Statistical Inference Assignement

by Bruno Berrehuel

November 14, 2016

Contents

I Exponantial Distribution **2**

1 Overview **2**

2 Simulations **2**

3 Sample mean and variance vs. theoricals **2**

4 Distribution **3**

5 Conclusion **3**

II ToothGrowth Data Analysis **4**

1 Data exploratory **4**

2 Probability tests **5**
2.1 Tests for H1 hypothesis 5
2.2 Tests for H2 hypothesis 5

3 Conclusions **6**
3.1 Is the vitamin C effective for tooth growth ? 6
3.2 What is the best delivery method ? 7

Part I
Exponantial Distribution

1 Overview

In this part I'll do some investigations of the exponential distribution, with the R function **rexp(n, λ)** and more precisely the distribution of averages of 40 exponentials distributions. In accordance with the Central Limit Theorem, I'll investigate that the sample mean μ and the sample variance s^2 are :

$$\mu = \frac{1}{\lambda} \quad \text{and} \quad s^2 = \frac{\sigma^2}{n}$$

Then I'll look at the distribution of 10 000 means of 40 exponential distributions, and verify that they follow a normal distribution $\mathcal{N}(\mu, s^2)$, as expected with the CLT.
For numerical purposes, λ is choosen as $\lambda = 0.2$, so $\mu = \sigma = \frac{1}{\lambda} = 5$.

2 Simulations

Calculate 10 000 means of 40 exponantials distributions with λ=$\frac{1}{5}$: for each i, calculate the mean of 40 exponantials distributions, then add it to the variables rexpMean and rexpVar. The rexpMean and rexpVar variables will contain 10 000 values each.

```
set.seed(12345)
rexpMean=NULL
rexpVar=NULL
for (i in 1:10000) {
    rexpMean = c(rexpMean, mean(rexp(40,0.2)))
    rexpVar = c(rexpVar, var(rexp(40,0.2)))
}
```

3 Sample mean and variance vs. theoricals

Calculate the means of the mean and the variance of the sample with R, and compare with mean and variance expected. Recall that the expected sample mean μ and the expected sample variance s are :

$$\mu = \frac{1}{\lambda} = 5 \quad \text{and} \quad s^2 = \frac{\sigma^2}{n} = \frac{5^2}{40} \simeq 0.625$$

The expected distribution variance is $\sigma^2 = 25$.

```
data.frame(mean=mean(rexpMean), sampleVariance=var(rexpMean),
           variance=mean(rexpVar))

##       mean sampleVariance variance
## 1 5.003857      0.6083614 24.92046
```

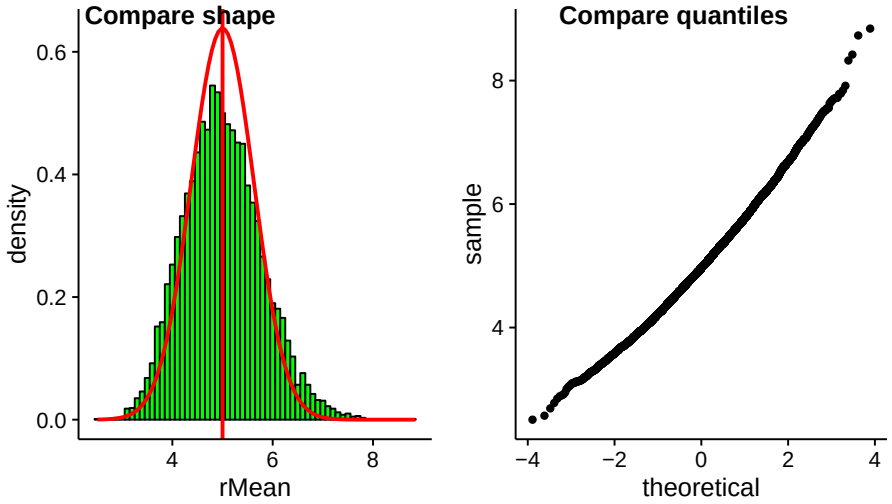
The values are very close to the expected values :

- 0.08% for the sample mean,
- 2.66% for the sample variance, and
- 0.32% for the distribution variance.

4 Distribution

Plot an histogram with the 10 000 previous calculation, and compare the shape with a normal distribution $\mathcal{N}(5, 0.625)$, then compare the distributions using a quantile-quantile diagram :

```
dfRexp  <- data.frame(rMean=rexpMean, rVar=rexpVar)
histoMean <- ggplot(dfRexp, aes(x=rMean)) +
    geom_histogram(aes(y=..density..), color="black",
                   fill="green", binwidth=0.1) +
    geom_vline(aes(xintercept=5), color="red", size=1) +
    stat_function(fun=dnorm, color="red", size=1,
                  args=list(mean=5, sd=0.625))
qqMean <- ggplot(dfRexp, aes(sample=rMean)) + stat_qq()
plot_grid(histoMean, qqMean, ncol=2, nrow=1,
          labels=c("Compare shape","Compare quantiles"))
```



5 Conclusion

As points are aligned in the qqnorm diagram and with the shape of the histogram, the distribution of the mean follows the normal distribution $\mathcal{N}$(5,0.625), as expected with the CLT.

