Financial Mathematics and Statistics

Msc Finance

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 the statements but also the corrections of the exercises treated during the different Lab sessions.

The exercises covered in the document range from elementary reminders in probability, through the construction of confidence intervals, hypothesis tests on parameters such as the mean or the proportion. This last notion is extended to tests of comparisons of means between several populations with the one or two factor ANOVA.

The last part of the course is devoted to simple and multiple linear models for the prediction of a real numerical variable. A small digression on logistic regression is also discussed.

Contents

1 Lab 1 **3**

1.1 Managing Ashland Multicomm Services . . . 3
1.2 Digital Case . . . 4
1.3 Nestlé Purina Petcare: part I . . . 9
1.4 Nestlé Purina Petcare: part II . . . 11
1.5 IT Consulting Firm . . . 13

2 Lab 2 **15**

2.1 Managing Ashland MultiComm Services . . . 15
2.2 Coffee Shop . . . 17
2.3 Managing Ashland Multicomm Services . . . 21
2.4 A Study of Different Systems . . . 25
2.4.1 Phase 1: Anova with one factor . . . 25
2.4.2 Phase 2: Anova with two factors . . . 28

3 Lab 3 **30**

3.1 Volatility of a stock . . . 30
3.2 Stock Prices . . . 31
3.3 Ashland MultiComm Services . . . 35
3.4 Digital Case . . . 40

4 Lab 4 **45**

4.1 Checking your Understanding . . . 45
4.2 Managing Ashland MultiComm Services . . . 48

1 Lab 1

1.1 Managing Ashland Multicomm Services

In this first part, the technical services department of AMS worked on a project to improve the service and the quality of access to Internet of these customers, for that they carried out measurements as for the speed of loading and have for objective a value of 1.0 in term of speed of loading. Their study is based on data from the previous year and they show that the data, which represent the loading speed, are distributed according to a normal law of mean $\mu = 1.005$ and standard deviation $\sigma = 0.10$. The loading speed is considered acceptable if the measure is between 0.95 and 1.05.

1. We assume that the distribution of upload speed remains the same as it was last year.

   I remind you that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $Z = \dfrac{X - \mu}{\sigma}$ is normally distributed with mean and variance respectively equal to 0 and 1, *i.e.* $Z \sim \mathcal{N}(0, 1)$.

   The exercise can be done using the Z-table or the function `NORM.INV` of Excel

   - $\mathbb{P}[X < 1] = 0.4801$
   - $\mathbb{P}[0.95 < X < 1.0] = \mathbb{P}[X < 1.0] - \mathbb{P}[X < 0.95] = 0.1889$
   - $\mathbb{P}[1 < X < 1.05] = \mathbb{P}[X < 1.05] - \mathbb{P}[X < 1] = 0.1936$
   - $\mathbb{P}[(X < 0.95) \cup (X > 1.05)] = 1 - \mathbb{P}[0.95 < X < 1.05] = 0.6175$

2. We now want to reduce the probability that the upload speed is below 1.0. We wonder if it is better to improve the mean upload speed or to reduce the standard deviation to the upload speed. To answer to this question, we will have to compare the probability that the upload speed takes values below 1 and select the process for which this probability is minimized

   - If we increase the value of μ to 1.05, we have:

$$\begin{aligned}
\mathbb{P}[X < 1] &= \mathbb{P}\left[\frac{X - \mu}{\sigma} < \frac{1 - 1.05}{0.1}\right], \\
&= \mathbb{P}\left[Z < -0.5\right], \\
&\quad \downarrow \text{ symmetry of the normal distribution} \\
&= \mathbb{P}\left[Z > 0.5\right],
\end{aligned}$$

$$
\begin{aligned}
&\downarrow \text{ complementary events} \\
&= 1 - \mathbb{P}\left[Z < 0.5\right], \\
\mathbb{P}[X < 1] &= 0.3085.
\end{aligned}
$$

- In the case we decrease the value of σ to 0.075, we have:

$$
\begin{aligned}
\mathbb{P}[X < 1] &= \mathbb{P}\left[\frac{X - \mu}{\sigma} < \frac{1 - 1.005}{0.075}\right], \\
&= \mathbb{P}\left[Z < -0.067\right], \\
&\downarrow \text{ symmetry of the normal distribution} \\
&= \mathbb{P}\left[Z > -0.067\right], \\
&\downarrow \text{ complementary events} \\
&= 1 - \mathbb{P}\left[Z < 0.067\right], \\
\mathbb{P}[X < 1] &= 0.4734.
\end{aligned}
$$

