Applied Statistics

MSc Digital Marketing & Data Science
Exercises

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

Abstract

This document contains a list of exercises designed to put into practice the various concepts covered in the course. They enable you to check that the concepts have been properly assimilated and to identify the points to be worked on. They are intended to be used in conjunction with the exercises carried out in class. Some exercises are straightforward applications, while others involve the study of data sets in order to carry out a much more complete study of a problem.

There are and will be no corrections for these exercises. Examples are already included in the course resources and should enable you to recognize whether or not you are applying the right method. However, if you have any questions about these exercises, please do not hesitate to contact me.

Contents

1 Generalities on Random Variables

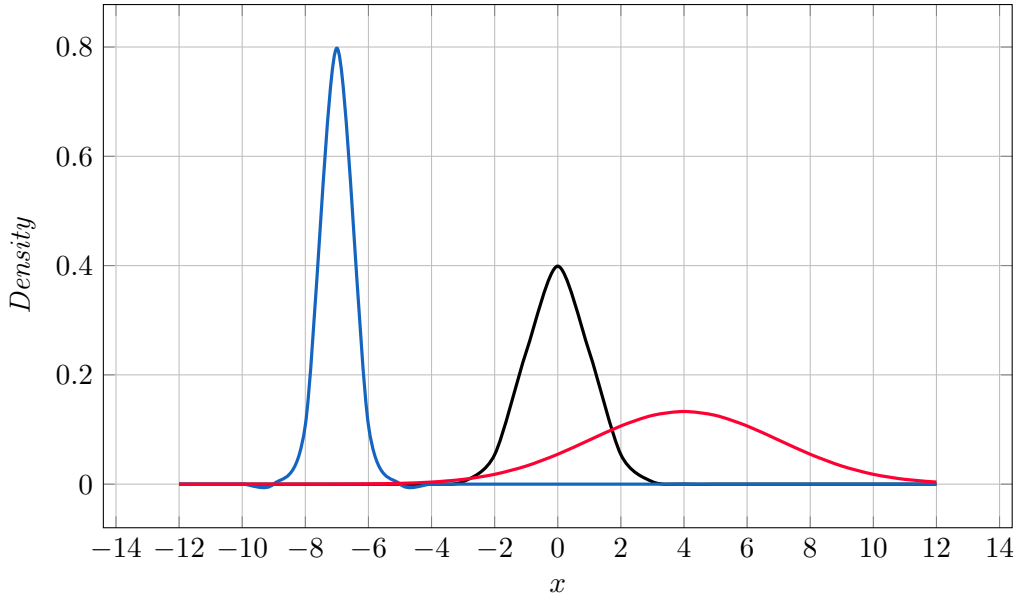
Exercise 1.1 (Generalities on Continuous Random Variables)**.**

We are working on the properties of the normal distribution.

1. *A continuous random variable X has a normal distribution with mean 73. The probability that X takes a value greater than 80 is 0.212. Find the probability that X takes a value less than 66 . Sketch the density curve with relevant regions shaded to illustrate the computation.*

2. *A continuous random variable X has a normal distribution with mean 60. The probability that X takes a value greater than 75 is 0.342. Find the probability that X takes a value less than 45. Sketch the density curve with relevant regions shaded to illustrate the computation.*

3. *The figure below represent 3 different normal distributions with mean values equal to $0, -7$ and 3 with standard deviations equal to $1, 3$ and 0.5.*



Associate each curve to its couple of parameters (μ, σ).

