Chapter 3

Descriptive measures for univariate distributions

There are four families of scale level-dependent standard measures one employs in Statistics to describe characteristic properties of univariate relative frequency distributions. We will introduce these in turn. In the following we suppose given from a survey for some 1-D statistical variable X either

(i) a raw data set $\{x_i\}_{i=1;\ldots;n}$ of n measured values, or

(ii) a relative frequency distribution $(a_j; h_j)_{j=1;\ldots;k}$ resp. $(K_j; h_j)_{j=1;\ldots;k}$.

3.1 Measures of central tendency

Let us begin with the measures of central tendency which intend to convey a notion of "middle" or "centre" of a univariate relative frequency distribution.

1. Mode
   The mode x_{mod} (nom, ord, metr) of the relative frequency distribution of any 1-D variable X is that value a_j in X's spectrum which occurred with the highest measured relative frequency in a statistical sample S. Note that the mode does not necessarily take a unique value.

   Definition 3.1.1 *$h_n(x_{mod}) \geq h_n(a_j)$ for all $j = 1, \ldots, k$.*

2. Median
   To determine the median $\tilde{x}_{0.5}$ (or Q_2) (ord, metr) of the relative frequency distribution of an ordinally or metrically scaled 1-D variable X, it is necessary to first bring the n observed values $\{x_i\}_{i=1,\ldots,n}$ into their natural hierarchical order, i.e., $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

   Definition 3.1.2 *For the sequentially ordered n observed values $\{x_i\}_{i=1,\ldots,n}$, at most 50% have a rank lower or equal to resp. are less or equal to the median value $\tilde{x}_{0.5}$, and at most 50% have a rank higher or equal to resp. are greater or equal to the median value $\tilde{x}_{0.5}$.*

   (a) *Discrete data* $\boxed{F_n(\tilde{x}_{0.5}) \geq 0.5}$

$$\tilde{x}_{0.5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right] & \text{if } n \text{ is even} \end{cases} \tag{3.1}$$

(b) *Binned data* $\boxed{F_n(\tilde{x}_{0.5}) = 0.5}$

The class interval K_i contains the median value $\tilde{x}_{0.5}$ if $\sum_{j=1}^{i-1} h_j < 0.5$ and $\sum_{j=1}^{i} h_j \geq 0.5$. Then

$$\tilde{x}_{0.5} = u_i + \frac{b_i}{h_i}\left(0.5 - \sum_{j=1}^{i-1} h_j\right) . \tag{3.2}$$

*Alternatively, the median of a statistical sample S for a continuous variable X with binned data $(K_j; h_j)_{j=1,\ldots,k}$ can be obtained from the associated empirical cumulative distribution function by solving the condition $\tilde{F}_n(\tilde{x}_{0.5}) = 0.5$ for $\tilde{x}_{0.5}$; cf. equation (**??**).*

Remark 3.1.1 Note that the value of the median of a relative frequency distribution is fairly insensitive to so-called outliers in a statistical sample.

3. α-Quantile
A generalisation of the median is the concept of the α-quantile $\tilde{x}_\alpha$ (ord, metr) of the relative frequency distribution of an ordinally or metrically scaled 1-D variable X. Again, it is necessary to first bring the n observed values $\{x_i\}_{i=1,\ldots,n}$ into their natural hierarchical order, i.e., $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Definition 3.1.3 *For the sequentially ordered n observed values $\{x_i\}_{i=1,\ldots,n}$, for given α with $0 < \alpha < 1$ at most $\alpha \times 100\%$ have a rank lower of equal to resp. are less or equal to the α-quantile $\tilde{x}_\alpha$, and at most $(1-\alpha) \times 100\%$ have a rank higher or equal to resp. are greater or equal to the α-quantile $\tilde{x}_\alpha$.*

(i) *Discrete data $F_n(\tilde{x}_\alpha) \geq \alpha$*

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \notin \mathbb{N}, k > n\alpha \\ \frac{1}{2}[x_{(k)} + x_{(k+1)}] & \text{if } k = n\alpha \in \mathbb{N} \end{cases} \tag{3.3}$$

(ii) *Binned data $\tilde{F}_n(\tilde{x}_{0.5}) = \alpha$*

The class interval K_i contains the α-quantile $\tilde{x}_\alpha$, if $\sum_{j=1}^{i-1} h_j < \alpha$ and $\sum_{j=1}^{i} \geq \alpha$. Then

$$\tilde{x}_\alpha = u_i + \frac{b_i}{h_i}\left(\alpha - \sum_{j=1}^{i-1} h_j\right) . \tag{3.4}$$

