

To cite this / Pentru citare:

Petrisor AI (2009), Statistics and Data Analysis. Unpublished course notes on the theory and methods, available at http://environmetrics.ro/Studenti/Sinteza_stat_ERASMUS.pdf

Notes on an Environmental Analysis and Environmental Impact Assessment Course

Alexandru-Ionut PETRISOR, PhD (Ecology), PhD (Geography)

Note: The text represents a synthesis of core methods taught in the courses *Urban Statistics and Environmental Analysis* and *Data Analysis Studio* designed for ERASMUS students. For any other uses, please request the permission of the author.

General concepts of statistics

The purpose of **statistics** is to study a set of observations on some objects of the same nature called **statistical units**, displaying **variable characteristics** (simply **variables**) susceptible to be *classed*, *ordered* or *measured*. The set is called **statistical series** or **string**.

There are two types of sets:

- **Populations** are sets of objects, individuals, phenomena, events, idea, opinions, numbers etc. focusing the interest of researchers. They are large (mostly infinite) and their exhaustive study is impossible or uneconomical.
- **Samples** are subsets of the *populations* drawn to obtain information on populations.

Studies made on *populations* by **descriptive statistics** produce certain results, while those made on *samples* by **inductive/inferential statistics** lead to uncertain results. The scientific expression of uncertainty is given in *inductive statistics* by statistical inference. **Statistical inference** represents the extrapolation of judgments from samples drawn through specific statistical-mathematical procedures to populations.

The attempt to explain one or more scientific observations is called **scientific hypothesis**. These hypotheses need to be sustained by data (experiments, observations) and statistics. **Statistical hypotheses** are statements concerning one or more *populations* made to check *scientific hypotheses*. A *scientific hypothesis* consists of a **null hypothesis** ("there are no differences") and a **alternative hypothesis** contradicting it and corresponding to the *scientific hypothesis*. After applying a **statistical test**, the *null hypothesis* is rejected when significant differences are detected or not, otherwise. **Significant differences** are too large, compared with a chosen **level of significance**, to be attributed to random fluctuations, but are due to a significant reason, *i.e.*, the *scientific hypothesis*.

Classification of scales: if A and B are two statistical units, and x some variable with the characteristics x_A and x_B ,

- **The nominal scale** makes a distinction: $x_A = x_B$ or $x_A \neq x_B$.
- **The ordinal scale** establishes an order. If $x_A \neq x_B$, then either $x_A > x_B$ or $x_A < x_B$.
- **The equal interval scale** provides a measure of the difference. If $x_A > x_B$, then A is greater than B with $x_A - x_B$.
- **The equal ratio scale** has in addition a absolute zero, implicitly a measure of the ratio of two values: A is x_A / x_B times greater than B .

