Statistical Inference Course Project

Daniel Pont

2018-09-30

Synopsis

This report is the final product of the Statistical Inference Course Project. It's made up of two parts :

- a simulation exercise, that investigates the distribution of averages of 40 exponentials in R and shows it behaves as predicted by the Central Limit Theorem
- a basic inferential data analysis of the ToothGrowth dataset that shows that vitamin C increases tooth growth in guinea pigs

A) Simulation of the Central Limit Theorem with exponential distribution averages

1) Let's simulate 1000 exponential samples of n=40 values with parameter lambda=0.2

For each one we calculate its average value. Hence we get a distribution of 1000 averages "rexpAvgsDist".

```
lambda <- 0.2
n  <- 40
rexpAvgsDist = NULL
for (i in 1 : 1000) rexpAvgsDist = c(rexpAvgsDist, mean(rexp(n = n, rate=lambda)))
```

2) Now let's compare theoritical and simulation values for both mean and variance :

- **The theoritical mean for the distribution of averages**
  as predicted by the Central Limit Theorem is the mean of the exponential :

```
1/lambda
```

```
## [1] 5
```

- **The simulation mean is** :

```
mean(rexpAvgsDist)
```

```
## [1] 4.990025
```

- **The theoritical variance for the distribution of averages**
  as predicted by the Central Limit Theorem is the variance of the exponential / n :

```
1/lambda^2/n
```

```
## [1] 0.625
```

- **The simulation variance is** :

```
var(rexpAvgsDist)
```

```
## [1] 0.6111165
```

To sum up :

Value	Theoritical	Simulation
Mean	5	4.99
Variance	0.625	0.611

- Theoritical and simulation values for both mean and variance are pretty close
- The mean of the averages distribution is the mean of the exponential
- The variance of the averages distribution is much smaller than the variance of the exponential (1/lambda^2/n = 0.625 vs 1/lambda^2 = 25).

3) The distibution of averages is normal

To show this, we simply plot the normal distribution over the distribution of averages with the mean and variance parameters calculated in point 2). We compare this plot with the histogram of large collection of random exponentials See Appendix A) Simulation Plot.

B) Basic Inferential Data Analysis of the dataset ToothGrowth

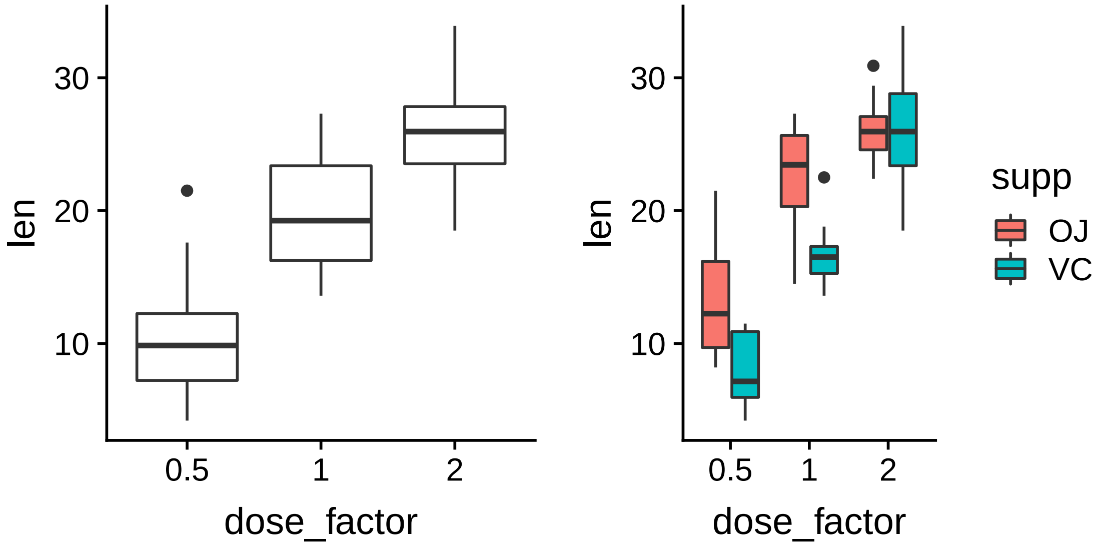
1) Exploratory Data Analysis

```r
library(cowplot)

ToothGrowth$dose_factor <- as.factor(ToothGrowth$dose)
p_all <- ggplot(ToothGrowth, aes(x=dose_factor, y=len)) +
    geom_boxplot()

p_by_dose <- ggplot(ToothGrowth, aes(x=dose_factor, y=len, fill=supp)) +
    geom_boxplot()

plot_grid(p_all, p_by_dose, ncol = 2, nrow = 1)
```



