GroupA_HM2

I.El Gataa, M.Munini, E.Ruoppolo, M.Tallone

2022-10-20

Contents

FSDS - Chapter 3

Ex 3.12

Simulate random sampling from a normal population distribution with several n values to illustrate the law of large numbers

Solution

The law of large numbers states that, given a sequence $X_1, X_2, \dots, X_n$ of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $var(x_i) = \sigma^2$, than, for any $\epsilon > 0$ we have that:

$$\lim_{n \to \infty} P\left(|\bar{X}_n - \mu| \geq \epsilon\right) = 0$$

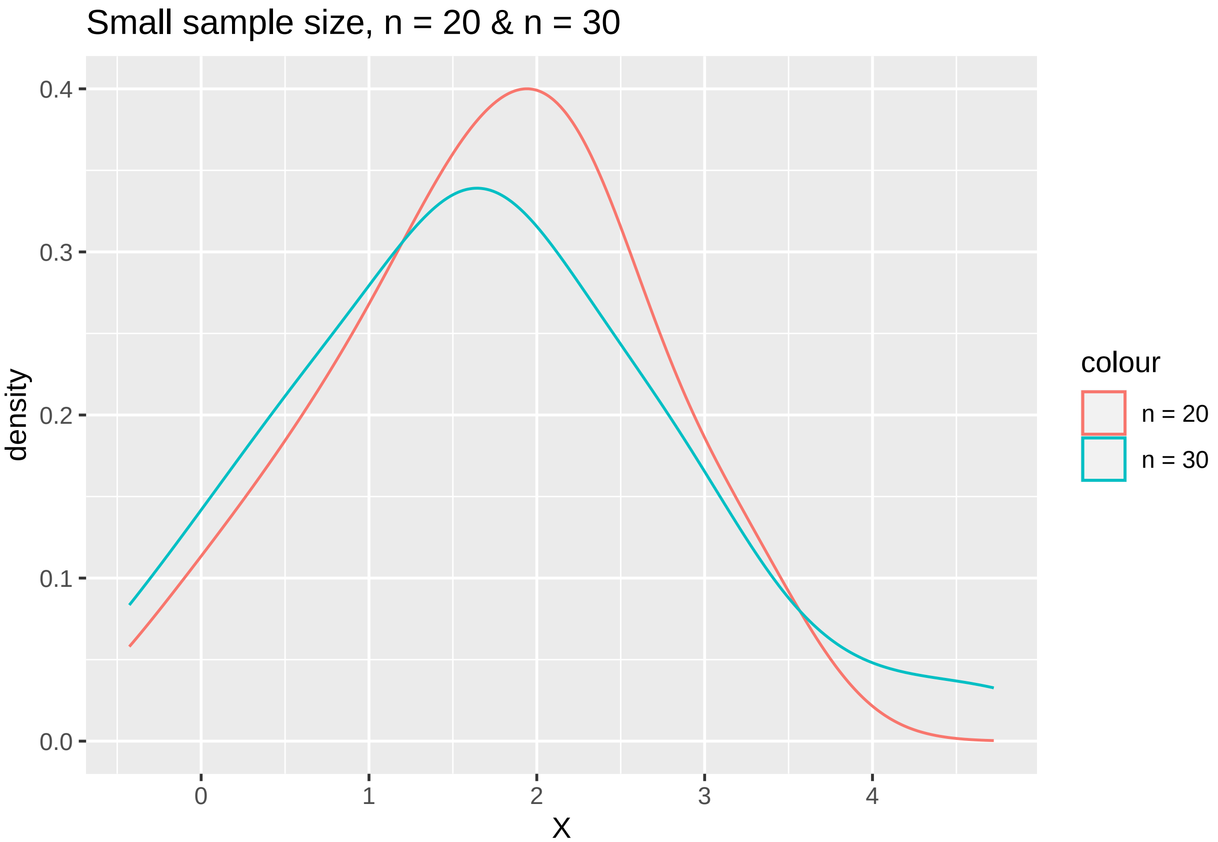
Where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

In this exercise, we have decided to simulate random samples from a Normal distribution, with $\mu = 2$ and $\sigma^2 = 1$. We will run eight different experiments, with respectively, 20 samples, 30, 100, 500, 10^3, 10^4,10^5 and 10^6.

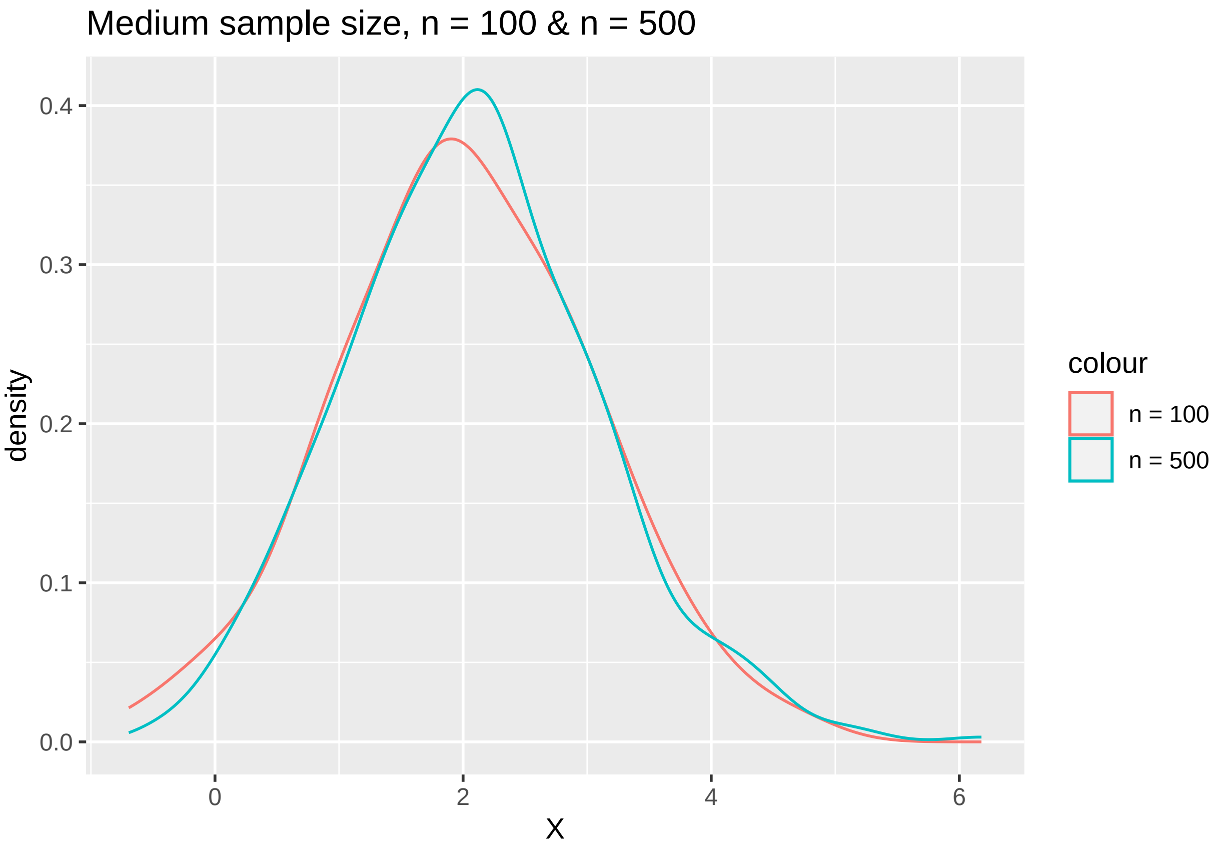
```r
n1 <- 20
n2 <- 30
n3 <- 100
n4 <- 500
n5 <- 1000
n6 <- 10000
n7 <- 100000
n8 <- 1000000
```

```
samples_n1 <- rnorm(n1, 2, 1)
samples_n2 <- rnorm(n2, 2, 1)
samples_n3 <- rnorm(n3, 2, 1)
samples_n4 <- rnorm(n4, 2, 1)
samples_n5 <- rnorm(n5, 2, 1)
samples_n6 <- rnorm(n6, 2, 1)
samples_n7 <- rnorm(n7, 2, 1)
samples_n8 <- rnorm(n8, 2, 1)
```

