Statistical Inference Course Project

Author: Saffet Gokcen Sen

The project consists of two parts: A simulation exercise and a basic inferential data analysis. In the simulation exercise, a sample average distribution is going to be obtained by sampling an exponential distribution. The properties of the sample average distribution are to be compared to the expected properties by virtue of the central limit theorem. In the basic inferential data analysis part, the ToothGrowth data is to be analyzed.

Sample Average Distribution of the Exponential Distribution

Let X be a random variable with an exponential distribution. Then, the probability density function (pdf) of this random variable is given by [1]

$$f(x) = \lambda e^{-\lambda x} \text{ for } \lambda > 0 \text{ and } 0 \leq x < \infty$$

λ is the rate parameter. The mean and the variance of the exponential random variable are given as follows:

$$\mu = \frac{1}{\lambda},\ \sigma^2 = \frac{1}{\lambda^2}$$

The lifetimes of electronic components and the waiting times between rare events are among the models built using the exponential distribution [2].

A sample plot for the pdf of the exponential distribution with $\lambda = 0.2$ is given in the Fig. 1.

As can be observed in Fig.1, the pdf has its maximum value equal to λ at $x = 0$. The function decays to zero at a rate given by λ. For the specific case of $\lambda = 0.2$, the pdf becomes practically zero at $x = 30$.

Now, the sample average distribution is to be created by sampling the exponential distribution. Each time, 40 exponentials will be sampled and the average of them will be taken to obtain a sample for the sample average distribution. A total of 1000 samplings will be carried out.

The sample average distribution is expected to be centered around the mean of the original exponential population with a mean of

$$\frac{1}{\lambda} = \frac{1}{0.2} = 5$$

The mean of the sample average distribution is 4.9725173 which is very close to the expected mean. This closeness can also be observed in the Fig.2. The sample average distribution mean and the exponentially distributed population mean are very near to each other, i.e. the red vertical line and the blue vertical line are almost overlapping.

In the Fig. 3, the points one sample standard deviation away from the sample mean are shown with red vertical lines. The points one theoretical standard deviation away from the sample mean are shown with blue vertical lines. They are almost overlapping. The closeness of the standard deviations observed in the Fig. 3 can be proven using numerical values of the variances as well. The theoretical variance of the sample average distribution is given by

$$\frac{\sigma^2}{n} = \frac{\frac{1}{\lambda^2}}{n}$$

The numerical value of the theoretical variance is 0.625. The variance of the sample average distribution is 0.6508124. They are close to each other.

Fig.1 exponential distribution, rate=0.2

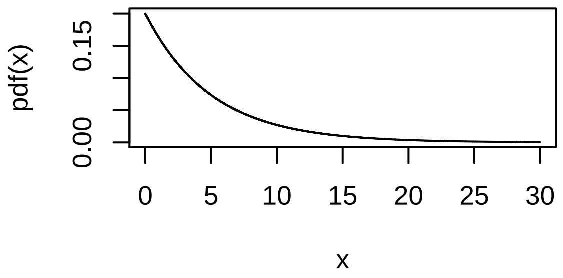


Fig.2 histogram, mean lines

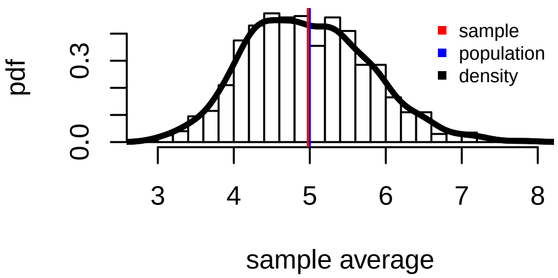


Fig.3 histogram, sd lines

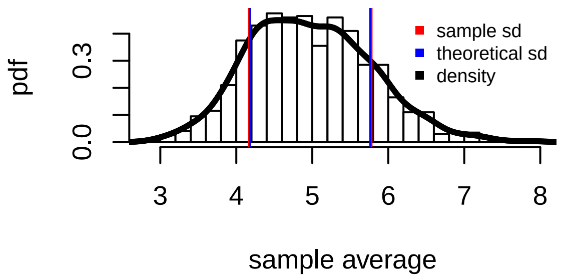
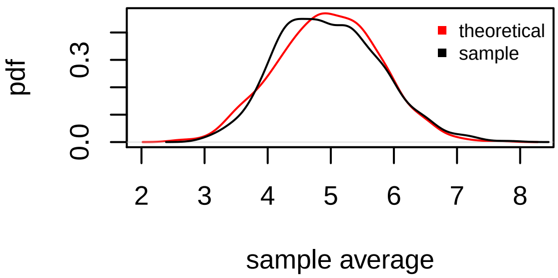