4. *Financial market prices are subject to wide variations over time, and it is assumed that the values taken by the latter over time are distributed according to a centered normal distribution whose mean is equal to 6800 points. We know that 22% of the values taken by the CAC 40 price are below 6500 points and that 10% of the values taken are between 7000 and 7100. In what proportion does the share price take values between 6500 and 7000?*

Exercise 1.2 (Compute probabilities)**.** *In the following Z is used to denote the Normal distribution with a mean value equal to $\mu = 0$ and a standard deviation $\sigma = 1$. X is used to denote a random normal distribution.*

1. *Compute the following probabilities using the Z-table*

   (a) $\mathbb{P}[Z \leq -2.3]$

   (b) $\mathbb{P}[Z \leq -1.72]$

   (c) $\mathbb{P}[Z \leq 1.85]$

2. *Compute the following probabilities using the Z-table*

(a) $\mathbb{P}[Z \geq 2.3]$

(b) $\mathbb{P}[Z \geq 1.3]$

(c) $\mathbb{P}[Z \geq -0.67]$

3. *Compute the following probabilities using the Z-table*

   (a) $\mathbb{P}[1.2 \leq Z \leq 3.43]$

   (b) $\mathbb{P}[-2.45 \leq Z \leq 2.45]$

   (c) $\mathbb{P}[-0.56 \leq Z \leq 2.78]$

4. *Compute the following probabilities using the Z-table*

   (a) $\mathbb{P}[Z \geq 0.5$ *or* $Z \leq -0.78]$

   (b) $\mathbb{P}[Z \geq -0.5$ *and* $Z \leq 2]$

   (c) $\mathbb{P}[Z \geq 0.5$ *and* $Z \leq -0.78]$

5. *X is a normally distributed random variable with mean* 57 *and standard deviation* 6*. Compute the following probabilities using the Z-table*

   (a) $\mathbb{P}[X \leq 50]$

   (b) $\mathbb{P}[X \geq 53]$

   (c) $\mathbb{P}[49 \leq X \leq 51]$

   (d) $\mathbb{P}[X \geq 52$ *or* $X \leq 59]$

   (e) $\mathbb{P}[X \geq 52$ *and* $X \leq 59]$

   X is a normally distributed random variable with mean 100 *and standard deviation* 10*. Compute the following probabilities using the Z-table*

   (a) $\mathbb{P}[X \leq 89]$

   (b) $\mathbb{P}[X \geq 108]$

   (c) $\mathbb{P}[81 \leq X \leq 115]$

   (d) $\mathbb{P}[X \geq 123$ *or* $X \leq 79]$

   (e) $\mathbb{P}[X \geq 89$ *and* $X \leq 120]$

Exercise 1.3 (Applications)**.** *We are now working on several applications with the normal distribution.*

1. *The systolic blood pressure X of adults in a region is normally distributed with mean equal* 112 *and standard deviation* 15*. A person is considered "prehypertensive" if his systolic blood pressure is between* 120 *and* 130*. Find the probability that the blood pressure of a randomly selected person is prehypertensive.*

2. *The length of time that the battery in Mathew cell phone will hold enough charge to operate acceptably is normally distributed with mean* 25.6 *hours and standard deviation* 0.32 *hour. Mathew forgot to charge his phone yesterday, so that at the moment she first wishes to use it today it has been* 26 *hours* 18 *minutes since the phone was last fully charged. Find the probability that the phone will operate properly.*

3. *Birth weights of full-term babies in a certain region are normally distributed with mean equal to* 7.125 *lb and standard deviation* 1.290 *lb. Find the probability that a randomly selected newborn will weigh less than* 5.5 *lb, the historic definition of prematurity.*

Exercise 1.4 (Find the parameters of a Normal Distribution)**.**

A company carried out an analysis of the various costs involved in developing new marketing strategies. It estimated that 10% *of development costs were in excess of* $\$\ 75,000$ *and that* 25% *were under* $\$\ 50,000$*. Assuming that* X *is a random variable following a normal distribution, determine the parameters of this distribution.*

2 Sampling Estimation

Exercise 2.1 (Basics on Sampling Estimation)**.**

Answer to the basic following questions:

1. *Random samples of size* 225 *are drawn from a population with mean 100 and standard deviation* 20*. Find the mean and standard deviation of the sample mean.*
2. *Random samples of size* 64 *are drawn from a population with mean* 32 *and standard deviation* 5*. Find the mean and standard deviation of the sample mean.*
3. *A population has mean* 75 *and standard deviation* 12*.*
   (a) *Random samples of size* 121 *are taken. Find the mean and standard deviation of the sample mean.*
   (b) *How would the answers to part (a) change if the size of the samples were* 400 *instead of* 121 *?*
4. *A population has mean* 128 *and standard deviation* 22 *.*
   (a) *Find the mean and standard deviation of* $\bar{X}$ *for samples of size* 36 *.*
   (b) *Find the probability that the mean of a sample of size* 36 *will be within* 10 *units of the population mean, that is, between* 118 *and* 138 *.*
5. *A population has mean* $1,542$ *and standard deviation* 246 *.*
   (a) *Find the mean and standard deviation of* $\bar{X}$ *for samples of size* 100 *.*
   (b) *Find the probability that the mean of a sample of size* 100 *will be within* 100 *units of the population mean, that is, between* $1,442$ *and* $1,642$ *.*

3 Confidence Interval

Exercise 3.1 (Basics). *We ask you to build confidence interval using the following information*

1. *A random sample is drawn from a population of known standard deviation* 11.3. *Construct a* 90% *confidence interval for the population mean based on the information given (not all of the information given need be used).*
   (a) $n = 36, \bar{x} = 105.2$ *and* $s = 11.2$
   (b) $n = 100, \bar{x} = 105.2$ *and* $s = 11.2$
2. *A random sample is drawn from a population of an unknown standard deviation. Construct a* 98% *confidence interval for the population mean based on the information given (not all of the information given need be used).*
   (a) $n = 225, \bar{x} = 92$ *and* $s = 8.4$
   (b) $n = 64, \bar{x} = 92$ *and* $s = 8.4$
3. *A random sample of size* 256 *is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are* $\bar{x} = 1011$ *and* $s = 34$.
   (a) *Construct a* 90% *confidence interval for the population mean* μ
   (b) *Construct a* 99% *confidence interval for the population mean* μ
   (c) *Comment on why one interval is longer than the other.*
4. *Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a* 90% *confidence interval for the population proportion.*
   (a) $n = 25$ *and* $\hat{p} = 0.7$.
   (b) $n = 50$ *and* $\hat{p} = 0.7$.
5. *Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a* 95% *confidence interval for the population proportion.*
   (a) $n = 325$ *and* $\hat{p} = 0.4$.
   (b) $n = 80$ *and* $\hat{p} = 0.4$.
6. *In a random sample of size* $1,100$ *we saw that* 338 *have the characteristic of interest.*
   (a) *Compute the sample proportion* $\hat{p}$ *with the characteristic of interest.*
   (b) *Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.*
   (c) *Construct an* 80% *confidence interval for the population proportion* p .
   (d) *Construct a* 90% *confidence interval for the population proportion* p.
   (e) *Comment on why one interval is longer than the other.*

Exercise 3.2 (Application 1).
The number of trips to a grocery store per week was recorded for a randomly selected collection of households, with the results shown in the table below:

2	2	2	1	4	2	3	2	5	4
2	3	5	0	3	2	3	1	4	3
3	2	1	6	2	3	3	2	4	4

Construct a 95% confidence interval for the average number of trips to a grocery store per week of all households.

Exercise 3.3 (Application 2)**.**

A factory specializing in cable construction wishes to verify the reliability of its products by evaluating the maximum mass that its cables can support.
To do this, we model the maximum mass, in tons, supported by a cable by a random variable X following a Normal distribution with unknown mean μ and standard deviation $\sigma = 0.5$. A study was carried out on a sample of 50 *cables. It was found that, on average, the maximum load supported by a cable is* 12.2 *tons.*

1. *Determine a confidence interval for μ at the* 0.99 *level.*
2. *Can we affirm that the machine, with a risk of error of* 1%*, produces cables capable of supporting a mass of at least* 11.5 *tons?*
3. *Determine the minimum size of the sample studied so that the length of the confidence interval at the* 99% *level is less than 0.2.*

Exercise 3.4 (Application 3)**.**

The lifetime of a light bulb, given in hours, is represented by a random variable X whose distribution is supposed to be Normal with a standard deviation $Sigma = 400$, the parameter of the mean Mu is unknown.

