Econometrics: Methods and Applications

Diego López Tamayo * Based on MOOC by Erasmus University Rotterdam

Contents

"There are two things you are better off not watching in the making: sausages and econometric estimates." -Edward Leamer

*El Colegio de México, diego.lopez@colmex.mx

Requirements

- You need some basic background in statistics and matrices.
- You can use any statistical package that is available to you, for example packages like R, Stata, EViews and other. The main requirement is that you can run regressions to get coefficients and standard errors.

Introduction

The following notes and code chunks are made in R statiscial package, you can also found the Do-File for Stata 16 in the download sections of my website. Both files follow the same structure and use the same data sets.

All the data sets are downlable from my Github repository

In the following notes we will cover: **simple regression, multiple regression, model specification, endogeneity, binary choice, and time series.**

For example:

Suppose you wish to predict the number of airplane passengers worldwide for next year.

- In **simple regression**, you use a single factor to explain airplane passenger traffic, for example, worldwide economic growth.
- In **multiple regression**, you use additional explanatory factors, such as the oil price, the price of tickets, and airport taxes.
- **Model specification** answers the question which factors to incorporate in the model, and in which way.
- **Endogeneity** is concerned with possible reverse causality. For example, if economic growth does not only lead to more air traffic, but reversely, increased air traffic also influences economic growth.
- **Binary choice** considers the micro level of individual decisions whether or not to travel by plane, in terms of factors like family income and the price of tickets.
- In **time series** analysis, you analyze trends and cycles in airplane passenger traffic in previous years, to predict future developments.

Building Blocks

Required background on matrices, probability and statitics:

Matrices Recommended: S.Grossman. *Elementary Linear Algebra*

- Matrix summation, matrix multiplication
- Square matrix, diagonal matrix, identity matrix, unit vector
- Transpose, trace, rank, inverse
- Positive and negative (semi)definite matrix
- Gradient vector, Hessian matrix
- First and Second Order Conditions for optimization of vector functions

Probability Recommended: Casella & Berger. *Statistical Inference*

- Univariate and multivariate random variables
- Probability density function (pdf)
- Cumulative density function (cdf)
- Expectation, expectation of functions
- Mean, variance, standard deviation
- Covariance, correlation
- Mean, variance, and covariance of linear transformations
- Independence
- Higher order moments, skewness, kurtosis

- Normal distribution, standard normal distribution
- Multivariate normal distribution
- Linear transformations of normally distributed random variables
- Chi-squared distribution, Student t-distribution, F-distribution

Statistics Recommended: J. Wooldridge *Introductory Econometrics: A Modern Approach*

- Statistic, estimator, estimate
- Standard error
- Confidence interval
- Unbiasedness
- Efficiency
- Consistency
- Sample mean, sample variance
- Hypothesis, null and alternative hypothesis
- Test statistic
- Type I and Type II error
- Size and power of a statistical test
- Significance level
- Critical value, critical region
- P-value
- T-statistic, Chi-squared statistic, F-statistic

Parameter Estimation

Suppose you have 26 observations of the yearly return on the stock market. We call a set of observations a sample. Bellow, you will see a histogram of the sample. The returns in percentages are on the x-axis and the y-axis gives the frequency. The sample mean equals 9.6%. What can we learn from this sample mean about the mean of the return distribution over longer periods of time? Can we be sure that the true mean is larger than zero?

Dataset S1

Contains 26 yearly returns based on the S&P500 index. Returns are constructed from end-of-year prices Pt as rt = (Pt – Pt-1)/Pt-1. Data has been taken from the public FRED database of the Federal Reserve Bank of St. Louis.

```
dataset_s1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset_s1.csv")
```

A simple stat description of our dataset:

```
summary(dataset_s1$Return)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.384858  0.007479  0.125918  0.096418  0.255936  0.341106
```

```
# mean,median,25th and 75th quartiles,min,max
```

