Statistical Interference Assignment Coursera

Michaela Spiegel

21 August 2018

Overview

Part 1: Exponential Distribution & Central Limit Theorem

This part covers the difference between the distribution of large collection of random exponentials vs. a distribution of large collections of averages of these random exponentials. In this simulation a large collection contains 1,000 values and the averages of exponentials are calculated as average over 40 exponentials.

The relation between these two distributions is given by the central limit theorem, which states when independent random variables are summed up, their distribution of sums tends towards a normal distribution, even if the original variables themselves are not normally distributed.

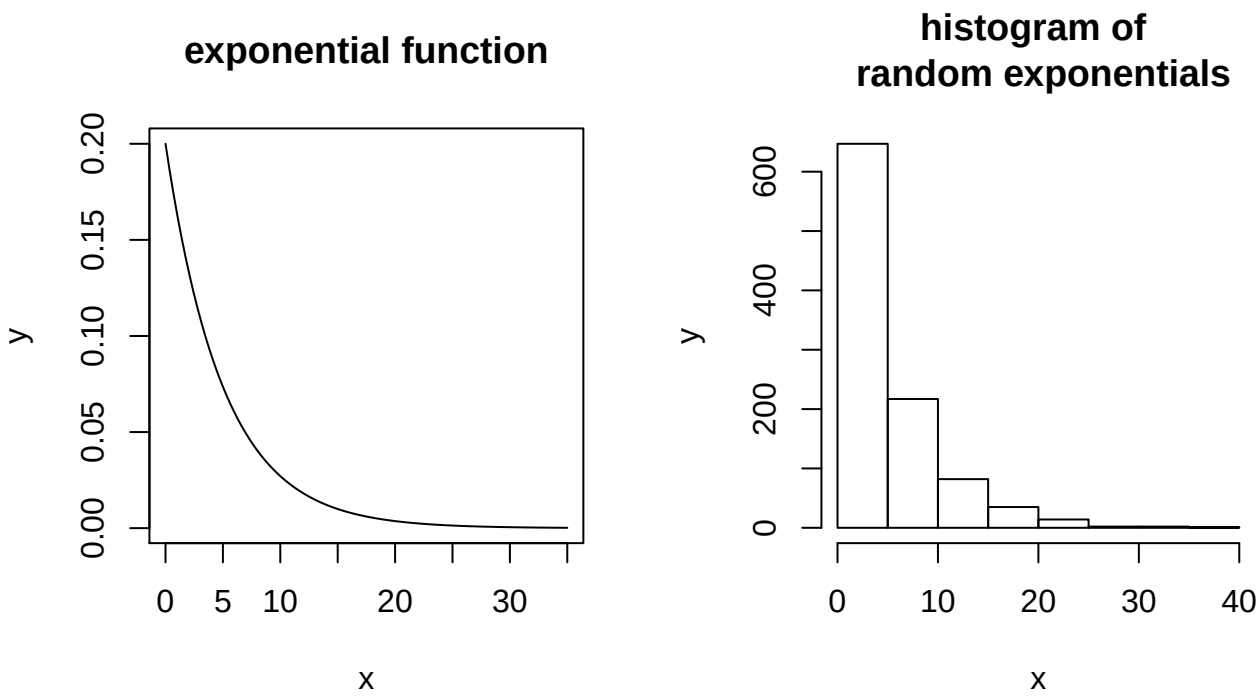
Part 2: Tooth Growth Data Set

Part 1: Exponential Distribution & Central Limit Theorem

Exponential distribution

The exponential function $\lambda * \exp(-\lambda x)$ in R is given by "rexp()" function. If we pick randomly from this distribution (1,000 times), we get a sample distribution which looks pretty much like the original exponential distribution. The random sample mirrors the underlying distribution, seen in the following plots.

```
lambda = 0.2
n=1000
eq = function(x){lambda*exp(-lambda*x)}
exponential_function <- rexp(n, lambda)
par(mfrow=c(1,2))
curve(eq, from=0, to=35, xlab="x", ylab="y", main = "exponential function")
hist(exponential_function, main = "histogram of \n random exponentials", xlab="x", ylab="y")
```