The measurements of the life span of a batch of 9 bulbs gave the following results:

$$2,000; 1,890; 3,180; 1,990; 2,563; 2,876; 3,098; 2,413; 2,596; 1,876$$

1. *Determine a confidence interval for the average life of a light bulb at the* 90% *level.*
2. *Can we affirm, with a risk of error of* 10%*, that the average life of a light bulb is equal to* $2,500$ *hours?*

Exercise 3.5 (Application 4)**.**
A biologist is studying a type of algae that attacks marine plants. The toxin contained in this alga is obtained as an organic solution. The biologist measures the amount of toxin per gram of solution which is modeled by a random variable X following a normal distribution whose expectation μ and variance σ^2 are unknown. He obtained the following nine measurements, expressed in milligrams:

$$1.2; 0.8; 0.6; 1.1; 1.2; 0.9; 1.5; 0.9; 1.0$$

1. *Calculate the mean $\bar{x}$ associated with this sample.*
2. *Give a confidence interval of level* 0.90 *of quantity of toxin.*

3. *Can we say, with a level of confidence of* 0.80 *that the quantity* μ *of toxin per gram of solution is equal to* 1.3 *mg.*

Exercise 3.6 (Application 5)**.**

A security feature on some web pages is graphic representations of words that are readable by human beings but not machines. When a certain design format was tested on 450 *subjects, by having them attempt to read ten disguised words,* 448 *subjects could read all the words.*

1. *Give a point estimate of the proportion* p *of all people who could read words disguised in this way.*
2. *how that the sample is not sufficiently large to construct a confidence interval for the proportion of all people who could read words disguised in this way.*
3. *Assume that the conditions are satisfied, build a* 95% *confidence interval on the proportion* p*.*

Exercise 3.7 (Application 6)**.**

In a random sample of 250 *employed people,* 61 *said that they bring work home with them at least occasionally.*

1. *Give a point estimate of the proportion of all employed people who bring work home with them at least occasionally.*
2. *Construct a* 99% *confidence interval for that proportion* p*.*

4 One Sample Test

Exercise 4.1 (Formulation of the assumptions)**.**

State the null H_0 and alternative hypotheses H_1 for each of the following situations. (That is, identify the correct number μ_0 and write the assumption H_0 and H_1)

1. *The average July temperature in a region historically has been $74.5F$. Perhaps it is higher now.*
2. *The average weight of a female airline passenger with luggage was 145 pounds ten years ago. The FAA believes it to be higher now.*
3. *The average stipend for doctoral students in a particular discipline at a state university is $\$14,756$. The department chairman believes that the national average is higher.*
4. *The average room rate in hotels in a certain region is $\$82.53$. A travel agent believes that the average in a particular resort area is different.*
5. *The average farm size in a predominately rural state was 69.4 acres. The secretary of agriculture of that state asserts that it is less today.*

Exercise 4.2 (Rejection Regions and p-values)**.**

Compute the statistical test, find the rejection regions and the p-value for each hypothesis testing.

1. $H_0 : \mu = 72.2$ *v.s.* $H_1 : \mu \geq 72.2$, *where* $\alpha = 0.05$.
   We assume that σ *is unknown,* $n = 55, \bar{x} = 75.1, s = 9.25$.
2. $H_0 : \mu = 58$ *v.s.* $H_1 : \mu \geq 58$, *where* $\alpha = 0.1$.
   We assume that $\sigma = 1.22, n = 40, \bar{x} = 58.5$ *and* $s = 1.29$.
3. $H_0 : \mu = -19.5$ *v.s.* $H_1 : \mu \leq -19.5$, *where* $\alpha = 0.01$.
   We assume that σ *is unknown,* $n = 30, \bar{x} = -23.2, s = 9.55$.
4. $H_0 : \mu = 805$ *v.s.* $H_1 : \mu \neq 805$, *where* $\alpha = 0.2$.
   We assume that $\sigma = 37.5, n = 75, \bar{x} = 818$ *and* $s = 36.2$.

