Quantitative Tools - Level 2 - Fall 2017
Louis de Charsonville

Lecture 1

Statistics Essential - Review of QT Level 1

There are three kinds of lies: lies, damned lies, and statistics.

B. Disraeli, British Prime Minister

This lecture assumes that lectures 1 to 4 of Quantitative Level 1 are known. It covers Lecture 5 to 7 on measure of central tendencies (mode, mean, median) as well as values of dispersion (standard deviation, variance, deciles) and concentration (Gini coefficient).

1 Introduction

This section discusses what are statistics and why there are important. It also provides a glossary of the few technical terms that we are going to use in this class.

"In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data."

Statistics
Statistics gives one means to properly evaluate the data and claims that one ready/listen/watchs on newspapers/radios/television. We can distinguish between two kinds of statistics:

- descriptive statistics: number and methods to summarize and describe data. Classic descriptive statistics include mean, minimum, maximum, standard deviation, median.
- inferential statistics: functions and tools to draw an inference regarding an hypothesis about a population parameter.

Distribution
The distribution of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur. When a distribution of categorical data is organized, you see the number or percentage of individuals in each group. When a distribution of numerical data is organized, they're often ordered from smallest to largest, broken into reasonably sized groups (if appropriate), and then put into graphs and charts to examine the shape, center, and amount of variability in the data.

Variables
Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as π that do not vary). We usually distinguish *qualitative* variables from *quantitative* variables.

- Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, gender. Qualitative variables can be split in two subcategories: ordered (sometimes referred as ordinal) or categorical (sometimes referred as nominal). An example of the former is educational experience while an example of the latter is the variable "religion".
- Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

2 Mode

Definition 2.1. Mode
The mode is the value that appears most often in a set of data. It is also the most likely value for a variable.

We distinguish two cases:

- Qualitative variables
- Quantitative variables

2.1 Qualitative variables & discrete quantitative variable

For these variables, finding the mode is straight-forward : this is the value which has the highest frequency.

Example 2.1. Student's grade in Statistics

Grades	7	8	9	10	11	12	13	14	16	19
Frequencies	1	2	4	6	9	10	8	7	1	2

Table 1: Grades

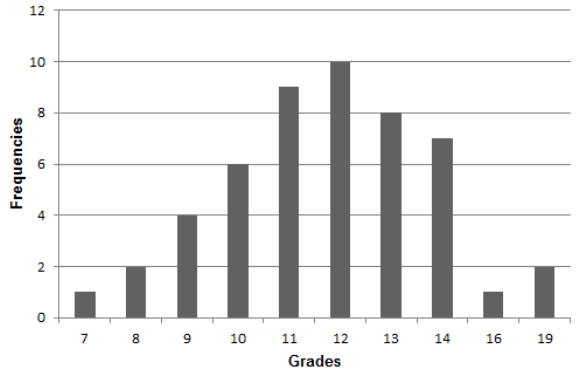


Figure 1: Bar plot of Grades

The Mode is 12.

2.2 Continuous quantitative variables

For continuous variables, it is a bit more tricky to find the mode since the strict definition of the mode does not make sense for continuous variable : two values will never be the same and each value will occur precisely once. In order to compute, the mode we discretize the quantitative variable, that is we divide it into classes. Then, finding the **modal class** is straight-forward : it is the class with the highest frequency.
The mode can then be defined as :

- Approximatively, the midpoint of the modal class.
- the estimated mode with the formula :

$$A = B^l + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} * w$$

$$with \begin{cases} B^L & \text{is the lower class boundary of the modal class} \\ f_{m-1}, f_m, f_{m+1} & \text{respectively the frequency of the group before the modal class, the frequency of the modal class, the frequency of the group after the modal group} \\ w & \text{is the group width} \end{cases}$$

But the mode of discretized continuous variable is class-sensitive in the sense that a different choice of bins could lead to a different mode. See the example below

Example 2.2. Wages in a firm Imagine that the distribution of the wages in a firm is the following

Wages	Frequencies	Corrected frequencies (or Densities)
10000 - 15000	4	4
15000 - 20000	6	6
20000 - 25000	10	10
25000 - 30000	20	20
30000 - 35000	15	15
35000 - 40000	12	12
40000 - 50000	4	2

Table 2: Wages in a firm

The modal class is the class for wages between \$25,000 and \$30,000. An approximative mode is \$27,500. Using the formula to compute the precise mode, we find :

$$\begin{aligned} Mode &= 25000 + 5000 * \frac{20 - 10}{20 - 10 + 20 - 15} \\ &= \$28,333 \end{aligned}$$

The mode can also be found graphically using the intersection of the two black dotted line in the graph below :