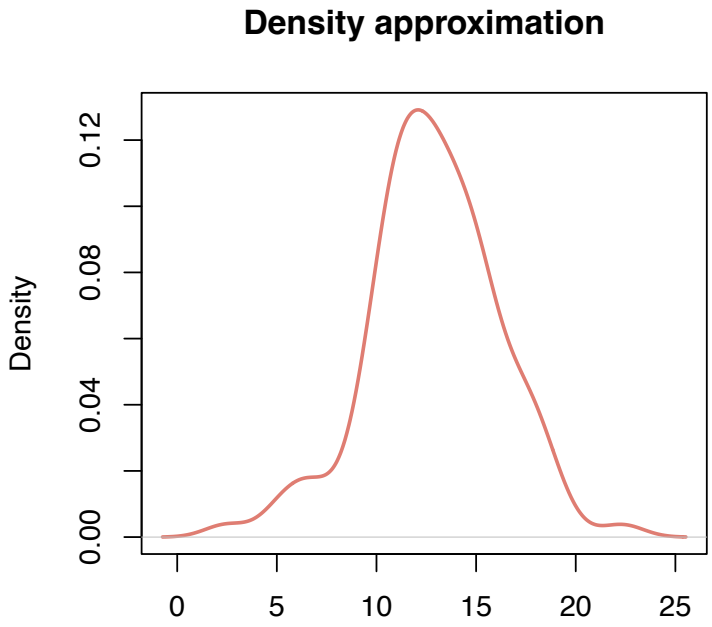
It it is then preferable to increase the mean upload speed to improve user experience.

1.2 Digital Case

The aim of this exercise is to study the veracity of a report on a study conducted in a company.

Our data have the following distribution

```
# Import the data
data = read.csv("data/DownloadTimes.csv", header = FALSE)
# We plot the density
d <- density(data$V1)
plot(d, lwd=2,col = "#DF7D72", xlab="", main = "Density approximation")
```



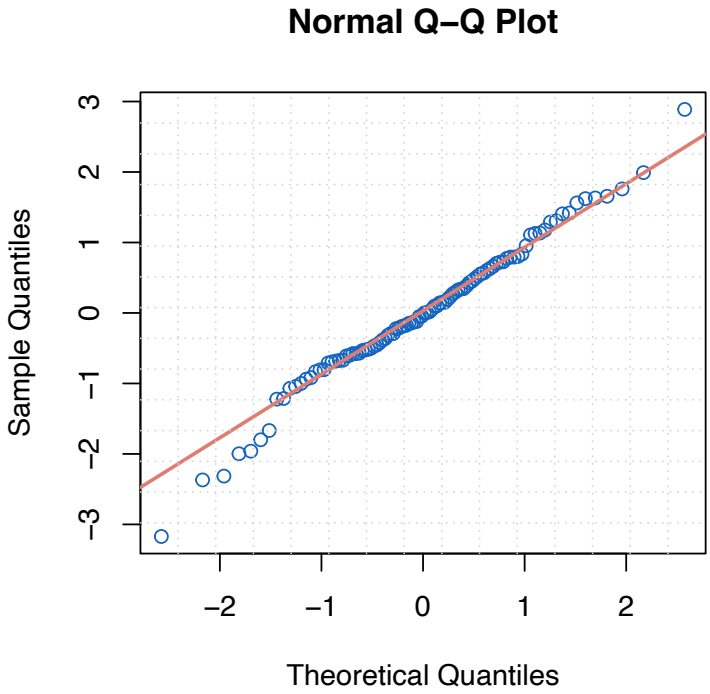
1. We start by looking if the data are normally distributed. To do this we will start by plotting the QQ-plot (also called the *Normal Probability Plot*, see Section 6.3 of your reference *Pearson*) of our data, this is to plot the empirical quantiles of the data vs. the empirical quantiles of the centered reduced normal distribution.

   If the points are aligned on a straight line, then we can say that our data are distributed according to a normal distribution. (See Excel). Attention, the only criterion *mean* = *median* = *mode* is not sufficient to say that data are from a normal distribution !

