Descriptive Statistics: Measure of Central Tendency, Variation and Position

Rosario Barone

Tor Vergata University of Rome

Statistical tools for decision making

Undergraduate Degree in Global Governance

A.Y. 2023/2024

Introduction

- When our data is very large, it is important to summarize the data. Summarizing the data really helps us in analyzing and extracting insights from the data.

- Descriptive statistics is a method of collecting data, processing data (summarizing and presenting), describing, and analyzing all data.

Measurements in descriptive statistics

- Descriptive statistics are essential tools for summarizing and interpreting data.
- Three key aspects of descriptive statistics:
  1. Measures of Central Tendency
  2. Measures of Variation
  3. Measures of Position

Measures of Central Tendency

- Central tendency measures provide a single value that represents the "center" of a dataset.
- The three main measures of central tendency are:
  1. Mean: The average of all values in the dataset.
  2. Median: The middle value when the data is sorted.
  3. Mode: The most frequently occurring value in the dataset.

Notation

- In the general definition, we indicate the mean with M.
- If the mean is calculated on the whole population, it is indicated with μ.
- If the mean is calculated on a sample, it is indicated with $\bar{x}$.

Arithmetic mean

We consider a dataset with n observations. More rigorously, we say:

Arithmetic mean

Let consider x_i with $i = 1, \ldots, n$ observations of X. We define the mean of the dataset as:

$$M_a(X) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The mean or average is the sum of the total values in the dataset divided by the number of values in the dataset. In general, when we talk about the mean, we refer to the arithmetic mean.

Arithmetic mean

- Provides information about the central value of the data distribution.
- Limits of the arithmetic mean: it does not provide any information on the trend of the phenomenon, nor on the real distribution of the data. Furthermore, the average is to be considered a good index of central position only if it is found in the area of the distribution with the greatest concentration of data. And that's not always the case.

Weighted mean

Weighted mean

Let consider x_i with $i = 1, \ldots, n$ observations of X. We define the weighted mean of the dataset as:

$$M_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

The weighted mean is a subset of the arithmetic mean. In the weight mean, we assume that each value has a certain weight, so to calculate the weight mean we must multiply the value by its respective weight first.

Weighted mean

- The weights can be specifically chosen to give greater importance to some values.
- Alternatively, the weights can also be the absolute frequencies of the values.
- if $w_i = 1$ for all $i = 1, \ldots, n$ then $M_w = M_a$.

Geometric Mean

Geometric mean

Let consider x_i with $i = 1, \ldots, n$ realization of a variable X. We define the geometric mean of the dataset as:

$$M_g = (x_1, \ldots, x_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

The geometric mean is calculated by multiplying all the values in the dataset and then taking the root with the power of the sum value in the dataset.

Geometric Mean

- The geometric mean is useful for some of its properties. In particular for calculating the average value of the ratio of two series x_i/y_i because of its property:

$$M_g\left(\frac{x_i}{y_i}\right) = \frac{M_g(x_i)}{M_g(y_i)}.$$

- It lends itself to the calculation of the average of the rates and of all the quantities which by their nature are multiplied together.
- The geometric mean is much more sensitive to the presence of very small values than the arithmetic mean.
- If one term of the distribution is equal to zero, the geometric mean vanishes completely.
- If the product of the values $\prod_{i=1}^{n} x_i$ is negative, the geometric mean does not exists, because a root with an even index has no real solutions if the radicand is negative.

Harmonic Mean

Harmonic mean

Let consider x_i with $i = 1, \ldots, n$ realization of a variable X. We define the harmonic mean of the dataset as:

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

The harmonic mean is calculated by dividing the number of values into the data set by the number of reciprocals of each value in the dataset.

Harmonic Mean

- The harmonic mean is used for the calculation of quantities that are inversely proportional to each other, for quantities defined as a ratio of other quantities (e.g. speed) or when the data distribution develops in arithmetic progression, i.e. when there is a constant difference between the consecutive terms.

- It is also widely used to calculate the purchasing power of a currency, as it is the reciprocal of the price of goods.

Quadratic Mean

Quadratic mean

Let consider x_i with $i = 1, \ldots, n$ realization of a variable X. We define the quadratic mean of the dataset as:

$$M_q = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}.$$

The quadratic mean is the square root of the arithmetic mean of the squares of the terms.

Quadratic mean

- The quadratic mean is more sensitive to the difference between the terms of the distribution $x_1, \ldots, x_n$ and the mean value.
- For example, it is useful for evaluating the deviation, deviation, or error of a data distribution from the arithmetic mean.
- One way to approximate the dispersion of a distribution from the mean value is to calculate the difference between the quadratic mean and the arithmetic mean

$$M_q - M_a \quad \text{or} \quad \frac{M_q - M_a}{M_a}$$

- The quadratic mean is equal to the arithmetic mean when all terms in the distribution are constant.

