Introduction to Statistics

Variables
Measurement scales

Prof. Giuseppe Verlato

Unit of Epidemiology & Medical Statistics
Department of Diagnostics & Public Health
University of Verona

Statistics

Discipline, whose substance is comprised of methods, i.e. by procedures useful to handle information dealing with variables assessed within groups.

In other words, Statistics deals with phenomena which vary within a population, made up of statistical units (usually patients)

Why is Statistics necessary in Health Care Professionals ?

1) To remain updated by reading current scientific literature, whose results are evaluated and expressed using statistical terminology.

2) To think over your patients, your series!

Why should Italian physicians think over their own patients ? To remain updated it is enough to read what has been discovered in the US, England, Germany, …

Italian physicians must think over their own series since:

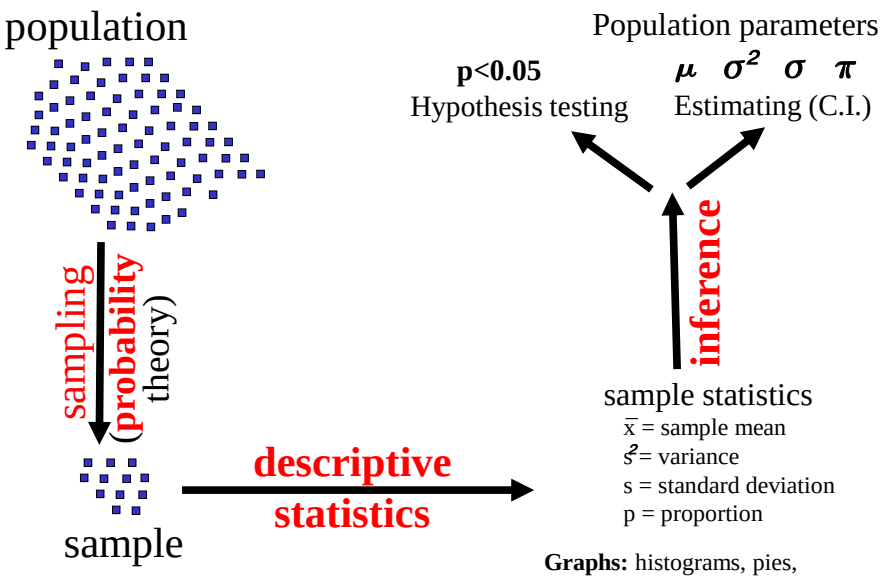
1) Patients often differ between countries: for instance, cardiovascular risk charts, which report 10 year risk of fatal cardiovascular disease, differ between Western Europe (low risk area) and Eastern Europe (high risk area) (http://www.escardio.org/Education/Practice-Tools/CVD-prevention-toolbox/SCORE-Risk-Charts).

2) Thinking over your actions usually improves the quality of your performance.

3) The scientific and technological gap between Italy and other industrialized countries should be prevented from further increasing.

4) In spite of lack of resources, Italian medicine is sometimes still teaching the rest of the world.

Syllabus of the course of Medical Statistics



EXAMPLE

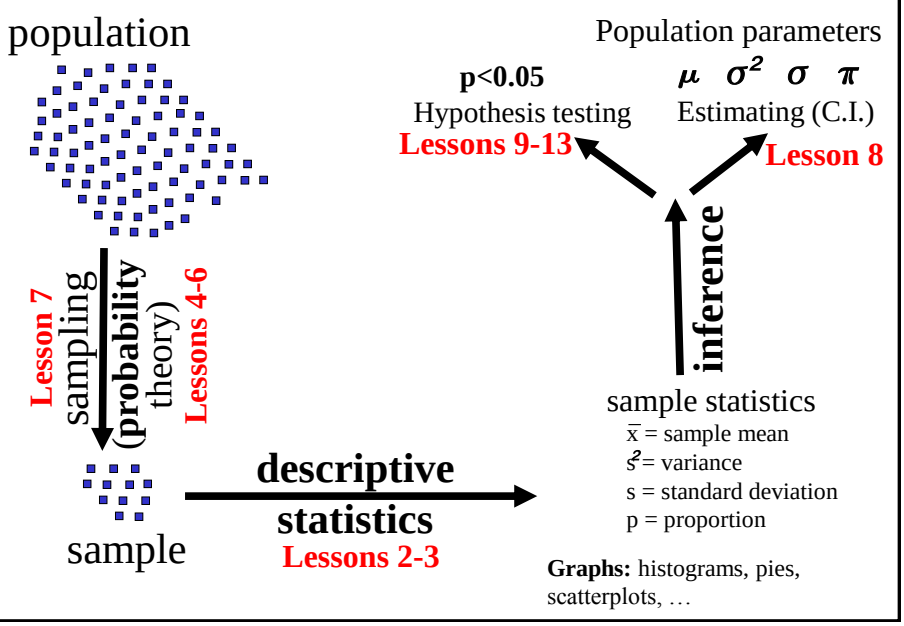
Knowing the political opinion of Italian electors, who are nearly 51 millions, requires a highly resource-consuming procedure. Surveys on the whole population are performed once every 5 years, in the frame of political elections.