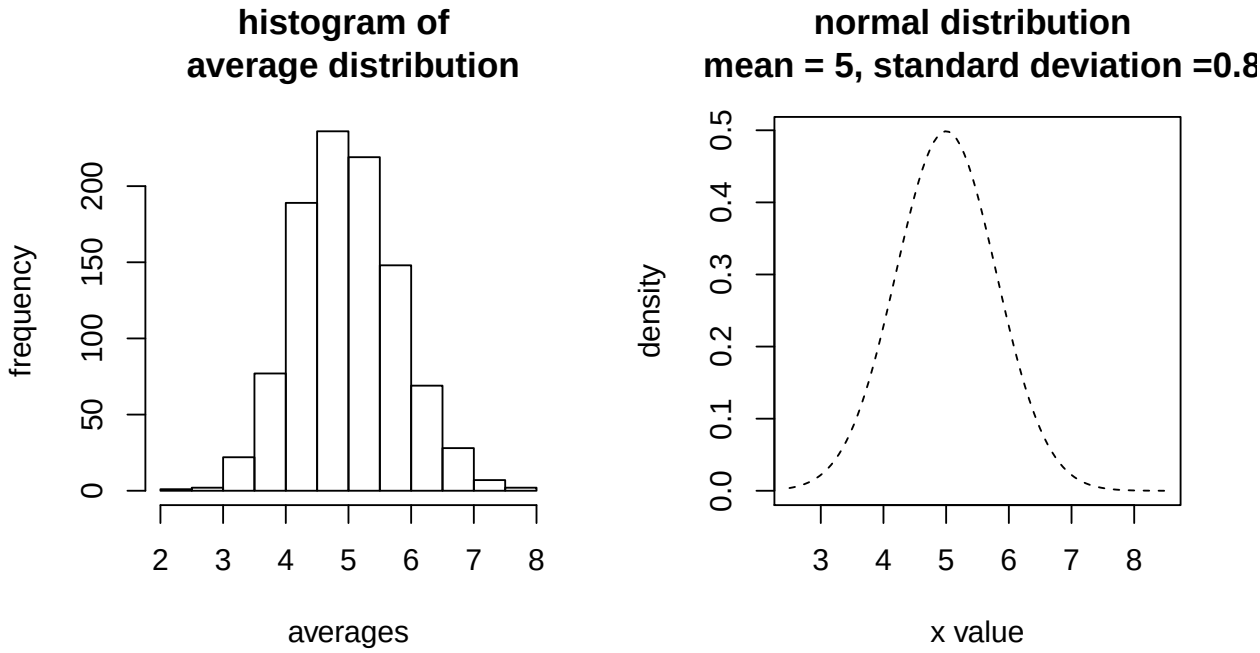
We calculate the theoretical mean and variance of the distibution and the experimental, from our random exponentials and see the values are pretty close.

```
sum_avgs <- data.frame("case" = c("theoretical", "experimental"),
                       "mean" = c(1/lambda, mean(exponential_function)),
                       "variance" = c(1/(lambda*lambda),var(exponential_function)))
sum_avgs
```

```
##           case    mean variance
## 1  theoretical 5.00000 25.00000
## 2 experimental 4.97085 25.28804
```

Now we pick randomly 40 numbers from the exponential function and calculate the average of these values. This step is repeated 1,000 times. The distribution of averages looks approximately normal distributed.

```
lambda = 0.2
n=40
average_distribution = c()
for (i in 1:1000)
  average_distribution[i] <- mean(rexp(n, lambda))
par(mfrow=c(1,2))
hist(average_distribution, main = "histogram of \n average distribution", xlab = "averages", ylab = "fr
x <- seq(2.5, 8.5, length=100)
hx <- dnorm(x, mean = 1/lambda, sd = 0.8)
plot(x, hx, type="l", lty=2, xlab="x value",
  ylab="density", main="normal distribution \n mean = 5, standard deviation =0.8")
```



From these distribution, we can estimate the average of the whole dataset, by calculating the mean value of our distribution, which we know is 5 and study the variance.

```
sum_avgs <- data.frame("case" = c("theoretical", "experimental"),
                       "mean" = c(1/lambda, mean(average_distribution)),
                       "variance" = c("?",var(average_distribution)))
sum_avgs
```

```
##           case     mean          variance
## 1  theoretical 5.000000                 ?
## 2 experimental 4.977854 0.642270150779969
```

Part 2: Tooth Growth Data Set

In part two the tooth growth data set is studied. The data set is explored via boxplots. From this point, hypothesis are constructed and then validated or rejected by t-testing.