Quadratic mean as deviation evaluator

Suppose we observe two groups of students and measure their heights:

- Group 1: $Z = \{1.58, 1.73, 1.86, 1.92, 1.67, 1.83, 1.74\}$
- Group 2: $Y = \{1.42, 1.47, 1.44, 2.05, 2.12, 1.98, 1.52\}$

Quadratic mean as deviation evaluator

Suppose we observe two groups of students and measure their heights:

- Group 1: $Z = \{1.58, 1.73, 1.86, 1.92, 1.67, 1.83, 1.74\}$
- Group 2: $Y = \{1.42, 1.47, 1.44, 2.05, 2.12, 1.98, 1.52\}$
- Calculating the Arithmetic Means we get :

$$M_a(Z) = 1.7614 \quad \text{and} \quad M_a(Y) = 1.7143.$$

Quadratic mean as deviation evaluator

Suppose we observe two groups of students and measure their heights:

- Group 1: $Z = \{1.58, 1.73, 1.86, 1.92, 1.67, 1.83, 1.74\}$
- Group 2: $Y = \{1.42, 1.47, 1.44, 2.05, 2.12, 1.98, 1.52\}$
- Calculating the Arithmetic Means we get :

$$M_a(Z) = 1.7614 \quad \text{and} \quad M_a(Y) = 1.7143.$$

- Calculating the Quadratic Means we get :

$$M_q(Z) = 1.7648 \quad \text{and} \quad M_q(Y) = 1.7394.$$

- $$\frac{M_q - M_a(Z)}{M_q(Z)} = 0.0019 \quad \text{and} \quad \frac{M_q - M_a(Y)}{M_q(Y)} = 0.0144$$

Median

- The median is the central value that divides an ordered distribution into two groups of equal size
- The first group is composed of terms less than or equal to the median.
- The second group is, however, composed of terms greater than or equal to the median.

How to calculate the Median

- Sort the distribution in increasing order.
- Calculate the median position depending on whether the ordered distribution has an even or odd number of terms:
  - If the distribution has an odd number of terms n, the median is the term that occupies the central position (c)

$$c = \frac{n+1}{2} \quad \text{and} \quad Me(X) = x_c$$

  - If the distribution has an even number of terms n, calculate the cenrtal positions (c_1) and (c_2)

$$c_1 = \frac{n}{2} \quad \text{and} \quad c_2 = \frac{n}{2} + 1$$

and

$$Me(X) = \frac{x_{c_1} + x_{c_2}}{2}.$$

How to calculate the Median

Let consider the distribution $X = \{4, 1, 7, 2, 6, 18, 12\}$:

- Sort the distribution in increasing order: $X = \{1, 2, 4, 6, 7, 12, 18\}$.
- To each term of the ordered distribution I assign an increasing index starting from 1 which identifies its position:
  $(x_1 = 1, x_2 = 2, x_3 = 4, x_4 = 6, x_5 = 7, x_6 = 12, x_7 = 18)$.
- Since $n = 7$, we get $c = \frac{7+1}{2} = 4$
- $Me(X) = x_4 = 6$

How to calculate the Median

Let consider the distribution $X = \{4, 1, 7, 2, 6, 18, 12, 3\}$:

- Sort the distribution in increasing order: $X = \{1, 2, 4, 4, 6, 7, 12, 18\}$.
- To each term of the ordered distribution I assign an increasing index starting from 1 which identifies its position:
  $(x_1 = 1, x_2 = 2, x_3 = 4, x_4 = 4, x_5 = 6, x_6 = 7, x_7 = 12, x_8 = 18)$.
- Since $n = 8$, we get $c_1 = \frac{8}{2} = 4$ and $c_2 = \frac{8}{2} + 1 = 5$.
- $Me(X) = \frac{x_4 + x_5}{2} = \frac{4+6}{2} = 5$.

Mode

- Mode is the only position summary index that can be calculated on nominal qualitative data as well as on quantitative data.
- In discrete distributions the mode is the value at which the frequencies (f_i or n_i) reaches its maximum value.
- In continuous distributions the mode is the value at which the density reaches its maximum value.
- In a distribution curve the mode is the maximum point of the graph. If the distribution is unimodal there is only one maximum point. Conversely, if the distribution is multimodal there are more maximum points.
- When the frequency of occurrence of a value in the data set is the same, it indicates that there is no mode.
- If there are two values that have the highest frequency of occurrence, it is called bimodal.

