To cite this / Pentru citare: Petrisor AI (2012), *Statistical analysis methods used in the research of territorial phenomena. Course outline*, available at http://www.environmetrics.ro/Studenti/Sinteza_SDU_Stat_EN.pdf

Statistical analysis methods used in the research of territorial phenomena

Alexandru-Ionuț PETRIŞOR, PhD (Ecology), PhD (Geography), Habil. (Urban Planning)

INFORMATION ON THE COURSE

Program of studies	Faculty of Urban Planning, Doctoral School of Urban Planning
Type of course	Required
Level of course	Doctoral
Number of ECTS credits	3
Hours / week	2c+1s
Competences to be developed	1) Understanding main concepts of statistics and data analysis 2) Knowledge of the statistical methods (including computer-assisted ones) 3) Interpretation of final results of analyses from the standpoint of their relevance to urban planning 4) Correct use of the technical jargon and specific concepts
Objectives	The course aims to familiarize students with the main concepts and methods used in the statistical analysis of research data, focusing on the study of territorial phenomena and stressing out the interpretation of results in urban planning terms.
Teaching methods	Lectures, PowerPoint presentations, computer demonstration (ArcView, Excel), discussions
Evaluation	Attendance of at least 75% results in the accumulation of 3 credits.
Bibliography	1) Petrişor A.-I. (2011), *Systemic theory applied to ecology, geography and spatial planning*, Lambert Academic Publishing GmbH & Co. KG, Saarbrücken, Germany, ISBN 978-3-8465-0260-0, 172 pp.

Course notes[*]

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.

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

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

[*] Based mostly on: Dragomirescu L. (1998), *Biostatistică pentru începători*, Editura Constelații, Bucharest, 220 pp.

Descriptive statistics

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, ..., X_i, ..., X_n. The sum receives an angle 360 degrees. X_1 receives X_1 * 360 / SUM (X_1, X_2, ..., X_i, ..., X_n) degrees etc.

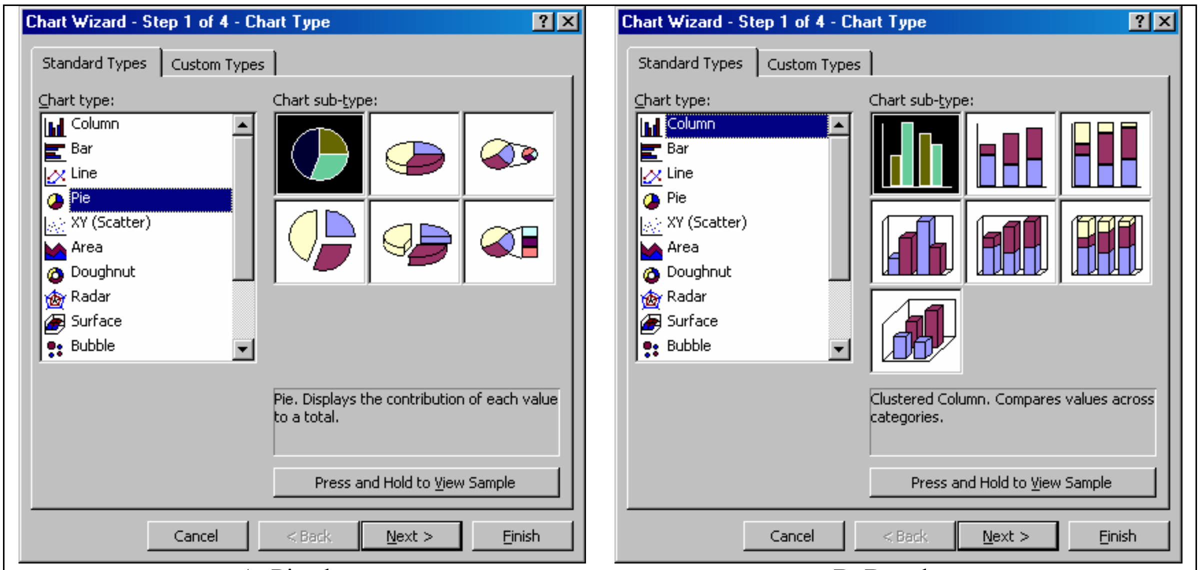


Chart Wizard - Step 1 of 4 - Chart Type

Standard Types | Custom Types

Chart type: Column, Bar, Line, **Pie** (selected), XY (Scatter), Area, Doughnut, Radar, Surface, Bubble

Chart sub-type:

Pie. Displays the contribution of each value to a total.