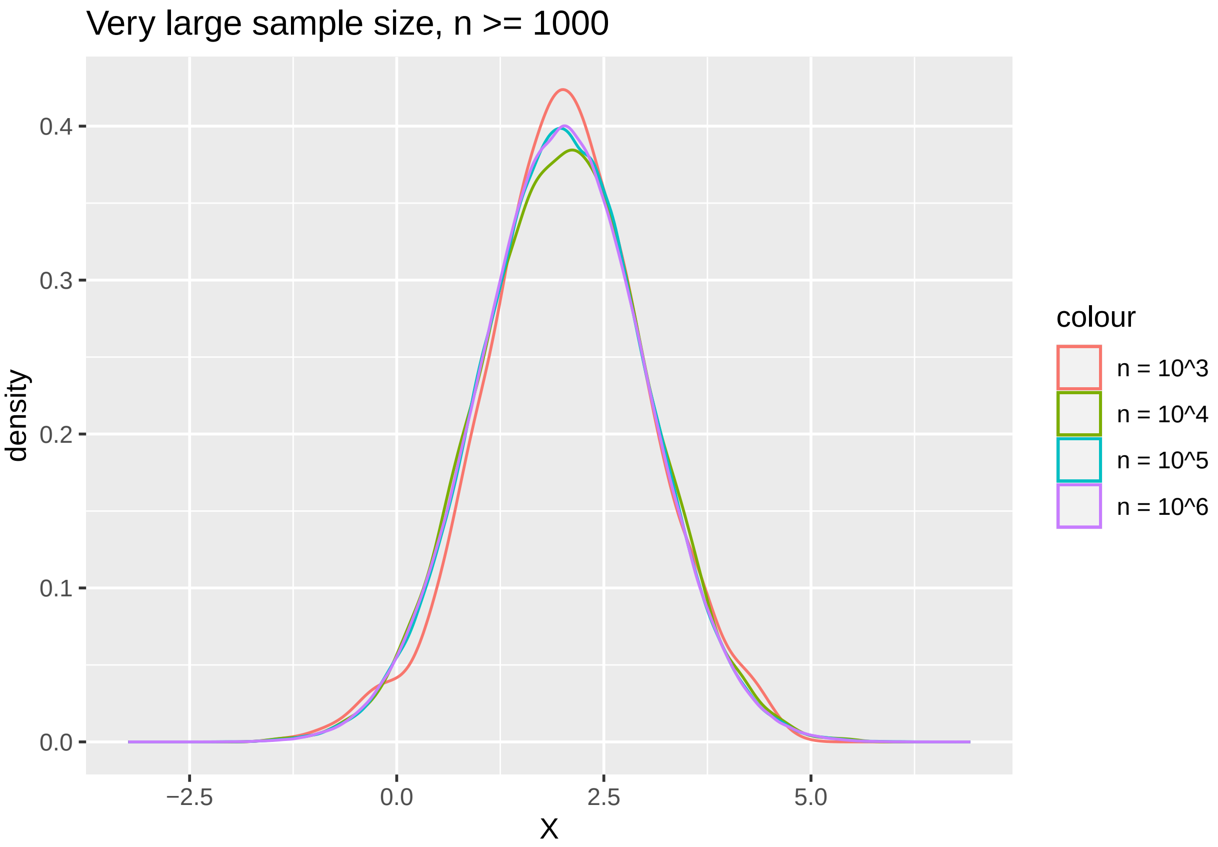
```
ggplot()+
  geom_density(aes(x = samples_n1, colour = "n = 20") )+
  geom_density(aes(x = samples_n2, colour = "n = 30") )+
  labs(title = "Small sample size, n = 20 & n = 30")+
  xlab("X")+
  ylab("density")
```



```
ggplot()+
  geom_density(aes(x = samples_n3, colour = "n = 100") )+
  geom_density(aes(x = samples_n4, colour = "n = 500") )+
  labs(title = "Medium sample size, n = 100 & n = 500")+
  xlab("X")+
  ylab("density")
```



```
ggplot()+
  geom_density(aes(x = samples_n5, colour = "n = 10^3") )+
  geom_density(aes(x = samples_n6, colour = "n = 10^4") )+
  geom_density(aes(x = samples_n7, colour = "n = 10^5") )+
  geom_density(aes(x = samples_n8, colour = "n = 10^6") )+
  labs(title = "Very large sample size, n >= 1000")+
  xlab("X")+
  ylab("density")
```



From the three density plots above it is possible to notice the effect of sample size. In particular, comparing the "Small sample size, n = 20 & n = 30" with "Medium sample size, n = 100 & n = 500" plot it is possible to notice that increasing n, the probability distributions' plots of normal random samples takes on the characteristics of a Gaussian distribution. When n is set to be equal or greater than 1000, then the probability distributions' plots of normal random samples are perfectly normal.

Ex 3.18

Sunshine City, which attracts primarily retired people, has 90,000 residents with a mean age of 72 years and a standard deviation of 12 years. The age distribution is skewed to the left. A random sample of 100 residents of Sunshine City has $\bar{y} = 70$ and $s = 11$.

- *Describe the center and spread of the (i) population distribution, (ii) sample data distribution. What shape does the sample data distribution probably have? Why?*
- *Find the center and spread of the sampling distribution of $\overline{\mathrm{Y}}$ for $n = 100$. What shape does it have and what does it describe?*
- *Explain why it would not be unusual to sample a person of age 60 in Sunshine City, but it would be highly unusual for the sample mean to be 60, for a random sample of 100 residents.*
- *Describe the sampling distribution of $\overline{\mathrm{Y}}$: (i) for a random sample of size $n = 1$; (ii) if you sample all 90,000 residents.*

Solution

- The problem introduces the distribution of the entire population of Sunshine City, made by $N = 90,000$ residents. We can imagine it as a given discrete probability distribution from which we extract random variables by sampling it. With this procedure indeed we are randomly taking some elements from this distribution. As the number n of the values increase we expect that their distribution starts to be similar to the one from which they are sampled. So, considering $n = 100$ a sufficient sampling value, considering a completely random sampling procedure, the sample should be distributed as the population, and its centre and spread should be similar to the *real* values, that are in this case $\mu = 72$ and $\sigma = 12$, in fact they are 70 and 11. Of course because of the randomness of the sampling procedure the values of the centre and the spread of the finite sample could not be the same of the true values, but as n approaches to N so will do the values of centre and spread.

- We are now considering the sampling distribution of $\bar{\mathrm{Y}}$. So we now have some different random samples each with $n = 100$ elements. The central limit theorem states that the distribution of the sample means, and in general of any sample statistics, will be approximately normal, no matter the shape of the population distribution. Clearly the samples for the applicability of the theorem it is required that the samples are all randomly taken from the population, that all the elements of each sample are independent and that the samples are large enough, and in our case $n = 100$ satisfies this condition. Since we suppose the other two condition to be true, the sample means will be approximately bell shaped, with expected value $\mu_{\bar{\mathrm{Y}}} = \mu = 72$ years and standard deviation $\sigma_{\bar{\mathrm{Y}}} = \sigma/\sqrt{n} = 1.2$ years. It says us that picking different random samples and each time computing the mean we expect that on average the mean will be 72 years, also these means will be normal distributed.
- It would not be strange to find a 60 years old person in Sunshine city. In fact the population distribution is left skewed, it means it has a long non-null "tail" on the left in which such an age, also considering the spread $\sigma = 12$ years that expresses the variability from the centre value, seems not strange. Anyway for a random sample of n finding a sample mean of 60 years would be really unusual for the reasons explained below. If the conditions of the CLT are all satisfied we expect a sample mean equal to the population mean with a low variability $\sigma_{\bar{\mathrm{Y}}} = 1.2$ years.
- In the first case we select every time a single value from the population. The mean is its self value, so practically we are randomly picking values from the population, they will be distributed approximately as the population itself. In the second case sampling the entire population means that each time the sample will always have mean value equal to the population one, so 72 years. Hence the sample distribution of means will not be a distribution at all since it will only have the same repeated value.

Ex 3.28

A survey is planned to estimate the population proportion π supporting more government action to address global warming. For a simple random sample, if π may be near 0.50, how large should n be so that the standard error of the sample proportion is 0.04?

Solution

The standard error for an estimate of a proportion π can be obtained with the following formula:

$$SE(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Since we are asking that $SE(\hat{\pi}) = 0.04$ and we are interested in the number of voters n, we can simply invert to find that:

$$n = \frac{\pi(1-\pi)}{(SE(\hat{\pi}))^2} = \frac{0.5 \cdot (1-0.5)}{(0.04)^2} = 156.25$$

Being the number of voters an integer number, it is possible to approximate the previous result to $n \approx 157$ in order to guarantee the desired standard error.

FSDS - Chapter 4

Ex 4.2

For a sequence of observations of a binary random variable, you observe the geometric random variable (Section 2.2.2) outcome of the first success on observation number $y = 3$. Find and plot the likelihood function.