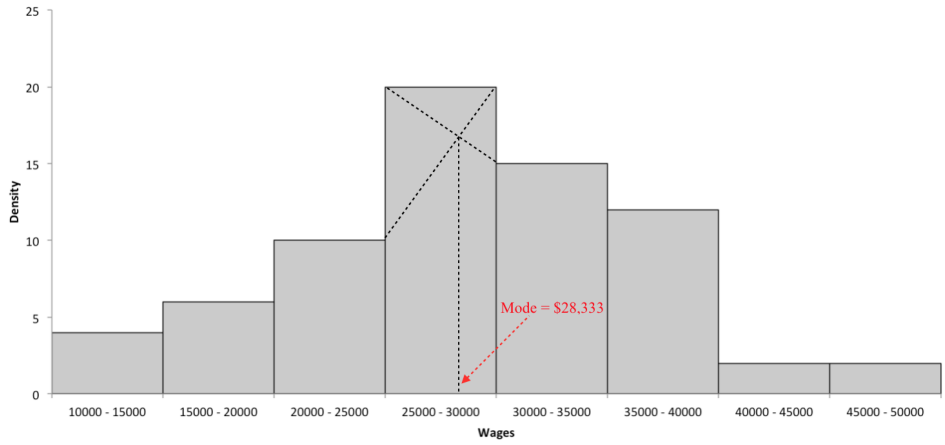


Figure 2: Graphical Determination of the Mode

A drawback of the Mode is that it is sensitive to the classes chosen, see the example below.

Example 2.3. Wages in a firm We use the same distribution of wages but just merge some bins. We got the following wages table :

Wages	Frequencies
10000 - 20000	10
20000 - 30000	30
30000 - 40000	27
40000 - 50000	4

Table 3: Wages in a firm

The modal class is the class for wages between \$20,000 and \$30,000.
We use the formula to compute the precise mode, and we get :

$$\begin{aligned} Mode &= 20000 + 10000 * \frac{30 - 10}{30 - 10 + 30 - 27} \\ &= \$28,695 \end{aligned}$$

3 Median

Definition 3.1. Median The median is the midpoint of a distribution, it is the value which splits the distribution of a variable in two equal parts. For a distribution of wages, for example, the median is the wage below which 50% of salaries are situated. Equivalently, it is the wage above which 50% of salaries are situated.

3.1 Discrete quantitative variable

Consider a series of N observations sorting in ascending order.

- If N is odd, then N can be written $N = 2 * k + 1$ finding the median is straightforward, the median is the $(k+1)^{th}$ number. Indeed, there is k values before and after the $(k+1)^{th}$ number.
- If N is even, then N can be written $N = 2 * k$, and no values of the series can be considered as the median. By convention we usually consider that the median is the midpoints between the k^{th} and $(k+1)^{th}$ numbers.

However, sometimes, the median may not exist (or cannot be found precisely). Consider the following example

Example 3.1. Student's grade in Statistics

Grades	7	8	9	10	11	12	13	14	16	19
Frequencies	1	2	4	6	9	10	8	7	2	2

Table 4: Grades

There are 51 students. So the median grade that splits the students distribution in two is the grade of the 26^{th} students. Here, the 26^{th} student got 12. But 19 students got more than 12 and 22 students got less than 12.

3.2 Continuous quantitative variables

The median for a continuous variable grouped in classes belongs to a class (or bin). The way to find it is the following :

- Find the class to which the median belong
- Derive the median from the boundaries of the class using the even distribution within the class hypothesis. Formally:

$$Median = B^l + \frac{n/2 - f_m}{f_m} * w$$

$$with \begin{cases} B^L & \text{is the lower class boundary of the modal class} \\ n & \text{the size of the distribution} \\ f_m & \text{the frequency of the modal class} \\ w & \text{is the group width} \end{cases}$$

Example 3.2. Wages in a firm

Wages bins	Frequencies	Cumulative frequencies
250 - 350	24	24
350 - 400	32	56
400 - 450	51	107
450 - 500	70	177
500 - 525	47	224
525 - 550	41	265
550 - 600	70	335
600 - 650	58	393
650 - 700	40	433
700 - 800	24	457
800 - 950	3	460

The median income is the income so that $\frac{460}{2} = 230$ workers earn less and 230 workers earn more. The median is the mean of the 230^{th} and 231^{th} workers' income The 230^{th} workers' income belong to the $525-550$ bin, so does the 231^{th} workers' income. We imagine that the wages are evenly distributed within the class. So that the 230^{th} worker's income is $525+\frac{6}{41}*(550-525) = 528.66$, and the 230^{th} worker's income is $525+\frac{7}{41}*(550-525) = 529.27$. The median income is therefore $\frac{528.66+529.27}{2} = 528.97$

3.3 Properties of the median

The median has two very interesting properties :

- It is indifferent to extreme values.
- The sum of absolute deviation of a series to a constant value is minimal when this constant value is the median.