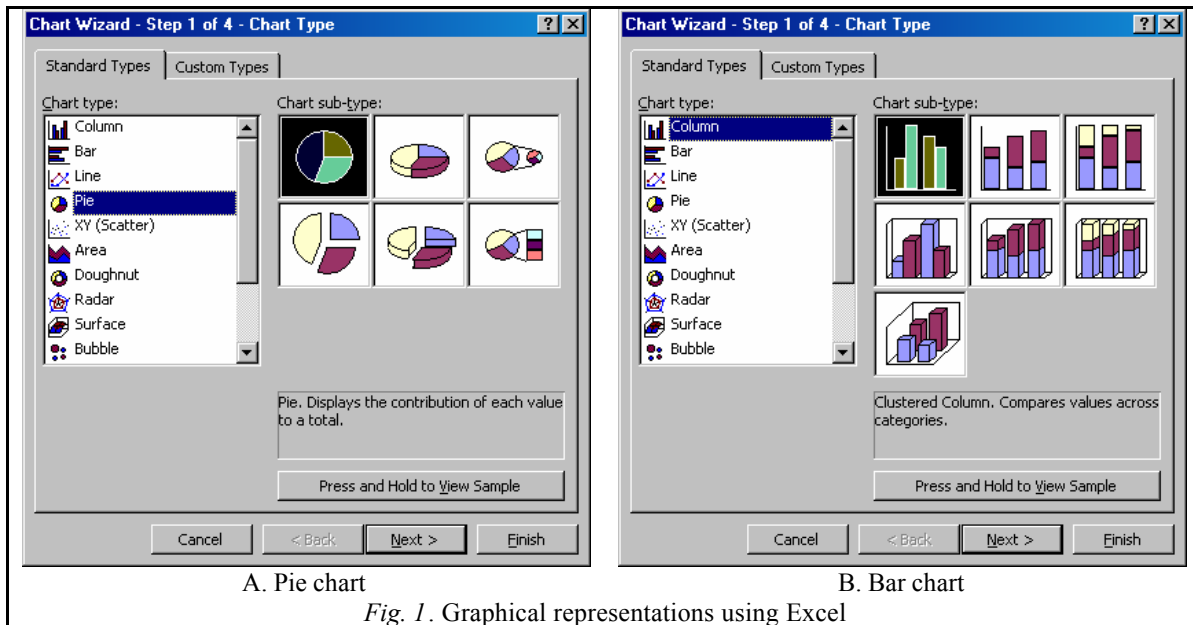
Classification of variables

- Qualitative – nominal scale, including binary variables (yes/no); they have variants that are classed
- Ranks – ordinal scale; they have values that are ordered
- Measures or dimensions – interval or ratio scales; they have values that are measured

Computing and interpreting results of the graphical synthesis of data

For the assignment, you will receive a data series: $X_1, X_2, \dots, X_i, \dots, X_n$. Using this data, you must be able to synthesize them graphically and numerically and interpret the results.

Represent your series as a pie chart (Fig. 1A). Write your data in an Excel column, pick the graph sign (📊), and then choose “pie”. You can do it by hand by summing $X_1, X_2, \dots, X_i, \dots, X_n$. The sum receives an angle 360 degrees. X_1 receives $X_1 * 360 / \text{SUM}(X_1, X_2, \dots, X_i, \dots, X_n)$ degrees etc.



Represent your series as a bar chart (Fig. 1B). Write your data in an Excel column, pick the graph sign (📊), and then choose “bar”. You can do it by hand by summing $X_1, X_2, \dots, X_i, \dots, X_n$. The sum receives an angle 360 degrees. X_1 receives $X_1 * 360 / \text{SUM}(X_1, X_2, \dots, X_i, \dots, X_n)$ degrees etc.

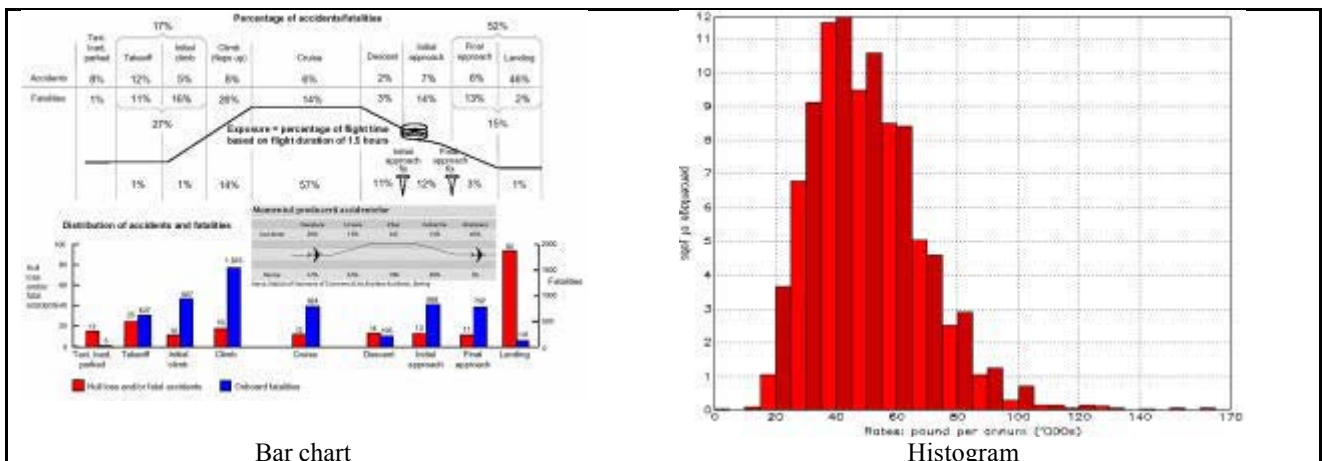
Other graphical representations of data

The following are not required for your assignment. However, they could be useful to know.