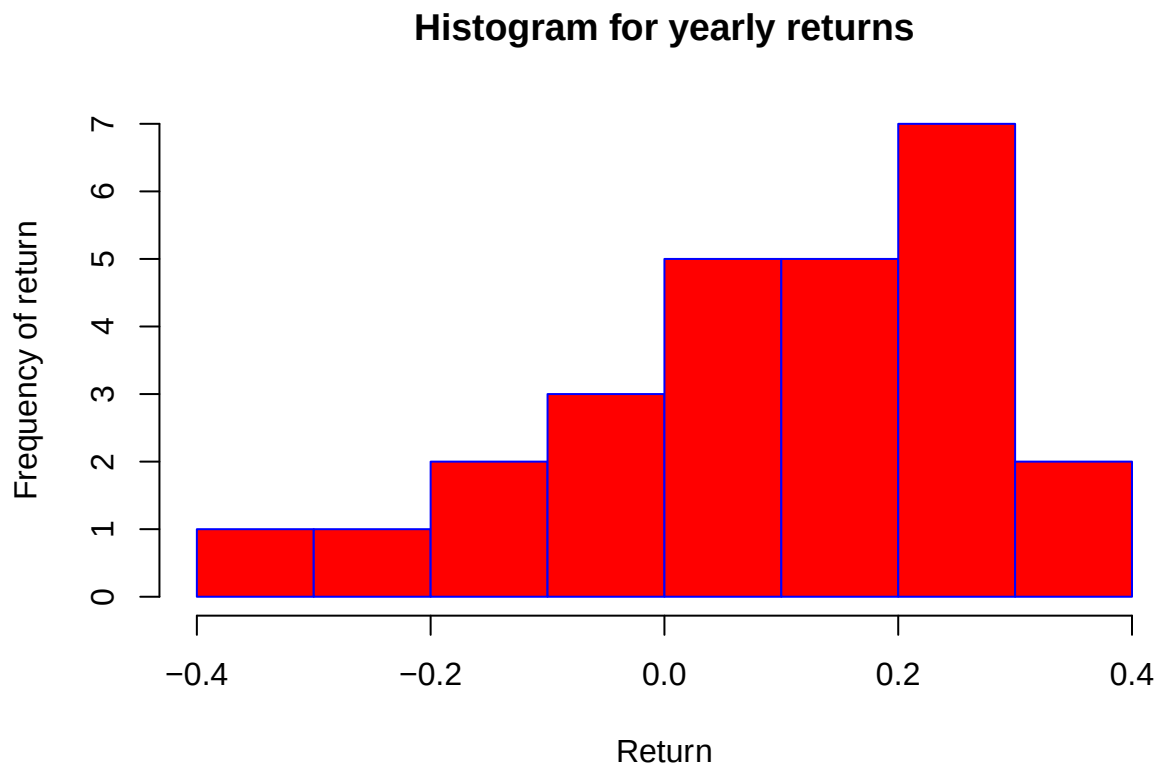
```
Hmisc::describe(dataset_s1$Return)
```

```
## dataset_s1$Return
##         n   missing  distinct      Info      Mean       Gmd       .05       .10
##        26         0        26         1   0.09642    0.2008 -0.207851 -0.115909
##       .25       .50       .75       .90       .95
##  0.007479  0.125918  0.255936  0.284259  0.306564
##
## lowest : -0.3848579 -0.2336597 -0.1304269 -0.1013919 -0.0655914
## highest:  0.2666859  0.2725047  0.2960125  0.3100818  0.3411065
```

```
# n, nmiss, unique, mean, 5,10,25,50,75,90,95th percentiles
# 5 lowest and 5 highest scores
```

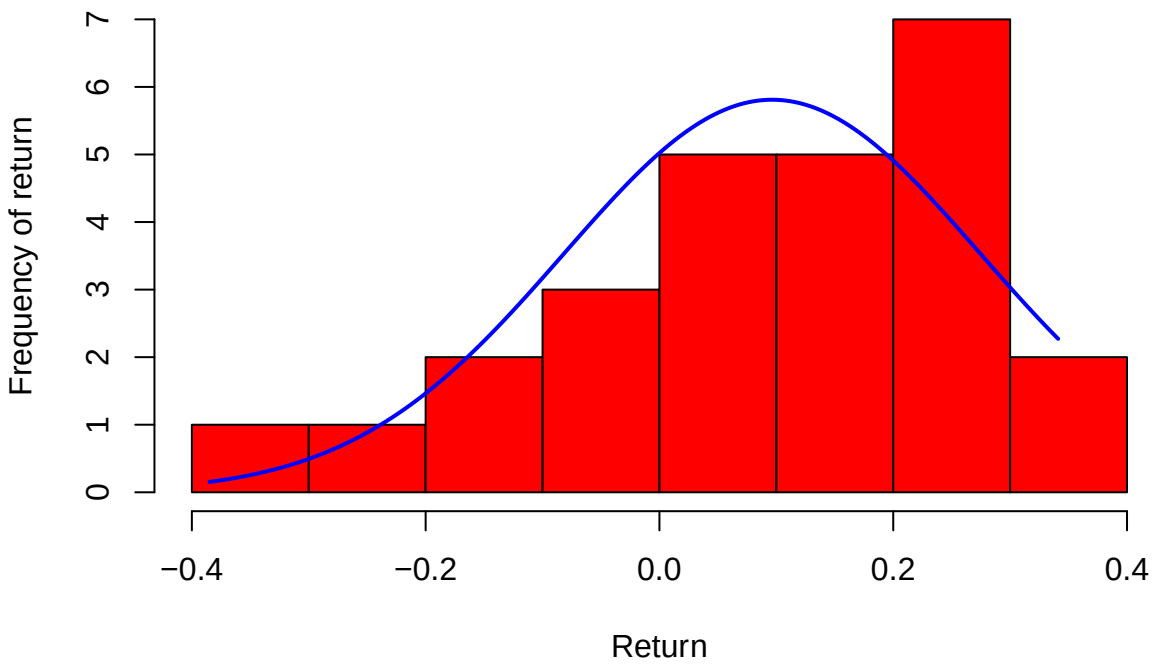
An histogram of the yearly returns on S&P500 index:

```
hist(dataset_s1$Return, main="Histogram for yearly returns",
     xlab="Return", ylab="Frequency of return",
     border="blue", col="red")
```



Let's add a Normal denisty curve on top of the distribution:

```
plotNormalHistogram(dataset_s1$Return, prob = FALSE, col = "red",
                    main = "Histogram for yearly returns with normal distribution overlay",
                    xlab="Return", ylab="Frequency of return",
                    linecol = "blue", lwd = 2)
```



Dataset Training Exercise S1 Uses 1000 simulated values from a normal distribution (mean 0.06, standard deviation 0.015).

```
trainexers_s1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexers1.csv")
```

You want to investigate the precision of the estimates of the mean return on the stock market. You have a simulated sample of 1000 yearly return observations $y_i \sim NID(\mu, \sigma^2)$.