   We assume

Exercise 4.3 (Application 1)**.**
A calculator has a built-in algorithm for generating a random number according to the standard normal distribution. Twenty-five numbers thus generated have mean 0.15 and sample standard deviation 0.94. Test the null hypothesis that the mean of all numbers so generated is 0 versus the alternative that it is different from 0, at the 20% level of significance. Assume that the numbers do follow a normal distribution.

Exercise 4.4 (Application 2)**.**

A wine merchant is interested in the capacity of the bottles of a producer suspected by some customers of fraud. He wants to make sure that this capacity respects on average the legal lower limit of 75 cl. To this end, he measures the contents of 10 bottles taken at random and obtains the following values:

$$73.2; 72.6; 74.5; 75; 75.5; 73.7; 74.1; 75.8; 74.8; 75$$

It is assumed that the filling process follows a Normal distribution $\mathcal{N}(\mu, \sigma^2)$ where $\sigma = 1$.

1. *Determine the average value of filling $\bar{x}$ on this sample.*
2. *Clearly formulate the choice of our null hypothesis H_0 and your alternative hypothesis H_1.*
3. *What type of test should be used to help him make a statistically correct decision? Specify whether it is a two-sided or one-sided (lower or upper) test and the law followed by the test statistic. Justify your answer.*
4. *Can the trader conclude, with a risk of error of 1%, that the producer respects the legal lower limit of 75 cl ?*
5. *The trader wants to be able to detect, with a high probability (99%), an average capacity of at least 74.8 cl while keeping the test at a risk of error of 1%. What should he do?*

Exercise 4.5 (Application 3)**.**

A credit risk engineer, employed in a company specialized in consumer credit, wants to verify the hypothesis that the average value of the monthly payments of his customers is the hypothesis that the average monthly payment of the clients in his portfolio is 200 euros. A random sample of 144 customers, taken at random from the database, gives an empirical average $\bar{x} = 193.74$ and an unbiased estimate of the standard deviation $s = 48.24$.

1. *What are the statistical hypotheses associated with the engineer's problem and what type of test should be implemented to help him make a statistically correct decision?*
2. *Can he conclude, at the risk of 5%, that the postulated average value of repayments is correct?*

Exercise 4.6 (Application 4)**.**

In a study of adolescents' learning ability, a sample of 30 adolescents is recruited for a series of tests. In order to have a relatively homogeneous sample in terms of IQ (Intelligence Quotient), the statistical protocol requires that the standard deviation of the IQ not exceed 20 points.

							QI							
131	*108*	*85*	*96*	*86*	*126*	*128*	*107*	*119*	*87*	*103*	*110*	*125*	*77*	*90*
109	*109*	*129*	*95*	*117*	*107*	*102*	*83*	*114*	*72*	*99*	*103*	*97*	*109*	*97*

1. *Estimate the mean and the standard deviation of the sample*
2. *Can we say that, on average, the mean QI is higher than 0.05 at a significance level of $\alpha = 0.05$?*

Exercise 4.7 (Application 5)**.**

An insurance company states that it settles 85% of all life insurance claims within 30 days. A consumer group asks the state insurance commission to investigate. In a sample of 250 life insurance claims, 203 were settled within 30 days.