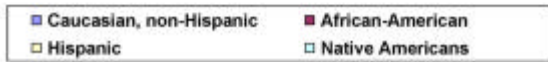
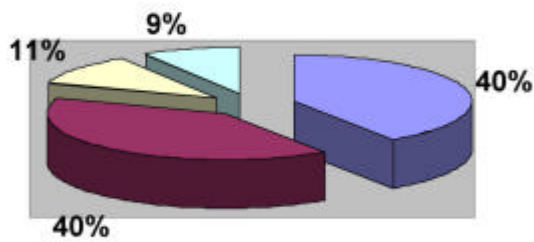
The histogram: unlike the previous ones, data are grouped. A convention requires each bar of a histogram to join the previous one in continuation (unlike the bars of a bar diagram). Histograms are drawn manually or using specialized software. We group data as an essential principle of statistics is to give up information in order to increase the relevance. More concretely, we want to see behind the histogram a theoretical curve, explained below.

Frequency polygons: join upper ends of bars within a bar chart. The line could also suggest some theoretical curve.

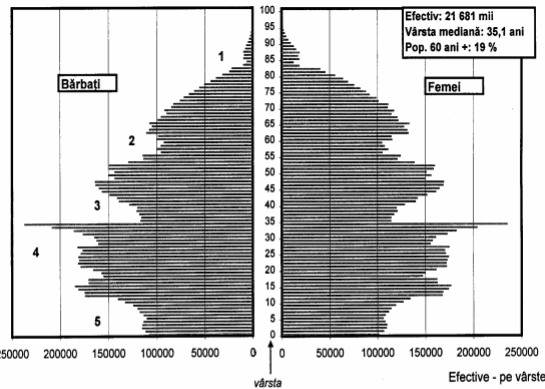
Population pyramids: draw two histograms with the age groups for each gender. Rotate them such that their bases are joined. These pyramids suggest the sizes of different population cohorts as a result of historical or environmental pressures.



Ethnic Background of U.S. Homeless Population

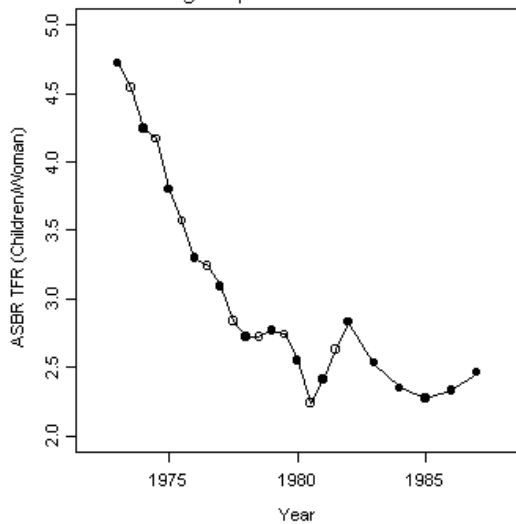


Pie chart

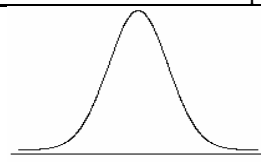


Population pyramid

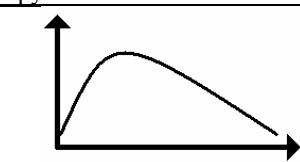
Figure 1
China Age-Specific Birth Rate TFRs