1. Construct a series of mean estimates m_i, where you use the first i observations, so $m_i = \frac{1}{i}\sum_{j=1}^{i} y_j$. Calculate the standard error for each estimate m_i. Make a graph of m_i and its 95% confidence interval, using the rule of thumb of 2 standar deviations.
2. Suppose that the standard deviation of the returns equals 15%. How many years of observations would you need to get the 95% confidence interval smaller than 1%?

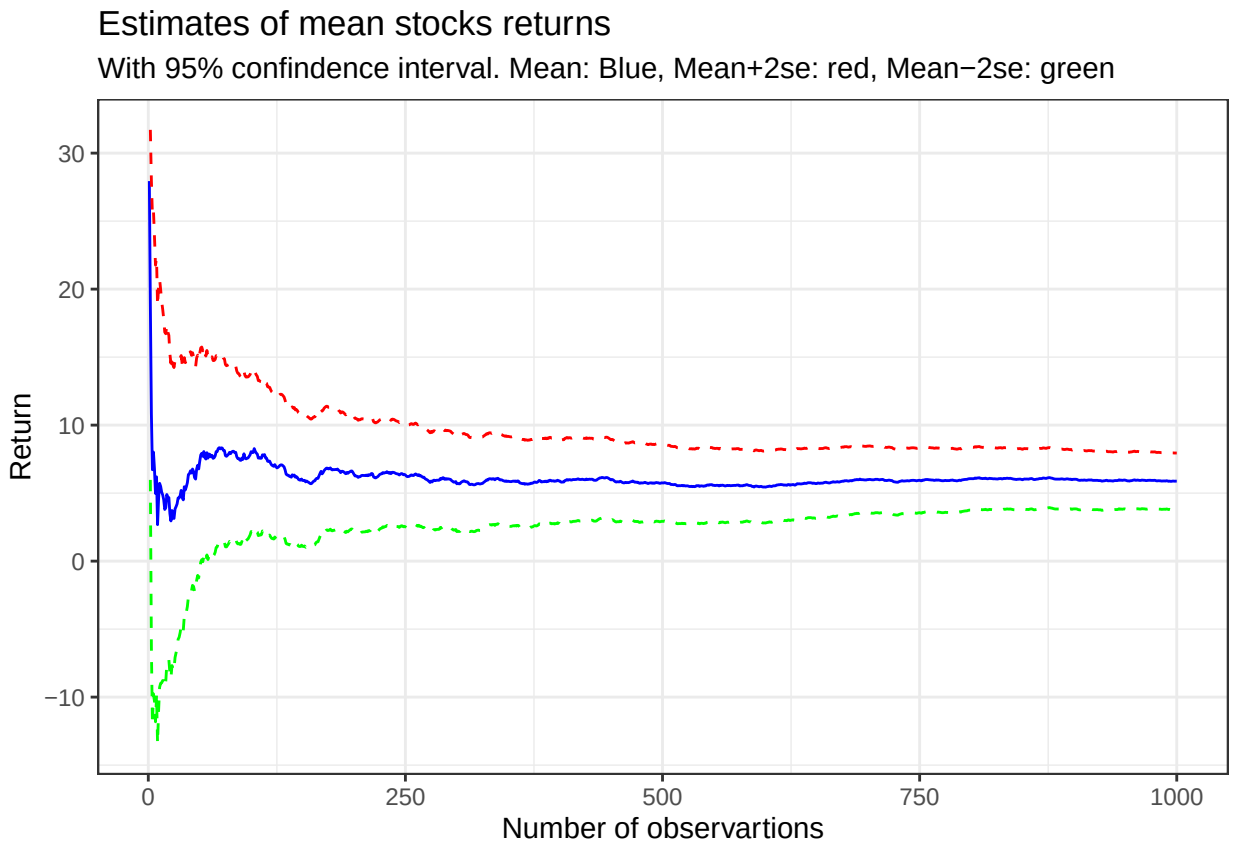
We know $se = \frac{\sigma}{\sqrt{n}}$. Solving $4\frac{\sigma}{\sqrt{n}} = 1 \Rightarrow n = 16\sigma^2$ therefore if $\sigma = 15\%$ yields $16(15^2) = 3,600$ years.