- Tooth growth seems to increase with vitamin C dose
- Tooth growth seems to increase more with OJ than with VC for dose < 2
- Tooth growth doesn't seem different with OJ or VC for dose = 2

2) Summary of the data

```
ToothGrowth %>% group_by(supp,dose) %>%
    summarize(mean=mean(len), sd=sd(len))
```

```
## # A tibble: 6 x 4
## # Groups:   supp [?]
##   supp   dose  mean    sd
##   <fct> <dbl> <dbl> <dbl>
## 1 OJ      0.5 13.2   4.46
## 2 OJ      1   22.7   3.91
## 3 OJ      2   26.1   2.66
## 4 VC      0.5  7.98  2.75
## 5 VC      1   16.8   2.52
## 6 VC      2   26.1   4.80
```

3) Comparison of tooth growth by supp and dose

According to 2) "Summary of the data", the variance of each group is different. We assume the groups are not paired. So we can perform t-tests with (paired=FALSE,var.equal=FALSE) for the following couple of datasets :

- dose = 2 vs dose = 0.5 (for all suppports)
- supp = OJ vs supp = VC (for dose <2)
- supp = OJ vs supp = VC (for dose =2)

See Appendix B) for the code. Results are :

Means difference	95% Confidence Interval	P-value
dose=2 - dose=0.5	[12.8,18.2]	4.4e-14
OJ - VC (dose <2)	[1.88,9.30]	0.0042
OJ - VC (dose =2)	[-3.80,3.64]	0.96

4) Conclusion

Under the following assumptions :

- the t-tests with default variance parameter (unequal variance) is relevant
- groups are not paired (different guinea pigs are used for different supp and dose combinations)

We found that :

- Tooth growth is significantly increased when dose is increased from 0.5 to 2
  (p-value << 0.05 and the 95% confidence interval doesn't contain 0)
- Tooth growth is significantly increased for dose <2 when OJ is used instead of VC (same reason than above)
- For dose =2, there's no significant difference in tooth growth between OJ and VC (p-value >> 0.05, confidence interval centered on 0)