4 Means

4.1 Arithmetic mean

4.1.1 Definitions

The arithmetic mean is the most known of measure of central tendency. It is simply the weighted sum of the numbers.

Definition 4.1. Simple arithmetic mean
The arithmetic mean of $x_1, x_2, ..., x_n$ values appearing once in a sample is :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Definition 4.2. Simple arithmetic mean
The arithmetic mean of $x_1, x_2, ..., x_n$ values appearing w_i times, called the weighted arithmetic mean is :

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i}$$

$$\bar{x} = \sum_{i=1}^{n} \alpha_i * x_i$$

$$\text{with } \alpha_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

4.1.2 Arthimetic Mean of a discretized quantitative variable

When we compute the mean for a quantitative variable which had been discretized, for instance like in the example of wages in the firm, we assume that the wages are identically distributed among the class and take the midpoints of the class. The weighted mean is then :

$$\bar{x} = \sum_{i=1}^{n} \alpha_i * x_i$$

$$where \begin{cases} \alpha_i & \text{is the weights of the class } i \\ x_i & \text{the midpoints of the class } i \\ n & \text{the number of class} \end{cases}$$

Example 4.1. Wages in a firm

Wage Range	Frequencies	Midpoints
250-350	24	300
350-400	32	375
400-450	51	425
450-500	70	475
500-525	47	512,5
525-550	41	537,5
550-600	70	575
600-650	58	625
650-700	40	675
700-800	24	750
800-950	3	875
Total	460	

To calculate the weights of each class, we divide the frequency by the total number of people in the firm, that is the sum of the frequencies. We then compute the weighted mean of the midpoints using the weights.

4.1.3 Properties

The arthimetic mean has very interesting properties :

- If all the values are equal then the arithmetic mean equals one of them

$$\text{if } x_1 = x_2 = ... = x_n \quad \text{then } \bar{x} = x_1$$

- The mean of a sum equals the sum of means

$$\bar{x + y} = \bar{x} + \bar{y}$$

- The sum of all deviations from mean equals to zero

$$\bar{x - \bar{x}} = \bar{x} - \bar{x} = 0$$

- If we add a number b to all x_i then the mean is also increased by b.

$$\bar{x + b} = \bar{x} + b$$

- If we multiply all the x_i by a number a then the mean is also multiplied by a

$$\bar{ax + b} = a * \bar{x} + b$$

- The sum of square deviations of the $x_i, i \in [1, n]$ from a number a reaches a minimal value when a is the mean of the $x_i, i \in [1, n]$. That is the minimum of $\sum_{i=1}^{n}(x_i - a)^2$ is obtained for $a = \bar{x}$. Or formally :

$$\arg\min_a \sum_{i=1}^{n}(x_i - a)^2 = \bar{x}$$

4.1.4 Structural effects & common pitfalls of means

The mean, while being, an used and useful indicator, has common drawbacks that it is important to keep in mind.

4.1.5 Structural effect

The mean is subjects to what we call structural effects : a change in the structure can affect the mean and distort the views on a dataset. Imagine the following distribution of wages in a firm

	Wages	Frequencies
Executives	50,000	20
Workers	10,000	80

Table 5: Wages distribution in 2000 in Mr.Wonka Factory

The mean is obviously :

$$\begin{aligned} Mean &= 0.2 * 50000 + 0.8 * 10000 \\ &= \$18,000 \end{aligned}$$

In 2001, due to a slowdown in world demand of chocolate, a 10% cut is decided on wages of all the employees and 34 workers are fired. The wages distribution is now :

	Wages	Frequencies
Executives	45,000	20
Workers	9,000	46

Table 6: Wages distribution in 2001 in Mr.Wonka Factory

We compute the mean :

$$\begin{aligned} Mean &= 0.2 * 45000 + 0.8 * 9000 \\ &= \$19,909 \end{aligned}$$

The mean wage has increased in the factory why all employees have seen their wages decreased. It is due to a change in the structure of the firm → we call it a structural effect.

4.2 Geometric mean

4.2.1 Definition

Definition 4.3. Geometric Mean
The geometric mean G of n, $(x_i)_{i=1}^n$ is the $1/n^th$ root of the product of the x_i.

$$G = (\Pi_{i=1}^n x_i)^{\frac{1}{n}}$$

With α_i[1] as the weight of x_i[2] :

$$G = (\Pi_{i=1}^n x_i^{\alpha_i})$$

4.2.2 Properties

Average Growth Rate : the geometric mean is very useful to compute average growth rate (as we have seen in the first lectures), or interest rates.