The Standard Error is $SE_i = \sqrt{var(m_i)} = \sqrt{\frac{1}{i-1}\sum_{j=1}^{i}(y_j - m_i)^2}$

```
# We create a new collumn for our estimates
trainexers_s1 <- trainexers_s1 %>% mutate(estimates=0)
# We add to each row the estimate with a for loop
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,3]=(1/i)*(sum(trainexers_s1[1:i,2]))
}

# We create a new collumn for our standard errors
trainexers_s1 <- trainexers_s1 %>% mutate(std_errors=0)
# We add to each row the standard error with a for loop
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,4]=sqrt(var(trainexers_s1[1:i,3]))
```

```
}

# We create the +- 2 Standar Errors
trainexers_s1 <- trainexers_s1 %>% mutate(plus2se=0,minus2se=0)
# We fill the rows with a for loop
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,5] = trainexers_s1[i,3]+2*trainexers_s1[i,4]
  trainexers_s1[i,6] = trainexers_s1[i,3]-2*trainexers_s1[i,4]
}
```

```
# We create the graph
plot1 <- ggplot(data=trainexers_s1, aes(x=Observation)) +
  geom_line(aes(y = estimates), color = "blue") +
  geom_line(aes(y = plus2se), color="red", linetype="dashed") +
  geom_line(aes(y = minus2se), color="green", linetype="dashed") +
  labs(title="Estimates of mean stocks returns  ",
       subtitle="With 95% confindence interval. Mean: Blue, Mean+2se: red, Mean-2se: green", y="Return"

plot1 + theme_bw()
```



Statistical Testing

We assumed an IID normal distribution for a set of 26 yearly returns on the stock market and calculated a sample mean of 9.6% and sample standard deviation of 17.9%. Suppose that you consider investing in the stock market. You then expect to earn a return equal to μ percent every year.

Of course, you hope to make a profit. However, a friend claims that the expected return on the stock market

is 0. Perhaps your friend is right. How can you use a statistical test to evaluate this claim?

A statistical hypothesis is an assertion about one or more parameters of the distribution of a random variable. Examples are that the mean mu is equal to 0, that it is nonnegative or larger than 5%, or that the standard deviation sigma is between 5 and 15%. We want to test one hypothesis, the null hypothesis against another one, the alternative hypothesis. We denote the null hypothesis by H0 and the alternative by H1. So H0 can be mu = 0 and H1, mu is unequal to 0.

A statistical test uses the observations to determine the statistical support for a hypothesis. It needs a test statistic t which is a function of the vector of observations y and a critical region C. If the value of the test statistic falls in the critical region, we reject the null hypothesis in favor of the alternative, if not we say that we do not reject the null hypothesis. Note that we do not say that we accept the null hypothesis. Suppose that we want to test the null hypothesis that mu is equal to 0, against the alternative that it is unequal to 0, with the variance sigma-squared known.

For a test statistic we use the sample mean. We define a critical region as the range below minus c and beyond c with c a positive constant. Small c is called the critical value. If the sample mean falls below minus c or beyond c, we reject the null hypothesis. The sample mean is then too far away from 0 for the null hypothesis to be true.

- If H null is false and the test rejects it, we call the outcome a true positive.
- If H null is true and the test does not reject it, we call it a true negative.
- If H null is true but a test rejects it, the outcome is a false positive or a type I error. If H null is false but a test does not reject it, the outcome is a false negative or type II error.

The probability of a type I error, so the probability to reject while the null hypothesis is true is called the **size of the test** or the significance level. The probability to reject while the null is false is called the **power of the test**. We prefer tests with small size and large power.

A smaller critical region means that we need larger deviations from the null hypothesis for a rejection. So the significance level decreases. However, this also means that the power of the test goes down. So in determining the critical region, we have to make a trade-off between size and power.

You can see an interactive hypothesis test calculator in my website

Example

Let's finish with the stock market example. The estimated mean and standard deviation were 9.6 and 17.9%.

The t statistic for the mean equal to 0 equals 2.75. The one-sided p-value = 0.54%. So for all significance levels beyond 0.54% we reject the null hypothesis in favor of the mean being positive.

The standard deviation of the stock market return is a measure for the risk of investing in the stock market. Suppose you want to limit your risk measured by the standard deviation to 25%. You test H0 that the standard deviation is equal to 25% against the alternative that it is smaller.

How would you decide?

The test statistic has a value of 12.74, which falls inside the critical region from 0 to 14.61. So we reject that the variance equals 25%. The p-value for a test equals 2.1%.

For more information look this website

```
# t test for mean = 0
t.test(trainexers_s1$Return, mu=0)
```

```
##
##  One Sample t-test
##
## data:  trainexers_s1$Return
## t = 12.424, df = 999, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.951096 6.808421
## sample estimates:
## mean of x
##  5.879759
```

```
ttest <- t.test(trainexers_s1$Return, mu=0)
```

Now, we want to determine how the sample size influences test statistics.

1. We want to test hypotheses of the form: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ Construct a series of statistics ti and corresponding p-values for $\mu_0 = 0\%$ and $\mu_0 = 6\%$ where t_i is the t-statistic based on the first i observations. Using the range $i = 5, 6...30$ make a table of t-statistics and p-values for both values of μ_0.

When calculating p-values we must take into account that the test is two sided. Then $p_i = 2\Psi_{i-1}(-|t_i|)$ where Ψ_n is the cumulative distribution function (CDF) of the t distribution function with n degrees of freedom.

