Descriptive Statistics
Formulary

I. Distribution Analysis

***Absolute frequency*:**

$F(a_j) :=$ "Number of cases in which the attribute a_j occurs", $j = 1, 2,..., k$.

Relative frequency:

$$f(a_j) := \frac{1}{n} \cdot F(a_j), \quad j = 1, 2,..., k .$$

n : Number of observations.

***Empirical distribution function of an ungrouped data*:**

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le a_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad a_j < x \le a_{j+1} \\ 1 & \text{for} \quad x > a_k \end{cases}$$

***Properties of the empirical distribution function*:**

1.

$F(x) = P(X < x)$ (P : proportion)

2.

$0 \le F(x) \le 1$

3.

$\forall\, x_1, x_2 : x_1 < x_2 \;\Rightarrow\; F(x_1) \le F(x_2)$

4.

$P(x_1 \le X < x_2) = F(x_2) - F(x_1)$

5.

$F(x)$ is at least left-sided continuous and has at most a finite number of jump discontinuities.

6.

$x \to -\infty \;\Rightarrow\; F(x) \to 0$

$x \to +\infty \;\Rightarrow\; F(x) \to 1$

Empirical distribution function of grouped data:

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le m_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad m_j < x \le m_{j+1} \\ 1 & \text{for} \quad x > m_k \end{cases}$$

m_i : Midpoint of the class C_i, $i = 1, 2, ..., p$.

II. Measures of Location

Arithmetic average (mean) of an ungrouped data:

$$\mu := \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \textit{(Simple arithmetic average for population data)}$$

$$= \frac{1}{N}\sum_{j=1}^{k} a_j \cdot F\left(a_j\right) \qquad \textit{(Weighted arithmetic average for population data)}$$

N : Population size.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{(Simple arithmetic average for sample data)}$$

$$= \frac{1}{n}\sum_{j=1}^{k} a_{ij} \cdot F\left(a_j\right) \qquad \textit{(Weighted arithmetic average for sample data)}$$

n : sample size.

Arithmetic average (mean) of a grouped data:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{(Simple arithmetic average for sample data)}$$

$$= \frac{1}{n}\sum_{j=1}^{k} a_j \cdot F\left(a_j\right) \qquad \textit{(Weighted arithmetic average for sample data)}$$

n : Sample size.

$$\mu := \approx \frac{1}{N} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot F_i \qquad \textit{(Population mean)}$$

$$\bar{x} := \approx \frac{1}{n} \cdot \sum_{i=1}^{p} \bar{x}_i \cdot F_i \qquad \textit{(Sample mean)}$$

(In case the informations $\bar{x}_i$ are not available, they will be replaced by the interval *midpoints*:

$$m_i := \frac{b_i + B_i}{2}$$

b_i, $i = 1, 2, ..., p$: lower bound of the class C_i

B_i, $i = 1, 2, ..., p$: upper bound of the class C_i.)

Estimations:

$$\frac{1}{N} \cdot \sum_{i=1}^{p} b_i \cdot F_i \le \mu \le \frac{1}{N} \cdot \sum_{i=1}^{p} B_i \cdot F_i$$

$$\frac{1}{n} \cdot \sum_{i=1}^{p} b_i \cdot F_i \le \bar{x} \le \frac{1}{n} \cdot \sum_{i=1}^{p} B_i \cdot F_i$$

Median of an ungrouped data:

$$Me := \begin{cases} x_{\left\lceil \frac{n+1}{2} \right\rceil} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left\lceil \frac{n}{2} \right\rceil} + x_{\left\lceil \frac{n}{2}+1 \right\rceil} \right) & \text{if } n \text{ is even} \end{cases}$$

Median of a grouped data:

Let $C_1, C_2, ..., C_p$ be given classes with the relative frequencies $f_1, f_2, ..., f_p$.

The *Median* of lies in the class C_i if

$$\sum_{j=1}^{i-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{i} f_j \ge 0.5$$

and will be defined by

$$Me := b_i + \frac{0.5 - \sum_{j=1}^{i-1} f_j}{f_i} \cdot w_i$$

α-Quantile of an ungrouped data:

Given a *ranked* data set, the, α-Quantile ($0 < \alpha < 1$) is defined by:

$$\tilde{x}_\alpha := \begin{cases} x_{[k]} & \text{if } n \cdot \alpha \text{ is not integer} \\ & (k \text{ is then the following integer}) \\ \frac{1}{2}\left(x_{[k]} + x_{[k+1]} \right) & \text{if } n \cdot \alpha \text{ is integer} \\ & (k = n \cdot \alpha) \end{cases}$$