Description

The response is the length of odontoblasts (cells responsible for 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).

[,1] len numeric Tooth length [,2] supp factor Supplement type (VC or OJ). [,3] dose numeric Dose in milligrams/day

Source

C. I. Bliss (1952) The Statistics of Bioassay. Academic Press.

Exploratory Data Analyses

First we take a look at the first lines of the data frame and summary.

```
library(datasets)
data(ToothGrowth)
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
head(ToothGrowth, n= 3)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
```

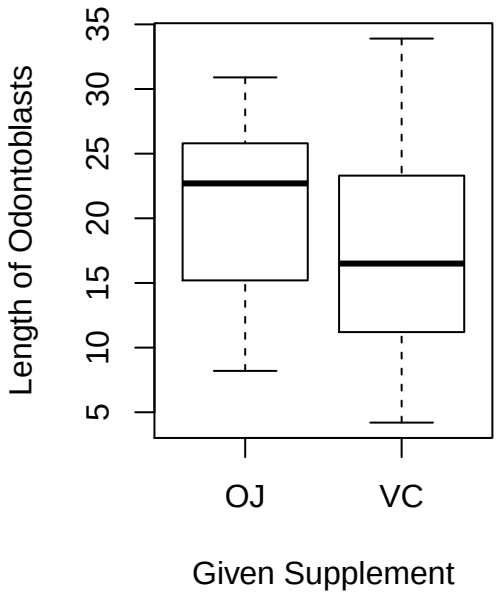
```
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

As the description already told us, we can see that we got 3 columns and 60 observables. The first column len, is a numerical value which gives us the length of the odontoblasts. The supplement is in 30 cases OJ and in the other 30 cases VC. The dose of the supplement can be either 0.5, 1 or 2.

First we compare the length of odontoblasts as a function of the given supplement.

```
boxplot(len~supp,data=ToothGrowth,
   xlab="Given Supplement", ylab="Length of Odontoblasts")
```

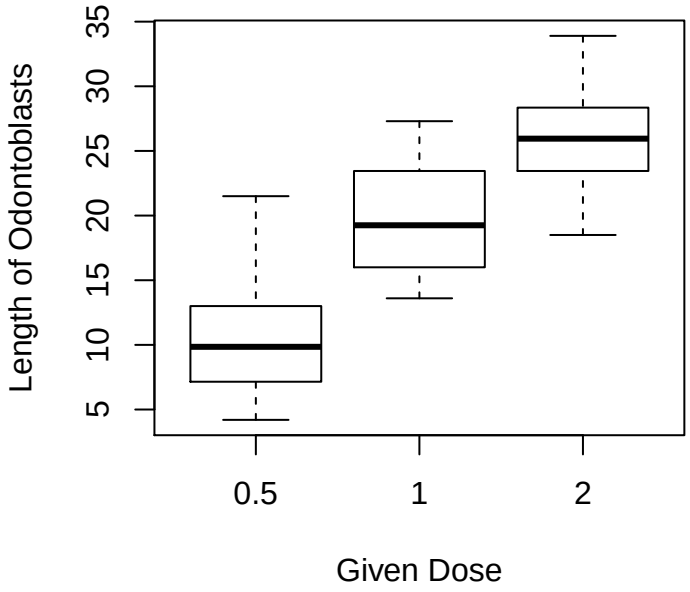


From this plot we can state a first hypothesis:

Hypothesis 1: The mean length of odontoblasts is higher, when the guinea pigs are given Vitamin C via Orange Juice (OJ), than via Adsorbic Acid (VC) (independent of dose).

Than we compare the length of odontoblasts as a function of the given dose.