   Your graph shall have the following form:

```
# We compute the Empirical Quantiles
data_plot = (data$V1 - mean(data$V1))/sd(data$V1)
# Plot the normal probability plot
qqnorm(data_plot, col = "#1565C0")
qqline(data_plot, col = "#DF7D72", lwd = 2)
grid(nx = 15, ny = 15, col = "lightgray", lty = "dotted",
```

```
    lwd = par("lwd"), equilogs = TRUE)
```



We can do something different and try to check if the data are normally distributed using hypothesis testing, *i.e.*, more precisely, using a **Shapiro Test**:

```
# We compute the Empirical Quantiles
shapiro.test(data$V1)

##
##  Shapiro-Wilk normality test
##
## data:  data$V1
## W = 0.98585, p-value = 0.3651
```

2. We will now check if the conclusion made by the experts are correct or not using the data (see the associated Excel File)

- ```r
  # Computation of the mean
  mean_sample = mean(data$V1)
  mean_sample
  ```

  ```
  ## [1] 12.8596
  ```

  ```r
  # Computation of the median
  median_sample = median(data$V1)
  median_sample
  ```

  ```
  ## [1] 12.785
  ```

  The mean value is accurate. On the other hand, we note that the median is slightly lower than the mean, however, this difference can only be explained by our sample size. Further analysis will be required to see if this difference is statistically different.

- This assertion is false, as a reminder the probability that a continuous law takes an exact value is zero. It would have been better to study the probability that a download time is less than a given value.

- The assertion is not good, removing extreme values contributes to strongly reduce the mean of our sample but especially its variance, so in general, we will find less data in the interval :

  $$\bar{x} - 3s; \bar{x} + 3s,$$

  where $\bar{x}$ stands for the empirical mean (the mean evaluated on the sample) and s for the standard deviation evaluated on the sample.

  The only thing we can say then is that the probability of observing loading times of less than 22.7 seconds will then increase.

- A quick analysis of our Excel spreadsheet shows us that this assertion seems to be true for the observed sample. On the other hand, be careful with the interpretation of this result. The data allowing us to say that the probability that we observe a loading time higher than the value of 17.06 is indeed 0.1, but this does not mean that the phenomenon will be observed in 10% of the cases, it is an *asymptotic* result.

  ```r
  quantile(data$V1,0.9)
  ```

  ```
  ##    90%
  ## 17.096
  ```

  It means that 90% of our sample values are lower or equal than 17.096.

- We have to check if 99% of the data lies in the interval
  ```

$$\bar{x} - 3s; \bar{x} + 3s,$$

to say that the process follows the Six Sigma benchmark for industrial quality. Thus, we only to see if 99% of the data belongs in this interval.

```
# Lower bound
L = mean(data$V1)-3*sd(data$V1)
L

## [1] 3.021766

# Upper bound
U = mean(data$V1)+3*sd(data$V1)
U

## [1] 22.69743

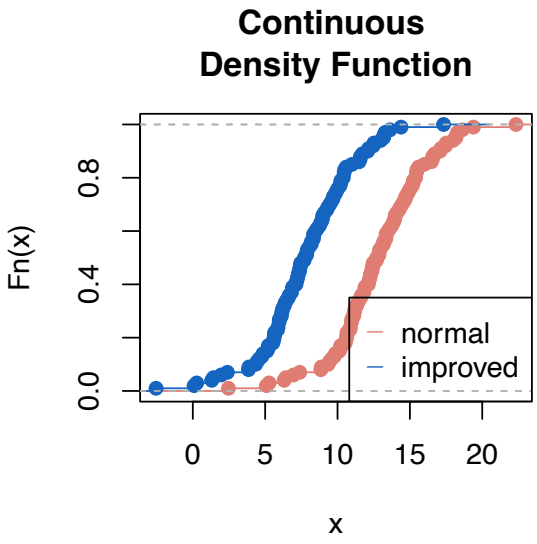
# Proportion of the data in the interval
Rate = mean( ((data$V1 >=L)&&(data$V1 <= U)))
Rate

## [1] 1
```

So the conditions are met with this sample, but it quite normal since the data are normally distributed.

3. Improving the loading time means reducing the average loading time. If we reason about the empirical distribution function, this will have the effect of increasing the observed probabilities for a given value of loading time. This can be seen by using the 15 seconds example used earlier.