Press and Hold to View Sample

Cancel | < Back | Next > | Finish

Chart Wizard - Step 1 of 4 - Chart Type

Standard Types | Custom Types

Chart type: **Column** (selected), Bar, Line, Pie, XY (Scatter), Area, Doughnut, Radar, Surface, Bubble

Chart sub-type:

Clustered Column. Compares values across categories.

Press and Hold to View Sample

Cancel | < Back | Next > | Finish

A. Pie chart B. Bar chart

Fig. 1. Graphical representations using Excel

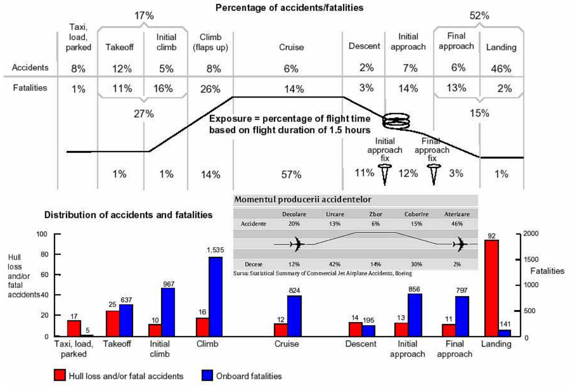
Represent your series as a bar chart (*Fig. 1B*). Write your data in an Excel column, pick the graph sign (📊), and then choose "bar". You can do it by hand by summing X_1, X_2, ..., X_i, ..., X_n. The sum receives an angle 360 degrees. X_1 receives X_1 * 360 / SUM (X_1, X_2, ..., X_i, ..., X_n) degrees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

Other graphical representations of data

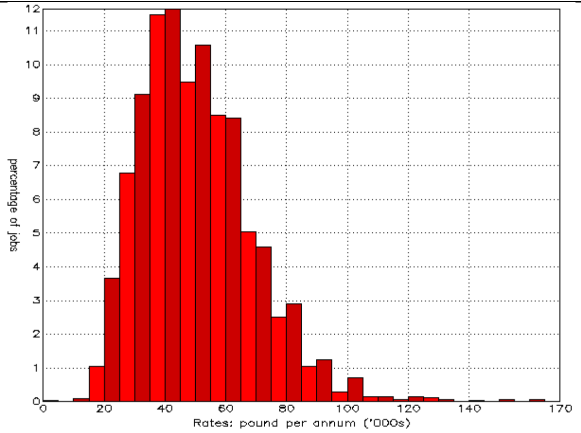
The histogram: unlike the previous ones, data are grouped. A convention requires each bar of a histogram to join the previous one in continuation (unlike the bars of a bar diagram). Histograms are drawn manually or using specialized software. We group data as an essential principle of statistics is to 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.