During the time between elections, political parties commission polls to specialized firms, which draw **representative samples** from the Italian population.

Selected subjects are interviewed, and results are synthesized through **descriptive statistics**, yielding summary tables and graphics.

Results obtained from the sample are generalized to the **source population** by **inference**. Usually, to take into account the uncertainty inherent to the procedure, the percentage of people declaring to support a particular party is expressed by intervals, named **confidence intervals**, rather than by single values.

Syllabus of the course of Medical Statistics



Descriptive statistics provide a concise summary of data, either numerical or graphical.

Probability theory deals with random events.

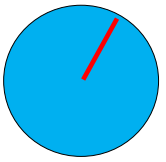
Inferential statistics allow one to generalize results obtained from the sample back to the source population.

Constants and Variables

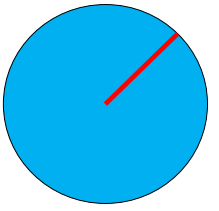
Constant = feature which does not change from one statistical unit to another

Variable = feature which does change from one statistical unit to another

Area = π * radius^2



Area = π * radius^2



Expected vital capacity

Men: vital capacity (l) = 5.76 * height (m) – 0.026 * age (years) – 4.34

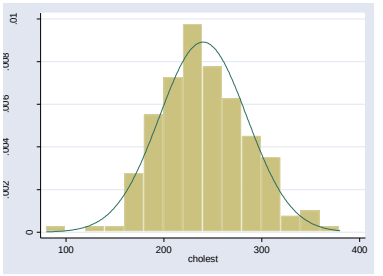
Women: vital capacity (l) = 4.43 * height (m) – 0.026 * age (years) – 2.89

Main types of measurement scales

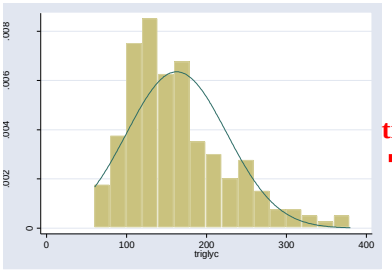
Types of scales	Operations allowed	Examples of variables	Descriptive statistics	Inferential statistics
Nominal	= ≠	Sex, hair color, marital status	Proportion (prevalence)	Chi-square, Mantel-Haenszel, logistic model
Ordinal	= ≠ < >	Pain intensity, coma depth, tumor stage	Median, range, interquartile range	Non-parametric tests
Ratio	= ≠ < >	**Asymmetrically-distributed variables:** Survival time, number of metastatic nodes	Median, range, interquartile range	Non-parametric tests, data normalization
	+ - * /	**Symmetrically-distributed variables:** glycaemia, systolic pressure, body mass index	Mean, standard deviation	t-test, ANOVA, regression

Identifying the correct scale is necessary to correctly choose descriptive and inferential statistics.

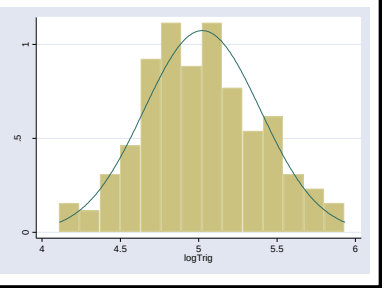
Distribution of total serum [cholesterol] (mg/dL) in 200 type 2 diabetic patients



Distribution of serum [triglycerides] in mg/dL in 200 type 2 diabetic patients



Logarithmic transformation



Usually statistical textbooks mention a fourth scale of measurement, **interval scale**.