Mean, Median and Mode

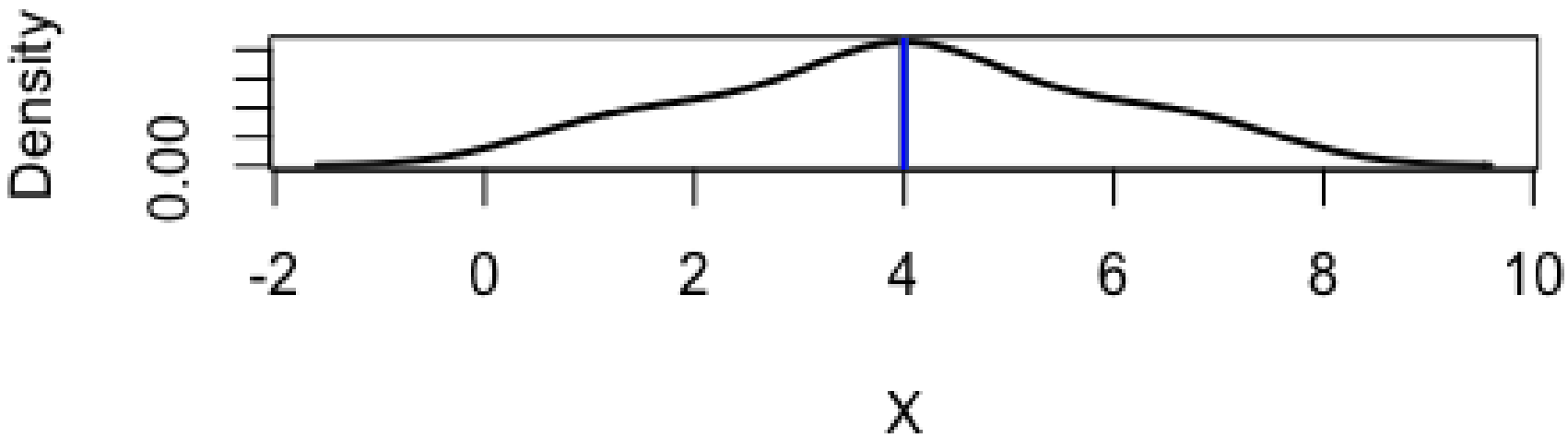
- Let consider $X = \{1, 2, 3, 4, 4, 4, 5, 6, 7\}$.
- Mode is equal to 4 because four is the term with highest frequency:

$$X = \{1, 2, 3, \color{red}{4, 4, 4}, 5, 6, 7\}$$

- Also the mean $M_a(X) = 4$
- Also the median $M_e(X) = 4$

Mean, Median and Mode

If a distribution is unimodal and symmetric the mode, mean and median are equal.



Mean, Median and Mode

If the distribution is asymmetric the mode, the mean and the median have different values:

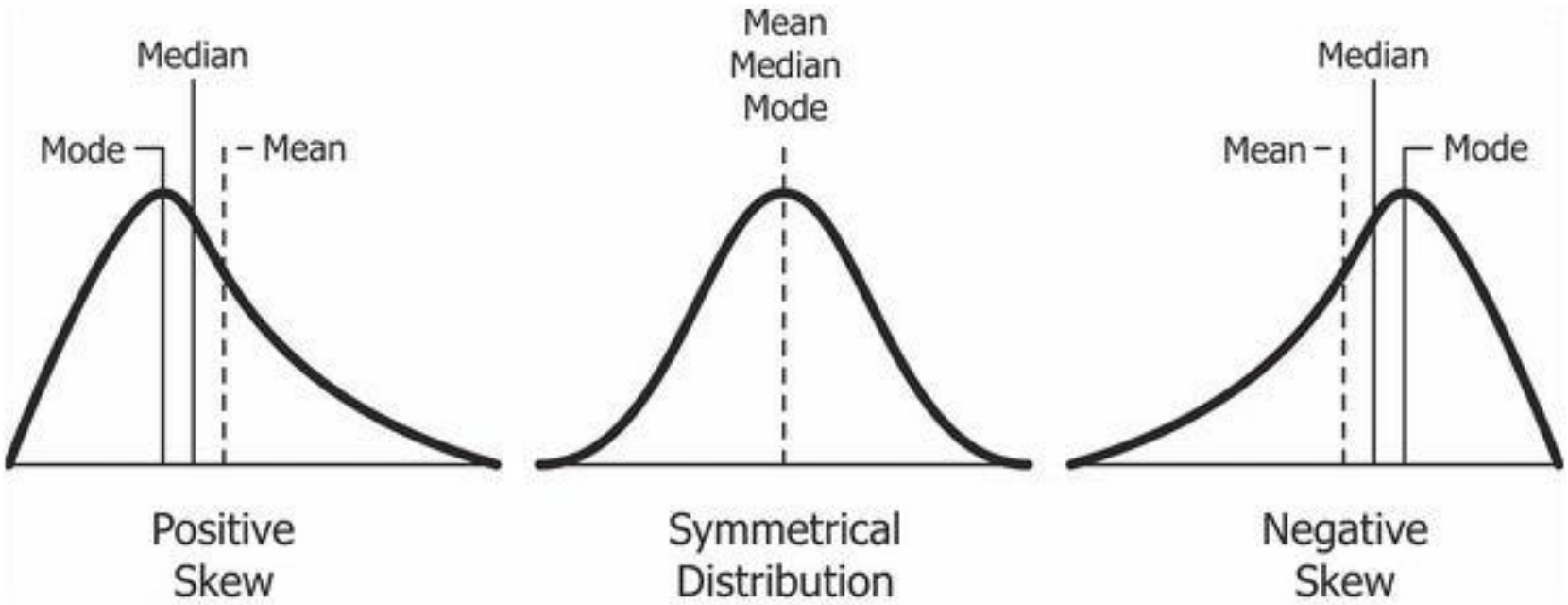
- **Positive Skewness**: if the frequency distribution is asymmetric and skewed to the left, the mode is lower than the median and mean:

$$Mo(X) < Me(X) < M_a(X)$$

- **Negative Skewness**: if the frequency distribution is asymmetric and skewed to the right , the mode is higher than the median and mean:

$$Mo(X) > Me(X) > M_a(X)$$

Mean, Median and Mode



Measures of Position

- Measures of position help us identify the relative position of a data point within a dataset.
- Quantiles are position indices that divide an ordered distribution into equal parts.
- The two common quantiles type are:
  1. Percentiles: Values that divide the data into 100 equal parts.
  2. Quartiles: Values that divide the data into four equal parts (Q1, Q2, Q3).

Quartiles

Quartiles

Quartiles are three position indices (quantiles) that divide a statistical distribution into four equal parts.

- The first quartile (Q_1) groups 1/4 of the elements (25%) of the distribution on the left.
- The second quartile (Q_2) groups 2/4 of the elements (50%) of the distribution on the left. It coincided with the Median (Me).
- The third quartile (Q_3) groups 3/4 of the elements (75%) of the distribution on the left

How to calculate quartiles

If the product k is not an integer, I get the quartile position by rounding k up to the next integer.

- Sort the distribution of values in ascending order
- multiply the number of elements of the series n by $p = 1/4$ in the case of Q_1, by $p = 2/4$ in the case of Q_2 and by $p = 3/4$ in the case of Q_3.

$$k = n \cdot p.$$

- Calculate the quartile position:
  - If the product k is an integer $Q_p = \frac{x_k + x_{k+1}}{2}$.
  - If the product k is not an integer, get the quartile position by rounding k up to the next integer.

Calculation of quartiles: Example

The distribution $X = \{9, 6, 11, 8, 4, 7, 10, 3, 5\}$ if composed of n=9 elements.

- sort the distribution of $X = \{3, 4, 5, 6, 7, 8, 9, 10, 11\}$.
- to calculate Q_1, first calculate $k = \frac{1}{4}9 = 2.25$. Then, since 2.25 is not an integer, round k up to the next integer: $k = 3$.
- then, we obtain $Q_1 = x_3 = 5$.
- calculate Q_2: $k = 0.5 * 9 = 4.5$, that is rounded up to the next integer: $k = 5$ and $Q_2 = 7$.
- calculate Q_3: $k = 0.75 * 9 = 6.75$, that is rounded up to the next integer: $k = 7$ and $Q_3 = 9$.

How to calculate quartiles in a frequency distribution

- Calculate the cumulative absolute frequencies of each class of the distribution
- Multiply the total cumulative frequencies ($f_{tot} = N$) by 1/4, by 2/4 and by 3/4. In this way I find the position of the first quartile (Q1), the second quartile (Q2) and the third quartile (Q3) in the cumulative frequencies.
- find the cumulative frequency ranges that include the Q1, Q2, and Q3 quartile positions. The respective frequency classes are the quartiles of the distribution.

Calculation of quartiles in a frequency distribution: Example

Grade	f_i	F_i	interval
18	2	2	[0,2[
20	7	9	[2,9[
21	4	13	[9,13[
22	3	16	[13,16[
24	6	22	[16,22[
25	8	30	[22,30[
26	4	34	[30,34[
27	3	37	[34,37[
28	2	39	[37,39[
30	1	40	[39,40[

Calculation of quartiles in a frequency distribution: Example

- Q_1: k=$N \cdot \frac{1}{4} = 40 \cdot \frac{1}{4} = 10$, that is in the interval 9-3, corresponding to te class 21. $Q_1 = 21$
- Q_2: k=$N \cdot \frac{1}{2} = 40 \cdot \frac{1}{2} = 20$, that is in the interval 16-22, corresponding to te class 24. $Q_2 = 24$
- Q_3: k=$N \cdot \frac{1}{3} = 40 \cdot \frac{3}{4} = 30$, that is in the interval corresponding to te class 26. $Q_3 = 26$.

Percentiles

Percentiles

Percentiles (or centiles) are 99 position indices that divide a statistical distribution into one hundred equal parts. Each part is a group with the same number of elements.

