitle("Tooth length by supplement type and dose")
```




```
ggplot(aes(x=supp, y=len), data=ToothGrowth) +
        geom_boxplot(aes(fill=supp)) +
        xlab("Supplement type") +
        ylab("Tooth growth length") +
        facet_grid(~ dose) +
        ggtitle("Tooth length by dose and supplement type")
```



P2: Compare tooth growth by supplement type and dose

P2: tooth growth by supplement type

```
## OJ and VC data
OJ <- filter(ToothGrowth, supp == 'OJ')$len
VC <- filter(ToothGrowth, supp == 'VC')$len

## t- test by supplement type OJ and VC
t.test(OJ, VC, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## 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 of x mean of y
##  20.66333  16.96333
```

Note: the p-value of this test is 0.06, which is greater than 0.05 and the confidence interval of the test contains zero. therefore, the supplement types seems to have no impact on tooth growth based on this test result.

P2: tooth growth by dose

```
##  data by dose: 0.5,1.0 and  2.0
dose_0.5 <- filter(ToothGrowth, dose == 0.5)$len
dose_1.0 <- filter(ToothGrowth, dose == 1.0)$len
dose_2.0 <- filter(ToothGrowth, dose == 2.0)$len

## t- test by dose 0.5 and 1.0
t.test(dose_0.5, dose_1.0, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_0.5 and dose_1.0
## 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 of x mean of y
##    10.605    19.735
```

```
## t- test by dose 0.5 and 2.0
t.test(dose_0.5, dose_2.0, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_0.5 and dose_2.0
## 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 of x mean of y
##    10.605    26.100
```

```
## t- test by dose 1.0 and 2.0
t.test(dose_1.0, dose_2.0, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose_1.0 and dose_2.0
## 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 of x mean of y
##    19.735    26.100
```

Note: from above three tests: the p-value of each test was essentially zero and the confidence interval of each test does not cross over zero (0). Therefore, the average tooth length increases with an inceasing dose.

P2: Conclusion

- Supplement type has no effect on tooth growth.
- The dose level increasing leads to increased tooth growth.