Biostatistics 28

Carlos Chavez de Paz and Allen Murga

Biostatistics and Epidemiology

Basic Statistical Concepts

- Statistics is a field of study concerned with the collection, organization, summarization, and analysis of data. When data analyzed is derived from biological science and medicine, we used the term biostatistics [1].
- Data may come from many sources: medical records, external sources, surveys, experiments.
- Descriptive Statistics: to describe and summarize the data.
- Inferential Statistics: to make inferences that can expand the data to a population.
- Variables: characteristics presented as values in different persons, places, or things.
  - Quantitative variable: can be measured or counted

C. C. de Paz
Department of Surgery, Cary Medical Center, Caribou, ME, USA
e-mail: cchavez@pineshealth.org

A. Murga (✉)
Department of Vascular Surgery, Loma Linda University,
Loma Linda, CA, USA
e-mail: amurga@llu.edu

A. Murga et al. (eds.), *The Vascular Surgery In-Training Examination Review (VSITE)*,
https://doi.org/10.1007/978-3-031-24121-5_28

    Discrete variable: the possible values are either finite or countable numbers (e.g. number of patients, length of hospital stay)
    Continuous variable: the possible values can take any value in a particular limit (e.g. height, weight)

  - Qualitative variable (categorical): can be placed in different categories distinguished by some characteristic or attribute (e.g. race, gender)
- Measurement scales: the first step in any statistical analysis is to determine the level of measurement [2], which influences the type of statistical analysis that can be performed on it.
  - Nominal scale: names or categories (e.g. sex, race, ethnicity)
  - Ordinal scale: ranked categories, classifications (e.g. CEAP (CVD [3]), Rutherford (PAD) [4])
  - Interval scale: ordinal scale, the differences between units of data can be defined, and there is no meaningful zero (e.g. temperature, years)
  - Ratio scale: interval scale, and there is a meaningful zero (e.g. age, weight)

Inferential Statistics

- Uses random samples of data and makes inferences (predictions) about the population.
- Uses sample data from the population to answer research questions (test hypothesis)
- Population: complete collection of all elements/subjects to be studied
- Sample: a subset of elements drawn from the population
- Methods of sampling: convenience, simple random, systematic, stratified random, and cluster.

Descriptive Statistics

- Results that summarize a given data set, usually a sample of a population
- Data may be distributed in different ways: skewed to the left, skewed to the right, or sometimes is disarranged without any particular shape (See Fig. 28.1)
- In some cases, the data tends to be around a central value (e.g. mean) with no bias left or right, resembling a “Normal Distribution” (See Fig. 28.1d)
- Descriptive statistics show their results as **measures of central tendency** (summary) and **measures of variability** (dispersion) [5]
- Measures of central tendency
  - Mean: the sum of scores in the data set divided by the total number of scores. Is strongly affected by extreme values

    $$\mu = \frac{3+3+7+9+10+10}{6} = 7; \quad \mu = \frac{3+3+7+9+10+100}{6} = 22$$

  - Median: the midpoint of the arranged $\frac{n+1}{2}$th observation of the dataset. Is not affected by extreme values

    From the set: 2, 5, 7, 16, 84; the midpoint is the $\left(\frac{5+1}{2}\right)$3rd observation = 7

  - Mode: the value in the data set that occurs most frequently. Sometimes there are more than one mode

    From the set 21, 45, 30, 25, 45, 21, 45; the mode = 45
- Measures of Variability: show the amount of dispersion present in a dataset; if the values are close to each other (small dispersion) or if they are widely scattered (greater dispersion)
  - Range: largest value − smallest value
  - Variance: measures dispersion relative to the scatter of the values about the mean
  - Standard deviation (SD): square root of the variance. In a normal distribution shows us the percentage of data that