1. *Test whether the true proportion of all life insurance claims made to this company that are settled within 30 days is less than 85% , at the 5% level of significance.*

2. *Compute the observed significance of the test, i.e. the p-value.*

Exercise 4.8 (Application 6)**.**

A report five years ago stated that 35.5% *of all state-owned bridges in a particular state were "deficient." An advocacy group took a random sample of* 100 *state-owned bridges in the state and found* 33 *to be currently rated as being "deficient." Test whether the current proportion of bridges in such condition is* 35.5% *versus the alternative that it is different from* 35.5% *, at the* 10% *level of significance.*

5 Two sample tests

Exercise 5.1 (Basics on Independent Populations)**.** *Construct the confidence interval for $\mu_1 - \mu_2$ for the level of confidence and the data from independent samples given. Remember that you shall first compare the variances with an F-test. Whatever the result of the F-test, we will consider that the variances of the two populations are equal.*

1. *a 90% confidence interval with:*

$$n_1 = 45,\ \bar{x}_1 = 27,\ s_1 = 2,$$
$$n_2 = 60,\ \bar{x}_2 = 22,\ s_2 = 3.$$

item a 99% confidence interval with:

$$n_1 = 30,\ \bar{x}_1 = -112,\ s_1 = 9$$
$$n_2 = 40,\ \bar{x}_2 = -98,\ s_2 = 4$$

Exercise 5.2 (Independent Populations: Application 1)**.**

In order to investigate the relationship between mean job tenure in years among workers who have a bachelor's degree or higher and those who do not, random samples of each type of worker were taken, with the following results.

	n	$\bar{x}$	s
Bachelor's degree or higher	155	5.2	1.3
No degree	210	5.0	1.5

1. *Construct a 99% confidence interval for the difference in the population means based on the above dataset.*
2. *Test, with a level of significance $\alpha = 0.01$, the claim that mean job among those with higher education is greater than among those without, against the default that there is no difference in the means.*
3. *Compute the p-value.*

Exercise 5.3 (Independent Populations: Application 2)**.**

Records of 40 used passenger cars and 40 used pickup trucks (none used commercially) were randomly selected to investigate whether there was any difference in the mean time in years that they were kept by the original owner before being sold. For cars the mean was 5.3 years with standard deviation 2.2 years. For pickup trucks the mean was 7.1 years with standard deviation 3.0 years.

1. *Construct the 95% confidence interval for the difference in the means based on these data.*
2. *Test the hypothesis that there is a difference in the means against the null hypothesis that there is no difference. Use the $\alpha = 0.01$ level of significance.*
3. *Compute the p-value of this test.*

Exercise 5.4 (Independent Populations: Application 3)**.**

The owner of a professional football team believes that the league has become more offense oriented since five years ago. To check his belief, 32 randomly selected games from one year's schedule were compared to 32 randomly selected games from the schedule five years later. Since more offense produces more points per game, the owner analyzed the following information on points per game (ppg).

	n	$\bar{x}$	s
ppg previously	32	20.62	4.17
ppg recently	32	22.05	4.01

Test, at the 10% level of significance, whether the data on points per game provide sufficient evidence to conclude that the game has become more offense oriented.
Is the conclusion changing if we consider that the two populations are related.

Exercise 5.5 (Basics on Related Populations)**.**

For all the datasets, we assume that the population of differences is normal.

1. *We consider the following dataset:*

Population 1	35	32	35	35	36	35	35
Population 2	28	26	27	26	29	27	29

   (a) *Compute the sample differences D and $\bar{x}_D$ and s_D.*
   (b) *Construct the 95% confidence interval for $\mu_1 - \mu_2$ from these data.*
   (c) *Test, at the 10% level of significance, the hypothesis that $\mu_1 - \mu_2 > 7$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = 7$.*

2. *We consider the following dataset:*

Population 1	103	127	96	110	90	118	130	106
Population 2	81	106	73	88	70	95	109	83

   (a) *Compute the sample differences D and $\bar{x}_D$ and s_D.*
   (b) *Construct the 90% confidence interval for $\mu_1 - \mu_2$ from these data.*
   (c) *Test, at the 1% level of significance, the hypothesis that $\mu_1 - \mu_2 < 24$ as an alternative to the null hypothesis that $\mu_1 - \mu_2 = 24$.*