Example 4.2. Of Mice and Means
In a scientific lab, there are 900 mices on the January, 1^{st}. 100 days after, the mice population is now 1,600 due to reproduction only. How many number of mices does the lab have on February 19^{th} (that's 50 days after January 1^{st}) ?
The arithmetic mean would not be appropriate since the newborn mouse are reproducing. The real number is likely to be the geometric mean of 900 and 1,600 :

$$G = (900 * 1600)^{1/2} = 1200$$

The geometric mean of the product is the product of the means : Let $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$, two set of values, G_x, G_y their geometric mean respectively, G_{xy} the geometric mean of the set $(x_i * y_i)_{i=1}^n$, and $G_{x/y}$ the geometric mean of the set $(\frac{x_i}{y_i})_{i=1}^n$ Then :

$$\begin{aligned} G_{xy} &= G_x * G_y \\ G_{x/y} &= \frac{G_x}{G_y} \end{aligned}$$

[1] $\sum_{i=1}^n \alpha_i = 1$.
[2] We can derive the unweighted mean from the weighted mean formula with $\alpha_i = \frac{1}{n}$

4.3 Harmonic mean

4.3.1 Definitions

Definition 4.4. Harmonic Mean
The harmonic mean H of a set of values $(x_i)_{i=1}^n$ is the inverse of the arithmetic mean of the inverse of the $(x_i)_{i=1}^n$.
Formally :

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$H = \frac{1}{\sum_{i=1}^{n} \frac{\alpha_i}{x_i}} \qquad \text{with } \alpha_i \text{ the weight of } x_i$$

4.3.2 Properties

The Harmonic mean is used to compute weights of ratio when both the denominator and the numerator are changing, like for doing the mean of speeds.

Example 4.3. Mean of Speed
Imagine a car travelling from Deauville to Paris at 100km/h and doing the return trip at 120km/h. What is the average speed of the car ?
An distracted reader could answer 110km/h. Unfortunately, that is wrong, since speed is the ratio of distance and time. If distance has not changed, that is not the case of time : the return trip was faster. The appropriate mean is here the harmonic mean.
Indeed, with $2 * d$ the total distance travelled, t the time and s the speed :

$$s_{mean} = \frac{2 * d}{t_{outward} + t_{return}}$$

$$s_{mean} = \frac{2}{\frac{t_{outward}}{d} + \frac{t_{return}}{d}}$$

$$s_{mean} = \frac{2}{\frac{1}{s_{outward}} + \frac{1}{s_{return}}}$$

$$s_{mean} = \frac{2}{\frac{1}{100} + \frac{1}{120}}$$

4.4 Quadratic Mean

Definition 4.5. Quadratic Mean
The quadratic mean Q of a set of n values $(x_i)_{i=1}^n$ is the square root of the arithmetic mean of the squares of the x_i.

Formally :

$$Q = \left(\frac{1}{N} \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}$$

The quadratic mean is used when we want the absolute values ($\Leftrightarrow$ get rid of the sign) and not the arithmetic ones. For instance, to compute errors in estimation, the value computed is usually the quadratic mean of errors.

Central values are usually necessary to characterise a set of values but not sufficient. In particular, they do not give any information about a variable is distributed around its central values, that is the statistical dispersion of the variable - how stretched or squeezed is its distribution.
For instance, the following variables have the mean and median (equals to 10) :

$$A = (6, 8, 10, 12, 14)$$
$$B = (2, 6, 10, 14, 18)$$

But their distribution is not the same : the distribution of the variable B is more *stretched* than the distribution of the variable A.

5 Values of dispersion

They are numerous values of dispersion - interquantiles ranges, absolute deviation and standard deviation -, that are based of the notion of distance. Relative values of dispersion (usually equal to a value of dispersion divided by the mean of the distribution) make comparisons between different variables and variables expressed in different currencies or units.

5.1 Interquantiles ranges, Deciles

Definition 5.1. q-Quantile
The q-Quantiles of a variable are the points that cut the distribution in q equal parts. There are $q-1$ q-Quantiles.

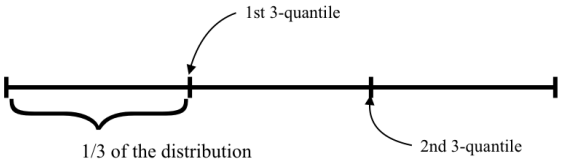


Figure 3: Example of a 3-quantile

The most quantile are :

- the 2-quantile is the point that cuts the distribution of a variable in two. It is known as the **median**.
- the 4-quantiles or **quartiles** are the three points that cut the distribution in 4 equal parts. Usually, we note them as Q_1, Q_2, Q_3 with $Q_1 < Q_2 < Q_3$.
- the 10-quantiles or **deciles** are the nine points that cut the distribution in 10 equal parts. Usually, we note them as D_1, D_2, ... , D_9 with $D_1 < D_2 < ... < D_9$
- the 100-quantiles or **percentiles** or **centiles** are the 99 points that cut the distribution in 100 equals parts.

For a continuous variable discretized in bins, the way to calculate a quantile is the same as for the median (see the next example for more details).