Solution

The *pmf* for the geometric distribution with parameter $\pi \in [0, 1]$ is given by

$$P(Y = y) = f(Y) = (1 - \pi)^{Y-1}\pi$$

In general, the likelihood function for a geometric distribution with parameter π is obtained as

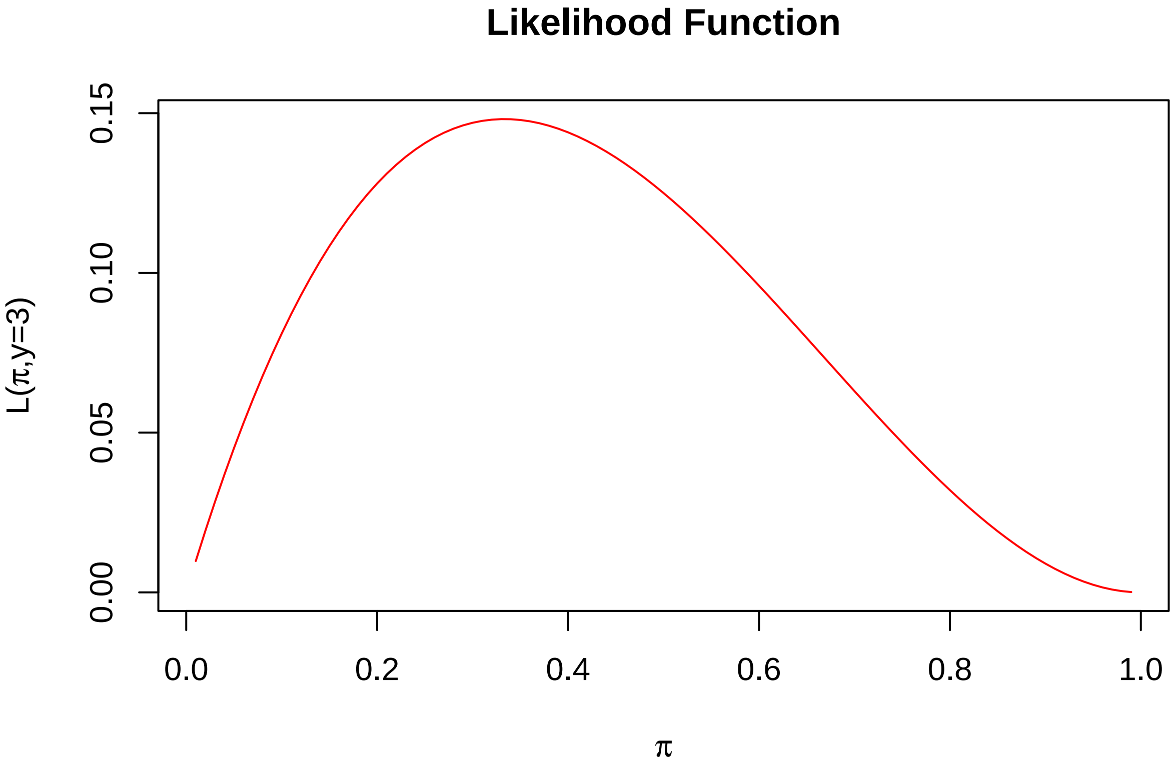
$$L(\pi|\mathbf{y}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n}(1 - \pi)^{y_i-1}\pi = \left[(1 - \pi)^{\sum_{i=1}^{n} y_i - n}\right](\pi^n)$$

In our case, since we only have one observation, that is $y = 3$, the Likelihood function simplifies down to:

$$L(\pi|y) = (1 - \pi)^2\pi$$

It is possible to plot this function in R by simply defining a sequence of values for π in its range $[0, 1]$ and calculating the relative values of the Likelihood function as done below.

```
# Define a range of pi values, where pv is a sequence from 0.01 to 0.99.
p_values <- seq(0.01, 0.99, by = 0.01)

# Calculate the likelihood for each pi value
likelihood <- (1 - p_values)^2 * p_values

# Plot the likelihood function
plot(p_values, likelihood, type = "l", col = "red",
     xlab = expression(paste(pi)),
     ylab = expression(paste("L(",pi,",y=3)")),
     main = "Likelihood Function")
```



Ex 4.4

For the `Students` *data file (Exercise 1.2 in Chapter 1) and corresponding population, find the ML estimate of the population proportion believing in life after death. Construct a Wald 95% confidence interval, using its formula (4.8). Interpret.*

Solution

```
Student <- read.table("https://stat4ds.rwth-aachen.de/data/Students.dat", header=TRUE)
Student$life <- as.factor(Student$life)
```

Upon an initial analysis of the `Students` dataset we can see that the possible answers for the question related to the belief of life after death were 3:

```
summary(Student$life)
```

```
##  1  2  3
## 31 13 16
```

where '1' stands for 'yes', '2' for 'no' while '3' for 'undecided'.

Since, it is asked to estimate the proportion of students who believe in life after death, we assume students undecided as students who do not believe. To achieved this the corresponding levels were changes as follows:

```
levels(Student$life) <- c(1, 0, 0)
Student$life <- as.numeric(Student$life)
```

```
length(Student$life)
```

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

The total number of observations in the dataset is 60.

ML theory shows that the log-likelihood function of a binomial distribution is maximized at $\hat{\pi} = \frac{\sum y_i}{n}$, so, in this case, it is the proportion of students who belives in life after death.

```
mle_pi <- sum(Student$life == 1)/length(Student$life)
mle_pi
```

```
## [1] 0.5166667
```

The MLE estimate is therefore 0.516.

The Wald confidence interval, at 95% of significance level is given by $\hat{\pi} \pm z_{\frac{\alpha=0.05}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$.

```
quant_z <- qnorm(0.975,mean=0,sd=1)
lower_b <- mle_pi - quant_z*sqrt((mle_pi*(1-mle_pi))/length(Student$life))
upper_b <- mle_pi + quant_z*sqrt((mle_pi*(1-mle_pi))/length(Student$life))
cat("Wald confidence interval: [",lower_b,",", upper_b,"]")
```

```
## Wald confidence interval: [ 0.3902218 , 0.6431115 ]
```

In conclusion, the Wald confidence interval is therefore $[0.3902; 0.6431]$.

Ex 4.38

For independent observations $y_1, \dots, y_n$ having the geometric distribution (2.1):

- *Find a sufficient statistic for π;*
- *Derive the ML estimator for π.*

Solution Let's first remind the geometric distribution:

$$f(y; \pi) = (1-\pi)^{y-1}\pi$$

- To find a sufficient statistic for the independent set $y_1, \dots, y_n$ from a geometric distribution we first need to write the likelihood function $L(\pi)$ that is:

$$L(\pi) = \prod_{i=1}^{n}(1-\pi)^{y_i-1}\pi = \pi^n \prod_{i=1}^{n}(1-\pi)^{y_i-1} = \left(\frac{\pi}{1-\pi}\right)^n \prod_{i=1}^{n}(1-\pi)^{y_i} = \left(\frac{\pi}{1-\pi}\right)^n (1-\pi)^{\sum_i^n y_i}$$

Since we can write it as a factorized function of type $g[T(\mathbf{y});\pi]$ where $T(\mathbf{y}) = \sum_i^n y_i$ and $L(\pi)$ depends on the data $\mathbf{y}$ only through this function that comprehends the entire data, and not single observations, we can say that $T(\mathbf{y}) = \sum_i^n y_i$ is a sufficient statistic.

- To find an ML estimator for π we have to find a value of π that maximize $\ell(\pi)$. First of all the log likelihood $\ell(\pi)$ is:

$$\ell(\pi) = \log L(\pi) = n[\log \pi - \log(1-\pi)] + \log(1-\pi)\sum_i^n y_i$$

So to find the ML estimator we shall find the value of π for which the derivative of ℓ is null:

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\ell(\pi) = n\left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) - \frac{1}{1-\pi}\sum_i^n y_i = 0$$

Hence:

$$\frac{1}{n}\sum_i^n y_i = \frac{1-\pi+\pi}{\pi} \implies \frac{1}{\pi} = \frac{1}{n}\sum_i^n y_i$$

Thus we find the ML estimator for π:

$$\hat{\pi} = \frac{n}{\sum_i^n y_i}$$

Ex 4.44

Refer to the previous two exercises. Consider the selling prices (in thousands of dollars) in the `Houses` *data file mentioned in Exercise 4.31.*

(a) *Fit the normal distribution to the data by finding the ML estimates of μ and σ for that distribution.*

(b) *Fit the log-normal distribution to the data by finding the ML estimates of its parameters.*

(c) *Find and compare the ML estimates of the mean and standard deviation of selling price for the two distributions.*

(d) *Superimpose the fitted normal and log-normal distributions on a histogram of the data. Which distribution seems to be more appropriate for summarizing the selling prices?*

Solution

(a) Assuming that the selling prices (x_i) found in the `Houses` dataset follow a normal distribution with mean μ and standard deviation σ, their log likelihood function is given by:

$$\begin{aligned}\ell(\mu,\sigma^2) &= \sum_{i=1}^n log\left(\frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right) \\ &= -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\end{aligned}$$