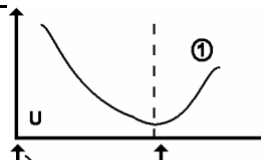
Frequency polygon



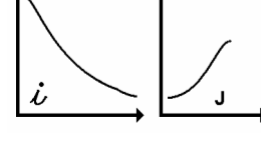
Gauss Bell



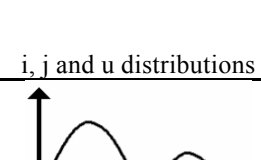
Unimodal, asymmetric to the left



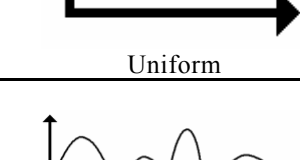
Unimodal, asymmetric to the right



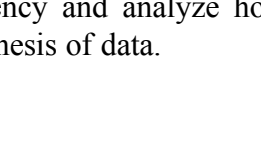
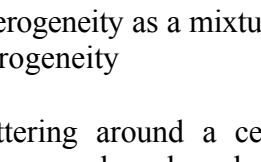
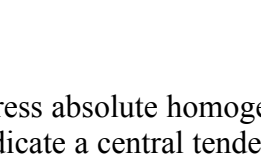
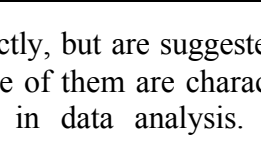
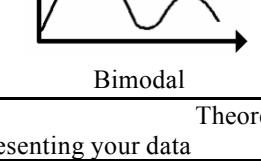
Uniform



Bimodal



Multimodal



Theoretical distributions

Fig. 2. Representing your data

Theoretical curves: they cannot be drawn directly, but are suggested by histograms and frequency polygons when the sample size increases. Some of them are characteristic to different phenomena and suggest what methods should be used in data analysis. The main representations are summarized in *Fig. 2* below.

Interpreting the theoretical curves

- Distributions concentrated in one point express absolute homogeneity
- Symmetrical distributions are the best to indicate a central tendency
- Bi – or multimodal distributions indicate heterogeneity as a mixture of homogenous distributions
- Uniform distributions express absolute heterogeneity

In statistics, variability is understood as scattering around a central tendency. To understand variability, we must identify the central tendency and analyze how scattered is the distribution around it. This is done using the numerical synthesis of data.

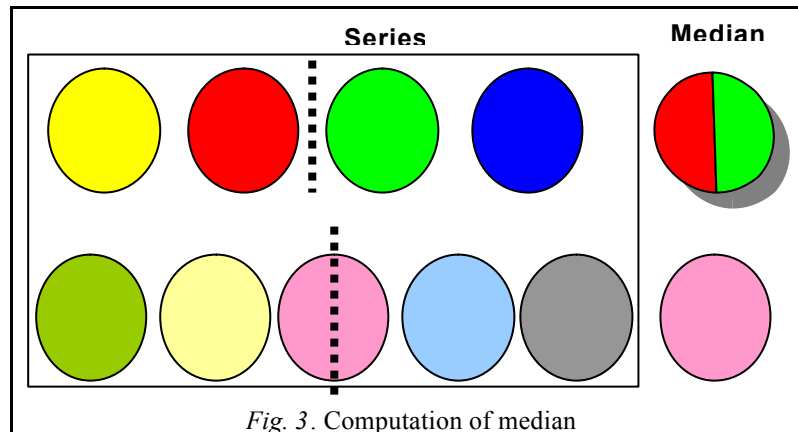
Computing and interpreting results of the numerical synthesis of data

1. Identify the central tendency

Compute M, the average

$$M = \frac{\sum_{i=1}^n X_i}{n}$$

Compute Me, the median: this is the value that splits a series into two. Arrange the values in increasing order and either choose the middle one or, if there are two, compute their average (Fig. 3).



Compute Mo, the mode and describe the series based on it: this is a value with maximum local frequency (is repeated more times than the values around in a series arranged in increasing order). A series can have one mode (this is called unimodal), two (bimodal) or more (multimodal). If all values have the same frequency (appear the same number of times), the distribution is uniform.

2. Look at scattering (variability)

Compute A, the range: this is the difference between the maximum and the minimum value.

Compute S^2 , the variance: subtract from each value the average, square the result (multiply it with itself), add all these results and divide the sum by the number of values. Please note that Excel divides the sum by the number of values minus one, so if you use Excel, you must adjust the result (multiply it with the number of values minus one and divide it by the sum by the number of values).

$$S^2 = \frac{\sum_{i=1}^n (X_i - M)^2}{n}$$

Compute S, the standard deviation: take the square root of the variance

$$S = \sqrt{S^2}$$

Compute CV, the coefficient of variation: divide the standard deviation by the average and multiply with 100 (express it as a percentage). Also, you must classify your series: if $CV < 10\%$, your series is homogenous; if $10\% < CV < 20\%$, the series is relatively homogenous; if $20\% < CV < 30\%$, the series is relatively heterogeneous; if $CV > 30\%$, the series is heterogeneous.

$$CV = 100 \times \frac{S}{M}$$