Remarks

- If q is even ($q = 2k$) then the k^{th} quantile is the median. For instance, for quartiles: $q = 4$ and consequently Q_2 is the median.
- When a number is not divisible by 4, the way to compute quartiles is not unique. In this course, we will follow this common rule : the set is divided in two equal parts by the median, Q_1 is the median of the lesser numbers, Q_3 is the median of the greater numbers.

Definition 5.2. Interquartile range
The interquartile range, or *midspread*, is the difference between the lower and the upper quartiles. Thus :

$$IQR = Q_3 - Q_1$$

The **relative interquartile range** equals to the interquartile divided by the unweighted arithmetic mean (or average) :

$$\text{Relative IQR} = \frac{Q_3 - Q_1}{\bar{X}}$$

The **midhinge** is the average of the lower and upper quartile. The midhinge is usually different from the median.

Definition 5.3. Interdecile range
The interdecile range is the difference between the lower and the upper decile. Thus :

$$IDR = D_9 - D_1$$

The **relative interdecile range** equals to the interdecile divided by the average :

$$\text{Relative IDR} = \frac{D_9 - D_1}{\bar{X}}$$

Definition 5.4. A measure of inequality : the ratio D9/D1
The ratio of the upper decile and the lower decile, that is D_9/D_1, is one of the measure of the inequality of a distribution. It evidences the difference between the top and the bottom of the distribution.

Example 5.1. Wages in the Grand Budapest Hotel

Wages	Frequencies	Cumulative Frequencies
100 - 200	5	5
200 - 300	8	13
300 - 400	7	20
400 - 500	8	28
500 - 600	9	37
600 - 700	8	45
700 - 800	9	54
800 - 900	4	58
900 - 1000	2	60

Table 7: Wages in Grand Budapest Hotel

There are 60 people working in the hotel.

- The **first decile** is the wage so that 10% of the employees - 6 here - are earning less than it. The income of the $\frac{60}{10} = 6^{th}$ employee is $200 + 100 * \frac{1}{8} = \$212.5$. The income of the 7^{th} is $200 + 100 * \frac{2}{8} = \$225$. The first decile is the arithmetic mean of the wages of the two : \$218.75.
- The **nineth decile** is the wage so that 90% of the employees are earning less than it (and 10% are earning more). It is the arithmetic mean of the income of the 54^{th} and 55^{th} employees of the firm. That is the arithmetic mean of \$800 and $800 + 100 * \frac{1}{4} = \$825$, which is \$812.5.
- The **ratio D9/D1** is $\frac{812.5}{218.75} = 3.71$ which means that the 10% of the most paid employees earn more than 3.71 times of what the 10% less paid employeers are earning.

5.2 Absolute deviation

5.2.1 Definitions

Definition 5.5. Absolute deviation
The absolute deviation, or **average absolute deviation**, of a set of values $(x_{i=1}^n)$ is the arithmetic mean of the absolute deviations from the mean[3]. That is :

$$\text{absolute deviation} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\text{absolute deviation} = \sum_{i=1}^{n}\alpha_i|x_i - \bar{x}| \qquad \text{with } \alpha_i \text{ the weight of } x_i$$

The absolute deviation is taking into account all the values of a set (unlike the ratio D9/D1 for instance). But one can make lump-sum transfers between the values of a set, affecting the distribution of the values, while not affecting the absolute deviation (see next example). That is why the standard deviation is usually preferred.

Example 5.2. Comparing the Absolute Deviation of two set of values
Here we compare two set of values. The second one differs from the first by an addition of 1 on the first value of the first set and a subtraction on the second value of the first set. (the 3 becomes a 4, and the 5 become a four) :

[3]For a continuous quantitative variable, the mean is usually noted μ

- First set : $\{3, 5, 7, 9, 11\}$
- Second set : $\{4, 4, 7, 9, 11\}$

The two sets share the same mean : 7. We then compute the absolute deviation of the two sets :

First Set		Second Set	
x_i	$\lvert x_i - \bar{x}\rvert$	x_i	$\lvert x_i - \bar{x}\rvert$
3	4	4	3
5	2	4	3
7	0	7	0
9	2	9	2
11	4	11	4
$\bar{x}$	Abs. Deviation	$\bar{x}$	Abs. Deviation
7	2.4	7	2.4

The absolute deviation is not affected by the lump sum transfer we've made while the distribution of the set had obviously been affected.

5.3 Standard deviation and variance

The standard deviation [4] is the most common value of the dispersion of a variable. It is usually referred as σ.

