What Every Radiology Resident Should Know About Biostatistics

Theodore Brown, MA, MPH; Adam Fonseca, MD; Viet Vu, MD;
Millie Yu, BS; Jeremy Nguyen, MD; Scott Beech, MD
Tulane University School of Medicine
Department of Radiology, 1430 Tulane Ave. #8654, New Orleans, LA. 70112

Introduction

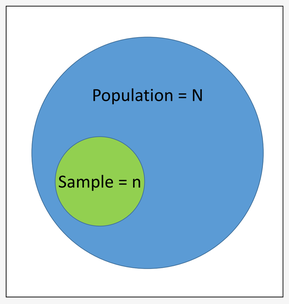
Biostatistics is defined as the application of statistical principles and techniques to biological, medical, and public health research. These principles include tools and techniques for the collection, summarization, and analysis of data or information and the subsequent interpretation of those results while accounting for uncertainty. Having an appropriate understanding of biostatistics is important part of working as a physician.

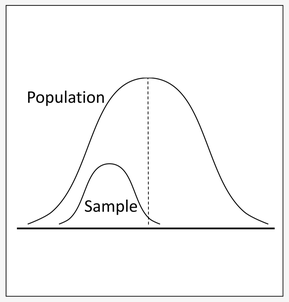
Biostatistics in Medical Research

Important concepts:
Variable: an alphabetic character which represents a number, called the *value* of the variable, which is either arbitrary, not fully specified, or unknown

- **Discrete variables:** variables which only take certain values and none in between. For example, the number of heart beats a patient has in one minute could be 80 or 81, but it cannot be anything between those two numbers
- **Continuous variables**: variables which can take any value. For example, a patient's temperature could be 37°C or 38°C or any number between them, such as 37.4°C.

Population: All members or cases of a group or category of scientific interest
Sample: A subset of cases selected from a larger population to gain knowledge about that population





Types of Variables:

Qualitative Variables:

- **Categorical** – variables with values that are unordered (ex. Hair colors)
- **Dichotomous variable** – A categorical variable with only two categories
- **Ordinal** – variables with values that follow an order but have uneven intervals (ex. Education level)

Quantitative Variables:

- **Interval** – variables with a constant interval size but no true zero (ex. Temperature in F or C)
- **Ratio** – variables with a constant interval size which have a non-arbitrary zero point (ex. temperature in K)

Screening Tests: No medical test is perfectly accurate. To determine how accurate a screening test is, we can use statistical measures of performance like sensitivity and specificity.

Sensitivity = True Positive Fraction = P(Positive Test | Has Disease)
Measures the proportion of positives that are correctly identified as such

Specificity = True Negative Fraction = P(Negative Test | No Disease)
Measures the proportion of negatives that are correctly identified as such

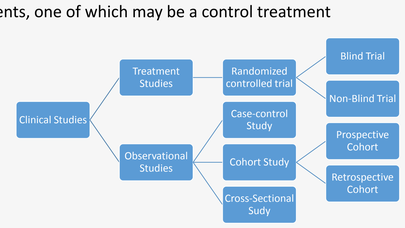
False Negative Fraction = (1 – Sensitivity) = P(Negative Test | Has Disease)
Percentage that tests negative but has the disease

False Positive Fraction = (1 – Specificity) = P(Positive Test | No Disease)
Percentage that test positive but doesn't have the disease

	Disease +	Disease -
Test +	+ + True Positive	+ - False Positive
Test -	- + False Negative	- - True Negative

Study Designs: Study designs are usually split into two types: observational and experimental. In an **observational study**, subjects are observed and variables are measured without assigning treatments to the subjects. In an **experimental study**, subjects are assigned to either a control or experimental group, treatments are applied to the experimental group, and then the effects on the variables are measured.

Randomized controlled (clinical) trial – subjects are randomized to one of two (or more) treatments, one of which may be a control treatment



Descriptive Statistics

Descriptive Statistics is the collection and summarization of data, the basis of quantitative analysis. It allows for presenting quantitative descriptions in a manageable form. Simplify large amounts of data, reduce data to a simpler summary. While this simplification as its limitations, it is a powerful tool for comparison.

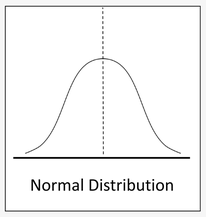
Univariate analysis involves examination across a **single variable** at a time. The three major characteristics are **frequency distribution**, **central tendency**, and **dispersion**.

