

Statistical Inference Course Project

Emmanouil Kalaitzakis

2021-07-02 11:47:03

This is a report for the final project of the Statistical Inference course offered by Johns Hopkins University on Coursera. This project is split into two parts:

- The first part involves a simulation exercise intended towards understanding the properties of the distribution of the sum of a number of independent and identically distributed exponential random variables.
- The second part involves the exploration of a data set, some statistical testing, and a discussion of the assumptions supporting the application of the methods.

Part 1: Simulation Exercise

To perform the simulation we create a random matrix where each row is a simulation iteration and each column is one sample.

```
n <- 40
lambda <- 0.2
mu <- 1/lambda
sigma <- 1/lambda

iter <- 1000

set.seed(1)
mc <- matrix(data = rexp(iter*n, lambda), nrow = iter, ncol = n, byrow = FALSE)

mus <- apply(mc, 1, mean)
```

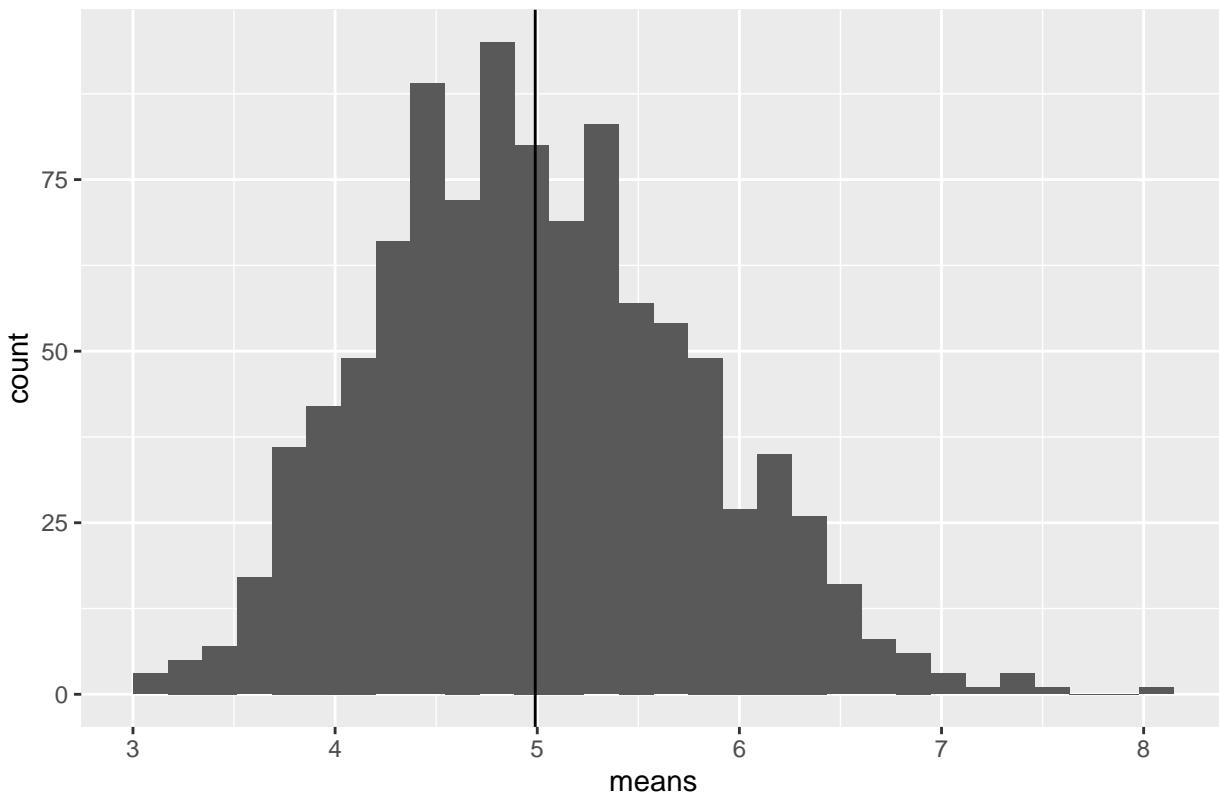
The sample mean of the distribution is 4.9900252 while the theoretical mean of the distribution is 5. The variance of the distribution is 0.6177072 while the theoretical variance of the distribution is 0.625. We can explore the distribution visually with the help of a histogram and a vertical line centered at the sample mean.

```
library(ggplot2)

ggplot(data.frame(means = mus), aes(means)) +
  geom_histogram() +
  geom_vline(xintercept = mean(mus)) +
  ggtitle("Distribution of sample means")

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

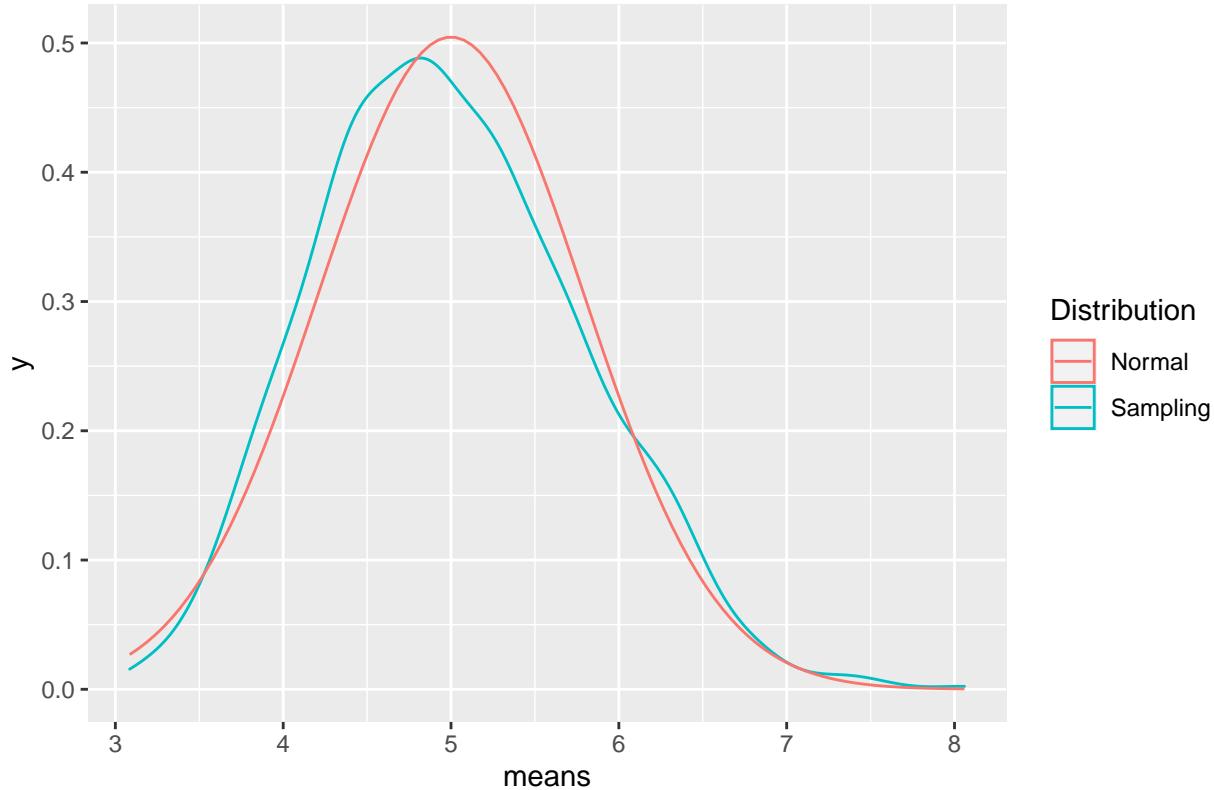
Distribution of sample means



To check whether the distribution is approximately normal, we plot it against a normal distribution with the theoretical mean and variance. We also perform a Shapiro-Wilk normality test to support our finding.