The p-values are based on the upper bounds of Critical Region and remember the t distribution is symmetric.

```
# We add the collumns for the t stat and p value for both mu_0 cases.
trainexers_s1 <- trainexers_s1 %>% mutate(t_stat_0=0,p_value_0=0,t_stat_6=0,p_value_6=0)
# We fill the columns using a for loop.
# These first two loops are for mu_0=0%
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,7]= (trainexers_s1[i,3]-0)/(trainexers_s1[i,4]/sqrt(i))
}
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,8]= 2*pt(-as.numeric(trainexers_s1[i,7]),i)
}
# The following two are for mu_0=6%
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,9]= (trainexers_s1[i,3]-6)/(trainexers_s1[i,4]/sqrt(i))
}
for (i in 1:length(trainexers_s1$Return)){
  trainexers_s1[i,10]= 2*pt(-as.numeric(trainexers_s1[i,9]),i)
}

# We select the sub-sample for observations 5-30
sample_s1 <- trainexers_s1[5:30,]
sample_s1 <- sample_s1 %>% select(Observation,t_stat_0,p_value_0,t_stat_6,p_value_6)
sample_s1
```

```
## # A tibble: 26 x 5
##    Observation t_stat_0 p_value_0 t_stat_6 p_value_6
##          <dbl>    <dbl>     <dbl>    <dbl>     <dbl>
##  1           5     2.02    0.0999   0.506      0.634
##  2           6     1.94    0.0998   0.225      0.830
##  3           7     1.57    0.161   -0.324      1.24
##  4           8     2.18    0.0610   0.0680     0.947
##  5           9     1.00    0.344   -1.24       1.75
##  6          10     1.87    0.0914  -0.565      1.42
##  7          11     2.53    0.0279  -0.123      1.10
##  8          12     2.63    0.0220  -0.242      1.19
##  9          13     2.61    0.0217  -0.475      1.36
## 10          14     2.67    0.0183  -0.622      1.46
## # ... with 16 more rows
```

Simple Regression.

A simple example concerning the weekly sales of a product with a price that can be set by the store manager.

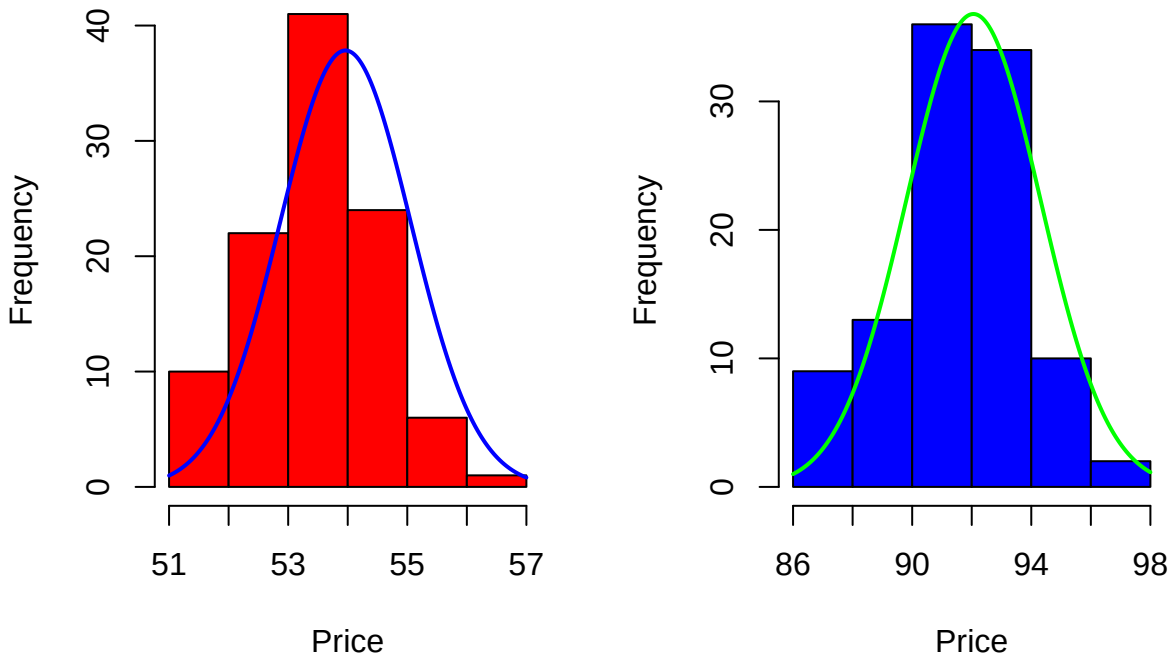
We'll use the following dataset:

Simulated price and sales data set with 104 weekly observations. - Price: price of one unit of the product - Sales: sales volume during the week

```
dataset1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/week1_dataset1.csv")
```

Let's look at our sample:

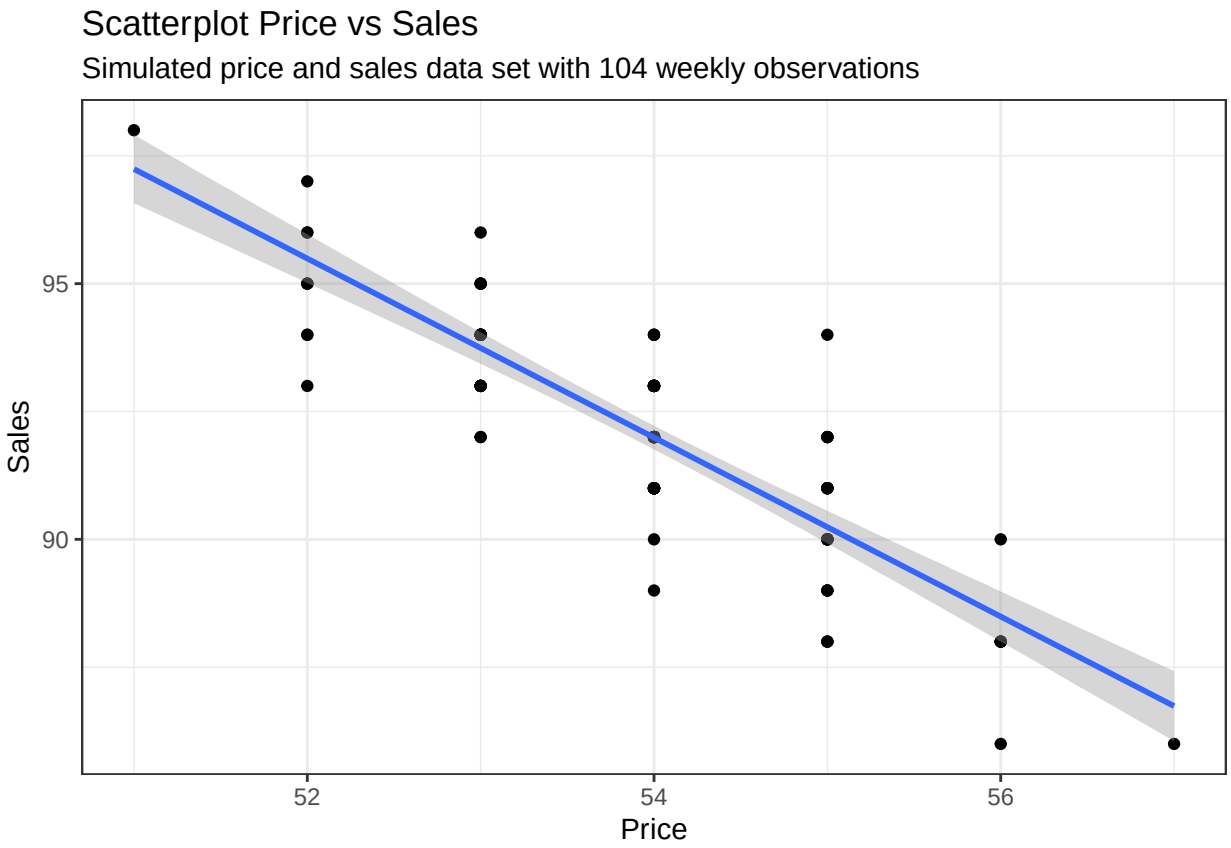
```
par(mfrow=c(1,2))
plotNormalHistogram(dataset1$Price, prob = FALSE, col = "red",
                    xlab="Price", ylab="Frequency",
                    linecol = "blue", lwd = 2)
plotNormalHistogram(dataset1$Sales, prob = FALSE, col = "blue",
                    xlab="Price", ylab="Frequency",
                    linecol = "green", lwd = 2)
```



We expect that lower prices lead to higher sales. The econometrician tries to quantify the magnitude of these consumer reactions to such price changes. This helps the store manager to decide to increase or decrease the price if the goal is to maximize the turnover for this product. Turnover is sales times price. You can see that the majority of weekly sales are somewhere in between 90 and 95 units, with a minimum of 86 and a maximum of 98. Sales of 92 and 93 units are most often observed, each 19 times. The store manager can freely decide each week on the price level, presented on the next slide.

When we plot sales against price that occur in the same week, we get the following scatter diagram.

```
plot2 <- ggplot(data=dataset1, aes(x=Price,y=Sales)) + geom_point() + geom_smooth(method='lm') +
  labs(title="Scatterplot Price vs Sales ",
       subtitle="Simulated price and sales data set with 104 weekly observations")

plot2 + theme_bw()
```



from the scatter plot of sales and price data, you see that different price levels associate with different sales levels. And this suggests that you can use the price to predict sales.

$$Sales = a + b \cdot Price$$

This equation allows us to predict the effects of a price cut that the store manager did not try before, or to estimate the optimal price to maximize **turnover.**(sales times price)

In simple regression, we focus on two variables of interest we denote by y and x, where one variable, **x**, is thought to be helpful to predict the other, **y**. This helpful variable x we call the regressor variable or the explanatory factor. And the variable y that we want to predict is called the dependent variable, or the explained variable.