```
boxplot(len~dose,data=ToothGrowth,
   xlab="Given Dose", ylab="Length of Odontoblasts")
```

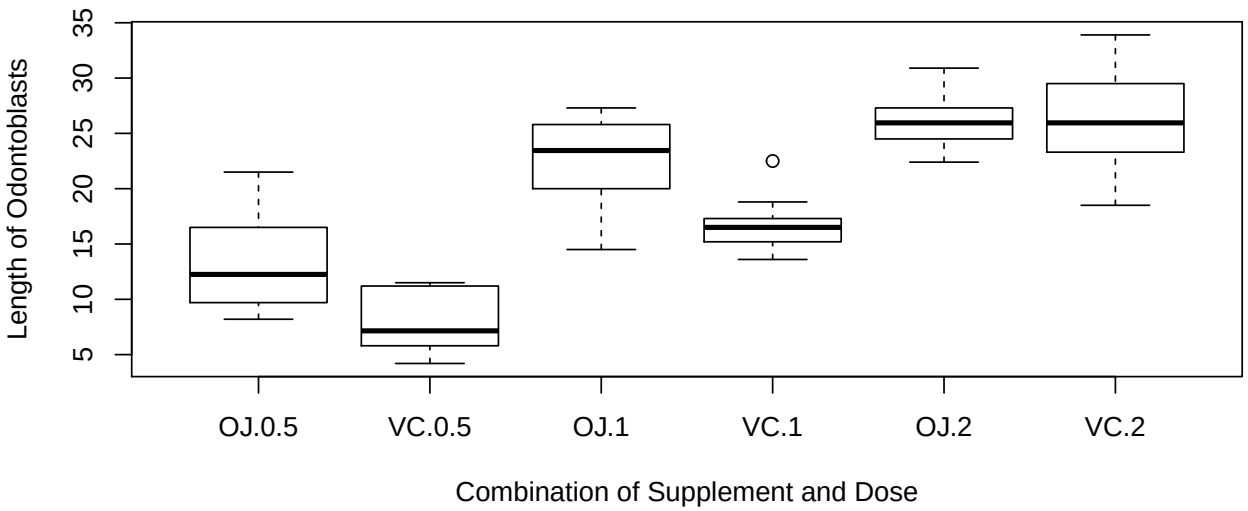


From this plot, we can state a new hypothesis:

Hypothesis 2: The mean length of odontoblasts is higher, when the guinea pigs are given higher doses of Vitamin C (independent of the supplement).

Now we plot the length of odontoblasts as a function of a combination of given dose and supplement.

```
boxplot(len~supp*dose, data=ToothGrowth,
  xlab="Combination of Supplement and Dose",
  ylab = "Length of Odontoblasts")
```



This plot now, gives us more insights. We can state new hypothesis, e.g.

Hypothesis 3: When we give a high dose (2) of Vitamin C, the length of Odontoblasts is independent of the supplement (either OJ or VC).

All three hypothesis are tested on significance.

Hypothesis Testing

Hypothesis 1: The mean length of odontoblasts is higher, when the guinea pigs are given Vitamin C via Orange Juice (OJ), than via Adsorbic Acid (VC) (independent of dose).

To test if there is a significant effect, the null hypothesis is tested, by a t-test, that there is no effect.

```
t.test(ToothGrowth$len[ToothGrowth$supp=="OJ"], ToothGrowth$len[ToothGrowth$supp=="VC"], paired=FALSE, v
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ"] and ToothGrowth$len[ToothGrowth$supp == "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
```

The p-Value is 6.3 %, this means with an alpha of 5% we can not reject the null hypothesis. So we can not (!) conclude that length of the odontoblasts depends on the supplement.

Hypothesis 2: The mean length of odontoblasts is higher, when the guinea pigs are given higher doses of Vitamin C (independent of the supplement).

When we test this hypothesis for dose of 0.5 vs dose 0f 0.1, wee get a p-Value smaller than 0.01, which indicates our hypothesis is true.

```
t.test(ToothGrowth$len[ToothGrowth$dose==0.5],
       ToothGrowth$len[ToothGrowth$dose==1], var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 0.5] and ToothGrowth$len[ToothGrowth$dose == 1]
## 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
```

The same holds for a dosage increase from 1 to 2.

```
t.test(ToothGrowth$len[ToothGrowth$dose==1],
       ToothGrowth$len[ToothGrowth$dose==2], var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 1] and ToothGrowth$len[ToothGrowth$dose == 2]
## 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
```

Hypothesis 3: When we give a high dose (2) of Vitamin C, the length of Odontoblasts is independent of the supplement (either OJ or VC).

This time, we calculate a p-value of 96.4%, which indicates, the probability that the means differ from each other is really low and we can throw away this hypothesis. This indicates the means are equal for this example

```
t.test(ToothGrowth$len[ToothGrowth$dose==2 & ToothGrowth$supp=="OJ"],
       ToothGrowth$len[ToothGrowth$dose==2 & ToothGrowth$supp=="VC"],
       var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 2 & ToothGrowth$supp == "OJ"] and ToothGrowth$len[ToothGr
## 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
```