Statistical Inference

Author: Vladislav Leonov

Description

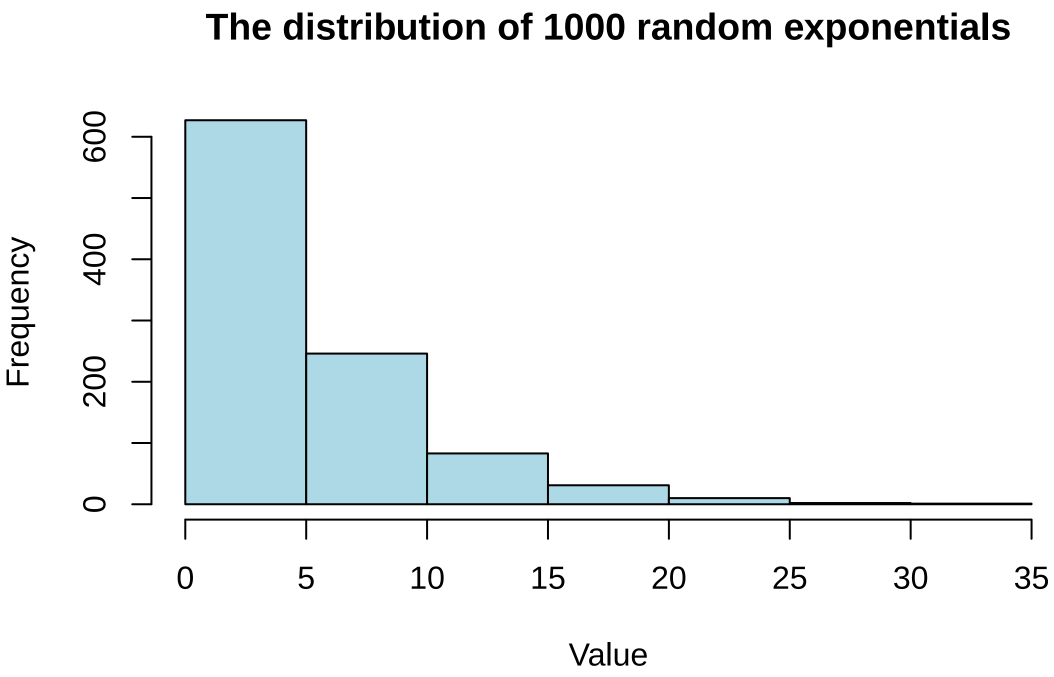
ToothGrowth data set contains the result from an experiment studying the effect of vitamin C on tooth growth in 60 Guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods (orange juice or ascorbic acid (a form of vitamin C and coded as VC)).

Part 1: Simulation Exercise

In this part of the project we will investigate the exponential distribution and compare it with the Central Limit Theorem with sample size of 40.

At first, let's look at the distribution of 1000 random exponentials:

```r
lambda <- 0.2
hist(rexp(1000, lambda),
     main = 'The distribution of 1000 random exponentials',
     xlab = 'Value',
     breaks = 10,
     col = 'lightblue')
```



Let's make sure that the distribution of means is approximately normal and compare sample mean and variance with their theoretical values.

Creating a sample of means:

```r
mns = NULL
for (i in 1:1000){
    mns = c(mns, mean(rexp(40, lambda)))
}
```

Calculating the sample mean and comparing it to the theoretical mean of the distribution:

```r
sample_mean <- mean(mns)
theor_mean <- lambda^(-1)
print(paste0('Sample mean: ', sample_mean, '; ',
             'Theoretical mean: ', theor_mean))
```

```
## [1] "Sample mean: 4.98024879415097; Theoretical mean: 5"
```

Means comparison:

```r
print(paste0('The absolute difference between means: ',
             abs(sample_mean - theor_mean)))
```

```
## [1] "The absolute difference between means: 0.0197512058490315"
```

Next, let's calculate the variance of the sample and compare it to the theoretical variance of the distribution:

```r
sample_var <- var(mns)
theor_var <- lambda^(-2)/40
print(paste0('Sample variance: ', sample_var, '; ',
             'Theoretical variance: ', theor_var))
```

```
## [1] "Sample variance: 0.629661275912715; Theoretical variance: 0.625"
```

Variances comparison:

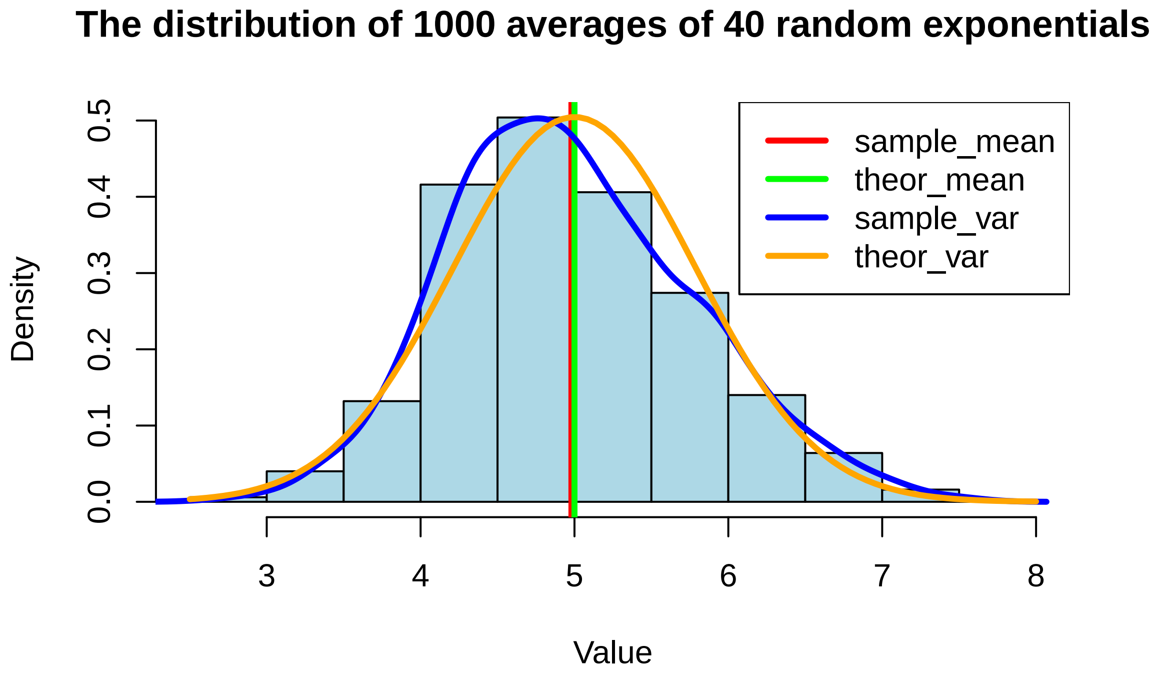
```r
print(paste0('The absolute difference between variances: ',
             abs(sample_var - theor_var)))
```

```
## [1] "The absolute difference between variances: 0.00466127591271526"
```

Now, look at the distribution of 1000 averages of 40 random exponentials with added means and KDEs:

```r
hist(mns,
     main = 'The distribution of 1000 averages of 40 random exponentials',
     xlab = 'Value',
     freq = FALSE,
     col = 'lightblue')
abline(v = sample_mean,col = 'red', lwd = 3)
abline(v = theor_mean, col = 'green', lwd = 3)
dens_sample <- density(mns)
lines(dens_sample, col = 'blue', lwd = 3)
x <- seq(2, 8, by = 0.02)
```

```
curve(dnorm(x, mean = theor_mean, sd = sqrt(theor_var)), add = TRUE, col = 'orange', lwd = 3)
legend("topright",
       legend = c("sample_mean", "theor_mean", "sample_var", "theor_var"),
       col = c("red", "green", "blue", "orange"),
       lwd = 3)
```



As you can see, it looks like a normal distribution (approximately). This is the result of the Central Limit Theorem in action. It says that the distribution of a large enough number of averages of samples tends to normal.

Part 2: Basic Inferential Data Analysis

In this part of the project we're going to analyze the ToothGrowth data in the R datasets package.

At first, let's load the ToothGrowth data and perform some basic exploratory data analyses.

The dataset dimensions:

```
library(datasets)
data("ToothGrowth")
dim(ToothGrowth)
```

```
## [1] 60  3
```

The first few rows:

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

The summary:

```
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
```

Here we can see how each substance influenced tooth growth (on the average):

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

```
##       OJ       VC 
## 20.66333 16.96333
```

So, it seems like orange juice had a stronger impact than vitamin C. We will check it further.