- The first percentile (P1) groups 1/100 of the elements (1%) of the distribution on the left.
- The second percentile (P2) groups 2/100 of the elements (2%) of the distribution on the left.
- $\vdots$
- The ninety-ninth percentile (P99) groups 99/100 of the elements (99%) of the distribution to the left

Measures of Variation

- Measures of variation help us understand the spread or dispersion of data.

- The variability of the data can be quantified in two different aspects

  - **Dispersion**: is the densification or not of values compared to an average value $M(X)$ (or other value). It can be measured through the deviation or the average deviation of the $x_i - M(X)$ values

  - **Inequality**: the difference in the value of the data. It can be measured by observing the differences in absolute value $|x_i - x_j|$ among all the elements of the population.

Measures of Variation

Variability indices are used in statistics to compare two or more statistical distributions or populations. For example, if I have two groups A and B with the same mean but group B has less dispersion, it means that the data from the second group is more concentrated around the mean value than the first group. Consider the distributions:

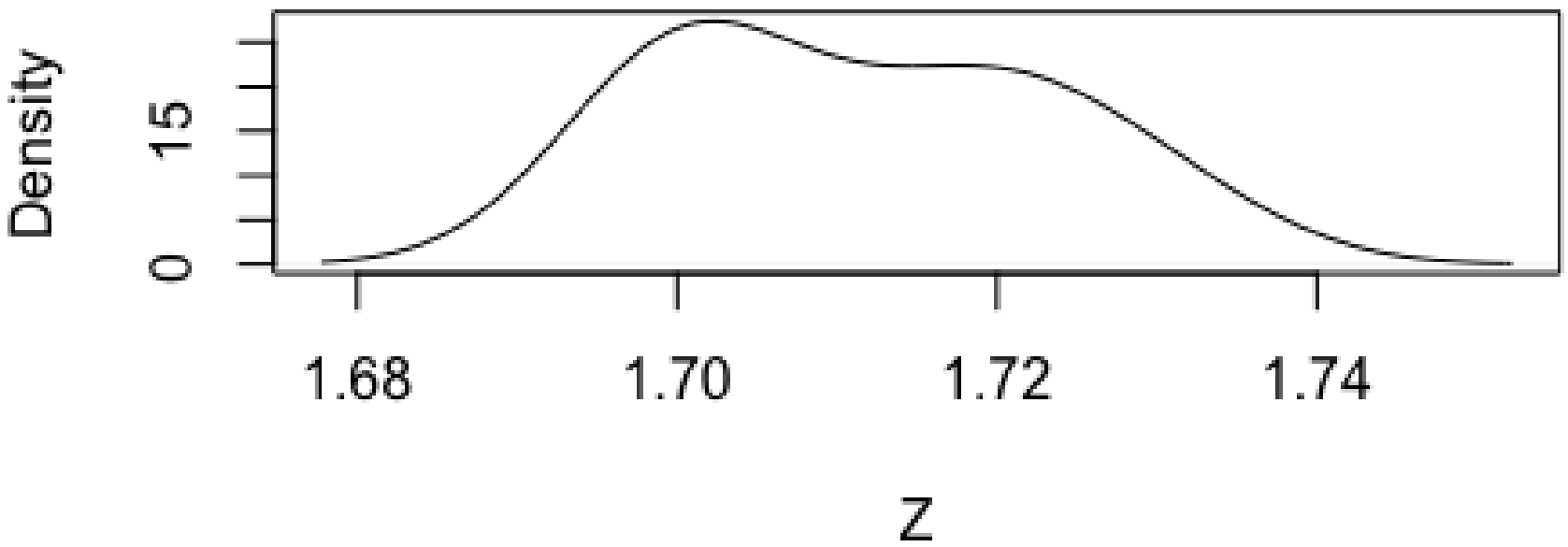
- Group 1: $Z = \{1.70, 1.73, 1.70, 1.70, 1.72, 1.71, 1.72\}$
- Group 2: $Y = \{1.42, 1.47, 1.44, 2.05, 2.12, 1.98, 1.52\}$