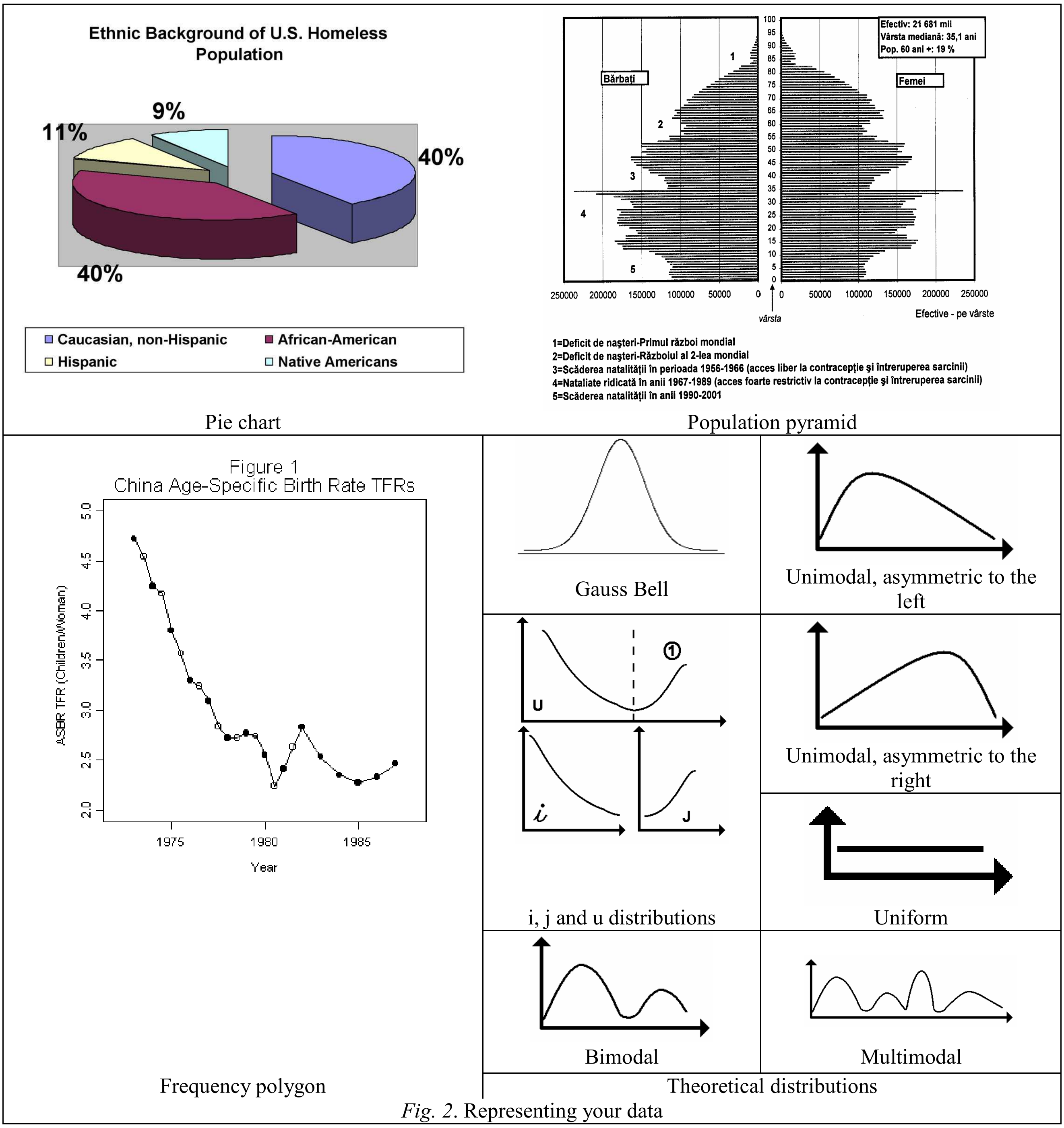
Population pyramids: draw two histograms with the age groups for each gender. Rotate them such that their bases are joined. These pyramids suggest the sizes of different population cohorts as a result of historical or environmental pressures.



Bar chart



Histogram



Pie chart

Population pyramid

Gauss Bell

Unimodal, asymmetric to the left

Unimodal, asymmetric to the right

i, j and u distributions

Uniform

Bimodal

Multimodal

Frequency polygon

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 the 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*).