Exercise 5.6 (Related Populations: Application)**.**

In order to cut costs a wine producer is considering using duo or $1 + 1$ corks in place of full natural wood corks, but is concerned that it could affect buyers's perception of the quality of the wine. The wine producer shipped eight pairs of bottles of its best young wines to eight wine experts. Each pair includes one bottle with a natural wood cork and one with a duo cork. The experts are asked to rate the wines on a one to ten scale, higher numbers corresponding to higher quality. The results are:

Duo Cork	*Wood Cork*
5.1	5.0
6.5	6.5
3.6	3.1
3.5	3.7
5.7	4.5
5.0	4.1
6.4	5.3
4.7	2.6
3.2	3.0
3.5	3.5
6.4	5.1

6 Linear Regression

Exercise 6.1 (Prediction using a linear model)**.**

Several companies have conducted a study in order to estimate the average return (Y) *of different stores of their brands according to their investment in marketing development* (X)*.*

For each of the companies, we aim to determine the average return using the learned model for the different investments x *(in thousand dollars)*

$$x = 12.3, \quad x = 7.1, \quad x = 3.9, \quad x = 17.6, \quad x = 5.2$$

1. *What are the variable* X *and* Y*.*
2. *The company A has the following model*

$$Y = 2.3X + 12$$

3. *The company B has the following model*

$$Y = -0.05X + 24$$

4. *The company C has the following model*

$$Y = 1.7X + 8.9$$

5. *The company D has the following model*

$$Y = 4.6X - 3.6$$

6. *Which company shall avoid to invest in marketing development?*
7. *For which company it sounds important to invest to increase its average return?*

Exercise 6.2 (Check your understanding: Part I)**.**

Based on the information given about a line, determine how Y *will change (increase, decrease, or stay the same) when* X *is increased, and explain. In some cases it might be impossible to tell from the information given.*

1. *The slope is positive.*
2. *The intercept is positive.*
3. *The slope is zero.*
4. *The intercept is negative.*
5. *The intercept is equal to* 0*.*
6. *The slope is negative.*

Exercise 6.3 (Check your understanding: Part II)**.**

We are studying different situations:

1. *A data consists of eight (x, y) pairs:*

$$(0, 12)\ (4, 16)\ (8, 22)\ (15, 28)$$
$$(2, 15)\ (5, 14)\ (13, 24)\ (20, 30)$$

   (a) *Plot the data in scatter plot.*
   (b) *Based on the plot, explain whether the relationship between X and Y appears to be deterministic or to involve randomness.*
   (c) *Based on the plot, explain whether the relationship between X and Y appears to be linear or not linear.*

2. *A data consists of eight (x, y) pairs:*

$$(3, 20)\ (6, 9)\ (11, 0)\ (14, 1)\ (18, 9)$$
$$(5, 13)\ (8, 4)\ (12, 0)\ (17, 6)\ (20, 16)$$

   (a) *Plot the data in scatter plot.*
   (b) *Based on the plot, explain whether the relationship between X and Y appears to be deterministic or to involve randomness.*
   (c) *Based on the plot, explain whether the relationship between X and Y appears to be linear or not linear.*

Exercise 6.4 (A simple application)**.**

The rate for renting a motor scooter for one day at a beach resort area is $\$25$ plus 60 cents for each mile the scooter is driven. The total cost Y in dollars for renting a scooter and driving it X miles is

$$Y = 0.60X + 25.$$

1. *Explain whether the relationship between the cost Y of renting the scooter for a day and the distance X that the scooter is driven that day is deterministic or contains an element of randomness.*

2. *A person intends to rent a scooter one day for a trip to an attraction 17 miles away. Assuming that the total distance the scooter is driven is 34 miles, predict the cost of the rental.*

Exercise 6.5 (Least Squares Regression)**.**

Answer to the following questions, for the two datasets used in exercise Check your understanding: Part II (feel free to use Excel for some of the computations):

1. *Estimate the coefficients of the regression.*

2. *Compute the residuals.*

3. *Compute SST, SSR and SSE and determine if a linear model is valid for the prediction task using a F-test.*

4. *Estimate the variance of the errors of the regression model.*

5. *Determine the quality by computing the* R^2.