Part II
ToothGrowth Data Analysis

The datas are about the length of odontoblast (cells responsible for tooth growth) for 60 pigs after an experimental threatment in vitamin C. Each animal received a dose of vitamin C, from 0.5 to 2 mg/day, by orange juice (OJ) or ascorbic acid (VC)[1]. Each experiment deals with 10 pigs.
I'll try to answer the two following questions :

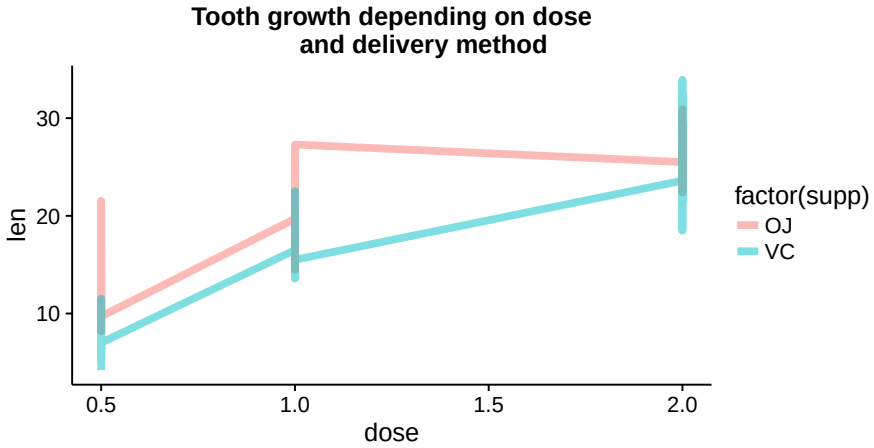
1. Q1 : is the vitamin C effective for tooth growth ?
2. Q2 : what is the best delivery method between orange juice or ascorbic acid ?

1 Data exploratory

First I'll take some informations about the means and the standard deviation of each dose, and plot the evolution of the tooth growth depending on the dose and the delivery method :

```
ToothGrowth %>% group_by(supp,dose) %>%
    summarise(mean(len), round(sd(len),3))

## Source: local data frame [6 x 4]
## Groups: supp [?]
##
##     supp  dose `mean(len)` `round(sd(len), 3)`
##   <fctr> <dbl>       <dbl>                <dbl>
## 1     OJ   0.5       13.23                4.460
## 2     OJ   1.0       22.70                3.911
## 3     OJ   2.0       26.06                2.655
## 4     VC   0.5        7.98                2.747
## 5     VC   1.0       16.77                2.515
## 6     VC   2.0       26.14                4.798
```



[1] Source : https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html

It seems that :

- H1 : odontoblasts are taller with a bigger dose of vitamin C, no matter of the delivery method.
- H2 : orange juice (OJ) has better results than ascorbic acid (VC).

2 Probability tests

I define some variables form the ToothGrowth datasets, in order to compare with the R function **t.test** :

- H1 : the mean of odontoblasts length at 0.5 mg/day dose with the mean at 2 mg/day
- H2 : the means of odontoblasts length for the 2 delivery method ar each vitamin C dose.

2.1 Tests for H1 hypothesis

First filter the ToothGrowth dataset depending on dose of vitamin C, for both of the delivery method, then do t.test to confirm that the mean at 2 mg/day is more important than the mean at 0.5 mg/day.

```
vita05 <- ToothGrowth %>% filter(dose==0.5)
vita2 <- ToothGrowth %>% filter(dose==2)
t.test(vita2$len-vita05$len, alternative="greater")

##
##  One Sample t-test
##
## data:  vita2$len - vita05$len
## t = 11.291, df = 19, p-value = 3.595e-10
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  13.12216      Inf
## sample estimates:
## mean of x
##    15.495
```

2.2 Tests for H2 hypothesis

First filter the ToothGrowth dataset depending on dose of vitamin C and delivery method, then compare them with the t.test function.

```
indice <- c(0.5,1,2)
OJ <- lapply(indice, function(i) {ToothGrowth %>% filter(supp=="OJ") %>%
              filter(dose==i)})
VC <- lapply(indice, function(i) {ToothGrowth %>% filter(supp=="VC") %>%
              filter(dose==i)})
t.test(OJ[[1]]$len-VC[[1]]$len, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  OJ[[1]]$len - VC[[1]]$len
## t = 2.9791, df = 9, p-value = 0.007736
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.019552      Inf
## sample estimates:
## mean of x
##      5.25
```

```
t.test(OJ[[2]]$len-VC[[2]]$len, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  OJ[[2]]$len - VC[[2]]$len
## t = 3.3721, df = 9, p-value = 0.004115
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.706401      Inf
## sample estimates:
## mean of x
##      5.93
```

```
t.test(OJ[[3]]$len-VC[[3]]$len)
```

```
##
##  One Sample t-test
##
## data:  OJ[[3]]$len - VC[[3]]$len
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of x
##     -0.08
```

We have the following results :

- we can reject $\mu_{OJ05} = \mu_{VC05}$ and accept $\mu_{OJ05} > \mu_{VC05}$
- we can reject $\mu_{OJ1} = \mu_{VC1}$ and accept $\mu_{OJ1} > \mu_{VC1}$
- we failed to reject $\mu_{OJ2} = \mu_{VC2}$

3 Conclusions

3.1 Is the vitamin C effective for tooth growth ?

As the t.test shows that the real mean of the 2 mg/day tooth growth is greater than the 0.5 mg/day tooth growth, the vitamin C threatment is effective for the pigs.

3.2 What is the best delivery method ?

The t.tests show that :

- the orange juice delivery method is more effective than the ascorbic acid method for small doses, as the real tooth growth means are greater for orange juice for 0.5 and 1 mg/day doses.
- the methods are equivalent for the dose 2 mg/day, as we failed to reject the null hypothesis that the means are equal.

I recommend the delivery of vitamin C, by orange juice, for the tooth growth of the guinea pigs.