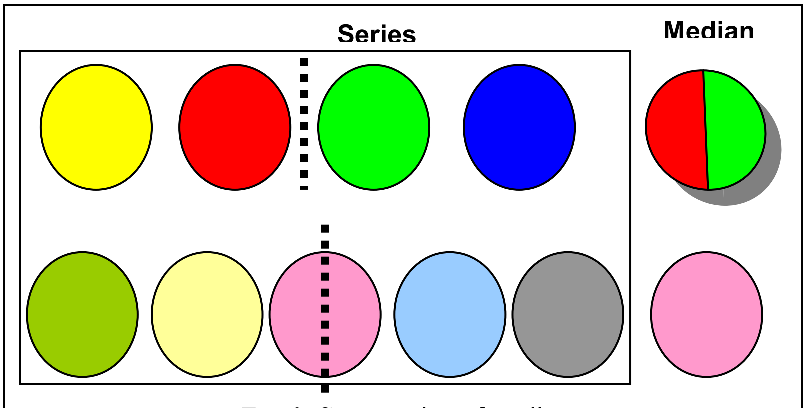


Fig. 3. Computation of median

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.

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

Inferential statistics

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.

Inferential statistics is a selective research, based on sampling, aiming to obtain most information on populations with a minimum effort – improving representativeness of samples, and includes the sampling theory, the estimation theory, testing statistical hypotheses, and experimental design.

Sampling theory

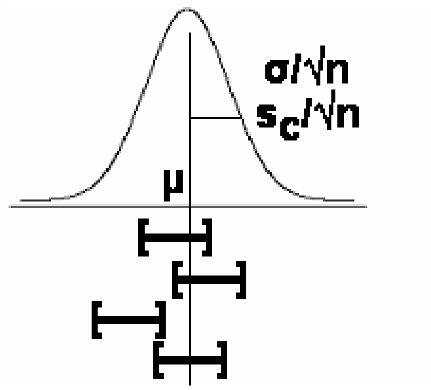
(a) Sampling selection

- **No return** – once included in a sample, a subject cannot be included in any other
- **Return** – a subject can be included in more samples

(b) Types of samples

- **Simple random sample** – selected using random numbers. They appear in random processes, such as radioactive decay. To select a sample, refer to random number tables. Pseudo-random numbers are generated by computers. Numbers provided by humans are the worst solution, as human mind cannot produce truly random results.
- **Systematic sample** – select k-th element of a frame – e.g., the a Y-th person on page Z, Z + 10, Z + 20 etc. in the phone book, Y and Z are random
- **Stratified sample** – simple random sample selected from each layer of a population
- **Cluster sample** – similar to the stratified one, but select clusters (groups) that form naturally instead of elements