6. *Study if the coefficients of the regression model are significant.*

Exercise 6.6 (A simple regression model)**.**
The volatility of a stock is often measured by its beta value. You can estimate the beta value of a stock by developing a simple linear regression model, using the percentage weekly change in the stock as the dependent variable and the percentage weekly change in a market index as the independent variable. The S&P 500 *Index is a common index to use. For example, if you wanted to estimate the beta value for Disney, a market model:*

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where Y *represents the* % *of weekly change in company,* X *represents the* % *of weekly change in S&P* 500 *Index.*

The least-squares regression estimate of the slope β_1 *is the estimate of the beta value for the studied company. A stock with a beta value of* 1.0 *tends to move the same as the overall market. A stock with a beta value of* 1.5 *tends to move* 50% *more than the overall market, and a stock with a beta value of* 0.6 *tends to move only 60% as much as the overall market. Stocks with negative beta values tend to move in the opposite direction of the overall market. The following table gives some beta values for some widely held stocks as of July* 11, 2015*:*

Company	β_1
Apple	1.13
Disney	1.42
American Eagles Mines	−0.64
Marriott	1.5
Microsoft	0.78
Procter & Gamble	1.0

1. *For each of the six companies, interpret the beta value.*

2. *How can investors use the beta value as a guide for investing?*

Exercise 6.7. *Refer to the discussion of beta values and market models. The S&P* 500 *Index tracks the overall movement of the stock market by considering the stock prices of* 500 *large corporations. The following variables are included:*

- *WEEK—Week ending on date given*
- *S$P—Weekly closing value for the S&P 500 Index*
- *GE—Weekly closing stock price for General Electric*
- *DISCA—Weekly closing stock price for Discovery Communications*

- *GOOG—Weekly closing stock price for Google*

The data are available in the file ***StockPrices****.*

1. *Estimate the market model for GE. (Hint: Use the percentage change in the S&P* 500 *Index as the independent variable and the percentage change in GE's stock price as the dependent variable.)*
2. *Interpret the beta value for GE.*
3. *Repeat the two previous questions for Discovery Communications.*
4. *Repeat the two previous questions for Google.*
5. *Write a brief summary of your findings.*

Exercise 6.8 (A Quadratic Model)**.**
We want to establish a model that predicts the purity (our variable Y*) of a liquid as a function of its filtration time (our variable* X*), using the* ***Purity*** *dataset. We will consider the two following models*

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

and

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon,$$

where X^2 *represents the squared value of the filtration time.*

1. *Estimate the coefficients of the regression models using the data.*
2. *We aim to determine which model is the best. Which criterion can be used for this purpose? What is its definition?*
3. *Which model is the best one?*

Exercise 6.9 (A Multiple Regression Model)**.**
A real estate broker in Dubai wants to develop a multiple regression model to predict house price,based on size, number of rooms, and type of house (whether it is new or up for resale). He collected data for the price of 50 *houses and stored it in* ***ResaleHouse****.*

1. *Build the regression model.*
2. *Analyze the regression results and write a report on the obtained model.*

Exercise 6.10 (A multiple Regression Model with a Categorical Variable)**.** *In its continuing study of the 3-For-All subscription solicitation process, a marketing department team wants to test the effects of two types of structured sales presentations (personal formal and personal informal) and the number of hours spent on telemarketing on the number of new subscriptions. The staff has recorded these data for the past 24 weeks.*

Analyze these data ***Subscription*** *and develop a multiple regression model to predict the number of new subscriptions for a week, based on the number of hours spent on telemarketing and the sales presentation type. Write a report, giving detailed findings concerning the regression model used and analyze your regression model.*