Using this function and solving the system of equations

$$\begin{cases}\frac{\partial\ell(\mu,\sigma^2)}{\partial\mu} = 0 \\ \frac{\partial\ell(\mu,\sigma^2)}{\partial\sigma^2} = 0\end{cases}$$

we find that the best estimators for the parameters μ and σ^2 are:

$$\begin{cases} \hat{\mu} = \bar{x} = \frac{1}{n}\sum_i x_i \\ \hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \end{cases}$$

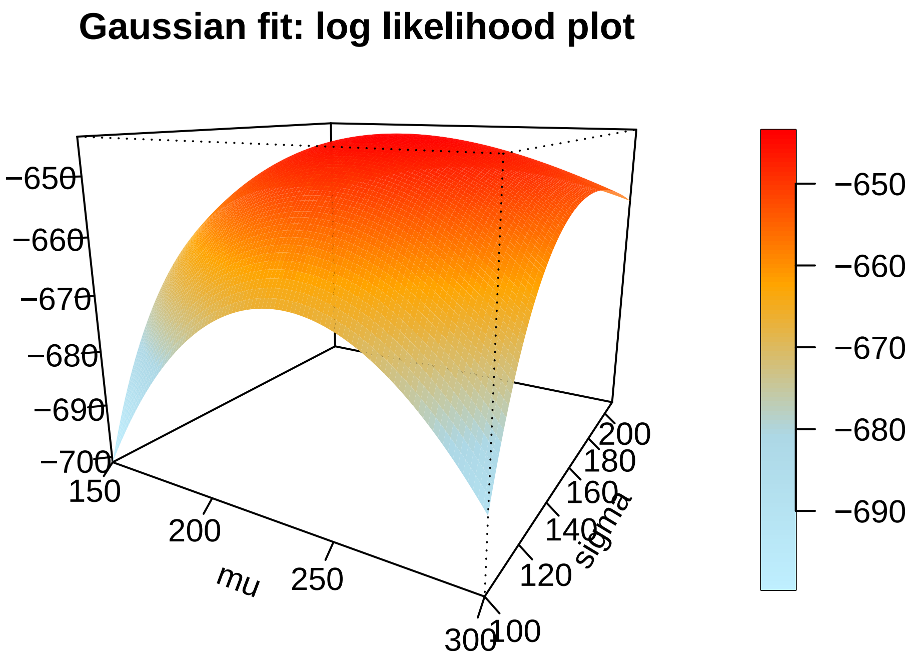
After importing the necessary dataset and libraries to conduct the fit in R studio, it's possible to define the Gaussian log likelihood function alongside the total number of observed prices in order to compute the point estimates directly.

```r
# Total number of observed prices
n <- length(houses$price)

# Gaussian log likelihood function
llike_gauss <- function(data, params){ sum(dnorm(x=data, mean=params[1], sd=params[2], log = TRUE)) }

# Estimates
mu_MLE <- mean(houses$price)
sigma2_MLE <- ((n-1)/n)*var(houses$price)

# Results
cat(" Estimated mean:", mu_MLE,"\n","Estimated standard deviation:",
    sqrt(sigma2_MLE),"\n","Value of Gaussian Log-Likelihood function: ",
    llike_gauss(houses$price, c(mu_MLE, sqrt(sigma2_MLE))))
```

```
##  Estimated mean: 232.9965
##  Estimated standard deviation: 151.1319
##  Value of Gaussian Log-Likelihood function:  -643.7092
```

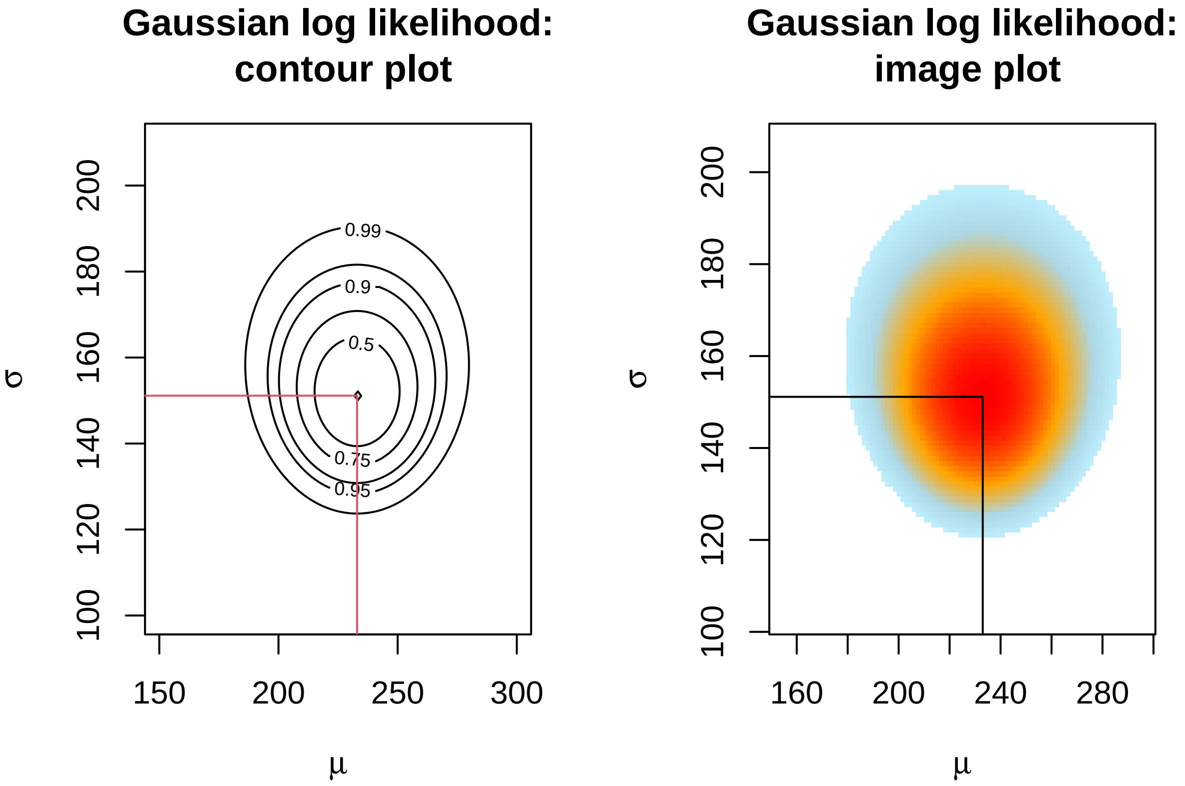
In addition, being the previously defined log likelihood a function of two parameters, it's possible to visualize the 3D plot of the function for values of the parameters in a neighborhood of the estimated ones, in order to visually verify that they actually maximize the function.

```r
# Define a parameter grid to plot the log-likelihood
mu_grid <- seq(150, 300, length=100)
sigma_grid <- seq(100, 210, length=100)
parvalues <- expand.grid(mu_grid, sigma_grid)

# Obtain the log-likelihood values for each point of the grid
n_llikvalues <- apply(parvalues, 1, llike_gauss, data = houses$price)
llikvalues <- matrix(n_llikvalues, nrow = length(mu_grid),
                     ncol = length(sigma_grid),byrow = FALSE)

# 3D plot with colormap
{
z<-llikvalues
nrz <- nrow(z)
ncz <- ncol(z)
# Create a function interpolating colors in the range of specified colors
jet.colors <- colorRampPalette( c("lightblue1","lightblue","orange", "red") )
# Generate the desired number of colors from this palette
nbcol <- 100
color <- jet.colors(nbcol)
# Compute the z-value at the facet centres
zfacet <- (z[-1, -1] + z[-1, -ncz] + z[-nrz, -1] + z[-nrz, -ncz])/4
# Recode facet z-values into color indices
facetcol <- cut(zfacet, nbcol)
persp(mu_grid,sigma_grid,llikvalues,
      main="Gaussian fit: log likelihood plot",
      zlab = "",
      xlab = expression(mu),
      ylab = expression(sigma),
      theta = 30, phi = 15,
      col = color[facetcol],
      expand=0.8,
      ticktype="detailed",
      border = NA)

image.plot(legend.only=T, zlim=range(zfacet), col=color)
}
```