```
data_normal = data$V1
data_improved = data$V1 - 5
plot(ecdf(data_normal),col = "#DF7D72",main="Continuous
Density Function", xlim=c(min(data_normal,data_improved)-0.1,
max(data_normal,data_improved)+0.1))
lines(ecdf(data_improved),col="#1565C0")
legend("bottomright",pch='_',c("normal","improved"),
col=c("#DF7D72","#1565C0"))
```



Another possibility, if we have access to the parameters of the distribution, is to compute the associated probabilities.

1.3 Nestlé Purina Petcare: part I

We consider the production of 85g cans. The process includes: filling in the cans, sealing, weighing and disposal if the weight is less than 85g (production waste), sorting, labeling. Even if the automatons are of very good quality, it is impossible to obtain cans weighing exactly 85g. There is a standard deviation in packing the cans related to the technical processes, which follows a normal distribution of the standard deviation equal to 0.8g.

1. We aim to determinate the mean value of packing the cans that has to be used so that the waste related to the weight of package is less than 2%.

   Let us denote as X the random variable used to describe the weight of a can. The technical processes follows a normal distribution $\mathcal{N}(\mu, 0.8^2)$ and we want to determine μ such that :

   $$P[X \leq 85] \leq 0.02.$$

   As usual we have to consider the standardization to solve the problem. It is also the only for us to ensure that the value μ appears.

$$P[X \leq 85] = P\left[\frac{X-\mu}{0.8} \leq \frac{85-\mu}{0.8}\right] = P\left[Z \leq \frac{85-\mu}{0.8}\right].$$

We now want to determine a value z such that

$$P\left[Z \leq z\right] \leq 0.02.$$

Note that is the same problem as finding z such that $P\left[Z \leq z\right] = 0.02$. Once the value z is found we have to use that fact that

$$z = \frac{85-\mu}{0.8}$$

to have the value of the mean. Using the z-table we found that z shall be equal to -2.06. Then μ is equal to

$$\mu = 85 - 0.8 \times z = 85 + 0.8 \times 2.06 = 86.648\text{g}.$$

We are now interested in the production of 100g packages. Once the mixture has been prepared, it is packaged by the automatons. The process includes: packing, sealing, weighing and disposal if the weight is less than 98.5g or bigger than 101.5g, sorting, labeling. The automatons are the same as those used for the 85g packages. The standard deviation in packing, related to the technical processes, follows, also for the 100g packages, a normal distribution of the standard error equal to 0.8g.

2. Knowing that the automatons are set to a mean $\mu = 100$g, we are interested in the percentage of packages that are then discarded as waste, statistically speaking.

   To answer this question we will apply the same process as before, we consider $X \sim \mathcal{N}(100, 0.8^2)$ and compute the following probabilities. Then the probability of being a reject is the sum of these two probabilities.

$$P[X \leq 98.5] \quad \text{and} \quad P[X \geq 101.5].$$

$$\begin{aligned}
\mathbb{P}[X < 98.5] &= \mathbb{P}\left[\frac{X-100}{0.8} < \frac{98.5-100}{0.8}\right], \\
&= \mathbb{P}\left[Z < -1.875\right], \\
&\quad \downarrow \text{ symmetry of the normal distribution} \\
&= \mathbb{P}\left[Z > 1.875\right], \\
&\quad \downarrow \text{ complementary events}
\end{aligned}$$

$$= 1 - \mathbb{P}[Z < 1.875],$$
$$\mathbb{P}[X < 98.5] = 0.03.$$

We also have $P[X \geq 101.5] = 0.03$ because the gaussian is centered at the position 100. Thus the probability of being a reject is equal to 0.06.

3. Would it be appropriate to set the machines to a mean $\mu = 100.5$?

   The answer is clearly no! And the justification is simply based on the fact that the value of a gaussian are centered around the mean, *i.e.* most of them are concentrated around the mean. Furthermore, the set of acceptable cans is also centered at the position 100, so if if we shift the distribution, we will necessarily increase the probability of being a reject for a can.
   This fact can also be shown by computing the probability.

1.4 Nestlé Purina Petcare: part II

To check the adjustment of the machine, it proceeds to a random sampling of cans on the production line. The company wants to check whether the packing mean is always equal to 86.65g.

For that reason, the company has sampled today 100 cans. The mean of that sample of 100 cans is 86.5g.

1. We first aim to give a confidence interval of the mean with a confidence rate equal to $1 - \alpha = 95\%$ in order the check the good behaviour of the process.
   We will see later another tool which will provide us and the answer.

   We still know that the filling process is normally distributed and we consider X the random variable which represents the weight of a can, *i.e.* $X \sim N(\mu, 0.8^2)$.

   We want a confidence interval **on the mean and we know the standard deviation of the distribution. Thus the later will be based on the Z-distribution.** Thus the confidence interval will have the following form:

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + \bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

   where n is the sample size, $\bar{x}$ is the mean value on the sample, σ is the standard deviation of the distribution and $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of a centered and reduced normal distribution. It remains to compute the bound using the information, $\alpha = 0.05$, thus $z_{0.975} = 1.96$