Fig.4 theoretical vs. sample densities



The histogram and the density estimate are similar to a normal distribution. The central limit theorem states that the sample average distribution is similar to a normal distribution if the sample size is big enough. Here, the sample average distribution looks like a normal distribution to the extent permitted by the sample size being equal to 40. As the sample size increases, the sample average distribution will look like a normal distribution more and more.

In the Fig.4, the similarity between the theoretical probability density function and the sample average probability density function can easily be detected.

Basic Inferential Data Analysis

In this part, the ToothGrowth data in the R datasets package will be analysed. The data will be explored and summed up first. Then, the effect of supplementaries and the dose on the tooth growth will be examined using confidence intervals and hypothesis tests.

After the duplicate rows are removed, the summary and the structure of the data set are obtained.

```
##       len        supp         dose      
##  Min.   : 4.20   OJ:28   Min.   :0.500  
##  1st Qu.:14.05   VC:27   1st Qu.:0.500  
##  Median :20.00           Median :1.000  
##  Mean   :19.05           Mean   :1.182  
##  3rd Qu.:25.35           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000
```

```
## 'data.frame':    55 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 5.2 7 16.5 ...
##  $ 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 1 ...
```

The unique dose values are 0.5, 1, 2. To have a quick overview of the correlations in the data, exploratory box plots are created.

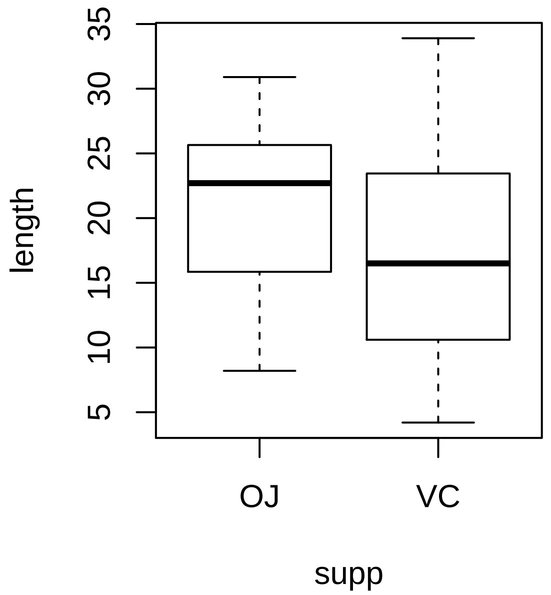


Fig.5 Variation of length with respect to supp

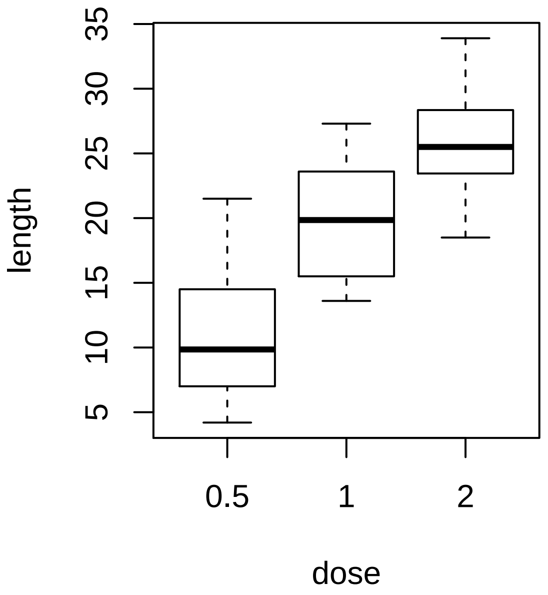


Fig.6 Variation of length with respect to dose

The boxplot in the Fig.5 suggests that the median tooth length for the OJ supp is greater than that for the VC supp. However, the VC supp has a wider range of tooth length. The boxplot in the Fig. 6 indicates that as the dose increases, the tooth length increases too.