Definition 5.6. Standard deviation
The standard deviation[5] of a set of values $(x_{i=1}^n)$ is the squared root of the average of the squares of the deviation from the mean. Thus :

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma_x = \sqrt{\sum_{i=1}^{n}\alpha_i(x_i - \bar{x})^2} \qquad \text{with } \alpha_i \text{ the weight of } x_i$$

Unlike the absolute deviation, the standard deviation is affected by lump-sum transfers (see the next example)

Example 5.3. Comparing the Standard Deviation of two set of values
We take the same set of values than in the previous example and compare the effect on the standard deviation of a lump-sum transfer of 1 between the two first values of the set :

First Set		Second Set	
x_i	$(x_i - \bar{x})^2$	y_i	$(y_i - \bar{y})^2$
3	16	4	9
5	4	4	9
7	0	7	0
9	4	9	4
11	16	11	16
$\bar{x}$	σ_x	$\bar{y}$	σ_y
7	2.83	7	2.76

The standard deviation is reduced by the lump-sum transfer while the distribution is indeed tightened.

The standard deviation outlines how much a variable varies. Applied to stocks in financial markets, it is used as a proxy of the risk of a stock.

A drawback of the standard deviation is that it is sensitive to the magnitude of the value. Thus, the comparison of the standard deviation between variables that don't have the same order of magnitude is pointless. To avoid this, the coefficient of variation is used.

[4] in French : "écart-type"
[5] in French : *écart-type*

Definition 5.7. Coefficient of variation
The coefficient of variation or **relative standard deviation** of a set of values $(x_{i=1}^n)$, usually expressed in percentages, is the ratio of the standard deviation by the mean of the $(x_{i=1}^n)$. Thus

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}}$$

The order of magnitude does not have an impact on the *coefficient of variation* : that means that doubling each values of a set does not affect the coefficient of variation of the set (see the next example).

Example 5.4. Comparing the Coefficient of Variation of two set of values
We use the same first set of values as in the previous examples, but this time, each values of the first set has been multiplied by two to form the second set of values. Imagine two stocks, the second one having prices two times bigger than the first one.

First Set		Second Set	
x_i	$(x_i - \bar{x})^2$	y_i	$(y_i - \bar{y})^2$
3	16	6	64
5	4	10	16
7	0	14	0
9	4	18	16
11	16	22	64
$\bar{x}$	σ_x	$\bar{y}$	σ_y
7	2.83	14	5.66
Coefficient of Variation		Coefficient of Variation	
0.40		0.40	

The standard deviation of the second set is two times the one of the first set while the coefficient of variation is unchanged.

Definition 5.8. Variance
The variance is the square of the standard deviation, noted as σ^2. Thus :

$$\text{variance} = \sigma^2$$

5.3.1 Properties

Translation The standard deviation of the $(x_{i=1}^n)$ is the same as the standard deviation of $(x_{i=1}^n + b)$

Product The standard deviation of the $(a * x_{i=1}^n)$, with a a constant real number, equals to a times the standard deviation of the $(x_{i=1}^n)$.

Formally :

$$\sigma_{x+b} = \sigma_x$$
$$\sigma_{ax} = a * \sigma_x$$

5.4 Box Plots

The boxplot is a chart created by John TUKEY in 1977 aiming at summing different values of dispersion.

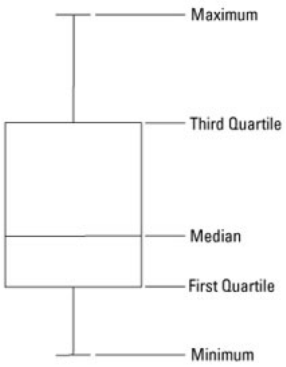


Figure 4: A Box-Plot or Box-and-Whisker Chart

6 Values of concentration

While dispersion values are aiming at exhibiting the stretching of a set of values (or the inequality), the values of concentration depicts the "concentration" or the dispersion of the division of a mass among entities. The concentration values are used for depicting how market shares are distribution among firms, wages among employees, etc.
The values of concentration can only be used with variables that can be summed and divided or shared : like wages, market share, revenues but not like heights.

6.1 Medial

Definition 6.1. Medial
The medial value [6] of a variable is the value that cut the total sum of the values of the variable in two. For instance, in a firm, if wages were given to employees from the less paid to the best paid and if there is $1,000 to be distributed to the employees; then the medial would be the wage given when half of $1,000 had already been distributed.

