Name of the course
"Statistical methods and data analysis"

Teacher
Stefano Siboni

Aim
The aim of the course is to illustrate the basic mathematical tools for the analysis and modeling of experimental data, particularly concering the main statistical methods. The idea is not simply to describe and discuss the "cooking recipes" for the statistical treatment of data, but also to introduce and analyze in a critical way the statistical models implemented by the computation procedures, stressing their limitations and the actual interpretation of the results.

Period of lectures
March – May 2021

Duration
36 hours (4.5 ECTS)

Assessment modality
Participants are required to pass a written examination consisting in the application of some statistical techniques (calculation of confidence intervals, hypothesis testing, calculation of correlation coefficients, ANOVA, linear regressions by chi-square or least squares fitting --- regression straight lines) presented in the course.

Tentative programme of the course

(1) Introduction to experimental measurements
Errors of measurement.
Systematic and random errors.
Absolute and relative errors.
Accuracy and precision of a measurement.
Postulate of the statistical population;
the result of a measurement as an outcome of a random variable.

(2) Random variables
Discrete and continuous random variables.
Cumulative frequency theoretical distribution of a random variable.
Frequency theoretical distribution (probability distribution) of a random variable, in the continuous and in the discrete case.
Mean, variance, skewness of a probability distribution.
Tchebyshev's theorem.

Examples of discrete probability distributions:
Bernoulli's, binomial, Poisson's.
Examples of continuous probability distributions:
uniform, normal or Gaussian, chi-square with n degrees of freedom (X^2), Student's with n degrees of freedom (t), Fisher's with n_1 and n_2 degrees of freedom (F).
Central Limit Theorem.
Multivariate random variables and related probability distributions;
(stochastically) dependent and independent random variables;
covariance matrix;
correlation matrix;
uncorrelated random variables;

stochastic independence as a sufficient but not necessary condition to the lack of correlation;
multivariate normal random variables,
stochastic independence as a condition equivalent to the lack of correlation for multivariate normal random variables.

(3) Functions of random variables
Functions of random variables and indirect measurements.
Linear combination of random variables:
calculation of mean and variance, case of uncorrelated random variables;
normal random variables, calculation of the joint probability distribution (the linear combination is again a normal random variable).

Quadratic forms of standard normal random variables:
Craig's theorem on the stochastic independence of positive semidefinite quadratic forms of independent standard normals;
theorem for the characterization of quadratic forms of independent standard normals which obey a chi-square probability distribution.
Illustrative examples.

Error propagation in the indirect measurements:
law of propagation of random errors in the indirect measurements (Gauss law)
example of application of Gauss law;
method of the logarithmic differential,
principle of equivalent effects,
example of application of the logarithmic differential;
error propagation in the solution of a linear set of algebraic equations, condition number,
example of application to the solution of a system of 2 equations in 2 variables;
probability distribution of a function of random variables by Monte-Carlo methods.

(4) Sample theory. Sample estimates of mean and variance
Estimate of the parameters of the statistical population (mean and variance).
Confidence intervals for the mean and the variance in the case of a normal population.
Example of calculation of the CI for mean and variance of a normal population.
Remark on the probabilistic meaning of the confidence intervals.
Large samples of an arbitrary statistical population:
approximate CI for the mean,
approximate CI for the variance,
example of calculation of the CI for the mean of a non-normal population by using a large sample.

(5) Hypothesis testing
Null hypothesis, type I errors, type II errors;
Tests based on the rejection of the null hypothesis.
Example illustrating the general concepts of hypothesis testing.
Summary of the main hypothesis tests and their typologies (parametric and non parametric).
X^2-test to fit a sample to a given probability distribution.
Example of X^2-test applied to a uniform distribution in the unit interval [0,1].
Example of X^2-test applied to a normal distribution of unknown parameters.
Kolmogorov-Smirnov test to fit a sample to a given continuous probability distribution.
Example of KS test applied to a uniform distribution in the unit interval [0,1].
Remark on the calculation of the KS test statistics.
Example of KS test applied to a standard normal distribution.
T-test to check if the mean of a normal population is equal to, smaller than or greater than an assigned value, respectively.
Example of application of the t-test on the mean of a normal population,

relation between the t-test on the mean and the CI for the mean.

X^2-test to check whether the variance of a normal population is equal to, smaller than or greater than a given value, respectively.