For instance, when temperature is measured in Celsius degrees, the zero value corresponds to melting ice rather than to the real zero temperature (-273.15 °C).

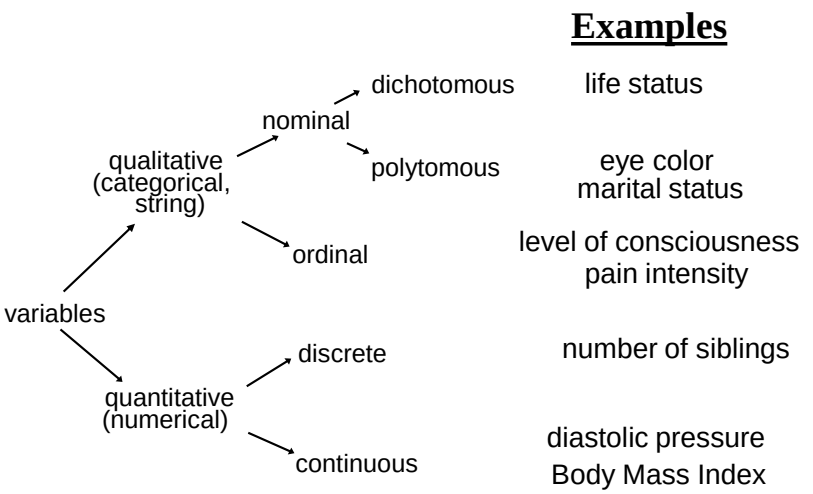
For this reason: **50 °C – 20 °C = 30 °C CORRECT**

but **~~50 °C / 20 °C = 2.5~~ WRONG**

50°C/20°C= (50+273.15)/(20+273.15)= 323.15/293.15= 1.10 CORRECT

However interval scale is rarely used in Medicine.

The tree of variables



Epidemiologic data come in many forms and sizes.

One of the most common forms is a **rectangular database** made up of rows and columns.

Each row contains information about one individual, and it is called a "record" or "**observation**."

Each column contains information about one characteristic such as sex, date of birth or disease, and it is called a "**variable**".

The first column of an epidemiologic database usually contains the individual's initials, or identification number which allows us to identify who is who. We can also use the name if this does not infringe the individual's right to privacy.

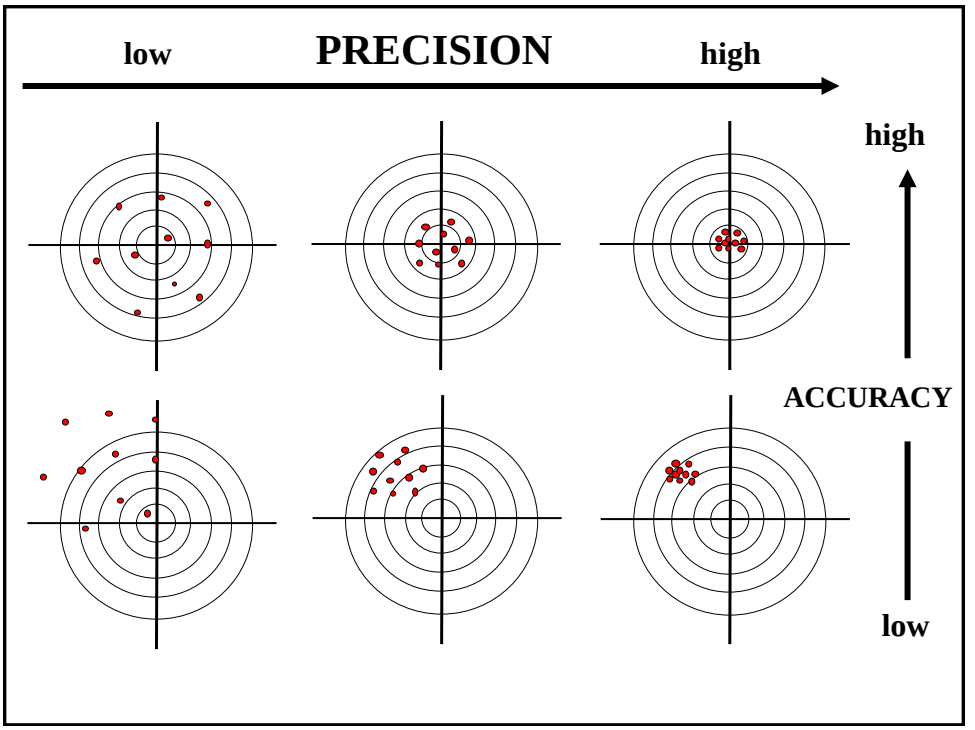
SEX	AGE years	Height (m)	Weight Kg	SMOKE	INFARCTION
M	42	1,70	58	F	N
M	48	1,84	90	N	I
M	51	1,66	70	F	I
M	54	1,78	76	F	I
M	58	1,74	72	N	N
M	60	1,76	85	N	I
M	62	1,64	62	F	I
M	64	1,90	88	F	I
M	65	1,72	69	N	N
M	70	1,77	77	N	N
M	75	1,68	73	F	I
M	81	1,74	75	F	I
F	45	1,68	59	F	N
F	49	1,58	55	N	N
F	51	1,62	68	N	N
F	53	1,65	64	F	I
F	60	1,72	70	N	N
F	63	1,69	65	F	I
F	68	1,70	73	N	I
F	75	1,66	52	N	N