```r
n = 100
z = 1.96
x = 86.5
sigma = 0.8
# Lower Bound
x-z*sigma/sqrt(n)

## [1] 86.3432

# Upper Bound
x+z*sigma/sqrt(n)

## [1] 86.6568
```

2. Can we conclude there is a miss-adjustment of the machines with a 5% error risk?

   The value $\mu = 86.65$ is in the previous interval, so we cannot say that the machine is miss-adjusted.

3. Let us now analyze the two previous result when we increase the sample size to $n = 1,000$.

   We apply the same process as the one used in question 1.. We simply replace $n = 100$ by $n = 1,000$.

```r
n = 1000
z = 1.96
x = 86.5
sigma = 0.8
# Lower Bound
x-z*sigma/sqrt(n)

## [1] 86.45042

# Upper Bound
x+z*sigma/sqrt(n)

## [1] 86.54958
```

   This time, we can conclude there is a miss-adjustment of the machines with a 5% error risk.

1.5 IT Consulting Firm

An information technology (IT) consulting firm specializing in health care solutions wants to study communication deficiencies in the health care industry. A random sample of 70 health care clinicians reveals the following:

- Time wasted in a day due to outdated communication technologies:

$$\bar{x} = 45 \text{ minutes} \quad \text{et} \quad s = 10 \text{ minutes}.$$

- Thirty-six health care clinicians cite inefficiency of pagers as the reason for the wasted time, *i.e.* $\bar{p} = \dfrac{36}{70} = \dfrac{18}{35}$

1. We first want to build a confidence interval to estimate for the population mean time wasted in a day due to outdated communication technologies. The rate of confidence equal uses for this interval is equal to $1 - \alpha = 0.99$, *i.e.* $\alpha = 0.01$.

   First we have to identify in which situation we are. **Here we want an estimation of a mean when the standard deviation at the population scale is unknown**, that is why the value s is provided.
   Our confidence interval will be based on the *Student* distribution and will have the following form

$$\left[\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \; ; \; \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right],$$

   where n is the sample size and $t_{1-\alpha/2,n-1}$ denotes the quantile of order $1-\alpha/2$ of student law where the number of degree of freedom is equal to $n-1$.

   In our context, our interval can be rewritten as:

$$\left[45 - t_{0.995,69}\frac{10}{\sqrt{70}} \; ; \; 45 + t_{0.995,69}\frac{10}{\sqrt{70}}\right],$$

   and $t_{0.995,69} \simeq 2.66$ using the t-table. It remains to compute the values.

```
x = 45
s = 10
n=70
t = 2.6489
# Lower Bound
x-t*s/sqrt(n)
```

```
## [1] 41.83396

# Upper Bound
x+t*s/sqrt(n)

## [1] 48.16604
```

2. We want now to apply the same process to construct a 95% confidence interval estimate for the population proportion of health care clinicians who cite inefficiency of pagers as the reason for the wasted time.

   We want to estimate proportion. Thus our confidence interval will be based on the center and reduced normal distribution and will have the following form:

$$\left[\bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\ ;\ \bar{p} + \bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right],$$

   where n is the sample size and $z_{1-\alpha/2}$ denotes the quantile of order $1-\alpha/2$ of center and reduced normal law.

   In our context, $1-\alpha = 0.95$, *i.e.* $\alpha = 0.05$ and $\bar{p} = \dfrac{18}{35}$ So our interval can be rewritten as:

$$\left[\frac{18}{35} - z_{0.975}\sqrt{\frac{\frac{18}{35}\left(1-\frac{18}{35}\right)}{70}}\ ;\ \frac{18}{35} + z_{0.975}\sqrt{\frac{\frac{18}{35}\left(1-\frac{18}{35}\right)}{70}}\right],$$

   and $z_{0.975} \simeq 1.96$ using the z-table. It remains to compute the values.

```
p = 18/35
n=70
z = 1.96
# Lower Bound
p-z*sqrt(p*(1-p)/n)

## [1] 0.3972011

# Upper Bound
p+z*sqrt(p*(1-p)/n)

## [1] 0.6313703
```