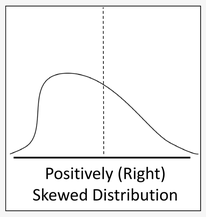
- **Frequency distribution:** the organization of data by listing all values in order, from least to greatest, and recording the frequency with which each group of scores occurs
- **Grouped frequency distribution:** individual scores are grouped, and each group of scores are given an equal class interval
- **Relative frequency distribution**: shows the percentage of all the elements that fall within each class interval

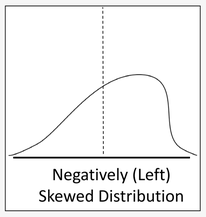
Ages of people in a study:
55, 59, 64, 60, 65, 35, 44, 54, 59, 63, 40, 45, 65, 49, 53, 58, 32, 44, 50, 63, 54, 31, 36, 42, 46, 58, 63, 51, 56, 61

Age Groups	Frequency	Relative Frequency
31-35	3	3/30=0.100
36-40	2	2/30=0.067
41-45	4	4/30=0.133
46-50	3	3/30=0.100
51-55	5	5/30=0.167
56-60	6	6/30=0.200
61-65	7	7/30=0.233
Total	**n=30**	

Distribution is best demonstrated as a graph or chart called a frequency distribution. Raw numbers or percentages can be used. The most common distribution is the **normal (Gaussian) distribution** which depends on the mean (μ) and standard deviation (σ). A normal distribution is symmetrical and bell-shaped but other types of distributions also exist. A positively skewed distribution (also called a right skewed distribution) and negatively skewed distribution (also called a left skewed distribution) can be seen below.


Normal Distribution


Positively (Right) Skewed Distribution


Negatively (Left) Skewed Distribution

Central tendency is an estimate of the center of a distribution of values. There are three major types of estimates of central tendency: mean, median, and mode.

- **Mean** is the average of values = (sum of values/number of values)
- **Median** is the value at the exact middle of the set of values.
- **Mode** is the most frequently occurring value within the set.

In a truly normal distribution (bell-shaped curve), mean, median, and mode will be equal to each other.

Values										
4	5	6	7	8	10	13	13	15	17	23
Mean = $\frac{\text{Sum of Values}}{\text{Number of Values}}$ = 11										
Median = 10										
Mode = 13										

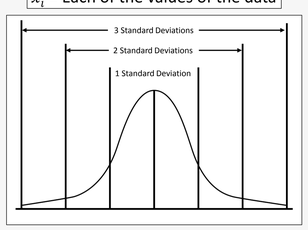
Dispersion refers to the spread of values around the central tendency. Two common measurements of dispersion are range and standard deviation.

- **Range** is the (highest value-lowest value).
- **Standard deviation** is a more accurate and detailed estimation of dispersion which accounts for outliers. While standard deviation is derived easily by computer, it is helpful to understand how it is calculated. The equation for standard deviation of a sample is shown at the right.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n-1}}$$

n = The number of data points
$\bar{x}$ = The mean of the x_i
x_i = Each of the values of the data

Each difference between each value of the mean is calculated using subtraction. To eliminated negative discrepancies from values below the mean, all differences are squared. These squares are added together to obtain the **Sum of Squares** (SS). The SS is divided by the total number of values minus one. This value is known as the **variance**. The standard deviation is the square root of the variance.



Inferential Statistics

Inferential statistics are used to apply statistical data obtained from a smaller sample size to a larger population. This is **inference**, or the use of a sample to draw conclusions about a population. In doing so, it is important to ensure that the sample group being tested, as closely as reasonably possible, represents the population to which one is trying to attribute the conclusions to. Invariably, different types of bias can be introduced when selecting a sample group. This is called **sampling error**, which is a natural, expected random variation that will cause the sample statistic to differ from the population parameter.

Central limit theorem: The central limit theorem forms the basis of inferential statistics. It states that:

1. The random sampling distribution of means will always to be normal, irrespective of the shape of the population distribution from which the samples were drawn.
2. The random sampling distribution of means will become closer to normal as the size of the sample increases

The extent to which a sample differs from the population can be measured using the **Standard Error the Mean (SE or SEM).**

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

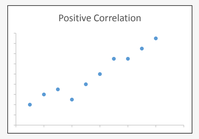
Hypothesis testing
To derive statistical conclusions from medical research, one must first construct two mutually exclusive hypotheses, usually termed the **null hypothesis** (H_0), which states that there is no difference or effect, for example, of a given treatment or drug, versus the **alternative hypothesis** (H_1), which posits that there is a difference/effect.

