Quantitative Tools - Level 1 - Fall 2015
Louis de Charsonville

Lecture 7

Values of dispersion and concentration

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.

1 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.

1.1 Interquantiles ranges, Deciles

Definition 1.1. 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.

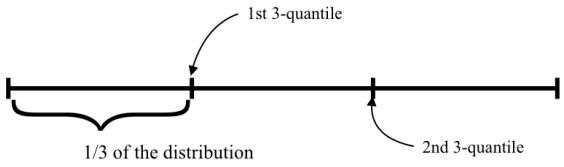


Figure 1: Example of a 3-quantile

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 1.2. 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 1.3. 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 1.4. 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 1.1. 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 1: 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.

1.2 Absolute deviation

1.2.1 Definitions

Definition 1.5. 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[1]. 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 1.2. 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) :

- 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.

1.3 Standard deviation and variance

The standard deviation [2] is the most common value of the dispersion of a variable. It is usually referred as σ.

Definition 1.6. Standard deviation
The standard deviation[3] 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)

[1] For a continuous quantitative variable, the mean is usually noted μ
[2] in French : "écart-type"
[3] in French : *écart-type*

Example 1.3. 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.

Definition 1.7. 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 1.4. 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 1.8. Variance
The variance is the square of the standard deviation, noted as σ^2. Thus :

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

1.3.1 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 :

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

1.4 Box Plots

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

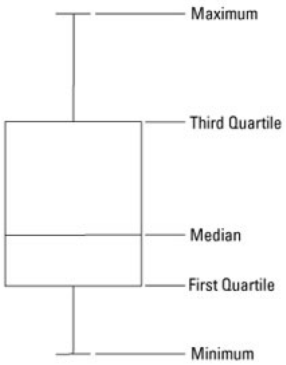


Figure 2: A Box-Plot or Box-and-Whisker Chart

2 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.

2.1 Medial

Definition 2.1. Medial
The medial value [4] 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 2.1. 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 2: Wages in Grand Budapest Hotel

[4] médiale in French

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[5] $500 + \frac{100}{2} * (\frac{30-28}{37-28} + \frac{31-28}{37-28}) = 528$

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 :

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

2.2 Gini coefficient and Lorenz curve

2.2.1 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.

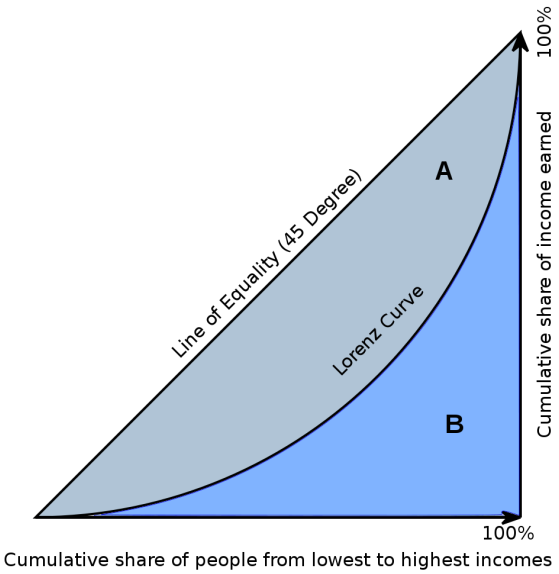


Figure 3: Lorenz Curve

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.