Estimation theory



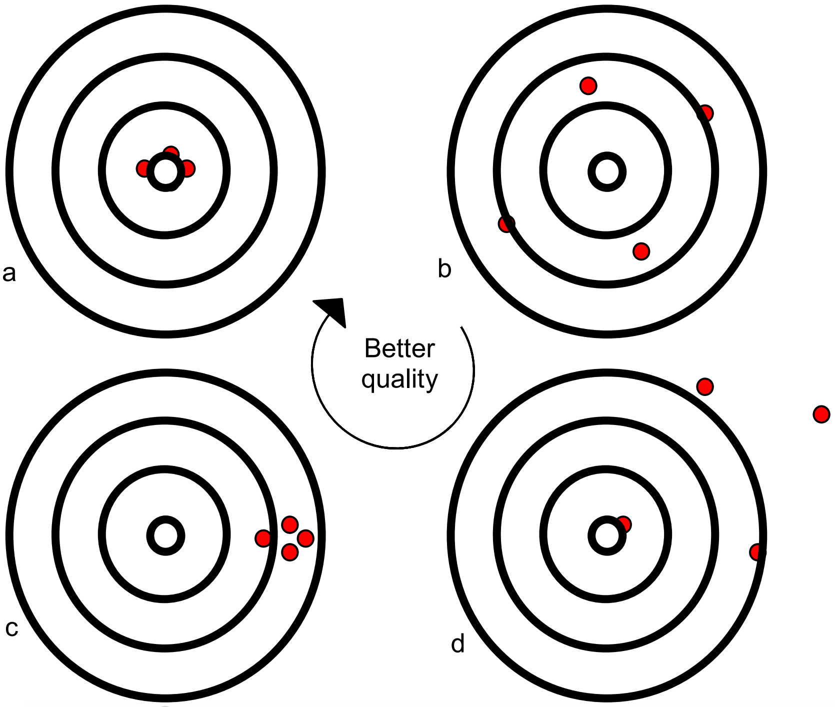
Properties of an estimator

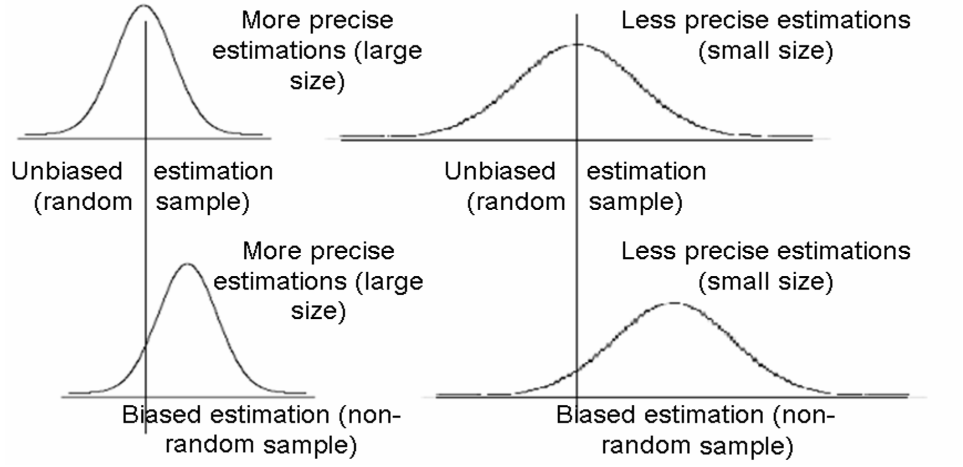
An estimator is **unbiased** if its mean coincides with the estimated parameter; otherwise it is called *biased*. In practice, a „correct" estimation is ensured by using an unbiased estimator and selecting a random sample.

An estimator is **more efficient** than another if it has a lower standard. The one with the lowest standard error is called **efficient estimator**.

Random sample	and unbiased estimator	"correct" estimation
Plus: large sample	and more efficient estimator	more precise estimation
Particularly: fixed size sample	and the most efficient estimator	the most precise estimation

A good estimator is like a good marksman!
A good marksman first checks if the weapon is "biased". Then he corrects the target to compensate it. Two unbiased marksmen differ by precision (c and d).
By analogy, an estimator must first be unbiased, and then efficient.





Testing statistical hypotheses

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*.

Steps in applying a statistical test:

1. Phrase the problem for which a decision must be taken
2. Identify the variable (type, scale), sample (selection, size), distribution of variable in population (Normal, transformable)
3. Determine the sample distribution of the **test statistic**.
4. Phrase hypotheses and determine the type of test (one-sided etc.).
5. Choose the significance level for rejecting the null hypothesis (**decision rule**)
6. Perform computation, obtain the decision by computing the test statistic or probability of rejection

Types of statistical tests:
1. Classification as binary tests
- two-sided (any difference)
- one-sided (left or right)
2. Classification based on how quantitative variables are analyzed
- As parameters – parametric tests
- As ranks or qualitative variables – non-parametric tests
3. Based on the hypotheses tested
- Conformity – compare a sample with a population based on a particular indicator
- Goodness of fit – compare a sample with a population based on maximum possible information
- Equality (homogeneity, comparison) – compare two or more populations using as many samples based on an indicator
- Independence versus association – study the association of characters, causal relationships, compare two methods etc.

In testing hypotheses, two types of errors correspond to two types of risks:

	H_0 accepted	H_0 rejected
H_0 true	Correct decision (probability: 1 – α)	1st kind error (α risk, "boozer's risk", furnisher's risk)
H_0 false	2nd kind error (β risk, "one-eyed man's risk", beneficiary's risk)	Correct decision (probability: 1–β)

Significance
- $\alpha < 1/1.000$ – highly significant
- $\alpha < 1/100$ – very significant
- $\alpha < 5/100$ – significant
- $\alpha \geq 5/100$ – not significant
- $10/100 < \alpha < 5/100$ – uncertainty