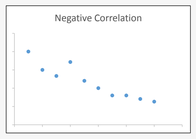
Reality	Null (H_0) not rejected	Null (H_0) rejected
Null (H_0) is true	Correct conclusion	Type 1 error
Null (H_0) is false	Type 2 error	Correct conclusion

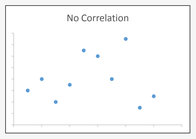
Type I Error α
A type I error or α, is the odds of saying there is a relationship, difference, gain, when in fact there is not. This is also known as the significance level of a test. In terms of false positives and false negatives, this would equate to a false positive. Along with this concept, and universally seen in research, medical or otherwise, is the p-value, which is the probability of making a Type I error based on the data obtained. This is typically set to <0.05, which means that there is a less than 5% chance that the data obtained is invalid, or to have occurred simply due to chance.
Type II Error β
The odds of saying there is no relationship, difference, gain, when in fact there is one. In terms of false positives and false negatives, this would equate to a false negative.
Power 1-β
The odds of saying that there is a relationship, difference, gain, when in fact there is one. In other words, the odds of confirming the alternative hypothesis/theory correctly. If the sample size increases (i.e. more data points), the power of the test/experiment increases.

Correlation

Correlation describes the relationship between two variables. The degree to which these two variables are related is called the **correlation coefficient**. The direction of this coefficient can be either positive or negative. In a positive correlation, the values of two variables, X and Y, move in the same direction. In a negative correlation, they move in opposite directions. If there is no correlation, then there is no discernable pattern.







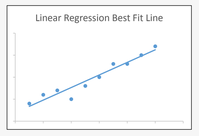
Correlation coefficients range from -1 to 1. Values closer to ±1 indicate a stronger relationship between X and Y while values closer to zero indicate a weaker relationship. Caution must be exercised when drawing conclusions as correlation does not imply causation. One of the most common types of correlational analysis used is the Pearson product-moment correlation which produces the **Pearson correlation coefficient** which is denoted by **r**.

$$r = \frac{\text{observed covariance}}{\text{max possible covariance}} \qquad r = \frac{SC_{XY}}{\sqrt{SS_X \times SS_Y}}$$

SS_X and SS_Y = Sum of squared deviates for X and Y values respectively
SC_{XY} = Sum of co-deviates for paired values of X and Y

Regression

Linear regression is an approach for modeling the predictive relationships between a predictor, or independent variable, and one or more outcomes, or dependent, variables. If there is a single independent variable and a single dependent variable, this approach is known as simple linear regression. This type of regression model attempts to find a linear function that predicts the dependent variable as a function of the independent variables.



- **Independent Variable (IV)** - A variable that is unaffected by the other variables being measured.
- **Dependent Variable (DV)** - A variable that is expected to change when the independent variable is changed.

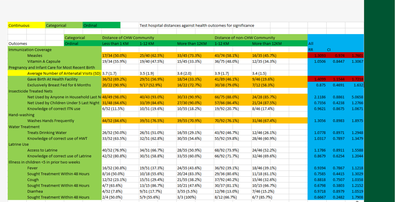
Example: Increasing the maternal age at first birth (an independent variable) will increase the number of diagnoses of autism in children within the first six months of life (a dependent variable). The mother's age will affect the number of diagnoses of autism but not the other way around.

ANOVA

ANOVA stands for "**analysis of variance**" which is a group of statistical models used to analyze differences between the means of two or more groups (or sets of data) and to determine if any differences seen are statistically significant. To use an ANOVA, it is necessary to have a categorical (or nominal) independent variable that has at least two independent groups or categories and a continuous (interval or ratio) dependent variable.

Electronic Spreadsheets

Electronic spreadsheets allow researchers to rapidly manipulate and analyze data. By building the formulas into the spreadsheet, new data can be automatically calculated whenever changes are made to the variables.



Take Home Points

Biostatistics play an important role in medical research by providing a powerful tool for both designing studies, and analyzing their results. This is accomplished by using descriptive statistics to analyze the frequency and distribution of data. These results can then be extrapolated to the population using inferential statistics.

The key to much of biostatistics is the central limit theorem, which states that the sampling distribution of the mean of any independent random variable will tend toward a normal distribution. This means that probabilistic and statistical methods that work for normal distributions can be applied to problems involving other types of distributions.

By combining the techniques described here, it is possible to test a hypothesis and determine if it is significant in an objective way.

Online educational information: http://learningbiostatistics.com

References

- L. M. Sullivan (2012). *Essentials of Biostatistics for Public Health*. 2nd Ed., Jones and Bartlett Publishers, Inc.
- J. H. Zar (2009). Biostatistical Analysis, 4th or 5th Edition. Prentice Hall
- B. G. Tabachnich and L. S. Fidell (2001). Using Multivariate Statistics, 4th Edition. Allyn and Bacon.