Examining a dose influence on tooth growth:

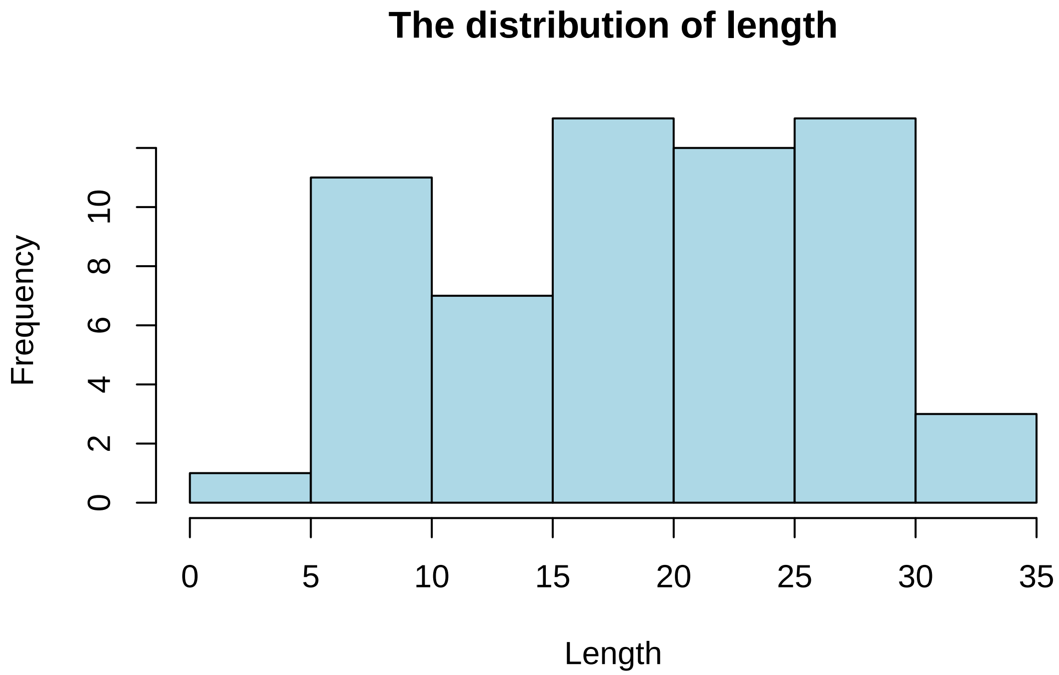
```
tapply(ToothGrowth$len,ToothGrowth$dose, mean)
```

```
##    0.5      1      2 
## 10.605 19.735 26.100
```

Well, here we see that a higher dose has a greater impact. We also need to check it later.

Let’s look at the distribution of tooth length:

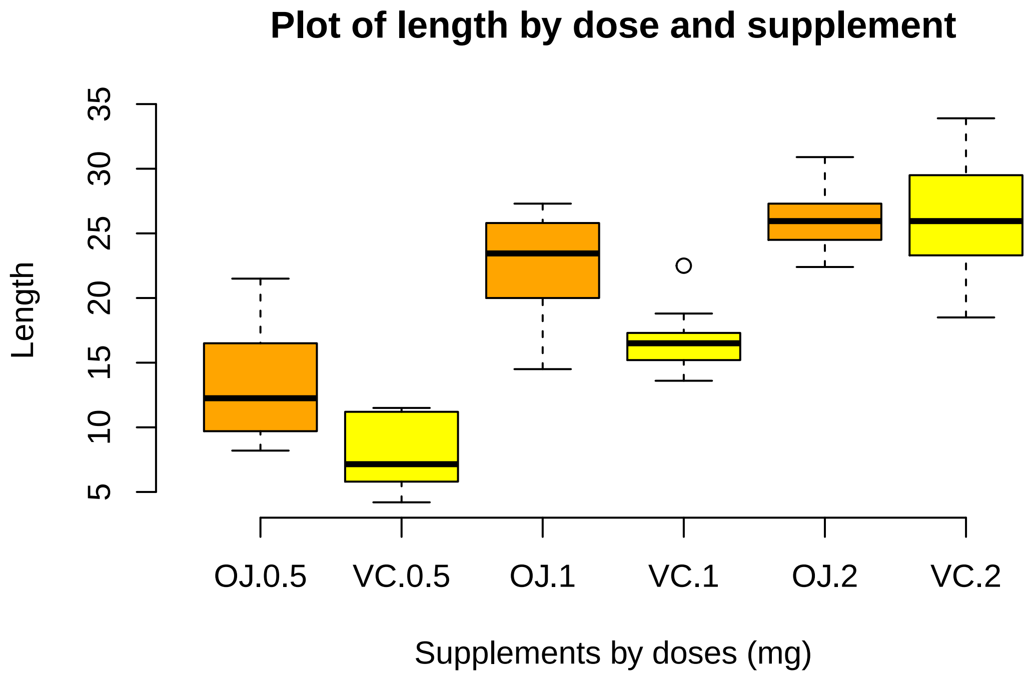
```
hist(ToothGrowth$len,
     main = 'The distribution of length',
     xlab = 'Length',
     col = 'lightblue')
```