*Alternatively, an α-quantile of a statistical sample S_Ω for a continuous variable X with binned data $(K_j, h_j)_{j=1,\ldots,}$ can be obtained from the associated empirical cumulative distribution function by solving the condition $\tilde{F}_n(\tilde{x}_\alpha) = \alpha$ for $\tilde{x}_\alpha$, cf. equation (**??**).*

Remark 3.1.2 The quantiles $\tilde{x}_{0:25}, \tilde{x}_{0:5}, \tilde{x}_{0:75}$ (also denoted by Q_1, Q_2, Q_3) have special status. They are referred respectively to as the first quartile, second quartile (median) and third quartile of a relative frequency distribution for an ordinally or a metrically scaled 1-D X and form the core of the five number summary of the respective distribution. Occasionally, α-quantiles are also referred to as percentile values.

4. Five number summary
The five number summary (ord, metr) of the relative frequency distribution of an ordinally or metrically scaled 1-D variable X is a compact compilation of information giving the

(i) lowest rank resp. smallest value,

(ii) first quartile,

(iii) second quartile or median,

(iv) third quartile, and

(v) highest rank resp. largest value

that X takes in a raw data set $\{x_i\}_{i=1,\ldots,n}$ from a statistical sample S, i.e.,

$$\{x_{(1)}; \tilde{x}_{0.25}; \tilde{x}_{0.5}; x_{\tilde{0.75}}; x_{(n)}\} \tag{3.5}$$

Alternative notation: $\{Q_0; Q_1; Q_2; Q_3; Q_4\}$.

A very convenient graphical method for transparently displaying distributional features of metrically scaled data relating to a five number summary is provided by a box plot.

All measures of central tendency which now follow are defined exclusively for characterising relative frequency distributions of metrically scaled 1-D variables X only.

5. Arithmetical mean
The best known measure of central tendency is the dimensionful arithmetical mean $\overline{x}$ (metr). Given adequate statistical data, it is defined by:

(i) From raw data set:

$$\overline{x} := \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.6}$$

(ii) From relative frequency distribution:

$$\overline{x} := a_1 h_n(a_1) + \ldots + a_k h_n(a_k) = \sum_{j=1}^{k} a_j h_n(a_j) \tag{3.7}$$

Remark 3.1.3 (i) The value of the arithmetical mean is very sensitive to outliers.

(ii) For binned data one selects the midpoint of each class interval K_i to represent the a_j (provided the raw data set is no longer accessible).

6. Weighted mean
In practice, one also encounters the dimensionful weighted mean $\overline{x}_w$ (metr), defined by

$$\overline{x}_w := w_1 x_1 + \ldots + w_n x_n = \sum_{i=1}^{n} w_i x_i; \tag{3.8}$$

the n weight factors $w_1, \ldots, w_n$ need to satisfy the constraints

$$0 \leq w_1, \ldots, w_n \text{ and } w_1 + \ldots + w_n = \sum_{i=1}^{n} = 1 \tag{3.9}$$

Exercise 1 On an interview for a job, the interviewer tells you that the average annual income of the company's 25 employees is \$60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes? Was the person telling you the truth?

$17,305	$478,320	$45,678	$18,980	$17,408
$25,676	$28,906	$12,500	$24,540	$33,450
$12,500	$33,855	$37,450	$20,432	$28,956
$34,983	$36,540	$250,921	$36,853	$16,430
$32,654	$98,213	$48,980	$94,024	$35,671

Exercise 2 Which measure of central tendency is the most representative of the data shown in each frequency distribution?

1.

1	2	3	4	5	6	7	8	9
7	20	15	11	8	3	2	0	15

2.

1	2	3	4	5	6	7	8	9
9	8	7	6	5	6	7	8	9

3.

1	2	3	4	5	6	7	8	9
6	1	2	3	5	5	4	3	0

Exercise 3 Find the lower and upper quartiles for the set:

34, 14, 24, 16, 12, 18, 20, 24, 16, 26, 13, 27

Exercise 4 Sketch a box-and-whisker plot for each set:

1. $27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66$
2. $82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99$
3. $11, 13, 13, 15, 17, 18, 20, 24, 24, 27$

3.2 Measures of variability

The idea behind the measures of variability is to convey a notion of the "spread" of data in a given statistical sample S , technically referred to also as the dispersion of the data. As the realisation of this intention requires a well-defined concept of distance, the measures of variability are meaningful for data relating to metrically scaled 1-D variables X only. One can distinguish two kinds of such measures:

(i) simple 2-data-point measures, and

(ii) sophisticated n-data-point measures.