α-Quantile of a grouped data:
Let $C_1, C_2, ..., C_p$ be given classes with the relative frequencies $f_1, f_2, ..., f_p$.
The *α-Quantile* $(0 < \alpha < 1)$ of the data set lies in the class C_i if

$$\sum_{j=1}^{i-1} f_j < \alpha \quad \text{and} \quad \sum_{j=1}^{i} f_j \geq \alpha$$

and will be defined by

$$\bar{x}_\alpha := b_i + \frac{\alpha - \sum_{j=1}^{i-1} f_j}{f_i} \cdot w_i$$

Mode of an ungrouped data:
The value of the item which occurs most frequently is called the *mode* of it is unique.
It will be denoted by *Mo*.

Mode of a grouped data:
The *mode* of a grouped data set lies in the class with the highest frequency. It will be defined by

$$Mo :\approx b_i + \frac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}} \cdot w_i$$

Geometric mean of an ungrouped data

$$\bar{x}_g := \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \qquad x_i > 0, \quad i = 1, 2, ..., n$$

III. Measures of Distribution

Range:

1. *Ungrouped*

   $$R := x_{\max} - x_{\min}$$

2. *Grouped*

   $$R \approx B_p - b_1$$

Interquartile range:

$$R_Q := x_{0.75} - x_{0.25}$$

Variance (ungrouped):

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \quad \textit{(population variance)}$$

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \textit{(sample variance)}$$

Standard deviation (ungrouped):

$$\sigma := \sqrt{\sigma^2}, \quad \sigma > 0 \quad \textit{(population standard deviation)}$$

$$s := \sqrt{s^2}, \quad s > 0 \quad \textit{(population standard deviation)}$$

Variance (grouped):

$$\sigma^2 := \approx \frac{1}{N}\sum_{i=1}^{p}\left(\bar{x}_i - \mu\right)^2 \cdot F_i \qquad \textit{(population)}$$

$$s^2 := \approx \frac{1}{n-1}\sum_{i=1}^{p}\left(\bar{x}_i - \bar{x}\right)^2 \cdot F_i \qquad \textit{(sample)}$$

In case the information $\bar{x}_i$ is not available, it will be replaced by m_i.

Standard deviation (grouped):

$$\sigma := \sqrt{\sigma^2}, \quad \sigma > 0 \quad \textit{(population)}$$

$$s := \sqrt{s^2}, \quad s > 0 \quad \textit{(sample)}$$

Chebyshev Theorem

For any number $k > 1$, at least $\left(1 - 1/k^2\right)$ of the data values lie within k standard deviations of the mean.

At least 50% of the observations in a data set lie in the interval

$$[\mu - \sigma, \quad \mu + \sigma], \qquad \textit{(population)}$$

$$\left[\bar{x} - s, \quad \bar{x} + s\right] \qquad \textit{(sample)}$$

Coefficient of Variation:

$$v := \frac{\sigma}{\mu} \qquad \textit{(population)}$$

$$v := \frac{s}{\bar{x}} \qquad \textit{(sample)}$$

A Rule of Thumb:
The arithmetic average of a data set is to be considered as representative only if its coefficient of variation is less than 0.5 (or 50%).

IV. Correlation and Regression

***Spearman's Rank Coefficient of Correlation*)**

$$\rho := 1 - \frac{6 \cdot \sum_{i=1}^{n} \left(R_i - R_i' \right)^2}{(n-1) \cdot n \cdot (n+1)} :$$

Here are:

$R_i, i = 1, 2, ..., n$: ranks of the characteristic X

$R_i', \ i = 1, 2, ..., n$: ranks of the characteristic Y,

Karl Pearson's Coefficient of Correlation:
Let X and Y be two variates.

$$r := \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \cdot \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}}, \qquad -1 \le r \le +1 .$$

Simple Linear Regression:
The coefficients of a simple linear regression function

$$y^* = a_0 + a_1 x$$

can be obtained by solving the following system of linear equations:

$$n \cdot a_0 + a_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i$$

Coefficients of Correlation and Determination:

1. The *coefficient of correlation of a simple linear regression function* is defined by

$$r := \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right) \left(n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}},$$

$$-1 \le r \le +1$$

2. r^2 $(0 \le r^2 \le 1)$ is called *coefficient of determination.*

Coefficient of Determination

The *coefficient of determination* in general is defined as:

$$r^2 := \frac{SSR}{SST}$$

where:

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$: sum of square due to regression

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$: total sum of squares

(Last revised: 22.10.2017)