**The small biostatistician**

**DESCRIPTIVE STATISTICS** **It describes the data.**

**Nominal (categorical) variable**: **categories** with distinct names, *among which it is not possible to establish a logical sorting criterion*.

**Ordinal variable**: **ordered series of categories**, *the difference among them cannot be considered constant*.

**Interval variables**: the possible values occur at **equal intervals**.

Bar chart: categorical data. Histogram: continuous. X = value of a variable for a subject. Sample size: number of subjects in a sample.

n = number of subjects in a group of the sample. N = number of subjects in the whole sample. Note: *there is not full agreement on these abbreviations.*

**INDICES OF CENTRAL TENDENCY** - "typical" values useful to summarize the data.

**Mean**: the (arithmetic) mean is the central tendency index for interval data (*sum of the values ​​of each element of the sample/number of elements*).

**Median**:central tendency index for ordinal data (the *value greater than 50% of the values and smaller than 50% of the values*).

**Mode**:tendency index for nominal data (*the category that occurs most frequently*) (*if there are 2 of them: bimodal distribution*).

**INDICES OF DISPERSION** - indicate how closely the data are clustered around the central tendency measures.

**Standard Deviation (SD)**: *the square root of the variance* (the **variance** is the average of the squared difference of data points from the mean of the data set).

**Minimum-maximum range**: *the difference between the maximum and the minimum value* - it is always a single number.

**Interquartile range**: (*midspread*): *the difference between upper and lower quartile, i.e. between 75th and 25th percentiles* (it includes the middle 50% of the data).

**INFERENTIAL STATISTICS**  **It establishes the probability that a deduction based on the data collected for a sample is correct.**

The **sample** consists of the subjects under observation. The **population** describes the hypothetical (and usually infinite) subjects of interest for the study.

*The key point is the quantification of the level of inaccuracy of the estimate based on the data obtained from the sample.*

**Null hypothesis**: there is no difference between the sample and the population.

**Standard error of the mean**: *standard deviation of the means of random samples (of a given size) extracted from the population.*

*P* (or *p*) = probability to be wrong stating that there is a real difference (i.e., rejecting the null hypothesis).

*p*=0,05: there is a difference; *or a coincidence has occurred that does not occurr more than one time in 20 experiments*.

By convention, if *p*<0.05, the difference is statistically significant; if *p*<0.01, the difference is highly significant.

A statistically significant difference may not be clinically significant.

Two-tailed test: it evaluates any difference among the groups, regardless of the *direction* of the difference.

Confidence interval (C.I.) around the mean: limits (lower and upper) within which the population mean is included with the (95)% probability.

**Type I Error**: stating that a significant difference *exists* when it is not true. Measured by the alpha probability (α). α=0.05 is often chosen

**Type II Error**: stating that there is no significant difference when it exists. Measured by the beta probability (β). β=0.20 is often chosen

**Power**: the probability of concluding that a difference exists when in fact, it exists. Power=1-

It is related to the *extent of the difference* to be detected and to the *sample size*.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test** **choice**: subjects? →  variables? ↓ | **2 groups**  **of different subjects** | **3 o more groups**  **of different subjects** | **1 treatment on the same subjects** | **>1 treatment on**  **the same subjects** | ***Association***  ***between 2 variables*** |  |
| **Interval variable**  (and normal distribution) | t-test for independent data (*Unpaired*) | Analysis of Variance (ANOVA) | t-test for paired data (*Paired*) | Repeated measures  ANOVA | Linear regression and Pearson correlation |  |
| **Nominal variable** | 2 (chi-square)  Fisher's exact test | 2 (chi-square) | McNemar's Test | Cochran's Q Test | -Relative risk (RR)  -*Odds ratio* (OR) | (Prospective)  (Retrospective) |
| **Ordinal variable** | Mann-Whitney Test | Kruskal-Wallis Test | Wilcoxon Test | Friedman Test | Spearman rank correlation |  |

**Sources** (website: <http://apollo11.isto.unibo.it/summa/it/metodo/stat.htm>)

Glantz SA. *Primer of Biostatistics*. McGraw-Hill, 2012 (7th Ed). Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. Pmph USA, 2014 (4th Ed).

Glantz SA. *Statistica per discipline biomediche*. McGraw-Hill, 2007 (6a Ed). Norman GR, Streiner DL. *Biostatistica. Quello che avreste voluto sapere...* CEA, 2015.