We begin with two examples belonging to the first category.

3.2.1 Range

For a raw data set $\{x_i\}_{i=1,\ldots,n}$ of n observed values for X, the dimensionful range R (metr) simply expresses the difference between the largest and the smallest value in this set, i.e.,

$$R := x_{(n)} - x_{(1)} \tag{3.10}$$

The basis of this measure is the ordered data set $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Alternatively, the range can be denoted by $R = Q_4 - Q_0$.

3.2.2 Interquartile range

In the same spirit as the range, the dimensionful interquartile range d_Q (metr) is defined as the difference between the third quantile and the first quantile of the relative frequency distribution for some X, i.e.,

$$d_Q := \tilde{x}_{0.75} - \tilde{x}_{0.25} \tag{3.11}$$

Alternatively, this is $d_Q = Q_3 - Q_1$.

3.2.3 Sample variance

The most frequently employed measure of variability in Statistics is the dimensionful n-data-point sample variance s^2 (metr), and the related sample standard deviation to be discussed below. Given a raw data set $\{x_i\}_{i=1,\ldots,n}$ for X, its spread is essentially quantified in terms of the sum of squared deviations of the n data points from their common mean $\bar{x}$. Due to the algebraic identity

$$(x_1 - \bar{x}) + \ldots + (x_n - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x}) = \left(\sum_{i=1}^{n} x_i\right) - n\bar{x} \equiv 0,$$

there are only $n-1$ degrees of freedom involved in this measure. The sample variance is thus defined by:

(i) From raw data set:

$$s^2 := \frac{1}{n-1}[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2] =: \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.12}$$

alternatively, by the shift theorem (using the algebraic identity $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1} nx_i^2 - \sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$):

$$s^2 = \frac{1}{n-1}[x_1^2 + \ldots + x_n^2 - n\bar{x}^2] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]. \tag{3.13}$$

(ii) From relative frequency distribution:

$$\begin{aligned} s^2 &:= \frac{n}{n-1}[(a_1 - \bar{x})^2 h_n(a_1) + \ldots + (a_k - \bar{x})^2 h_n(a_k)] \\ &=: \frac{n}{n-1}\sum_{j=1}^{k}(a_j - \bar{x})^2 h_n(a_j). \end{aligned} \tag{3.14}$$

Alternatively,

$$\begin{aligned} s^2 &= \frac{n}{n-1}[a_1^2 h_n(a_1) + \ldots + a_k^2 - \bar{x}^2] \\ &= \frac{n}{n-1}\left[\sum_{j=1}^{k} a_j^2 h_n(a_j) - \bar{x}^2\right]. \end{aligned} \tag{3.15}$$

Remark 3.2.1 (i) We point out that the alternative formulae for a sample variance provided here prove computationally more efficient.

(ii) For binned data, when one selects the midpoint of each class interval K_j to represent the a_j (given the raw data set is no longer accessible), a correction of Eqs. (3.14) and (3.15) by an additional term $\frac{1}{12}\frac{n}{n-1}\sum_{j=1}^{k} b_j^2 h_j$ becomes necessary, assuming uniformly distributed data within each class intervals K_j of width b_j.

3.2.4 Sample standard deviation

For ease of handling dimensions associated with a metrically scaled 1-D variable X, one defines the dimensionful sample standard deviation s (metr) simply as the positive square root of the sample variance, i.e.,

$$s := +\sqrt{s^2}, \tag{3.16}$$

such that a measure for the spread of data results which shares the dimension of X and its arithmetical mean $\overline{x}$.

