Quantitative Tools - Level 1 - Fall 2015
Louis de Charsonville

Lecture 5

Core Values

---

1 Mode

Definition 1.1. 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.

1.1 Qualitative variables & discrete quantitative variable

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

Example 1.1. 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

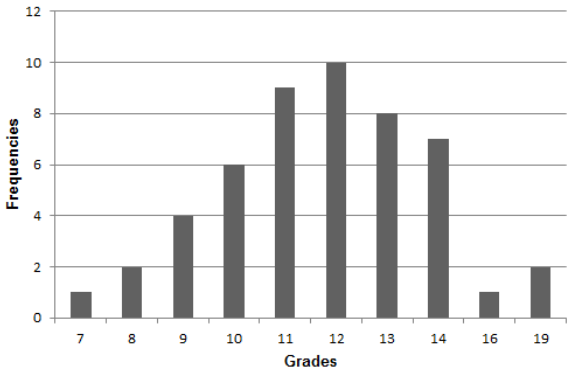

Figure 1: Bar plot of Grades

The Mode is 12.

1.2 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 1.2. 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 :

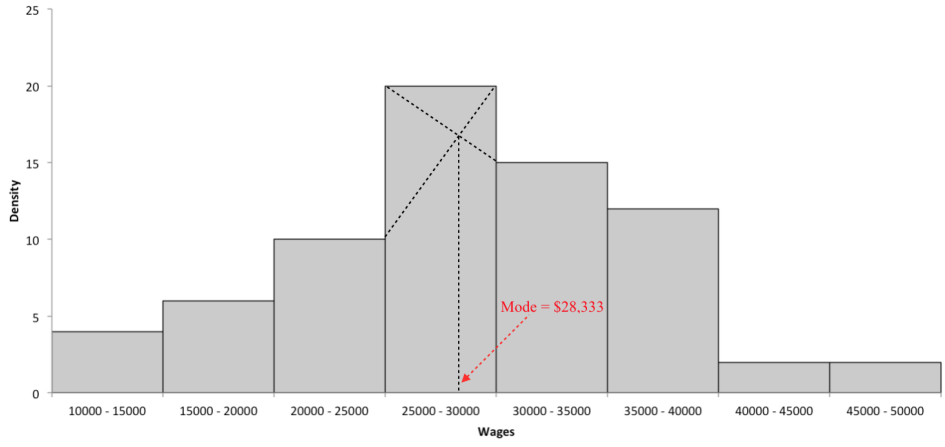

Figure 2: Graphical Determination of the Mode

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

Example 1.3. 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 :

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

2 Median

Definition 2.1. Median The median 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.

2.1 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 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	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.

2.2 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.

Example 2.2. 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 $525 + \frac{6}{41} = 528.66$, and the 230^{th} worker's income is $525 + \frac{7}{41} = 529.27$.
The median income is therefore $\frac{528.66+529.27}{2} = 528.97$

2.3 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.

3 Means

3.1 Arithmetic mean

3.1.1 Definitions

The arithmetic mean is the most known of core values and it is also used very frequently.

Definition 3.1. 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 3.2. Simple arithmetic mean
The arithmetic mean of $x_1, x_2, ..., x_n$ values appearing w_i times, called the weighted arithmetic mean is :

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

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

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

3.1.2 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 3.1. 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.

3.1.3 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

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

- The sum of all deviations from mean equals to zero

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

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

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

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

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

- 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 $x_i, i \in [1, n]$.

x

3.1.4 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.

3.1.5 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 $\rightarrow$ we call it a structural effect.

3.2 Geometric mean

3.2.1 Definition

Definition 3.3. 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] :

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

3.2.2 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 3.2. 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 :

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

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

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 :

$$G_{xy} = G_x * G_y$$
$$G_{x/y} = \frac{G_x}{G_y}$$

3.3 Harmonic mean

3.3.1 Definitions

Definition 3.4. 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 :

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

3.3.2 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 3.3. 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 :

$$s_{mean} = \frac{2 * d}{t_{outward} + t_{return}}$$
$$s_{mean} = \frac{2}{\frac{t_{outward}}{d} + \frac{t_{return}}{d}}$$
$$s_{mean} = \frac{2}{s_{outward} + s_{return}}$$
$$s_{mean} = \frac{2}{\frac{1}{100} + \frac{1}{120}}$$

3.4 Quadratic Mean

Definition 3.5. 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.

* * *