Now, these exploratory observations will be checked using hypothesis testing. It is assumed that the population underlying the tooth length data has a normal (Gaussian) distribution. The OJ supp group and the VC supp group will be compared to each other using a two-sided t test. The two groups are assumed to be independent and to have different variances.

```
##
##  Welch Two Sample t-test
##
## data:  len_OJ and len_VC
## t = 1.781, df = 47.907, p-value = 0.08126
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4732938  7.8103308
## sample estimates:
## mean of x mean of y
##  20.85000  17.18148
```

The null hypothesis and the alternative hypothesis of the two-sided t test for the supp groups are as follows:

$$H_0 : \text{true difference in means is equal to zero.}$$

$$H_a : \text{true difference in means is not equal to zero.}$$

The p-value of the two-sided t test is 0.0812592. This p-value is greater than 0.05. Then, the alternative hypothesis is rejected. The two-sided t test states that the mean lengths of the OJ supp group and VC supp group are not significantly different from each other. Hence, it is concluded that the supp type does not have a significant effect on the tooth length.

Now, the effect of the dose on the tooth length is to be examined. First, dose 1.0 and dose 0.5 are to be compared to each other to determine if the mean tooth length differs between these two doses.

```
##
##  Welch Two Sample t-test
##
## data:  len_1 and len_05
## t = 6.0799, df = 33.939, p-value = 3.416e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.805623      Inf
## sample estimates:
## mean of x mean of y
##  20.05000  10.62222
```

The null hypothesis and the alternative hypothesis of the one-sided t test for the dose 1.0 and dose 0.5 groups are as follows:

$$H_0 : \text{true difference in means is equal to zero.}$$

$$H_a : \text{true difference in means is greater than zero.}$$

The p-value of the two-sided t test is 3.4161029×10^{-7}. This p-value is significantly smaller than 0.05. Hence, the null hypothesis is rejected. The dose 1.0 mean length is significantly greater than the dose 0.5 mean length.

Now, dose 2.0 and dose 1.0 are to be compared to each other to determine if the mean tooth length differs between these two doses.

```
##
##  Welch Two Sample t-test
##
## data:  len_2 and len_1
## t = 4.3299, df = 33.467, p-value = 6.391e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.676641      Inf
## sample estimates:
## mean of x mean of y
##  26.08421  20.05000
```

The null hypothesis and the alternative hypothesis of the one-sided t test for the dose 2.0 and dose 1.0 groups are as follows:

$$H_0 : \text{true difference in means is equal to zero.}$$

$$H_a : \text{true difference in means is greater than zero.}$$

The p-value of the two-sided t test is 6.3907414×10^{-5}. This p-value is significantly smaller than 0.05. Hence, the null hypothesis is rejected. The dose 2.0 mean length is significantly greater than the dose 1.0 mean length.

Hence, the t tests conclude that the dose factor has a significant effect on the tooth length. Specifically, as the dose increases, the mean tooth length increases too.