The same graph can also be plotted as follows with contour lines for different confidence levels, where the vertical and horizontal lines are traced in correspondence of the estimated parameters.

```
# Define the confidence levels
{
conf_levels <- c(0, 0.5, 0.75, 0.9, 0.95, 0.99)
par(mfrow = c(1, 2))

# contour plot
contour(mu_grid, sigma_grid, llikvalues - max(llikvalues),
levels = -qchisq(conf_levels, 2)/2,
xlab = expression(mu), ylab = expression(sigma),
labels = as.character(conf_levels))
segments(0, sqrt(sigma2_MLE), mu_MLE, sqrt(sigma2_MLE), col=2, lwd = 1)
segments(mu_MLE, 0, mu_MLE, sqrt(sigma2_MLE), col=2, lwd = 1)
title("Gaussian log likelihood:\n contour plot")

# image plot
image(mu_grid, sigma_grid, llikvalues - max(llikvalues),
zlim = c(-6, 0), col = color,
xlab = expression(mu), ylab = expression(sigma))
segments(0, sqrt(sigma2_MLE), mu_MLE, sqrt(sigma2_MLE), col=1, lwd = 1)
segments(mu_MLE, 0, mu_MLE, sqrt(sigma2_MLE), col=1, lwd = 1)
title("Gaussian log likelihood:\n image plot")
}
```



(b) In this case, assuming the the house prices (x_i) are distributed following a Log-Normal distribution, we can find new estimates for the parameters using the relative log likelihood function that can be written as:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} log\left(\frac{1}{x_i \cdot \sqrt{2\pi\sigma^2}} exp\left[-\frac{(log(x_i - \mu)^2}{2\sigma^2}\right]\right)$$

$$= -\frac{n}{2} log(2\pi x_i^2\ \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (log(x_i) - \mu)^2$$

Solving the same system of equations of the previous point with this function we obtain that the best estimators for the parameters are:

$$\begin{cases} \hat{\mu} = \frac{1}{n} \sum_i log(x_i) \\ \hat{\sigma}^2 = \frac{1}{n} \sum_i [log(x_i) - \hat{\mu}]^2 \end{cases}$$

Following the same procedure in R Studio, we obtain as estimates:

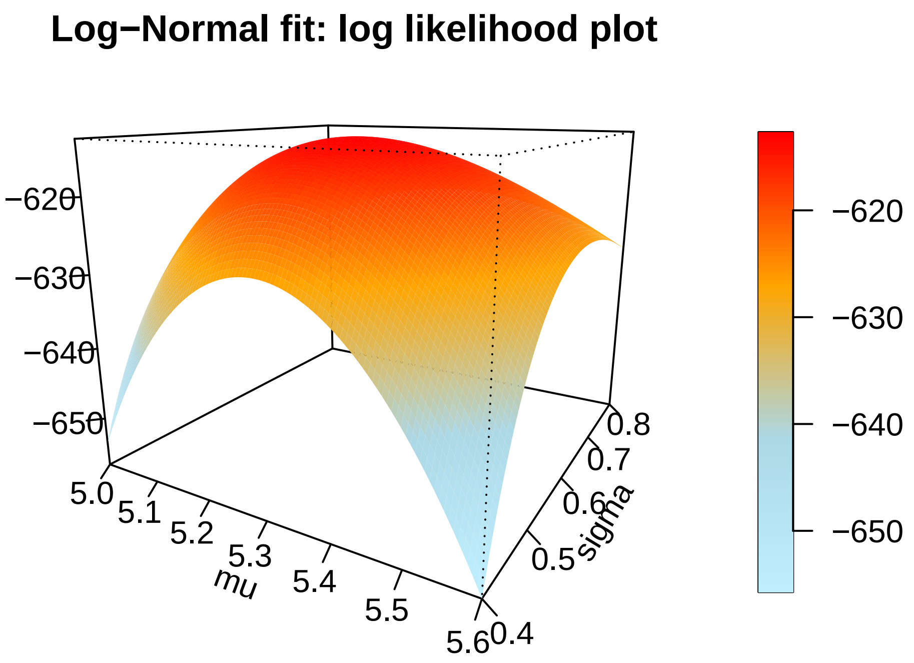
```
# Log-Likelihood of the Log-Normal distribution
llike_lnorm <- function(data, params){ sum(dlnorm(x=data, meanlog=params[1],
                                                  sdlog=params[2], log = TRUE)) }

#Estimates
lmu_MLE <- mean(log(houses$price))
```

```
lsigma2_MLE <- ((n-1)/n)*var(log(houses$price))

cat(" Estimated mean:", lmu_MLE,"\n","Estimated standard deviation:",
    sqrt(lsigma2_MLE),"\n","Value of Log-Likelihood function: ",
    llike_lnorm(houses$price, c(lmu_MLE, sqrt(lsigma2_MLE))))
```

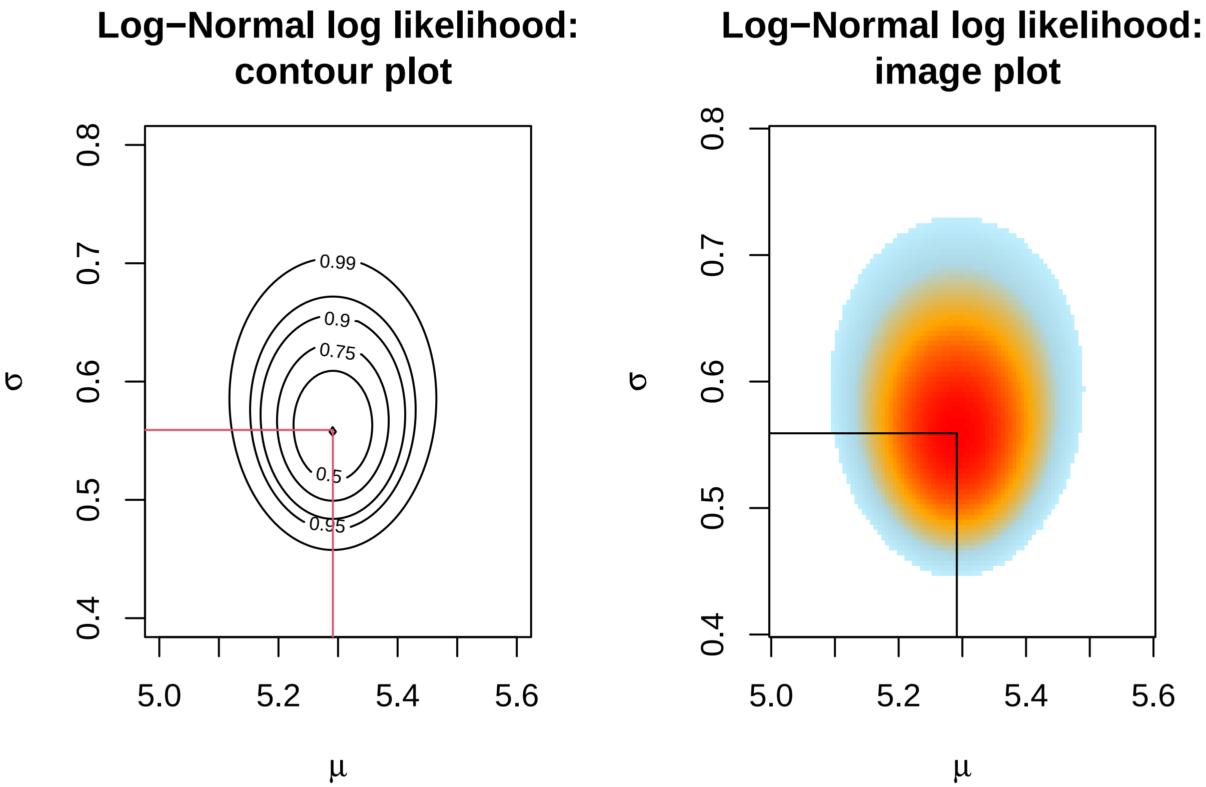
```
##  Estimated mean: 5.291323
##  Estimated standard deviation: 0.5591038
##  Value of Log-Likelihood function:  -612.8841
```