We can observe that both the distributions have (approximately) the same mean

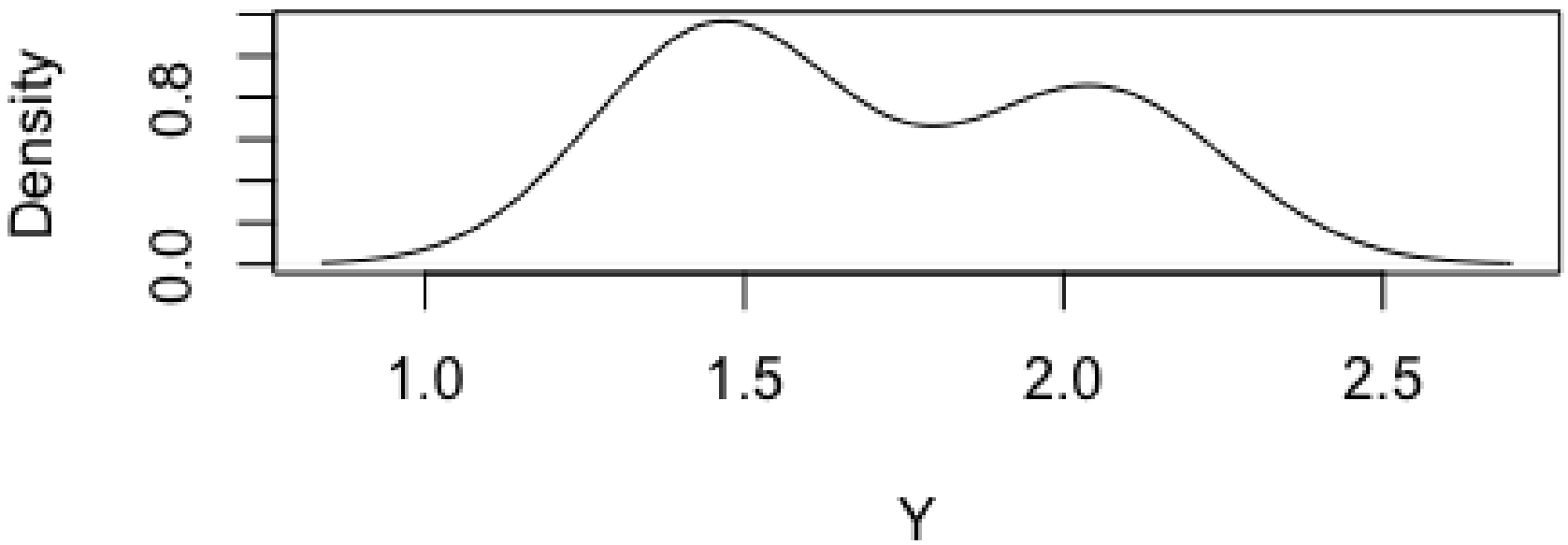
$$M(Z) = 1.7114 \quad \text{and} \quad M(Y) = 1.7143.$$

The mean is a summary index that tells me nothing about the variability of the data.

Density



Density



Measures of Variation

- Range
- Interquartile range
- Variance
- Standard Deviation
- Deviance
- Coefficient of Variation

Range

Range

The range is an index of variability obtained by measuring the difference between the maximum value and the minimum value of a statistical distribution:

$$\omega = x_{max} - x_{min}.$$

- it is an absolute index and provides little information because it only measures the dispersion between the extremes of the distribution.
- it is strongly influenced by the presence of outliers in the data.

Range

Range

The range is an index of variability obtained by measuring the difference between the maximum value and the minimum value of a statistical distribution:

$$\omega = x_{max} - x_{min}.$$

- it is an absolute index and provides little information because it only measures the dispersion between the extremes of the distribution.
- it is strongly influenced by the presence of outliers in the data.

Example:

$$\omega(Z) = 1.73 - 1.7 = 0.03 \quad \text{and} \quad \omega(Y) = 2.12 - 1.42 = 0.7$$

Interquartile range

Interquartile range

The interquartile range is an absolute variability index that measures the difference between the third and the first quartiles

$$\delta_Q = Q_3 - Q_1$$

- Less sensitive to outliers and extreme values of a distribution than the range because it only considers the central part of the distribution.
- Quite easy to calculate.
- Less complete index of variability, because it considers only 50% of the elements of the statistical. distribution.

Interquartile range

Interquartile range

The interquartile range is an absolute variability index that measures the difference between the third and the first quartiles