```
ggplot(data.frame(means = mus), aes(means)) +  
  geom_density(aes(colour = "Sampling")) +  
  stat_function(fun = dnorm, args = list(mean = mu, sd = sigma/sqrt(n)), aes(colour = "Normal")) +  
  labs(colour = "Distribution") +  
  ggtitle("Visual comparison with a normal distribution")
```

Visual comparison with a normal distribution



```
shapiro.test(mus)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: mus  
## W = 0.99157, p-value = 1.759e-05
```

We observe that the sampling distribution is quite similar to the normal. The Shapiro-Wilk test supports this finding further since the p-value is much smaller than 0.1 which is threshold for the distribution to be considered approximately normal.

Part 2: Basic Inferential Data Analysis

The first step in the second part is to load the `ToothGrowth` dataset and perform some basic exploratory analysis using the `str` and `summary` functions.

```
library(datasets)  
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 ...
```

```
summary(ToothGrowth)
```

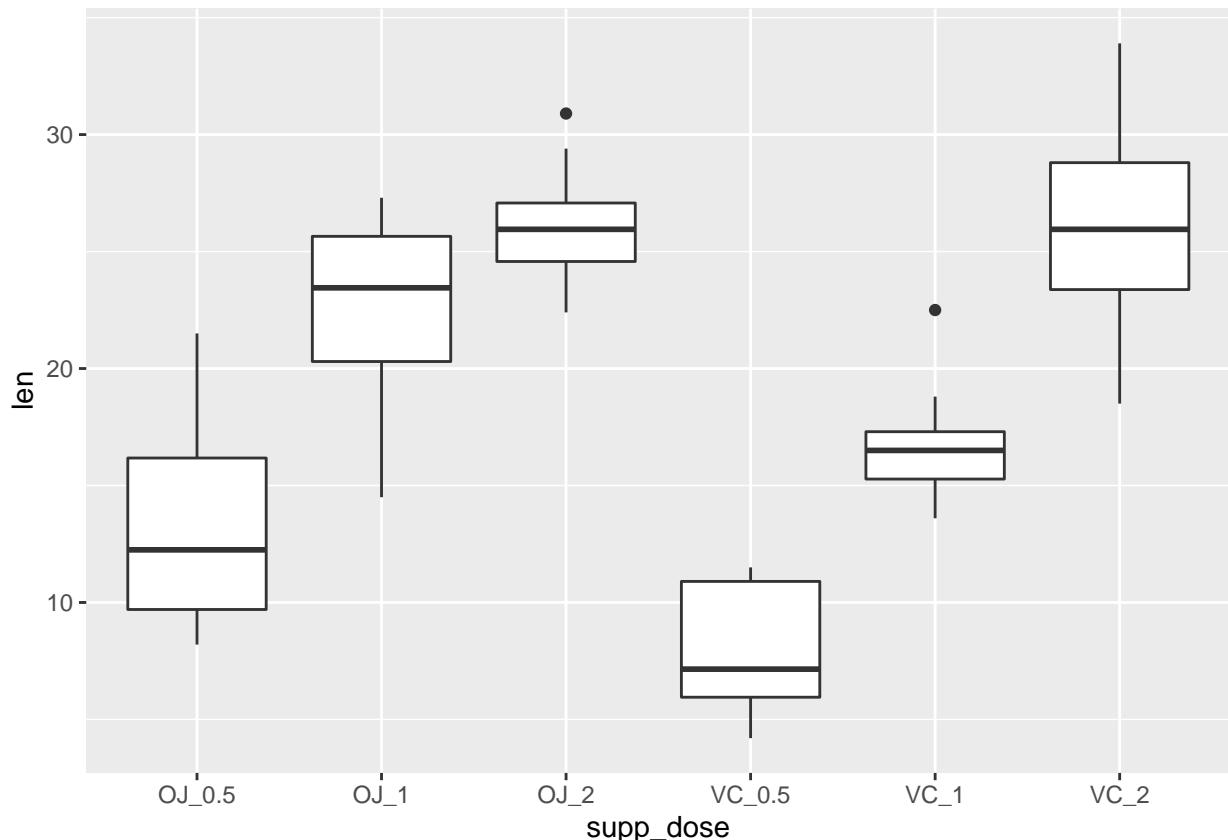
```
##      len      supp      dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07  VC:30   1st Qu.:0.500
##  Median :19.25          Median :1.000
##  Mean   :18.81          Mean   :1.167
##  3rd Qu.:25.27          3rd Qu.:2.000
##  Max.   :33.90          Max.   :2.000
```

The dataframe consists of 3 variables:

- The response variable `len` which is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs,
- `supp`, one of two delivery methods of Vitamin C, orange juice or ascorbic acid (a form of vitamin C and coded as `VC`),
- and `dose`, the dose levels of vitamin C (0.5, 1, and 2 mg/day) received.

We are interested in investigating how the various dose levels effect the odontoblast length controlling for supplement type. We create all variable combinations and plot the boxplots.

```
library(tidyr)
ggplot(ToothGrowth %>% unite(supp_dose, supp:dose), aes(y = len)) +
  geom_boxplot(aes(x = supp_dose))
```



We observe that the dose level is critical for both delivery methods but it is super-critical for the VC method. Next we are going to test whether the difference in group means of the `len` variable is statistically significant for the two `supp` groups and the three `dose` groups. Let's take a look to the group means and standard deviations.

```
tapply(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), mean)
```

```
##      0.5      1      2
## OJ 13.23 22.70 26.06
## VC  7.98 16.77 26.14
```

```
tapply(ToothGrowth$len, list(ToothGrowth$supp, ToothGrowth$dose), sd)
```

```
##      0.5      1      2
## OJ 4.459709 3.910953 2.655058
## VC 2.746634 2.515309 4.797731
```

We observe that VC performs at par with OJ at the 2 mg/day level while it is inferior in smaller doses. We proceed with the `len ~ supp` t.test controlling for the dose. The second table indicates why it is fit for our analysis to assume unequal variances.

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 0.5,])
```

```
##
##  Welch Two Sample t-test
##
## data: len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##           13.23            7.98
```

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 1,])
```

```
##
##  Welch Two Sample t-test
##
## data: len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##           22.70            16.77
```

```
t.test(len ~ supp, ToothGrowth[ToothGrowth$dose == 2,])  
  
##  
## Welch Two Sample t-test  
##  
## data: len by supp  
## 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 in group OJ mean in group VC  
## 26.06 26.14
```

Since the p-value is smaller than 0.05 and the 95% confidence interval excludes 0, we reject the null hypothesis that states that the means of the two groups are equal in favor of the alternative hypothesis which states that the OJ group mean is different (larger) than the VC group mean for the smaller two doses. As suspected this is not the case for the 2 mg/day dose where the two supplements perform in a similar manner.

Based on evidence we have assumed that variances are not equal between groups. We are also assuming that the observations are not paired since we have no such information. Finally, we are assuming that the group means are independent and identically distributed and they follow a (approximately) normal distribution.