And plotting as before the log likelihood graphs we get the following:

```
# Define a parameter grid to plot the log-likelihood
mu_grid <- seq(5, 5.6, length=100)
sigma_grid <- seq(0.4, 0.8, length=100)
parvalues <- expand.grid(mu_grid, sigma_grid)

# Obtain the log-likelihood values for each point of the grid
n_llikvalues <- apply(parvalues, 1, llike_lnorm, data = houses$price)
llikvalues <- matrix(n_llikvalues, nrow = length(mu_grid),
                     ncol = length(sigma_grid),byrow = FALSE)

# 3D plot
{
z<-llikvalues
nrz <- nrow(z)
ncz <- ncol(z)
# Create a function interpolating colors in the range of specified colors
jet.colors <- colorRampPalette( c("lightblue1","lightblue","orange", "red") )
# Generate the desired number of colors from this palette
nbcol <- 100
color <- jet.colors(nbcol)
# Compute the z-value at the facet centres
zfacet <- (z[-1, -1] + z[-1, -ncz] + z[-nrz, -1] + z[-nrz, -ncz])/4
# Recode facet z-values into color indices
facetcol <- cut(zfacet, nbcol)

persp(mu_grid,sigma_grid,llikvalues,
      main="Log-Normal fit: log likelihood plot",
      zlab = "",
      xlab = expression(mu),
      ylab = expression(sigma),
      theta = 30, phi = 15,
      col = color[facetcol],
      expand=0.8,
      ticktype="detailed",
      border = NA)

image.plot(legend.only=T, zlim=range(zfacet), col=color)
}
```



```
# Define the confidence levels
{
conf_levels <- c(0, 0.5, 0.75, 0.9, 0.95, 0.99)
par(mfrow = c(1, 2))

# contour plot
contour(mu_grid, sigma_grid, llikvalues - max(llikvalues),
levels = -qchisq(conf_levels, 2)/2,
xlab = expression(mu), ylab = expression(sigma),
labels = as.character(conf_levels))
segments(0, sqrt(lsigma2_MLE), lmu_MLE, sqrt(lsigma2_MLE), col=2, lwd = 1)
segments(lmu_MLE, 0, lmu_MLE, sqrt(lsigma2_MLE), col=2, lwd = 1)
title("Log-Normal log likelihood:\n contour plot")

# image plot
image(mu_grid, sigma_grid, llikvalues - max(llikvalues),
zlim = c(-6, 0), col = color,
xlab = expression(mu), ylab = expression(sigma))
segments(0, sqrt(lsigma2_MLE), lmu_MLE, sqrt(lsigma2_MLE), col=1, lwd = 1)
segments(lmu_MLE, 0, lmu_MLE, sqrt(lsigma2_MLE), col=1, lwd = 1)
title("Log-Normal log likelihood:\n image plot")
}
```



(c) In the first case, since the data were fitted using a Gaussian distribution, the mean of the the selling prices and the standard deviation are the ones estimated for the Gaussian distribution itself:

$$\begin{cases} \hat{\mu}_{gauss} \approx 233 \\ \hat{\sigma}_{gauss} \approx 151 \end{cases}$$

In the second scenario, since the distribution used to fit the data is the Log-Normal distribution, the estimated parameters μ and σ give a mean and a standard deviation for the data in a logarithmic scale. Therefore, in order to compare the results with the ones from the first fit we should apply the formulas for the expected value (μ_{log}) and the standard deviation (σ_{log}) of the Log-Normal distribution:

$$\mu_{log} = exp\left(\mu + \frac{\sigma^2}{2}\right) \quad , \quad \sigma_{log} = \sqrt{(e^{\sigma^2} - 1) \cdot exp\,(2\mu + \sigma^2)}$$

We can get these value in R with the following code.

```
lmean <- exp(lmu_MLE+0.5*lsigma2_MLE)
lsigma <- sqrt((exp(lsigma2_MLE)-1)*exp(2*lmu_MLE+lsigma2_MLE))
cat(" Mean exponential:", lmean,"\n","Standard deviation exponential:", lsigma,"\n")
```

```
##  Mean exponential: 232.2052
##  Standard deviation exponential: 140.6655
```

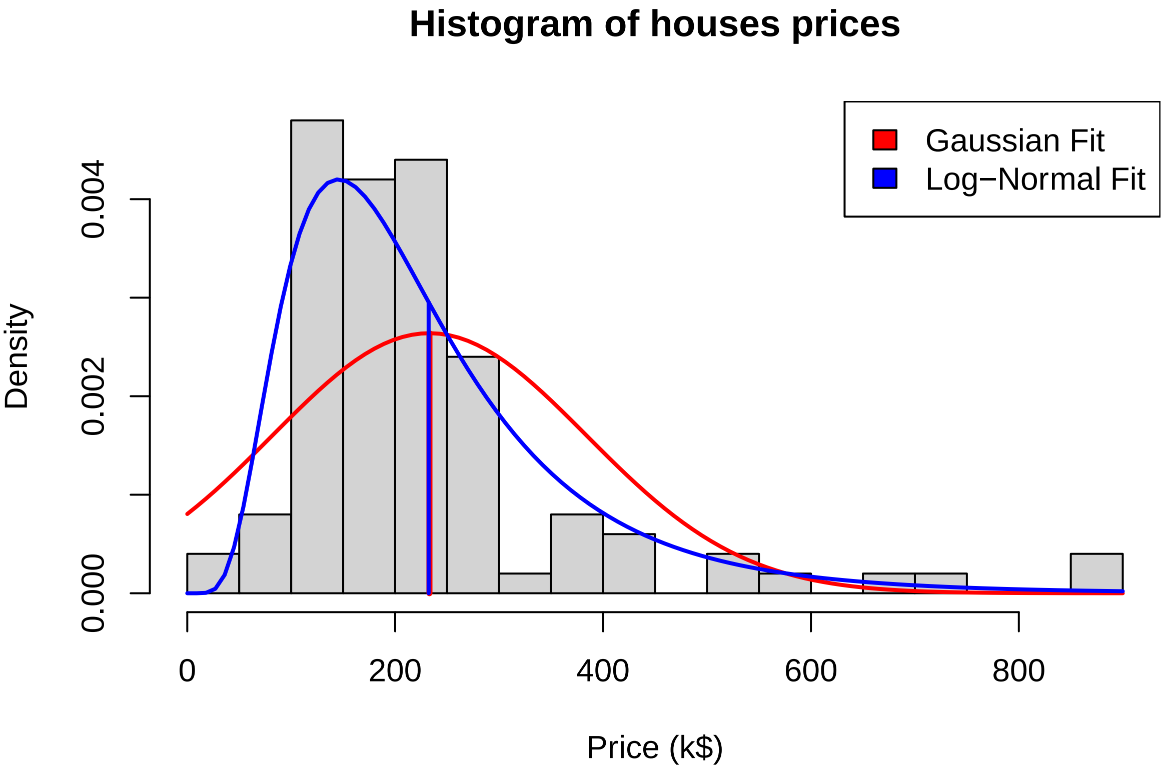
Which means that, approximately:

$$\begin{cases} \hat{\mu}_{log} \approx 232 \\ \hat{\sigma}_{log} \approx 140 \end{cases}$$

Having obtained comparable values, it's now possible to see that the mean and the standard deviation estimated for the selling prices do not seem to change significantly from one fit to the other. By only looking at the estimated mean and standard deviation the two models seem to describe the same phenomenon.

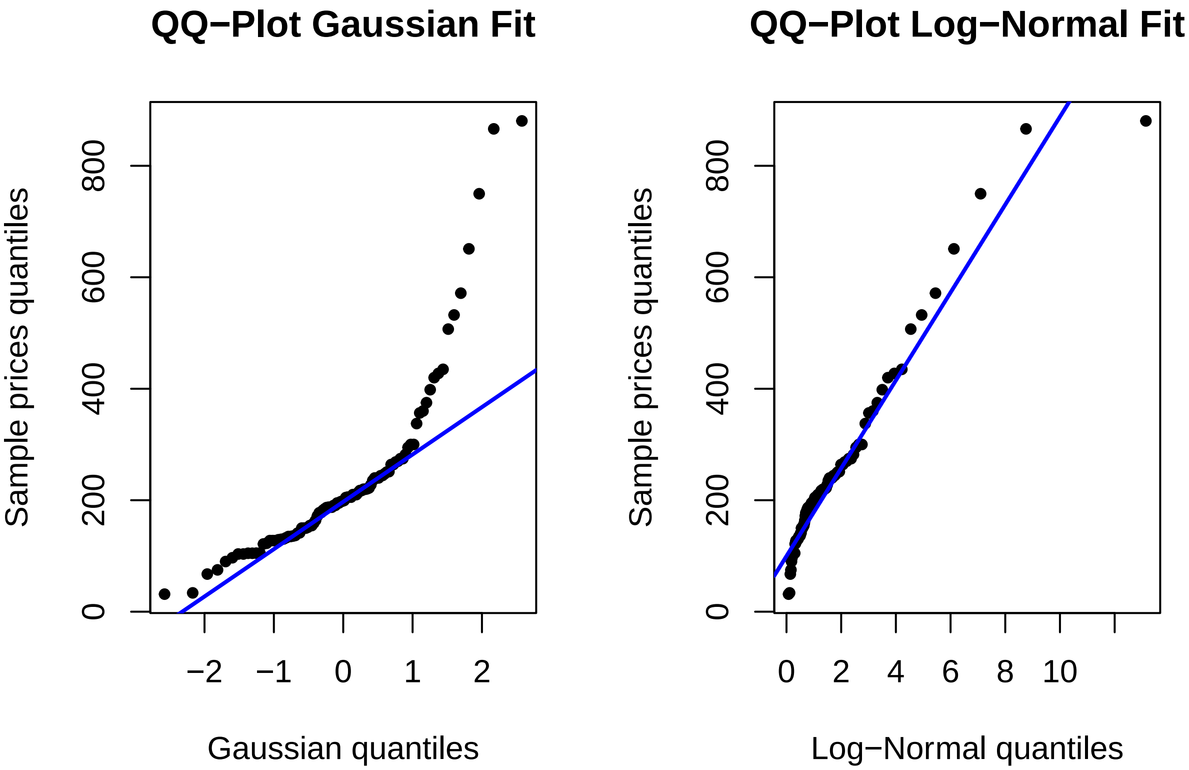
(d) However, by superimposing the two distribution to the histogram of the prices it is possible to better visualize the difference between the two models.

```
# Superimpose Gaussian plot
{hist(houses$price, prob=TRUE, nclass = 15, main = "Histogram of houses prices",
      xlab="Price (k$)", ylab="Density")

curve(dnorm(x, mean = mu_MLE, sd = sqrt(sigma2_MLE)), lwd = 2, col="red", add = TRUE)
segments(mu_MLE, 0, mu_MLE, dnorm(mu_MLE, mu_MLE, sqrt(sigma2_MLE)), col="red", lwd = 3)

curve(dlnorm(x, mean = lmu_MLE, sd = sqrt(lsigma2_MLE)), lwd = 2, col="blue", add = TRUE)
dlmean <- dlnorm(lmean, lmu_MLE, sqrt(lsigma2_MLE))
segments(lmean, 0, lmean, dlmean, col="blue", lwd = 2)

legend(x = "topright", legend=c( "Gaussian Fit", "Log-Normal Fit" ),
       fill = c(col ="red",col ="blue"))
}
```