3.2.5 Sample coefficient of variation

For ratio scaled 1-D variables X, a dimensionless relative measure of variability is the sample coefficient of variation v (metr: ratio), defined by

$$v := \frac{s}{\overline{x}}, \text{ if } \overline{x} > 0. \tag{3.17}$$

3.2.6 Standardisation

Data for metrically scaled 1-D X is amenable to the process of standardisation. By this is meant a transformation procedure $X \to Z$, which generates from data for a dimensionful X, with mean $\overline{x}$ and sample standard deviation $s_X > 0$, data for an equivalent dimensionless variable Z according to

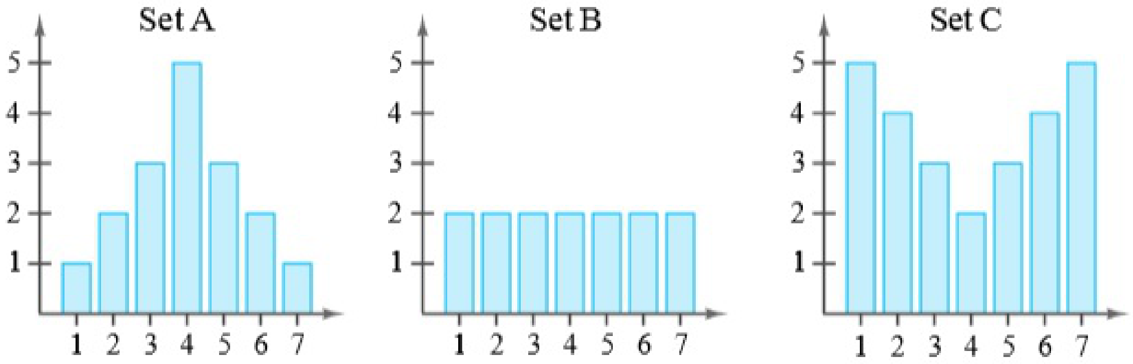
$$x_i \to z_i := \frac{x_i - \overline{x}}{s_X} \text{ for all } i = 1, \ldots, n \tag{3.18}$$

For the resultant Z-data, referred to as the z-scores of the original metrical X-data, this has the convenient consequence that the corresponding arithmetical mean and the sample variance amount to

$$\overline{z} = 0 \text{ and } s_Z^2 = 1$$

respectively.

Exercise 5 Consider the three sets of data represented by the bar graphs in Figure below.



1. Which set has the smallest standard deviation? Which has the largest?
2. Find the standard deviation of each set.

3.3 Measures of relative distortion

The third family of measures characterising relative frequency distributions of data $\{x_i\}_{i=1,\ldots,n}$ for metrically scaled 1-D variables X, having specific mean $\overline{x}$ and standard deviation s_X, relate to the issue of shape of a distribution. These measures take a Gaussian normal distribution with parameter values equal to mean $\overline{x}$ and s_X as a reference case. With respect to this reference distribution, one defines two kinds of dimensionless measures of relative distortion as described in the following.

3.3.1 Skewness

The skewness g_1 (metr) is a dimensionless measure to quantify the degree of relative distortion of a given frequency distribution in the horizontal direction. Its implementation in the software package EXCEL employs the definition

$$g_1 := \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right)^3 \quad \text{for } n > 2 \tag{3.19}$$

wherein the observed values $\{x_i\}_{i=1,\ldots,n}$ enter standardised according to Eq. (3.18). Note that $g_1 = 0$ for an exact Gaussian normal distribution.

Exercise 6 Data frames are the typical R representation of data sets. Here we create a data frame "hand" to become familiar with data frames.
Use the command `data.frame()` to create a data frame students with the following entries

Name	Degree	mat.nr
Leonie	Master	1111
Luka	Master	1112
Lea	Bachelor	1113
Leon	Bachelor	1114
Laura	Bachelor	1115
Luis	Bachelor	1116

1. Get an overview of students with the commands `names()`, `str()` and `summary()`.
2. Which command returns the fifth element of the vector 'mat.nr' ?
3. Check existence of the variable `Degree` by entering it into the R command line. Now copy students into the search path with the command `attach()`. Check again whether `Degree` is a known variable.
4. Define a new data frame named 'ba.students' which consists of all students with degree Bachelor (without using the command `data.frame()`). As all students in ba.students have degree Bachelor the variable `Degree` is not needed in ba.students.
5. Write the data frame students into the file 'studentsfile.txt'. Then read the data frame from this file into the new variable students2. If you used the right command, then students and students2 are identical. Check this using the command `all()`.
6. Write the data frame students into the fille 'studentsfile.txt'. Then read the data frame from this file into the new variable students2. If you used the right command, then students and students2 are identical. Check this using the command `all()`.

3.3.2 Excess kurtosis

The excess kurtosis g_2 (metr) is a dimensionless measure to quantify the degree of relative distortion of a given frequency distribution in the vertical direction. Its implementation in the softwarepackage EXCEL employs the definition

$$g_2 := \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \quad \text{for } n > 3 \tag{3.20}$$

wherein the observed values $\{x_i\}_{i=1,\ldots,n}$ enter standardised according to Eq. (3.18). Note that $g_2 = 0$ for an exact Gaussian normal distribution.