Next, we will use confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

At first, we should see the graphical representation of dependency of tooth growth on supplement type and dose:

```
boxplot(len ~ supp*dose, data = ToothGrowth,
        col = c("orange", "yellow"), frame = FALSE,
        main = "Plot of length by dose and supplement",
        xlab = "Supplements by doses (mg)", ylab = "Length")
```



Look at 2 mg doses. It seems like at this dosage OJ and VC influence tooth growth at the same power. This needs to be checked.

So, our null hypotheses are:

1. the type of a supplement doesn't influence tooth growth;
2. the dose of a supplement doesn't influence tooth growth;
3. it doesn't matter what supplement to use with 2 mg dose for a greater tooth growth.

Alternative hypothesis are (respectively):

1. OJ supplement influences tooth growth more than VC supplement;
2. the greater a dose of a supplement, the greater tooth growth is;
3. there is a difference between tooth growth influence of supplements at 2 mg dosage.

Our p-value threshold is 0.05.

Checking the first hypothesis:

```
OJ = ToothGrowth$len[ToothGrowth$supp == "OJ"]
VC = ToothGrowth$len[ToothGrowth$supp == "VC"]
```

```
p1 <- t.test(OJ, VC, alternative = "greater", conf.level = 0.95)$p.value
print(paste0('The p-value for the 1st hypothesis: ', p1))
```

```
## [1] "The p-value for the 1st hypothesis: 0.030317253940467"
```

The p-value is approximately 0.03, which is less than 0.05. Thus, we reject the null hypothesis and consider that OJ has a greater impact on tooth growth than VC.

Now, let's check the next hypothesis about dosage:

```
dose_0.5 = ToothGrowth$len[ToothGrowth$dose == 0.5]
dose_1 = ToothGrowth$len[ToothGrowth$dose == 1]
dose_2 = ToothGrowth$len[ToothGrowth$dose == 2]
```

Testing 0.5 mg dose and 1 mg dose:

```
p2_1 <- t.test(dose_1, dose_0.5, alternative = "greater", conf.level = 0.95)$p.value
print(paste0('The p-value for the 2nd hypothesis (first test): ', p2_1))
```

```
## [1] "The p-value for the 2nd hypothesis (first test): 6.34150360086923e-08"
```

We see that the p-value is less than our threshold. We need to make another test with 1 mg and 2 mg doses and prove that we can reject the null hypothesis:

```
p2_2 <- t.test(dose_2, dose_1, alternative = "greater", conf.level = 0.95)$p.value
print(paste0('The p-value for the 2nd hypothesis (second test): ', p2_2))
```

```
## [1] "The p-value for the 2nd hypothesis (second test): 9.53214756835902e-06"
```

Again, the p-value is less than the threshold (0.05), so, we reject the null hypothesis and consider that a higher dose of a supplement have a greater impact on tooth growth.

The last hypothesis is about the effectiveness of supplements at 2 mg dosage:

```
OJ_2 = ToothGrowth$len[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2]
VC_2 = ToothGrowth$len[ToothGrowth$supp == "VC" & ToothGrowth$dose == 2]
```

```
p3 <- t.test(OJ_2, VC_2, alternative = "two.sided", conf.level = 0.95)$p.value
print(paste0('The p-value for the 3rd hypothesis: ', p3))
```

```
## [1] "The p-value for the 3rd hypothesis: 0.963851588723373"
```

The p-value is much greater than our threshold, so, we don't reject the null hypothesis and consider that it doesn't matter what supplement to use with 2 mg dose for a greater tooth growth.

Conclusion

- In the first part of the project we checked the Central Limit Theorem and made sure that the distribution of a large enough number of averages of samples tends to normal.
- In the second part we made some inferential data analysis and concluded that:
  - OJ has a greater impact on tooth growth than VC;
  - a higher dose of a supplement have a greater impact on tooth growth;
  - it doesn't matter what supplement to use with 2 mg dose for a greater tooth growth.