Even if the two distributions produced similar means and standard deviations, among them the Log-Normal one seems to be the better choice for fitting the data due to their high skewness. This is also supported by the qq-plots obtainable for the two fits.

```
{par(mfrow = c(1, 2))
qqPlot(houses$price, dist="norm", pch=20, main="QQ-Plot Gaussian Fit",
       xlab="Gaussian quantiles", ylab="Sample prices quantiles",
       envelope=FALSE, grid=FALSE, id=FALSE)

qqPlot(houses$price, dist="lnorm", pch=20, main="QQ-Plot Log-Normal Fit",
       xlab="Log-Normal quantiles", ylab="Sample prices quantiles",
       envelope=FALSE, grid=FALSE, id=FALSE)
}
```



As it's possible to see on the qq-plot for the Gaussian model on the left, the fact that the last points detach form the line suggests that there is not great agreement between the measured data and the theoretical Gaussian distribution and also indicates that the distribution of the data is skewed to the right as we've already deduced from the previous histogram.

To complete our analysis, we have decided to perform the Kolmogorov-Smirnov test, a non parametric test used for goodness-of-fit. In particular, we will run two different one-sample KS test, to compare the data from the dataset with a reference probability distribution, in this case the Gaussian and Log-Normal distributions.

```r
ks.test(houses$price, "pnorm", mean = 233, sd = 151 )
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  houses$price
## D = 0.2019, p-value = 0.0005759
## alternative hypothesis: two-sided
```

Considering a significance level of 5%, the p-value of this test is much lower than 0.05, so the data does not provide enough evidence to accept H_0, so there is not evidence that the prices are distributed as a Gaussian, with $\mu = 233$ and $\sigma = 151$.

```r
ks.test(houses$price, "plnorm", meanlog = 5.2913, sdlog = 0.5591 )
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  houses$price
## D = 0.091335, p-value = 0.3746
## alternative hypothesis: two-sided
```

Considering a significance level of 5%, the p-value of this test is much greater than 0.05, so the data does provide enough evidence to accept H_0, so there is evidence that the prices are distributed as a Log-Normal distribution, with $\mu_{log} = 5.2913$ and $\sigma_{log} = 0.5591$.

Ex 4.48

For a simple random sample of n subjects, explain why it is about 95% likely that the sample proportion has error no more than $1/\sqrt{n}$ in estimating the population proportion. (Hint: To show this " $1/\sqrt{n}$ rule", find two standard errors when $\pi = 0.50$, and explain how this compares to two standard errors at other values of π .) Using this result, show that $n = 1/M^2$ is a safe sample size for estimating a proportion to within M with 95% confidence.

Solution

Let's define the function to compute the standard error for the estimate of proportion:

```
se_prop <- function(prop, len_n)  {
  val <- (((prop)*(1-prop))/len_n)**0.5
  return(val)
}
```

Let's understand what happens in the following example:

```
quant_zz <- qnorm(0.975,mean=0,sd=1)
quant_zz*se_prop(0.5,20)
```

```
## [1] 0.2191306
```

```
1/(20**0.5)
```

```
## [1] 0.2236068
```

Here, $\sqrt{1/20}$ is approximately the same value of $z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$.

Looking another example:

```
quant_zz*se_prop(0.5,60)
```

```
## [1] 0.1265151
```

```
1/(60**0.5)
```

```
## [1] 0.1290994
```

This happens because $z_{0.975} = 1.96$, which is approximately 2.

If $\pi = 0.5$, then $\sqrt{\frac{\pi(1-\pi)}{n}} = \frac{\pi}{\sqrt{n}}$, so $z_{0.975}\frac{\pi}{\sqrt{n}} = 2\frac{0.5}{\sqrt{n}} = \frac{1}{\sqrt{n}}$. This $\frac{1}{\sqrt{n}}$ is called margin of errors. In particular, this is the margin of errors at 95% of confidence.

Similarly, it is approximately the same if π is not 0.5, because $\sqrt{\pi(1-\pi)}$ is approximately 0.5.

```
quant_zz*se_prop(0.7,60)
```

```
## [1] 0.115953
```

```
1/(60**0.5)
```

```
## [1] 0.1290994
```

When designing an experiment, in this case the estimation of class's proportion in a population, it is possible that the researchers defined the significance level a priori. Also, it can be defined the margin of error a priori. In this way it is possible to define in advance the necessary sample size, n. When doing this, if there is no prior knowledge about teh proportion to estimate, is is possible to set $\hat{\pi} = 0.5$, since it is the most non informative choice.

Let's $M = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$. From this equation it is possible to define the sample size as $n = \frac{{z_{1-\frac{\alpha}{2}}}^2}{M^2}\hat{\pi}(1-\hat{\pi})$.

Then, as shown before, $z_{0.975} = 2$ approximately, so $z_{0.975}^2 = 4$. The $\hat{\pi}(1-\hat{\pi})$ is 0.25.

So, $n = \frac{1}{M^2}$ approximately.

FSDS - Chapter 5

Ex 5.2

When a government does not have enough money to pay for the services that it provides, it can raise taxes or it can reduce services. When the Florida Poll asked a random sample of 1200 Floridians which they preferred, 52% (624 of the 1200) chose raise taxes and 48% chose reduce services. Let π *denote the population proportion of Floridians who would choose raising taxes.Analyze whether this is a minority of the population* $(\pi < 0.50)$ *or a majority* $(\pi > 0.50)$ *by testing* $H_0 : \pi = 0.50$ *against* $H_\alpha : \pi \neq 0.50$ *. Interpret the P-value. Is it appropriate to "accept"* H_0 *? Why or why not?*

Solution

Let $n = 1200$ the number of Floridians which answered the poll, $n_1 = 624$ the number of people who voted for raising the taxes, which corresponds to the $\hat{\pi} = 52\%$ and $n_2 = 576$ those who voted for reducing the services, $1 - \hat{\pi} = 48\%$. Therefore the parameter π corresponds to the proportion of people wo wuold like t raise taxes. The two hypothesis are:

$$\begin{cases} H_0 : \pi = \pi_0 = 0.50 \\ H_\alpha : \pi \neq 0.50 \end{cases}$$

Where H_0 is the null hypothesis, or status quo, so the hypothesis we'd like to verify, and H_α is the alternative hypothesis. The sampling size n is sufficiently large so that we can assume the sampling distribution of $\hat{\pi}$ being normal under H_0 and we used the two tailed test because we don't know which direction the distribution takes.