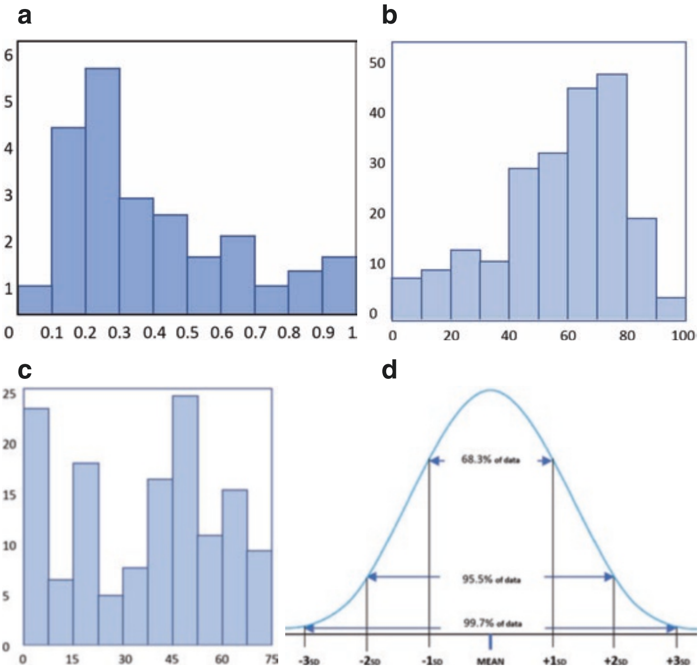

Fig. 28.1 (**a**) Distribution bar chart with frequencies skewed to the left. (**b**) Distribution bar chart with frequencies skewed to the right. (**c**) Distribution bar chart with frequencies disarranged. (**d**) Normal distribution of a dataset with the corresponding standard deviations from the mean

falls between 1, 2, or 3 SDs (68.3%, 95.5%, and 99.7%, respectively) (Fig. 28.1d).

- Coefficient of variation: standard deviation divided by the mean, used to compare dispersion between two or more groups.

Probability

- Probability is the likelihood (chance) of the occurrence of an event
- Observational probability will calculate probabilities from a sample using relative frequencies

- The **Law of large numbers**: summary results that are based upon a large number of independent observations (trials) which are less susceptible to the effects of variance (random error) when compared to results derived from fewer observations
- In probability, the **central limit theorem** (CLT) states that the means of a large number of independent random samples, each with a finite mean and variance, will approach a **normal distribution** (usually if the sample size is >30) (See Fig. 28.1d).
- The normal distribution can be used to model the distribution of many **variables** that are of interest. This allows us to answer probability questions about these random variables.

Estimating Population Parameters

- Estimate: process of using the data available from the sample to **estimate** the unknown value of the population parameter (statistic from the population)
- We can compute two types of estimates: a **point estimate** and an **interval estimate**.
- Point estimate: a single value used to estimate a population parameter.
- Interval estimate: a range of values that includes the parameter being estimated

Hypothesis Testing

- Null hypothesis (**$H_Ø$**): hypothesis to be tested, is a statement of status quo (no difference)
- Alternate hypothesis (**H_A**): hypothesis that competes with the **$H_Ø$**, is a statement of what we believe is true if the sample data results in rejecting the null hypothesis
- Significance level (**α level**): is the probability of rejecting the **$H_Ø$** when it is true. It is suggested that a probability of 1 in 20 (0.05), is a convenient cutoff level to reject the null hypothesis, but the significance level can change according to specific circumstances

- *p*-value: smallest value of $\boldsymbol{\alpha}$ for which the $\mathbf{H}_{\boldsymbol{\emptyset}}$ can be rejected, or said in other words, the probability of obtaining a result more extreme than the result actually obtained when the null hypothesis is true [6]
- Confidence interval: shows an estimated range of values which is likely to include an unknown population parameter
  - The narrower the confidence interval is, the more precision it has
  - A wide interval may indicate that more data should be collected before making assumptions about the parameter
- Test statistic: is a value computed from the sample data that is used in making the decision about the rejection of the null hypothesis
  - Helps to evaluate whether the test statistic falls within the rejection region
  - Helps to decide if we reject or fail to reject the $\mathbf{H}_{\boldsymbol{\emptyset}}$ at the pre-specified α level and make a conclusion
  - There are different hypothesis tests which use different test statistics based on the probability model assumed in the null hypothesis. The most common are:

    Z-test, to determine differences between two population means, used if the data has a known normal distribution

    t-test, similar to a *Z*-test but used if an unknown normal distribution, used for continuous or ordinal scales

    ANOVA, F statistic, similar to a *t*-test and can compare means of more than two groups

    Chi-square test, X^2 statistic, used for categorical variables (counts or frequency data)
- Type I error (α error): reject the $\mathbf{H}_{\boldsymbol{\emptyset}}$ when it is true. The probability of committing a type I error is the same as α (the significance level)
- Type II error (β error): failing to reject the null hypothesis when it is false (See Table 28.1)
- Power: probability of rejecting the $\mathbf{H}_{\boldsymbol{\emptyset}}$ when it is false [7]. This is defined by 1-β. Decreasing $\boldsymbol{\alpha}$ makes it harder to reject the null hypothesis and thus lowers the power (See Fig. 28.2)

Table 28.1 Hypothesis testing table

	Truth	
Decision	$\mathbf{H_\emptyset}$ is true	$\mathbf{H_\emptyset}$ is false
Reject $\mathbf{H_\emptyset}$	Type I error (α)	Correct
Fail to reject $\mathbf{H_\emptyset}$	Correct	Type II error (β)