We can say from our histogram

$$Sales \sim N(\mu, \sigma^2)$$

This notation means that the observations of sales are considered to be independent draws from the same Normal distribution, with mean mu and variance sigma squared, abbreviated as NID. Note that we use the Greek letters mu and sigma squared for parameters that we do not know and that we want to estimate from the observed data. The probability distribution of sales is described by just two parameters, the mean and the variance. On this slide you see the graph of a standardized normal distribution with mean 0 and variance 1. And if you wish, you can consult the Building Blocks for further details on the normal distribution.

For a normal distribution with mean mu, the best prediction for the next observation on sales is equal to that mean mu. An estimator of the population mean mu is given by the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_n$, where y subscript i denotes the i-th observation on sales. The sample mean is called an unconditional prediction of sales, as it does not depend on any other variable.

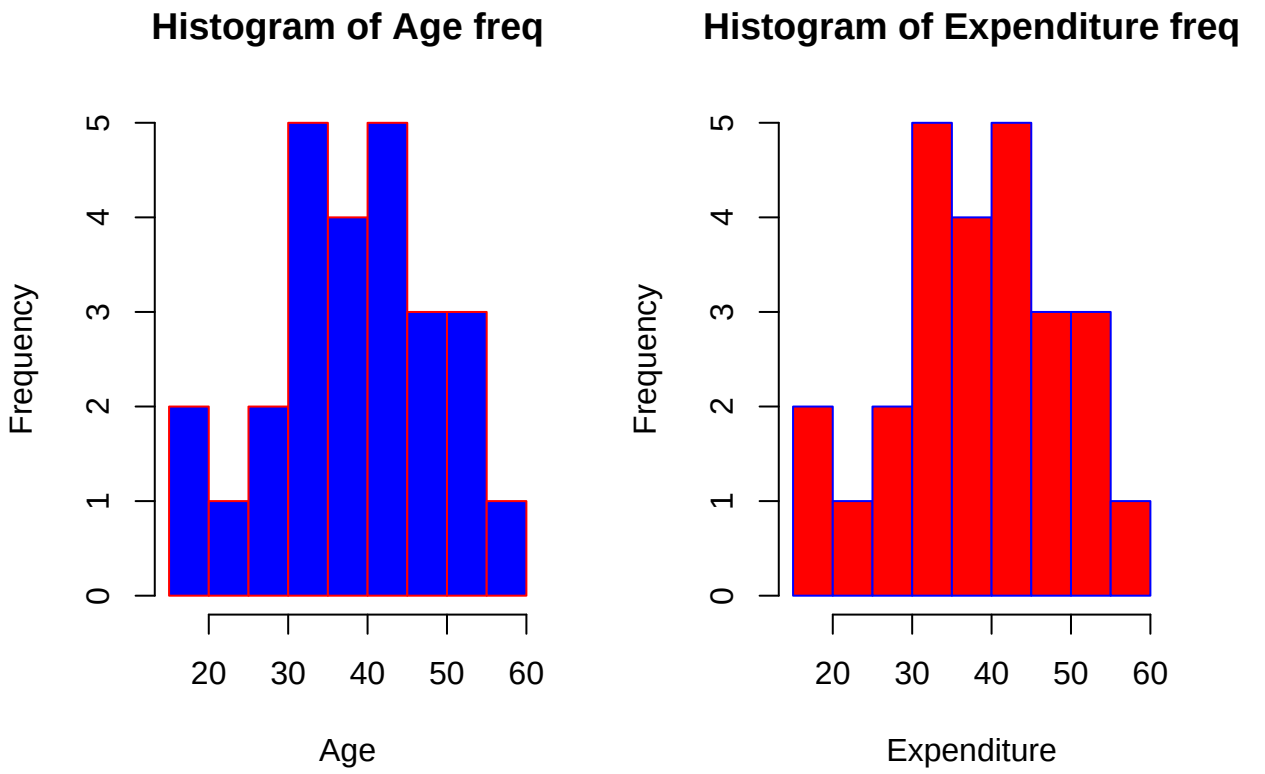
Example:

TrainExer1_1 Simulated data set on holiday expenditures of 26 clients. - Age: age in years - Expenditures: average daily expenditures during holidays