The test statistic used to conduct the test is

$$z(\pi) = \frac{\pi - \pi_0}{SE(\pi_0)} = \frac{\pi - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

First of all we calculated the value of the test statistic and the p-value obtained with the give observations:

```r
n <- 1200
n1 <- 624
p <- n1/n
p0 <- 0.5

#computing std error
se0 <- sqrt((p0*(1-p0))/n)

#computing test statistic
Z <- (p-p0)/se0
pv <- 2*pnorm(Z, lower.tail =  F)
cat(" Value of the test statistic: ", Z, "\n Value of the p-value: ", pv)
```

```
##  Value of the test statistic:  1.385641
##  Value of the p-value:  0.1658567
```

Then, the `prop.test()` function has been used to conduct the test in R.

```r
prop.test(x=624, n=1200, p=0.5, alt="two.sided", conf.level = 0.95, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  624 out of 1200, null probability 0.5
## X-squared = 1.92, df = 1, p-value = 0.1659
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4917142 0.5481581
## sample estimates:
##    p
## 0.52
```

To double check the results, we have also done the hypothesis test *"by hand"* using probability tables rather than using R libraries. In this case the value we obtained for the test statistic is $z(\hat{\pi}) = 1.4284$ and from the normal quantile tables we got the p-value corresponding to a significance level of $\alpha = 0.05$:

$$p = F(z < -1.4184) + F(z > +1.4184) \approx 0.0778 + 0.0778 = 0.1556$$

The p-value obtained in R and the one we computed *"by hand"* differ slightly due to different approximation adopted, but are both not very small, so there is not a strong evidence against the null hypothesis $H_0 : \pi = \pi_0 = 0.50$. We can therefore conclude that it's plausible that the true value of the parameter is $\pi = 0.5$. In other words, there is insufficient evidence to determine whether the majority of the people wants to raise taxes ($\pi > 0.5$) or not ($\pi < 0.5$).

Ex 5.12

The example in Section 3.1.4 described an experiment to estimate the mean sales with a proposed menu for a new restaurant. In a revised experiment to compare two menus, on Tuesday of the opening week the owner gives customers menu A and on Wednesday she gives them menu B. The bills average \$22.30 for the 43 customers on Tuesday (s = 6.88) and \$25.91 for the 50 customers on Wednesday (s = 8.01). Under the strong assumption that her customers each night are comparable to a random sample from the conceptual population of potential customers, show how to compare the mean sales for the two menus based on (a) the P-value of a significance test, (b) a 95% confidence interval. Which is more informative, and why? (When used in an experiment to compare two treatments to determine which works better, a two-sample test is often called an A/B test.).

Solution

a) To compare the mean sales for the two menus with a significance test, given the assumption made by the problem we can use the standard t-test with the pooled standard deviation. First of all we estimate a pooled value s for the standard deviation:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

This is possible since s_1 and s_2 do not result greatly different. So in this case the estimated standard error is:

$$se = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The test consider as null hypothesis $H_0 : \mu_1 = \mu_2$, so $\mu_1 - \mu_2 = 0$, otherwise $H_\alpha : \mu_1 \neq \mu_2$. Hence the test statistic is:

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - 0}{SE}$$

and df is $df = n_1 + n_2 - 2$. So considering $(\overline{Y}_1, s_1, n_1) = (22.30, 6.88, 43)$ and $(\overline{Y}_2, s_2, n_2) = (25.91, 8.01, 50)$ we can compute:

```r
y1 <- 22.30
s1 <- 6.88
n1 <- 43
y2 <- 25.91
s2 <- 8.01
n2 <- 50

degf = n1+n2-2
s = sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(degf))
se = s * sqrt(1/n1+1/n2)
t = (y1-y2)/se
pvalue = 2*pt(q= t, df = degf, lower.tail = T)


cat(" Standard error: ", se,
    "\n Degrees of freedom: ", degf,
    "\n Value of the test statistic: ", t,
    "\n Value of the p-value: ", pvalue)
```

```
##  Standard error:  1.561853
##  Degrees of freedom:  91
##  Value of the test statistic:  -2.311357
##  Value of the p-value:  0.02307139
```

We can estimate the p-value as $p = 0.02$. This low value suggests that we should reject H_0 hence $\mu_1 \neq \mu_2$ and the mean sales between the two menus are different.

b) Now we use a 95% confidence interval to evaluate the difference between μ_1 and μ_2. We now repeat the procedure already used above. The confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,\, n_1+n_2-2}(se)$$

In our case $\alpha/2 = (1 - 0.95)/2 = 0.025$, $df = n_1 + n_2 - 2 = 91$ thus we can compute the quantities:

```r
y1 <- 22.30
s1 <- 6.88
n1 <- 43
y2 <- 25.91
s2 <- 8.01
n2 <- 50
s = sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(degf))
se = s * sqrt(1/n1+1/n2)
m_dif=y1-y2
t = qt(0.025, 91, lower.tail = F)
dif = t*se
m_dif
```

```
## [1] -3.61
```

```r
dif
```

```
## [1] 3.102429
```

The result is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025,91}(se) = -3.61 \pm 3.1 \text{ or } (-6.71, -0.51)$$

The interval not including 0 can be interpreted as sign that the two means are not equal. Also the interval covering only negative values may mean that $\mu_2 > \mu_1$.

So, by these two different analysis we can find that the hypothesis test only suggests that the hypothesis of having $\mu_1 = \mu_2$ should be rejected because of the low $p-$value, while the confidence interval analysis adds to the statement of incompatibility that the mean sale of menu b is higher than the one of menu a.

Ex 5.68

Explain why the confidence interval based on the Wald test of $H_0 : \theta = \theta_0$ is symmetric around $\hat{\theta}$ (i.e., having center exactly equal to $\hat{\theta}$. This is not true for the confidence intervals based on the likelihood-ratio and score tests.) Explain why such symmetry can be problematic when θ and $\hat{\theta}$ are near a boundary, using the example of a population proportion that is very close to 0 or 1 and a sample proportion that may well equal 0 or 1.

Solution

The Wald test statistic is defined as

$$W_e(\theta) = (\hat{\theta} - \theta)^2 \cdot J(\hat{\theta}) = \frac{(\hat{\theta} - \theta)^2}{(SE(\hat{\theta}))^2}$$

which, in the null hypothesis $H_0 : \theta = \theta_0$, approximately follows a chi-squared distribution with 1 degree of freedom($W_e(\theta) \dot{\sim} \chi^2_1$).

With this definition it's possible to build the $(1 - \alpha) \times 100\%$ confidence interval, that is formally defined as:

$$\{ \ \theta \ : \ W_e(\theta) \leq \chi^2_{1;1-\alpha} \ \}$$

where $\chi^2_{1;1-\alpha}$ is the α quantile of a chi-squared distribution with 1 degree of freedom. Using the definition of the test statistic, this interval can be rewritten as:

$$\begin{aligned}
\{ \ \theta \ : \ W_e(\theta) \leq \chi^2_{1;1-\alpha} \ \} &= \left\{ \ \theta \ : \ \frac{(\hat{\theta} - \theta)^2}{(SE(\hat{\theta}))^2} \leq \chi^2_{1;1-\alpha} \ \right\} \\
&= \left\{ \ \theta \ : \ (\hat{\theta} - \theta)^2 \leq (SE(\hat{\theta}))^2 \cdot \chi^2_{1;1-\alpha} \ \right\} \\
&= \left\{ \ \theta \ : \ -SE(\hat{\theta}) \cdot \sqrt{\chi^2_{1;1-\alpha}} \leq \hat{\theta} - \theta \leq +SE(\hat{\theta}) \cdot \sqrt{\chi^2_{1;1-\alpha}} \ \right\} \\
&= \left\{ \ \theta \ : \ \hat{\theta} - SE(\hat{\theta}) \cdot \sqrt{\chi^2_{1;1-\alpha}} \leq \quad \theta \quad \leq \hat{\theta} + SE(\hat{\theta}) \cdot \sqrt{\chi^2_{1;1-\alpha}} \ \right\}
\end{aligned}$$

This demonstrated that the interval is indeed symmetric around $\hat{\theta}$, with width equal to $2 \cdot SE(\hat{\theta}) \cdot \sqrt{\chi^2_{1;1-\alpha}}$.

To explain why this symmetry can be problematic when θ and $\hat{\theta}$ assume values close to the boundary we can consider the example of a population proportion $\pi \in [0, 1]$. Since the standard error for a proportion π can be obtained with

$$SE(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

then, the Wald test statistic in this case becomes:

$$W_e(\theta) = \frac{n(\hat{\pi} - \pi)^2}{\hat{\pi}(1 - \hat{\pi})}$$

This implies that the previously introduced confidence interval can be expressed as

$$\left\{ \ \pi \in [0, 1] \ : \ \hat{\pi} - \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n} \chi^2_{1;1-\alpha}} \leq \quad \pi \quad \leq \hat{\pi} + \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n} \chi^2_{1;1-\alpha}} \ \right\}$$

and it's immediately possible to notice that the width of the interval, given by $2 \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} \chi^2_{1;1-\alpha}}$, goes to zero for any given confidence level as the sample proportion $\hat{\pi}$ approaches either 0 or 1 (boundary values). This behavior results problematic when conducting a hypothesis test since it most likely leads to the rejection of the null hypothesis (in favor of the alternative one) independently of both the value assumed by the test statistic $W_e(\theta)$ and the chosen significance level to conduct the test.