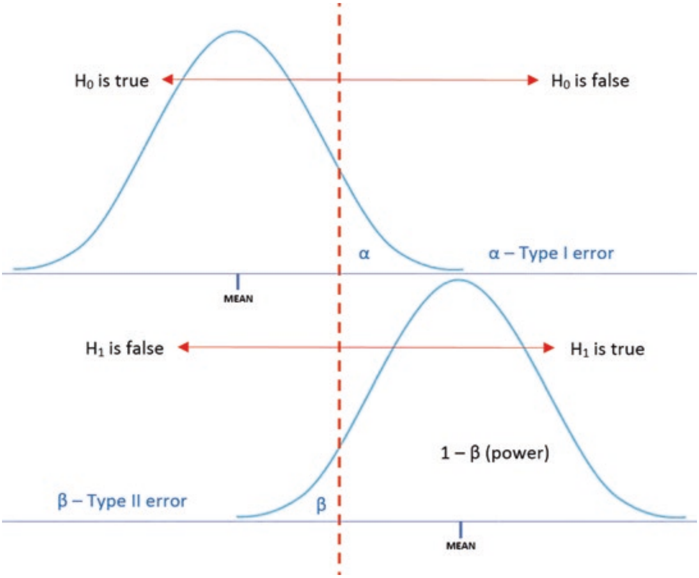

Fig. 28.2 Hypothesis testing graph: distributions of the null (H_0) and the alternate (H_A) hypothesis

Evaluating the Relationship Between Variables

- Correlation: measure of association between two continuous variables (strength). The direction and strength of the linear relationship are measured by the correlation coefficient (*r*).
  - If the variables *X* and *Y* have nonlinear relationship, it will not provide a valid measure of association
  - Correlation does not imply causation

- Simple linear regression: determines the linear relationship between two continuous variables and determines the equation of the best line that fits through the data
  - The regression equation is then used to predict the value of the dependent variable Y given the independent variable X. The dependent variable is continuous.
- Multiple linear regression: introduces two or more predictor variables into the prediction model
  - Multicollinearity: occurs when the predictor variables are so highly intercorrelated that they produce instability problems [8]
  - Overfitting: the inclusion of too many variables in the equation can lead to an equation that does not predict well the outcome [9]
  - Stepwise regression, forward selection, and backward elimination help with overfitting and multicollinearity problems
- Logistic Regression: method for predicting binary outcomes on the basis of one or more predictor variables. The dependent variable is binary (dichotomous). Measure of effect is the Odds ratio.
- Poisson Regression: uses a count dependent variable, is suitable for rate data. Measure of effect is Incidence Rate ratio.
- Proportional Hazards Regression: models the relationship between survival of a patient and a set of independent variables (e.g. age, comorbidity index, BMI). Measure of effect is Hazards ratio.

Epidemiology

- Definition: The study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control of health problems [10]
- Descriptive Epidemiology: Study of the amount and distribution of disease within a population by person, place, and time.
  - Person variables: age, sex, race, social status

  - Place variables: natural boundaries, urban/rural differences, international comparisons
  - Time variables: secular trends, cyclic trends
- Types of studies
  - Observational: does not manipulate the exposure and does not randomize subjects

    Descriptive: useful to generate hypothesis, e.g. case reports, cross-sectional surveys, ecologic studies

    Analytic: generate and test hypothesis, suggest causality
    - Case-control: used for rare diseases. Less expensive. Start from the outcome and look for the exposures. High selection and information bias
    - Cohort: used for common diseases (outcomes). Start from the exposure and follow the development of an outcome. Can calculate incidence and risk ratio. Confounding and loss to follow-up may occur.
  - Interventional (clinical trial)

    Randomized controlled trials: experimental studies. The researcher manipulates the exposure and randomly assigns subjects to the exposed and unexposed. Eliminates selection bias. Strongest prove of cause and effect. Highest cost.
- Measures of effect: summarize the strength of the association between exposures and outcomes [11]
  - Rate: probability of occurrence of some particular event (outcome) in relation to a population and a measure of time.

    $$\frac{\text{number of events}^*}{\text{population at risk}}\ \text{time specification}$$
  - Incidence:

    $$\frac{\text{number of new cases}}{\text{total population at risk}}\left(\text{at a given point in time}\right)$$
  - Prevalence:

    $$\frac{\text{number of existing cases}}{\text{total population at risk}}\left(\text{at a given point in time}\right)$$
  - Crude Mortality rate:

    $$\frac{\text{number of all deaths}\left(\text{at a defined period of time}\right)}{\text{total population in the same period of time}}$$

- Relative risk:

$$\frac{\text{Probability of the event in the exposed group}}{\text{Probability of the event in the unexposed group}}$$