Example of application of the X^2-test on the variance of a normal population,

alternative form of the test based on the CI for the variance.

Z-test to compare the means of two independent normal populations of known variances.

Example of application of the Z-test to compare the means of two independent normal populations of known variances.

F-test to compare the variances of two independent normal populations.

Example of application of the F-test to compare the variances of two independent normal populations.

Unpaired t-test to compare the means of two independent normal populations with unknown variances:

case of equal variances;

case of unequal variances.

Example of application of the unpaired t-test for the comparison of the means of two normal populations (equal unknown variances).

Example of application of the unpaired t-test for the comparison of the means of two normal populations (unequal unknown variances).

Paired t-test to compare the means of two normal populations of unknown variances.

Example of application of the paired t-test to compare the means of two normal populations.

A remark about the Excel implementation of the paired t-test: relation with Pearson’s linear correlation coefficient.

Sign test for the median of a population.

Sign test to check if two independent samples belong to the same unknown statistical population.

Example of application of the sign test to check if two samples belong to the same unknown population.

Detection of outliers non belonging to a statistical normal population by Chauvenet criterion.

Example of application of the Chauvenet criterion.

(6) Pairs of random variables

Pairs of random variables, covariance and linear correlation coefficient (Pearson r).

Probability distributions for the linear correlation coefficient r and test to check the hypothesis of stochastic independence based on r:

case of independent random variables with convergent moments of sufficiently high order, for sufficiently large samples (z-test);

case of normal random variables

- approximation for large samples, Fisher transformation (z-test),
- independent normal random variables (t-test).

Example of application of the linear correlation coefficient to check the stochastic independence of two normal random variables.

Remark: linear correlation coefficient versus regression analysis.

A. F-test for the comparison of means. Introduction to ANOVA.

Typical problem addressed by (1-factor) ANOVA and its matematical formulation.

Factors and levels: 1-factor ANOVA, 2-factor ANOVA without replication, 2-factor ANOVA with replication, 2-factor ANOVA, etc.

General mathematical formulation of ANOVA.

A.1 1-factor ANOVA (one-way ANOVA)

A.1.1 Basic definitions

A.1.2 Statistical model of 1-factor ANOVA
Fundamental theorem of 1-factor ANOVA and related remarks.
F-test for the dependence of the means on the group.
Example of application of 1-factor ANOVA.
Remark. 1-factor ANOVA for groups with unequal numbers of data.
A.1.3 Confidence intervals for the difference of two means: Student’s approach and Scheffé’s approach.

A.2 2-factor ANOVA without interaction.
A.2.1 Basic definitions of 2-factor ANOVA without interaction.
A.2.2 Statistical model of 2-factor ANOVA without interaction.
Fundamental theorem of 2-factor ANOVA without interaction and related remarks.
F-test for the dependence of the means on the index i (first factor).
F-test for the dependence of the means on the index j (second factor).
Example of application of the 2-factor ANOVA without interaction.

A.3 2-factor ANOVA with interaction
A.3.1 Basic definitions of 2-factor ANOVA with interaction.
A.3.2 Statistical model of 2-factor ANOVA with interaction.
Fundamental theorem of 2-factor ANOVA with interaction and related remarks (hints).
F-test for the dependence of the means on the index i (first factor).
F-test for the dependence of the means on the index j (second factor).
F-test for the occurrence of interactions between the factors.
Example of application of the 2-factor ANOVA with interaction.

R. Data modeling. Introduction to regression analysis
General setup of the problem: models, model parameters.
Fitting of model parameters to the data of the small sample.
Merit function.
Calculation of the best values of the model parameters (best-fit).
Maximum likelihood method for the definition of the merit function.
Some important cases:
least-squares method;
weighted least-squares method, or chi-square method;
robust fitting methods for non-Gaussian data,
general notion of outlier.
Linear regression by the chi-square method.
Linear regression by the least-squares method as a particular case.
Determination of the best-fit parameters by the normal equation.
The best-fit parameters as normal random variables.
Mean and covariance matrix of the best-fit parameters.
The normalized sum of squares of residuals around the regression (NSSAR) as a chi-square random variable.
Stochastic independence of the estimated parameters.
Goodness of fit Q of the regression model.
Extreme values of the goodness of fit Q.
Case of equal and *a priori* unknown standard deviations,
estimate of the standard deviations of the data by the chi-square method.
F-test for the goodness of fit (in the presence of an additive parameter).
F-test with repeated observations:
case of unequal, known variances;
case of equal, possibly unknown variances.
Incremental F-test on the introduction of a further fitting parameter.
Self-consistency tests of the regression model for homoskedastic systems with unknown variance:

coefficient of determination and adjusted coefficient of determination,
line fit plot,
residuals and residual plot,
standardized residuals and normal probability plot.
Confidence intervals for the regression parameters.
Confidence intervals for predictions.
Important case. Regression straight line:
regression straight line, basic model;
regression straight line, alternative model;
confidence intervals for the regression parameters, basic model;
confidence intervals for the regression parameters, alternative model;
confidence interval for predictions, alternative model, confidence region;
confidence interval for predictions, basic model.
Example of regression straight line in a homoscedastic model, confidence intervals for the intercept and the slope, confidence region, confidence interval for the prediction at a given value of the abscissa, goodness of fit for an assigned value of the variance.
Maple implementation of the linear regression.
Excel implementation of the linear regression.
Excel implementation of the linear regression for the case of no additive parameter.
Linear regression with more variables.
Example of multiple linear regression.

Topics treated in the lecture notes:
Chi-square fitting by SVD.
Singular Value Decomposition of a matrix.
SVD and covariance matrix of the estimated best-fit parameters.
Example of SVD by Maple.

N. Nonlinear regression (on having enough time, very unlikely)
N.1 Nonlinear models
N.2 Maple implementation of nonlinear regression.
Example of a nonlinear regression model reducible to a linear one.
N.3 Example of a nonlinear regression model in one dependent variable.
N.4 Example of a nonlinear regression model with two independent variables.
N.5 Robust fitting.
N.5.1 Estimate of the best-fit parameters of a model by local M-estimates.
N.5.2 Excel and Maple implementation of robust fitting
N.5.3 Example of robust regression straight line
N.6 Multiple regression

Principal Component Analysis (PCA) (on having enough time, i.e. almost impossible)
Basic ideas.
Principal component model.
Principal component model for predictions.
Relationship between PCA and SVD.
Use of PCA for the grouping of data.
PCA for nonb-centred data (General PCA).
Implementation of PCA by R.
Illustrative example of application of PCA.
Further example of PCA.

References:

[1] E. Lloyd Ed., Handbook of applicable mathematics, Volume VI- Part A: Statistics, John Wiley

& Sons, New York, 1984, p. 57

[2] Handbook VI-A (op. cit.), p. 41

[3] C. Capiluppi, D. Postpischl, P. Randi, Introduzione alla elaborazione dei dati sperimentali, CLUEB, Bologna, 1978, pagg. 89-98 (media) e 100-102 (varianza)

[4] T.H.Wannacott, R.J.Wannacott, Introduzione alla statistica, Franco Angeli, Milano, 1998, p. 205-228

[5] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T.Vetterling, Numerical Recipes, Cambridge University Press, Cambridge, 1989, p. 470-471 and
Capiluppi (op. cit.), p. 115-117

[6] C. Capiluppi (op.cit.), p. 120-121

[7] C. Capiluppi (op.cit.), p. 121-122

[8] Numerical Recipes (op. cit.), p. 468, Handbook VI-A, p. 264

[9] C. Capiluppi (op.cit.), p.123-124 (equal variances)
and Numerical Recipes (op. cit.), p. 466-467 (different variances)

[10] http://ishtar.df.unibo.it/stat/avan/misure/criteri/chauvenet.html

[11] Numerical Recipes (op. cit.), p. 484-487
and for further reference, in the case of nonzero p, see Handbook VI-A (op. cit.), p. 53-54

General references:

- John R. Taylor, Introduzione all'analisi degli errori, Zanichelli, Bologna, 1986
- Elena S. Ventsel, Teoria delle probabilità, Ed. MIR, 1983
- Sheldon M. Ross, Introduction to probability and statistics for engineers and scientists, 3rd edt., Academic Press, New York, 2004 (very good!)
- Sheldon M. Ross, Probabilità e statistica per l'ingegneria e le scienze, APOGEO, Milano, 2003
- http://ishtar.df.unibo.it/stat/avan/temp.html (in Italian)
- http://onlinestatbook.com/index.html (in English, very good!)