Exercise 7

1. Find the kurtosis of eruption duration in the data set `faithful`.
2. Find the kurtosis of eruption waiting period in `faithful`.

3.4 Measures of concentration

Finally, for data $\{x_i\}_{i=1,\ldots,n}$ relating to a ratio scaled 1-D variable X, which has a discrete spectrum of values $\{a_j\}_{j=1,\ldots,k}$ or was binned into k different categories $\{K_j\}_{j=1,\ldots,k}$ with respective midpoints a_j, two kinds of measures of concentration are commonplace in Statistics; one qualitative in nature, the other quantitative. Begin by defining the total sum for the data $\{x_i\}_{i=1,\ldots,n}$ by

$$S := \sum_{i=1}^{n} x_i = \sum_{j=1}^{k} a_j o_n(a_j) = n\bar{x} \tag{3.21}$$

where $(a_j; o_n(a_j))_{j=1,\ldots,k}$ is the absolute frequency distribution of the observed values (or categories) of X. Then the relative proportion that the value a_j (or the category K_j) takes in S is

$$\frac{a_j o_n(a_j)}{S} = \frac{a_j h_n(a_j)}{\bar{x}} \tag{3.22}$$

3.4.1 Lorenz curve

From the elements introduced in Eqs. (3.21) and (3.22), the US-American economist Max Otto Lorenz (1876-1959) constructed cumulative relative quantities which constitute the coordinates of a so-called Lorenz curve representing concentration in the distribution of the ratio scaled 1-D variable X. These coordinates are defined as follows:

- Horizontal axis:

$$k_i := \sum_{j=1}^{i} \frac{o_n(a_j)}{n} = \sum_{j=1}^{i} h_n(a_j) \quad (i = 1, \ldots, k), \tag{3.23}$$

- Vertical axis:

$$\ell_i := \sum_{j=1}^{i} \frac{a_j o_n(a_j)}{S} = \sum_{j=1}^{i} \frac{a_j h_n(a_j)}{\bar{x}} \quad (i = 1, \ldots, k), \tag{3.24}$$

The initial point on a Lorenz curve is generally the coordinate system's origin, $(k_0; l_0) = (0; 0)$, the final point is $(1; 1)$. As a reference to measure concentration in the distribution of X in qualitative terms, one defines a null concentration curve as the bisecting line linking $(0; 0)$ to $(1; 1)$. The Lorenz curve is interpreted as stating that a point on the curve with coordinates $(k_i; l_i)$ represents the fact that $k_i \times 100\%$ of the n statistical units take a share of $l_i \times 100\%$ in the total sum S for the ratio scaled 1-D variable X. Qualitatively, for given data $\{x_i\}_{i=1,\ldots,n}$, the concentration in the distribution of X is the stronger, the larger is the dip of the Lorenz curve relative to the null concentration curve. Note that in addition to the null concentration curve, one can define as a second reference a maximum concentration curve such that only the largest value a_k (or category K_k) in the spectrum of values of X takes the full share of 100% in the total sum S for $\{x_i\}_{i=1,\ldots,n}$.

3.4.2 Normalised Gini coefficient

The Italian statistician, demographer and sociologist Corrado Gini (1884-1965) devised a quantitative measure for concentration in the distribution of a ratio scaled 1-D variable X. The dimensionless normalised Gini coefficient G_+ (metr: ratio) can be interpreted geometrically as the ratio of areas

$$G_+ := \frac{\text{(area enclosed between Lorenz and null concentration curves)}}{\text{(area enclosed between maximum and null concentration curves)}} \tag{3.25}$$

Its related computational definition is given by

$$G_+ := \frac{n}{n-1}\left[\sum_{i=1}^{k}(k_{i-1} + k_i)\frac{a_i o_n(a_i)}{S} - 1\right]. \tag{3.26}$$

Due to normalisation, the range of values is $0 \leq G_+ \leq 1$. Thus, null concentration amounts to $G_+ = 0$, while maximum concentration amounts to $G_+ = 1$.

Exercise 8 After having downloaded the data ``Airpassengers'' (these data being the monthly totals of international airline passengers, from 1949 to 1960 and are thus relevant enough for a concentration analysis), compute the Gini index of the distribution and display the associated Lorenz curve.