- Risk difference (excess risk, or attributable risk): $\text{Risk}_{\text{exposed}} - \text{Risk}_{\text{unexposed}}$
- Number needed to treat (NNT): measurement of the impact of a therapy by estimating the number of patients that need to be treated in order to have an impact on one person

$$\text{NNT} = \frac{1}{\text{Risk}_{\text{unexposed}} - \text{Risk}_{\text{exposed}}}$$

- Screening Tests: widely used in medicine to assess the likelihood that members of a defined population have a particular disease (see Table 28.2).
  - Sensitivity: screening test's ability to correctly identify those individuals who truly have the disease.

$$\text{Sensitivity }\% = \frac{a}{a+c} \times 100 = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

  - Specificity: screening test's ability to correctly identify those individuals who truly do not have the disease.

$$\text{Specificity }\% = \frac{d}{b+d} \times 100 = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

  - Positive Predictive value (PPv): screening test's ability to identify correctly those individuals who truly have the disease (true positive) among all individuals whose screening tests are positive [12]. PPv increases with increasing disease preva-

Table 28.2 Screening test for a disease

	Disease		
Test result	Present	Absent	Total
Positive	TP (a)	FP (b)	$a + b$
Negative	FN (c)	TN (d)	$c + d$
Total	$a + c$	$b + d$	n

lence, therefore high-risk populations are the best targets for screening programs. It is a critical measure of the performance of a diagnostic method, as it reflects the probability that a positive test reflects the underlying condition being tested for.

$$\text{PPv } \% = \frac{a}{a+b} \times 100 = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

- Negative Predictive value (NPv): screening test's ability to identify correctly those individuals who truly do not have the disease (true negative) among all individuals whose screening tests are negative [12]. NPV decreases with increasing disease prevalence.

$$\text{NPv } \% = \frac{d}{d+c} \times 100 = \frac{\text{TN}}{\text{TN}+\text{FN}}$$

Questions and Answers

1. A surgeon has designed a study in which he will compare the mean length of stay in days after the use of three different endovascular devices for the treatment of type A and type B aorto-iliac disease. The best statistic test is:
   (a) Correlation coefficient
   (b) Chi-square
   (c) ANOVA
   (d) Paired *T*-test
   (e) *Z*-test
2. Specificity is determined by:
   (a) False negatives/False negatives + True negatives
   (b) False positives/False positives + True positives
   (c) True negatives/True negatives + False negatives
   (d) False negatives/False negatives + True positives
   (e) True negatives/True negatives + False positives
3. The following type of study includes the outcome variable (e.g. severity of internal carotid occlusion) and tries to estimate the exposure variable:
   (a) Correlation
   (b) Case-control

(c) Cohort
(d) Randomized control trial
(e) Logistic regression

4. A research group is studying the effect on survival after BKA for complicated type II diabetes. They obtained a couple of thousand patients from an established prospectively collected database. After stepwise elimination of variables, they would like to perform a multivariate analysis, the best statistical test would be:
   (a) Logistic regression
   (b) Cox proportional Hazards regression
   (c) Poisson regression
   (d) Simple linear regression
   (e) Meta-analysis
5. To analyze the data of a new screening tool for detecting skin perfusion of the lower limbs in patients with moderate to severe claudication, the following statistical test has a direct relationship with the incidence of the disease in a population:
   (a) Specificity
   (b) Sensitivity
   (c) Odds ratio
   (d) Positive predictive value
   (e) Number needed to treat

Answers: 1 (c), 2 (e), 3 (b), 4 (b), 5 (d)

References

1. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. New York: Wiley; 2018.
2. Mertler CA, Reinhart RV. Advanced and multivariate statistical methods: practical application and interpretation. London: Routledge; 2016.
3. Eklöf B, Rutherford RB, Bergan JJ, Carpentier PH, Gloviczki P, Kistner RL, et al. Revision of the CEAP classification for chronic venous disorders: consensus statement. J Vasc Surg. 2004;40(6):1248–52.
4. Hardman RL, Jazaeri O, Yi J, Smith M, Gupta R. Overview of classification systems in peripheral artery disease. Semin Intervent Radiol. 2014;31:378–88.

5. Marshall G, Jonker L. An introduction to descriptive statistics: a review and practical guide. Radiography. 2010;16(4):e1–7.
6. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011;24(1):18–23.
7. Krzywinski M, Altman N. Points of significance: power and sample size. London: Nature Publishing Group; 2013.
8. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology. 2016;6(2):227.
9. Ivanescu AE, Li P, George B, Brown AW, Keith SW, Raju D, et al. The importance of prediction model validation and assessment in obesity and nutrition research. Int J Obesity. 2016;40(6):887.
10. Porta M. A dictionary of epidemiology. Oxford: Oxford University Press; 2014.
11. Tripepi G, Jager K, Dekker F, Wanner C, Zoccali C. Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. Kidney Int. 2007;72(7):789–91.
12. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Front Public Health. 2017;5:307.