[5] with the same hypothesis of an equipartition of the wages inside the bin

2.2.2 The Gini coefficient

Definition 2.2. 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 2.2. 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 3: Gini coefficient among different countries in the world

2.2.3 Calculating the Gini Coefficient

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

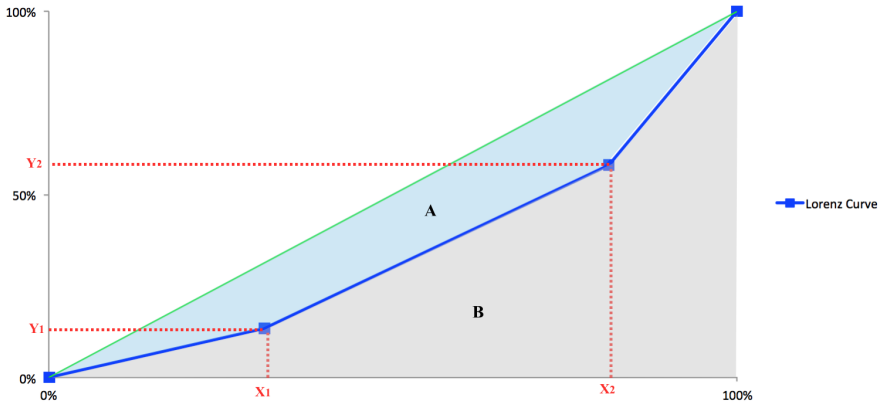


Figure 4: Lorenz Curve

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 2.3. 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 4: 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 :

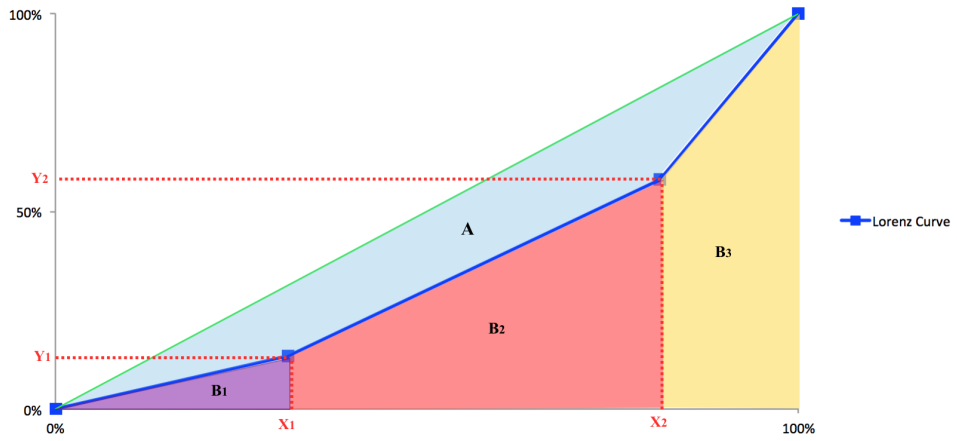
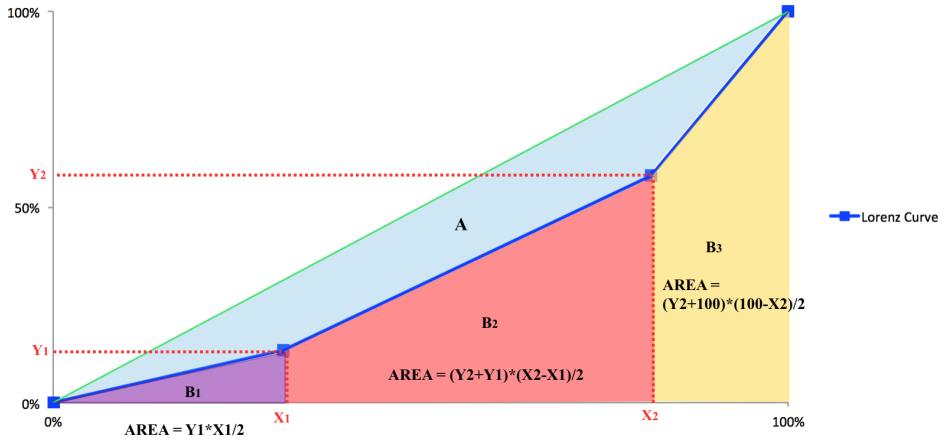


Figure 5: Lorenz Curve in the Grand Budapest Hotel

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

* * *