Main properties of a measure

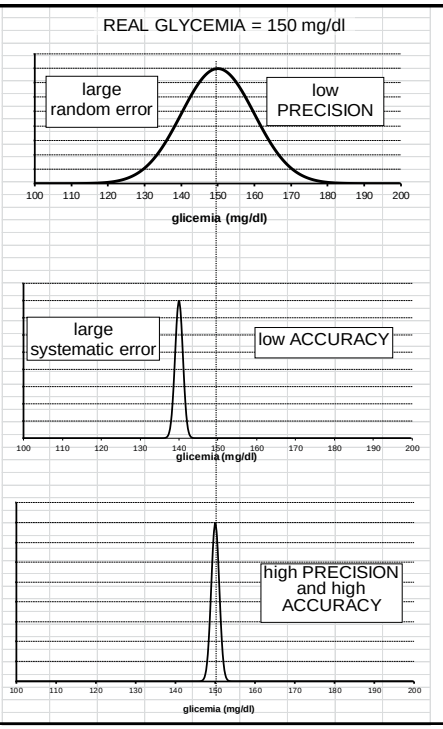
- ACCURACY (unbiassedness) = capacity to avoid/minimize systematic error (bias). Does the mean value differ from the real value ?

- PRECISION = capacity to reduce/minimize random variability of values; it is often evaluated by the coefficient of variation. Repeatibility = capacity to obtain the same value under the same conditions. Reproducibility = capacity to obtain the same value under different conditions.

- VALIDITY = extent to which a measurement reflects the real phenomenon which is intended to measure.



Clinical example



Untrained personnel

Large delay between blood withdrawal and measurement

VALIDATION of a MEASUREMENT TOOL

The measurement tool is compared with a "**gold standard**".

For instance self-reported smoking habits, assessed by a postal questionnaire, can be compared with blood carbon monoxide (CO) or serum cotinine (the main metabolite of nicotine) [Olivieri et al, 2002].

Likewise self-reported asthma, assessed by postal questionnaire, is usually validated using as gold standard clinical diagnosis, atopy assessed by skin prick tests/IgE levels, lung function assessed by spirometry [de Marco et al, 1998].

References

de Marco R, Cerveri I, Bugiani M, Ferrari M, Verlato G (1998) An undetected burden of asthma in Italy: the relationship between clinical and epidemiological diagnosis of asthma. *Eur Respir J*, 11: 599-605

Olivieri M, Poli A, Zuccaro P, Ferrari M, Lampronti G, de Marco R, Lo Cascio V, Pacifici R (2002) Tobacco smoke exposure and serum cotinine in a random sample of adults living in Verona, Italy. Arch Environ Health, 57: 355-359