Appendix

The R code which is used to prepare this report is documented in this appendix.

```r
# x values for the pdf f(x)
x <- seq(0, 30, length.out=50000)
fx <- dexp(x, rate = 0.2) # pdf of the exponential distribution
# 1000*40 samples will be taken from the exponential distribution
num_of_samples <- 1000*40
lambda <- 0.2 # rate of the exponential distribution
# 1000*40 sample vector from the exponential distribution
exp_vector <- rexp(n=num_of_samples, rate=lambda)
num_rows <- 1000 # number of rows
num_columns <- 40 # number of columns
# the vector is cast into a 1000x40 matrix
sampling_matrix <- matrix(exp_vector, num_rows, num_columns)
# take the average of each row of the matrix
sample_avg_dist <- apply(sampling_matrix, 1, mean)
par(mfrow=c(2, 2)) # 2x2 plot window
the_cex <- 0.70 # legend size
main_cex <- 0.90 # title size
xlab <- "x"
ylab <- "pdf(x)"
# plot the exponential pdf
plot(x, fx, xlab=xlab, ylab=ylab, type="l")
title("Fig.1 exponential distribution, rate=0.2", cex.main=main_cex)
# the estimate to the sample average distribution density
the_density <- density(sample_avg_dist)
library(MASS) # the library MASS is loaded
# the histogram of the sample average distribution is produced
truehist(sample_avg_dist, nbins="FD", col="white", xlab="sample average", ylab="pdf")
# the density estimate is added to the plot
points(the_density, pch=20, cex=0.5)
# a red vertical line at the mean of the sample average distribution
abline(v=mean(sample_avg_dist), col="red")
# a blue vertical line at the mean of the exponentially distributed population
abline(v=5, col="blue")
legend("topright", legend=c("sample", "population", "density"),
col=c("red", "blue", "black"), pch=15, cex = the_cex, bty="n")
title("Fig.2 histogram, mean lines", cex.main=main_cex)
# the histogram of the sample average distribution is produced
truehist(sample_avg_dist, nbins="FD", col="white", xlab="sample average", ylab="pdf")
# the density estimate is added to the plot
points(the_density, pch=20, cex=0.5)
# a red vertical line at sample mean + sample sd
abline(v=mean(sample_avg_dist) + sd(sample_avg_dist), col="red")
# a red vertical line at sample mean - sample sd
abline(v=mean(sample_avg_dist) - sd(sample_avg_dist), col="red")
# a blue vertical line at sample mean + theoretical sd
abline(v=mean(sample_avg_dist) + sqrt(1/(lambda^2)/num_columns), col="blue")
# a blue vertical line at sample mean - theoretical sd
abline(v=mean(sample_avg_dist) - sqrt(1/(lambda^2)/num_columns), col="blue")
legend("topright", legend=c("sample sd", "theoretical sd", "density"),
       col=c("red", "blue", "black"), pch=15, cex = the_cex, bty="n")
```

```r
title("Fig.3 histogram, sd lines", cex.main=main_cex)
# the normal distribution predicted by the central limit theorem
theoretical_normal <- rnorm(n=1000, mean=5, sd=((1/lambda)/sqrt(num_columns)))
# the normal densiy predicted by the central limit theorem
theoretical_density <- density(theoretical_normal)
# the maximum value of the y axis
y_max <- max(max(theoretical_density$y), max(the_density$y))
# plot the theoretical and the sample average densities for comparison
plot(theoretical_density, pch=20, col="red", ylim=c(0, y_max),
     xlab="sample average", ylab="pdf", main = "")
points(the_density, type="l", col="black")
legend("topright", legend=c("theoretical", "sample"),
       col=c("red", "black"), pch=15, cex = the_cex, bty="n")
title("Fig.4 theoretical vs. sample densities", cex.main=main_cex)
library(dplyr) # load the package dplyr
data_set <- distinct(ToothGrowth) # remove duplicate rows
summary(data_set) # data summary
str(data_set) # data structure
# boxplots for checking correlations among length, dose and supp
with(data_set,{
     par(mfrow=c(1,2))
     boxplot(len ~ supp, xlab="supp", ylab="length", main="Fig.5 Variation of length
             with respect to supp", cex.main=0.7)
     boxplot(len ~ dose, xlab="dose", ylab="length", main="Fig.6 Variation of length
             with respect to dose", cex.main=0.7)
     })
# OJ supp group is created
group_OJ <- filter(data_set, supp == "OJ")
# VC supp group is created
group_VC <- filter(data_set, supp == "VC")
len_OJ <- group_OJ$len # OJ group lengths
len_VC <- group_VC$len # VC group lengths
# two-sided t test is applied
t_test_supp <- t.test(len_OJ, len_VC, alternative="two.sided", paired=FALSE,
                      var.equal=FALSE)
t_test_supp
group_05 <- filter(data_set, dose==0.5) # dose 0.5 group
group_1 <- filter(data_set, dose==1.0) # dose 1.0 group
len_05 <- group_05$len # 0.5 group lengths
len_1 <- group_1$len # 1.0 group lengths
# one-sided t test is applied
t_test_dose <- t.test(len_1, len_05, alternative="greater", paired=FALSE, var.equal=FALSE)
t_test_dose
group_2 <- filter(data_set, dose==2.0) # dose 2.0 group
len_2 <- group_2$len # 2.0 group lengths
# one-sided t test applied
t_test_dose2 <- t.test(len_2, len_1, alternative="greater", paired=FALSE, var.equal=FALSE)
t_test_dose2
```

References

[1] Montgomery, Douglas C. and Runger, George C. (2014), *Applied Statistics and Probability for Engineers*, 6th ed., Wiley.

[2] Wasserman, Lary (2003), *All of Statistics: A Concise Course in Statistical Inference*, Springer.