APPENDIX

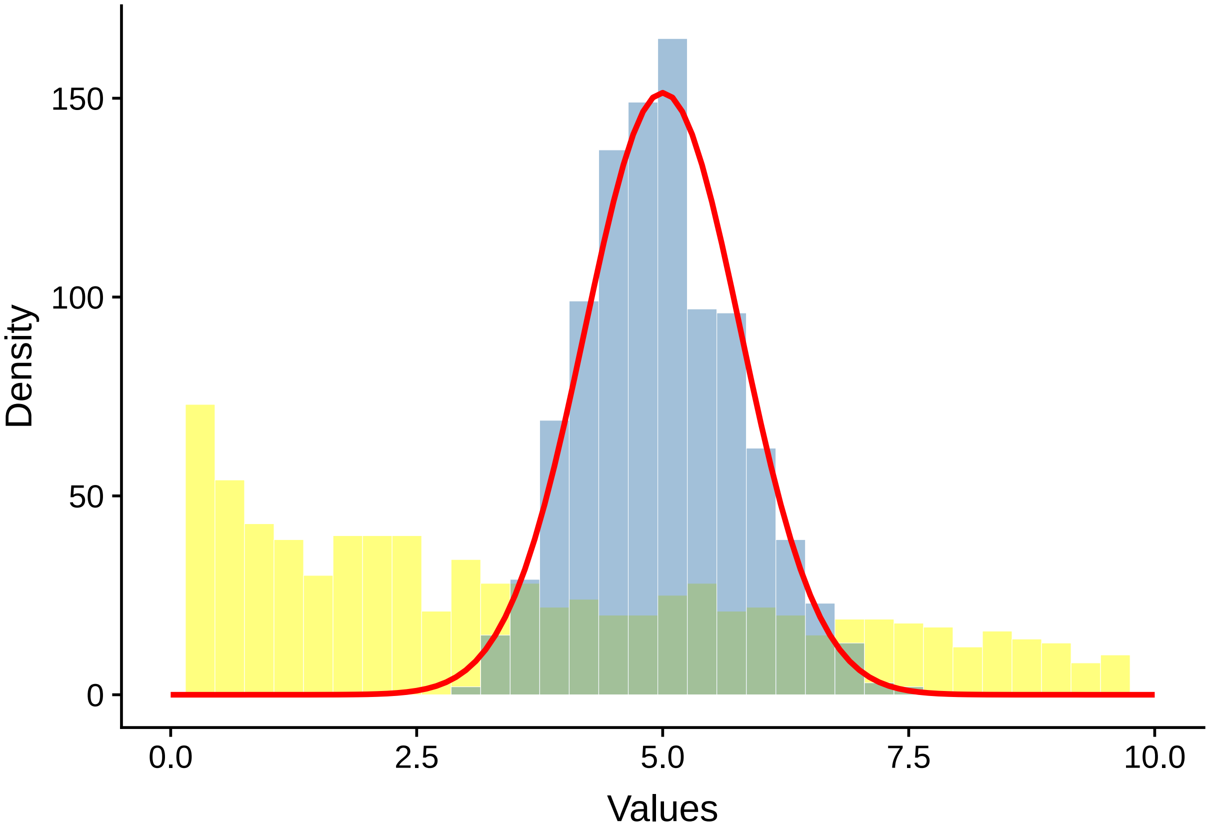
A) Simulation plot

```r
nbPoints <- length(rexpAvgsDist) # 1000

rexpRandomDist <- rexp(n = nbPoints, rate=lambda)

df <- data_frame(rexpAvgsDist,rexpRandomDist)

binwidth <- 0.3
ggplot(df) +
        scale_x_continuous(limits = c(0, 10)) +
        geom_histogram(binwidth = binwidth, aes(x=rexpRandomDist),
                       colour = "white", fill = "yellow",  size = 0.1,alpha=0.5)+
        geom_histogram(binwidth = binwidth, aes(x=rexpAvgsDist),
                       colour = "white", fill = "steelblue", size = 0.1,alpha=0.5)  +
        stat_function(fun = function(averages)
            dnorm(averages, mean = 5, sd = sqrt(0.625)) * nbPoints*binwidth,
            color= "red",size=1 ) +
        labs(x = "Values",y ="Density")
```



- The histogram in yellow represents the distribution of 1000 random exponentials
- The histogram in blue represents the distribution of 1000 averages of 40 exponentials
- The red curve is the normal distribution :
  - centered on the mean of the exponentials : 1/lambda = 5
  - with standard deviation = sqrt(1/lambda^2/n) = sqrt(0.625)

Visually the distribution of averages clearly matches the normal distribution. On the other hand the distribution of random exponential doen't look at all like a normal distribution since it has not a bell shape.

B) Data analysis : ToothGrowth T-Tests

```
d0.5 <- ToothGrowth %>% filter(dose==0.5) %>% select(len)
d2.0 <- ToothGrowth %>% filter(dose==2) %>% select(len)

t.test(d2.0,d0.5,paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  d2.0 and d0.5
## 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:
##  12.83383 18.15617
## sample estimates:
## mean of x mean of y
##    26.100    10.605
```

```
OJ.low.dose <- ToothGrowth %>% filter(supp=="OJ",dose<2) %>% select(len)
VC.low.dose <- ToothGrowth %>% filter(supp=="VC",dose<2) %>% select(len)

t.test(OJ.low.dose,VC.low.dose,paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ.low.dose and VC.low.dose
## 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 of x mean of y
##    17.965    12.375
```

```
OJ.high.dose <- ToothGrowth %>% filter(supp=="OJ",dose==2) %>% select(len)
VC.high.dose <- ToothGrowth %>% filter(supp=="VC",dose==2) %>% select(len)

t.test(OJ.high.dose,VC.high.dose,paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ.high.dose and VC.high.dose
## 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 of x mean of y
##     26.06     26.14
```