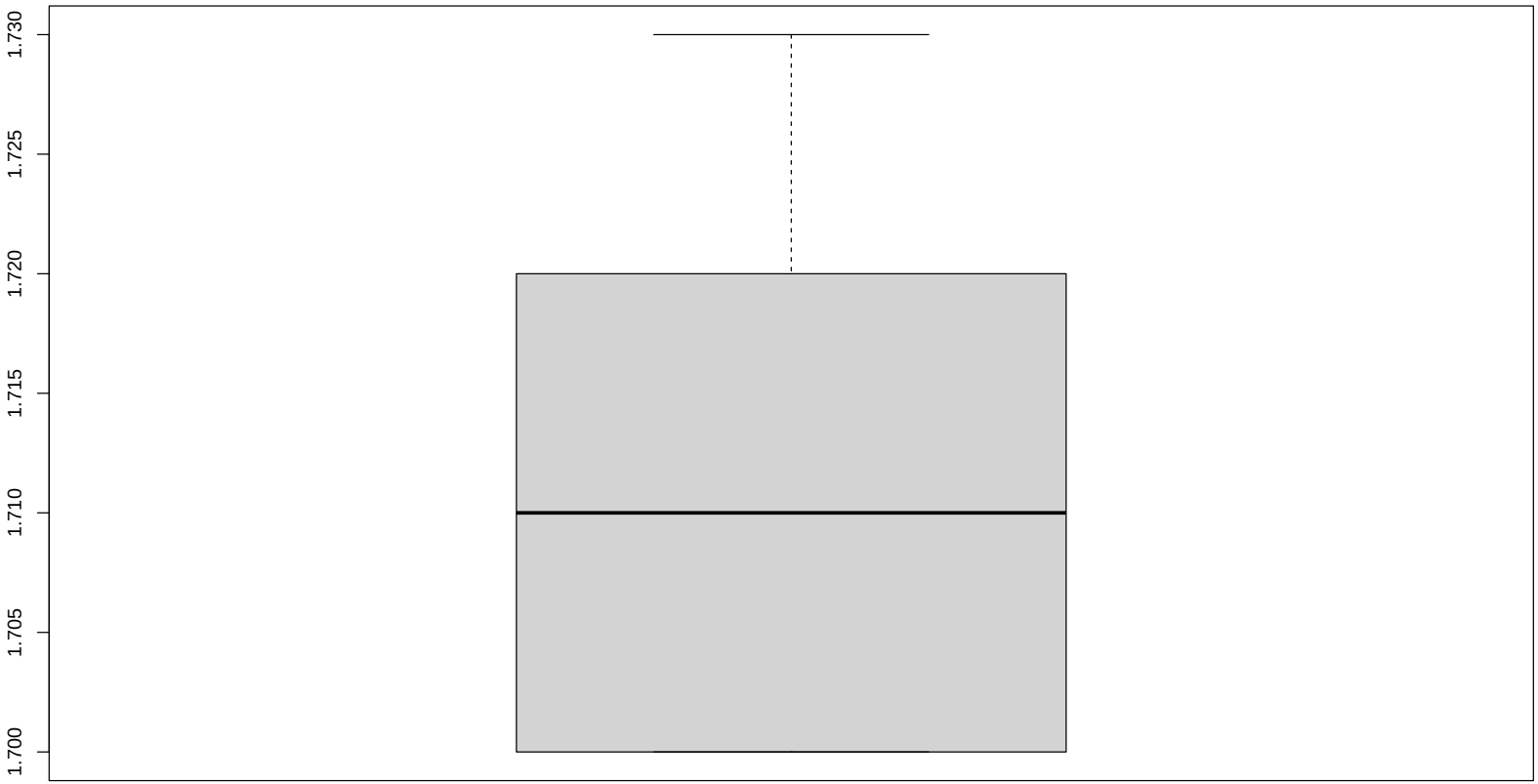
$$\delta_Q = Q_3 - Q_1$$

- Less sensitive to outliers and extreme values of a distribution than the range because it only considers the central part of the distribution.
- Quite easy to calculate.
- Less complete index of variability, because it considers only 50% of the elements of the statistical. distribution.

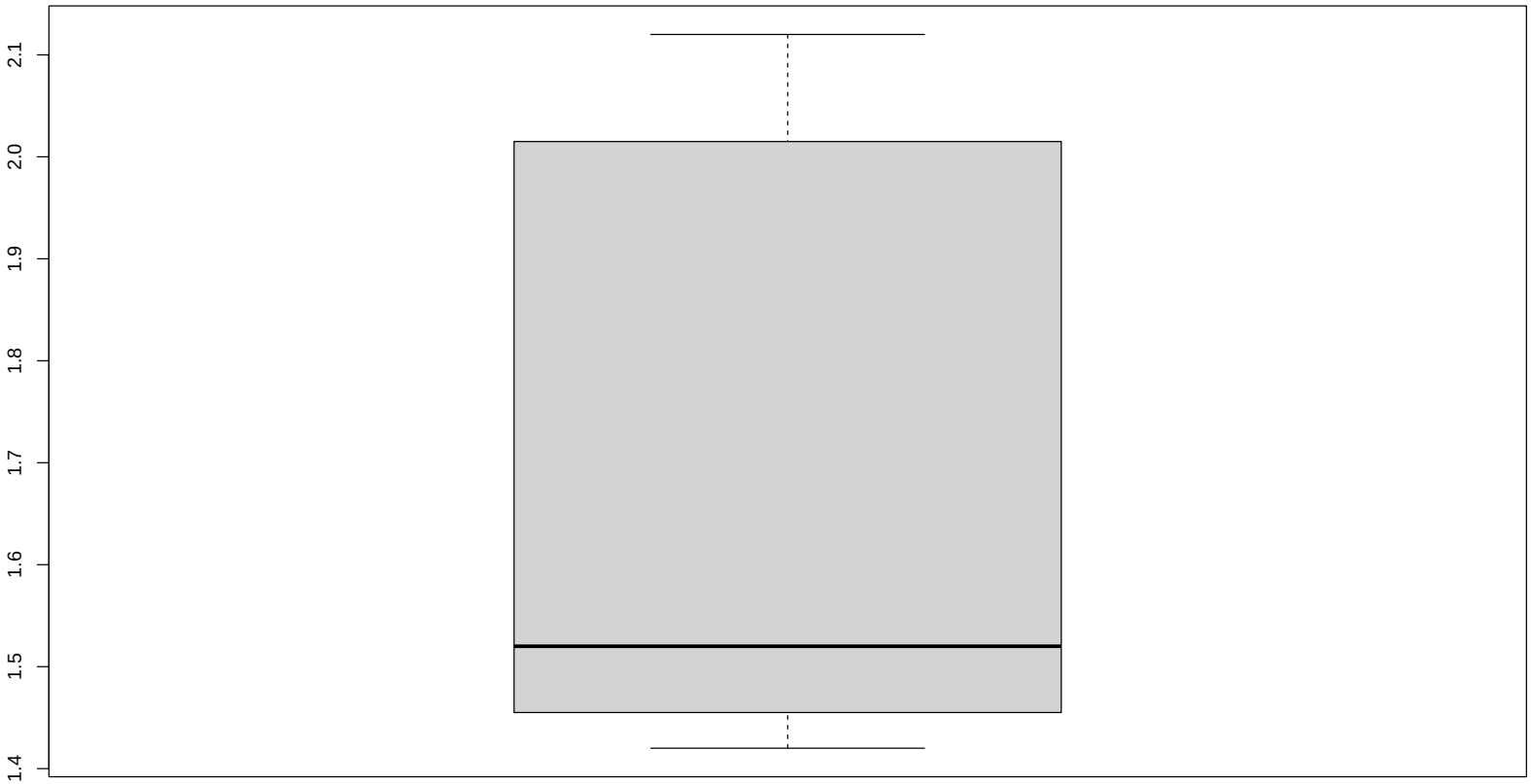
Example:

$$\delta_Q(Z) = 1.72 - 1.7 = 0.02 \quad \text{and} \quad \delta_Q(Y) = 2.015 - 1.455 = 0.56$$

Boxplot of Z



Boxplot of Y



Variance

Variance

Variance is an indicator of the dispersion of a variable or statistical distribution obtained by calculating the average of the squares of the deviations of the arithmetic mean M:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2.$$

In the case of the Frequency distributions, the variance is calculated as:

$$\sigma^2 = \frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=1}^{n}(x_i - M)^2 f_i$$

where:

- n is the total number of values.
- f_i is the absolute frequency of each value.

Variance

- Variance does not have the same unit of measurement as the observed phenomenon The variance is equal to the square of the unit of measurement of the observed phenomenon. For example, if the phenomenon is measured in meters (m), the variance of the phenomenon is measured in meters squared (m^2).
- Alternative way of calculating the variance:

$$\sigma^2 = (M_q)^2 - (M_a)^2.$$

- We will uniderline the difference between *population* variance and *sample* variance.

Variance

- Variance does not have the same unit of measurement as the observed phenomenon The variance is equal to the square of the unit of measurement of the observed phenomenon. For example, if the phenomenon is measured in meters (m), the variance of the phenomenon is measured in meters squared (m^2).
- Alternative way of calculating the variance:

$$\sigma^2 = (M_q)^2 - (M_a)^2.$$

- We will uniderline the difference between *population* variance and *sample* variance. Example:

$$\sigma^2(Z) = 0.0002 \quad \text{and} \quad \sigma^2(Y) = 0.1012$$

Standard Deviation

Standard Deviation

Standard Deviation is an indicator of the dispersion obtain as the square root of the Variance σ^2:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2}.$$

In the case of the Frequency distributions, the variance is calculated as:

$$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=1}^{n}(x_i - M)^2 f_i}$$

where:

- n is the total number of values.
- f_i is the absolute frequency of each value.

Standard Deviation

- Is a measure of how widely the data is spread around the mean
- Generally about $2/3$ of the elements of a distribution are included in the interval $(M - \sigma, M + \sigma)$. Almost all elements of the distribution are included in the interval $(M - 3\sigma, M + 3\sigma)$.

Standard Deviation

- Is a measure of how widely the data is spread around the mean
- Generally about $2/3$ of the elements of a distribution are included in the interval $(M - \sigma, M + \sigma)$. Almost all elements of the distribution are included in the interval $(M - 3\sigma, M + 3\sigma)$. Example:

$$\sigma(Z) = 0.0122 \quad \text{and} \quad \sigma(Y) = 0.3181$$

Deviance

Deviance is an indicator of the dispersion of a variable or statistical distribution obtained by calculating the sum of the squares of the deviations of the data of a distribution from the mean M:

$$D = \sum_{i=1}^{n} (x_i - M)^2.$$

In the case of the Frequency distributions, the variance is calculated as:

$$D = \sum_{i=1}^{n} (x_i - M)^2 f_i$$

where:

- n is the total number of values.
- f_i is the absolute frequency of each value.ù

Example:

$$D(Z) = 0.0918 \quad \text{and} \quad D(Y) = 62.9434$$