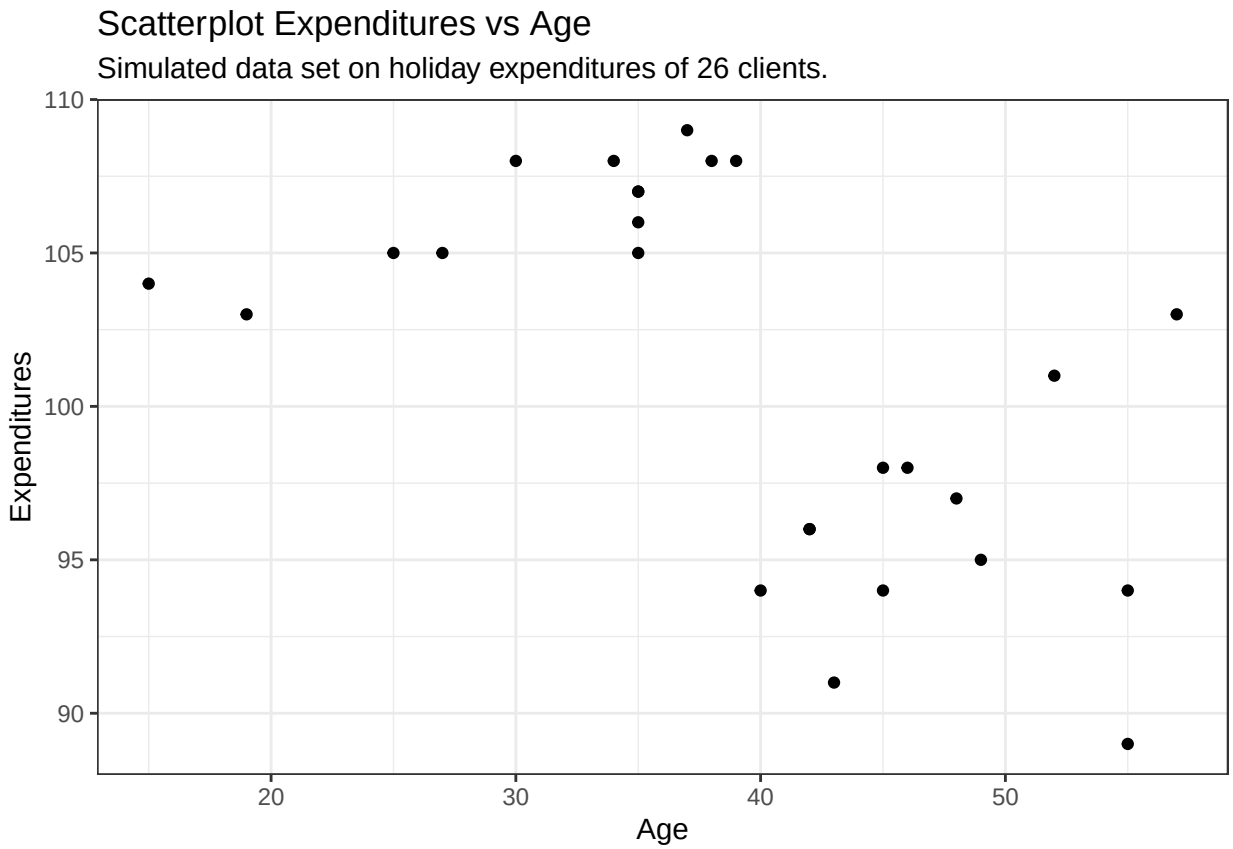
```
dataset2 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer1_1.csv")
```

1. Make two histograms, one of expenditures and the other of age. Make also a scatter diagram with expenditures on the vertical axis versus age on the horizontal axis.

```
par(mfrow=c(1,2))
hist(dataset2$Age,xlab = "Age",col = "blue",border = "red",
     main = "Histogram of Age freq")
hist(dataset2$Age,xlab = "Expenditure",col = "red",border = "blue",
     main = "Histogram of Expenditure freq")
```




```
plot3 <- ggplot(data=dataset2, aes(x=Age,y=Expenditures)) + geom_point() +
  labs(title="Scatterplot Expenditures vs Age ",
       subtitle="Simulated data set on holiday expenditures of 26 clients.")

plot3 + theme_bw()
```



The points in the scatter doesn't associate with a single line, there appears to be two groups in the samples, a group of people younger than 40 and another group older than 40 years old.

- In what respect do the data in this scatter diagram look different from the case of the sales and price data discussed in the last section? Propose a method to analyze these data in a way that assists the travel agent in making recommendations to future clients.

The scatter diagram indicates two groups of clients. Younger clients spend more than older ones. Further, expenditures tend to increase with age for younger clients, whereas the pattern is less clear for older clients.

```
summary(dataset2$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   35.00   39.50   39.35   45.75   57.00
```

```
summary(dataset2$Expenditures)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    89.0    96.0   103.0   101.1   106.8   109.0
```

- Compute the sample mean of expenditures of all 26 clients.

```
#dataset2_descr <- psych::describe(dataset2)
#as_data_frame(dataset2_descr)
# item name ,item number, nvalid, mean, sd,
# median, mad, min, max, skew, kurtosis, se
print(paste("The mean of the expenditures of clients is ",
            mean(dataset2$Expenditures)))
```

```
## [1] "The mean of the expenditures of clients is  101.115384615385"
```

- Compute two sample means of expenditures, one for clients of age forty or more and the other for clients of age below forty.

```
dataset2_over40 <-dataset2 %>% filter(Age>=40)
dataset2_below40 <-dataset2 %>% filter(Age<40)
print(paste("The mean of the expenditures of clients over 40 is ",
            mean(dataset2_over40$Expenditures)))
```

```
## [1] "The mean of the expenditures of clients over 40 is  95.8461538461538"
```

```
print(paste("The mean of the expenditures of clients below 40 is ",
            mean(dataset2_below40$Expenditures)))
```

```
## [1] "The mean of the expenditures of clients below 40 is  106.384615384615"
```

- What daily expenditures would you predict for a new client of fifty years old? And for someone who is twenty-five years old?

Someone of fifty (in older that 40 group) is expected to spend (unconditional prediction) $95.84, someone of twenty-five (in younger that 40 group) is expected to spend (unconditional prediction) $ 106.38