Example 6.1. Example of a Medial Here are the wages in a firm :

Wages	Frequencies	Cumulative Frequencies	Middle of the bin	Wages	Cumulative Wages
100 - 200	5	5	150	750	750
200 - 300	8	13	250	2000	2750
300 - 400	7	20	350	2450	5200
400 - 500	8	28	450	3600	8800
500 - 600	9	37	550	4950	13750
600 - 700	8	45	650	5200	18950
700 - 800	9	54	750	6750	25700
800 - 900	4	58	850	3400	29100
900 - 1000	2	60	950	1900	31000

Table 8: Wages in Grand Budapest Hotel

The firm's payroll is $31,000, half of it is $15,500. The medial belongs to the bin $600 - 700$. As for the median, we suppose the equipartition of the wages inside a bin and we do a linear interpolation :

$$\begin{aligned} \text{Medial} &= 600 + 100 * \frac{15,500 - 13,750}{18,950 - 13,750} \\ &= 634 \end{aligned}$$

The median is the mean of the wage of the 30^{th} employee and of the 31^{th}, both belonging to the bin $500 - 600$ and equals to[7] $500 + \frac{100}{2} * (\frac{30-28}{37-28} + \frac{31-28}{37-28}) = 528$

[6] médiale in French
[7] with the same hypothesis of an equipartition of the wages inside the bin

Medial - Median : The difference between the medial and the median is a statistical measure of the concentration of the distribution of the variable. The higher the difference is, the more concentrated the variable is. Usually, the Medial - Median difference is divided by the median.
In the previous example, we have :

$$\begin{aligned}\frac{\text{Medial - Median}}{\text{Median}} &= \frac{634-528}{528}\\ &= 0.20\end{aligned}$$

6.2 Gini coefficient and Lorenz curve

6.2.1 The Lorenz Curve

The Lorenz curve is a curve developed by Max Lorenz in 1905 for representing inequality of the wealth distribution with :

- On the x-axis : percentage of the cumulative frequencies ($\Leftrightarrow$ % of the employees)
- On the y-axis : percentage of the cumulative sum of the values ($\Leftrightarrow$ % of cumulative wages)

The curve is often compared to the hypothetical distribution where $x\%$ of the employees earn $x\%$ of the payroll.

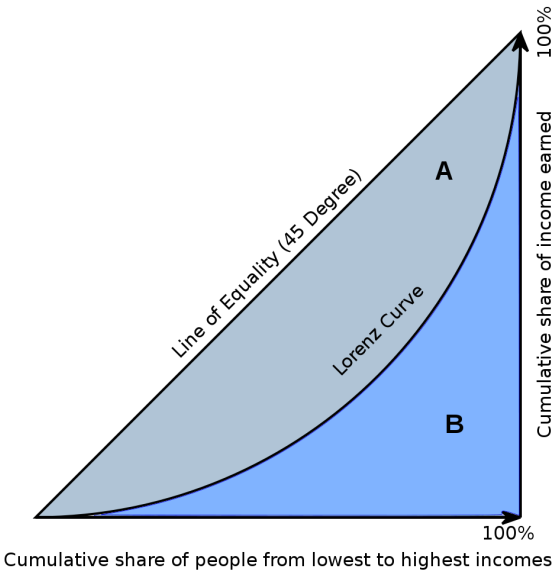


Figure 5: Lorenz Curve

Properties

- The more the distribution is unequal the more the distance between the two curves - the Lorenz and the line of perfect equality - is important (see graph above).
- The less equal distribution would be a Lorenz Curve that follows the x-axis (the "B" area being zero) from the origin to 100% and then be a vertical bar.

6.2.2 The Gini coefficient

Definition 6.2. The Gini coefficient is the ratio of the area between the Lorenz curve and the "line of equality" and between the "line of perfect equality" and "the line of perfect inequality".
With the notation of the figure above :

$$\begin{aligned}\text{Gini} &= \frac{A}{A+B}\\ \text{Gini} &= 2*A\\ &= 1-2*B\end{aligned}$$

The Gini coefficient belongs to the range $[0,1]$ with :

- 1 in case of perfect inequality
- 0 in case of perfect equality

Example 6.2. Gini coefficient in the World (World Bank data, latest data available)

	Gini coefficient
Argentina	0.423
China	0.421
France	0.331
Germnay	0.301
Norway	0.259
Russia	0.409
UK	0.326
USA	0.411

Table 9: Gini coefficient among different countries in the world

6.2.3 Calculating the Gini Coefficient

With only a few points are available, the Lorenz Curve looks like Figure 4.

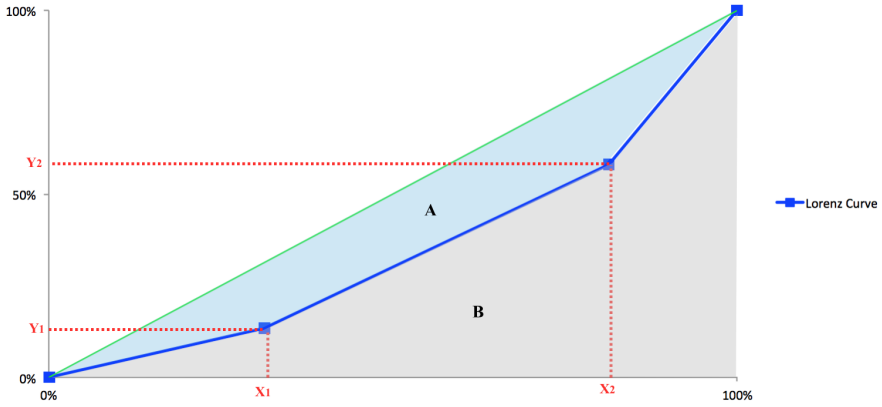


Figure 6: Lorenz Curve

To compute the Gini Coefficient, we need to calculate area A or B. Area B can be easily calculated, noting that it is composed of one triangle and two right-angled trapezoid (there is always at least one triangle, the number of right-angled trapezoid equals the number of available points) and using two basic area's formula :