In ecology, frequently α = 10/100, even 20/100 is significant.
Relationship between α and β: α ↑, β ↓
p value: probability to obtain due to sampling fluctuations a value of the statistical test equal to the one obtained or even more extreme if H0 is true – measure of uncertainty due to the study based on samples

Exploratory statistics. **Exploratory statistics** (also called **exploratory data analysis**) starts with experimental data, identifying by simple arithmetic computations and graphic representations some structures within, offering partial descriptions, finding what is "behind“ , testing hypotheses in a reduced number of cases and with serious extrapolation caveats. The instrument allows for identifying models starting from samples, allowing the specialist to test the models by other models when noticing that the structures and patterns repeat.

Parametric vs. nonparametric statistics. There are **parametric** and **nonparametric** statistical methods. Parametric methods are more powerful, but less robust; those nonparametric are less powerful, but more robust. **Robustness** refers to the validity of results when the application conditions are not met. Most often conditions refer to the Normal distribution. Normality tests are used to look at a distribution and choose the appropriate method. Regardless of the field, few researchers use nonparametric methods, even for non-Normally distributed data.

Correlation. The link between two quantitative variables is called ***correlation***, and between two qualitative variables, ***association***. Causality determines the classification of relationships in ***false (apparent)*** (exist in data, are not sustained by phenomenology and could be ***artifactual (fictive)*** or ***indirect***) and ***causal (etiological)***; the *artifactual* are due to sampling fluctuations, and those *indirect* due to confounding. Causality is analyzed based on: time sequence, strength of association, specificity, diverse analogies, existence of a dose-response relationship, replication of results, biological plausibility, considering alternative explanations, evolution at and after the cease of exposure, and correlation with other studies.

Phenomenological correlation (scientific reasons) present	→	Data correlation (if the experimental design is correct) present
Data correlation present	→	Phenomenological correlation possible (not mandatory – see false correlations)
No data correlation	→	No phenomenological correlation

Factorial analysis and PCA. There are exploratory and confirmatory factor analyses. The first aims to identify the nature of patterns influencing a set of response variables, and the second tests whether a set of patterns influences in a pre-established manner the answers. In all cases the model is similar to regression, but dependent variables are not measured and occasionally unknown. Therefore, in a model like **Y(j)=f(Xj)**, factor analysis focuses almost exclusively on **Xi**. Factor analysis aims to identify the factors influencing **Xi**, analyzing their correlations according to the idea that the correlation of variables is due to some common factors **Fk**, and for uncorrelated variables, assigned to unique factors **Ei**. The principle of factor analysis is similar to the method of **Principal Component Analysis** (PCA), but the later aims for data reduction, *i.e.*, obtain a set of variables explaining most of the observed variability. There are two differences between the two methods: (1) while exploratory factor analysis assumes that the measured variables are influenced by factors, PCA assumes that the components are intrinsic to measured variables and (2) exploratory factor analysis assumes that the variation can be decomposed in the one explained by common factors and the other explained by unique factors, while the principal components are linear combinations of the measured variables, containing both common and unique variation.

3. Experimental design

Steps:

1. Phrase statistical hypotheses
2. Determine experimental conditions (independent variables), external ones (noise) to be controlled
3. Specify sample size and define population
4. Specify how subjects are assigned to samples (experimental conditions)
5. Determine measurements (dependent variables) and statistical analysis methods

Determining the sample size depends on the standard error and π (α). In estimation – if the variability of character in population (σ) is known, define a certain standard error ($\sigma/\sqrt{n}$) and determine sample size, n, based on it; if the variability is unknown, select a sample and estimate variability using corrected variance (sc2) and use the same algorithm. If testing statistical hypotheses, size can be determined in several situations choosing a certain confidence level and knowing or estimating the variability of character in population: z and t tests, comparing proportions and ANOVA. Financial constraints: first determine size needed for a certain standard error or power, then estimate possible size based on cost and determine the standard error or power than could be obtained based upon it.

Note. This document is a course outline produced for students and professors of "Ion Mincu" University of Architecture and Urban Planning. For any other uses, the author's consent must be obtained in advance.