- Right-angled triangle's area $= \dfrac{base * height}{2}$
- Right-angled trapezoid's area $= \dfrac{(a+b) * height}{2}$ with a and b the two parallel sides.

Example 6.3. Gini coefficient in the Grand Bupapest Hotel We use the same example but with three bins only for the sake of simplicity. The table of wages begin :

Wages	Frequencies	Cumulative Frequencies	Middle of the bin	Wages per bin	Cumulative Wages
100 - 200	5	5	150	750	750
200 - 300	8	13	250	2000	2750
300 - 400	3	16	350	1050	3800

Table 10: Wages in Grand Budapest Hotel

We first compute the cumulative frequencies and wages in percentage of the number of employee and payroll. We get the following table

Cumulative Frequencies	Cumulative Wages	Cumulative Wages w. Perfect Equality
0.00%	0.00%	0.00%
31.25%	19.74%	31.25%
81.25%	72.37%	81.25%
100.00%	100.00%	100.00%

We then draw the Lorenz Curve :

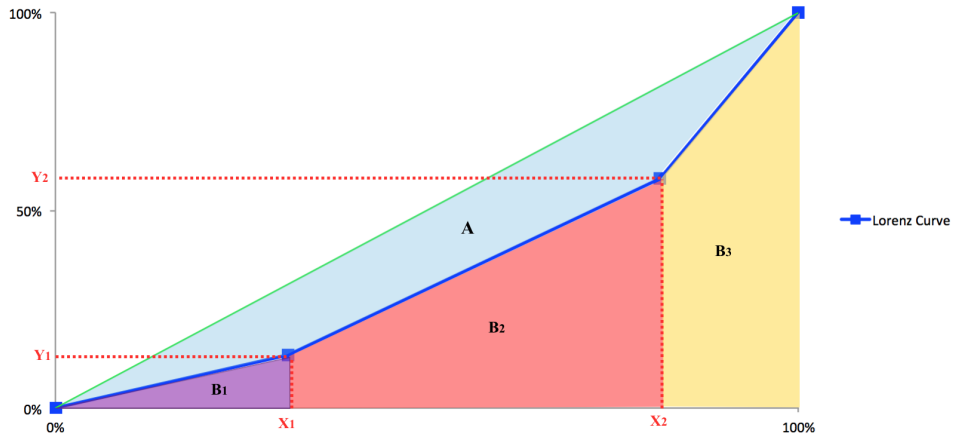
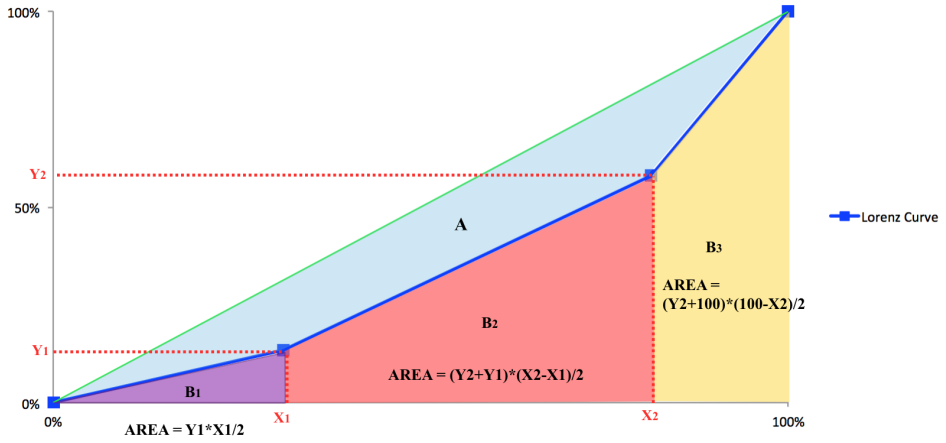


Figure 7: Lorenz Curve in the Grand Budapest Hotel

We then apply the formula to get the areas of the triangle and of the trapeze.



We find :

X1	31.25%
X2	81.25%
Y1	19.74%
Y2	72.37%
Area B1	0.031
Area B2	0.230
Area B3	0.162
Area B	0.423
Gini Coef.	0.155

* * *