### 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 37.5°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

![Population and Sample Diagram](image)

**Types of Variables:**

**Qualitative Variables:**
- Categorical – variables with values that are unordered (ex. Hair colors)
- Dichotomous variable – A categorical variable with only two categories

**Quantitative Variables:**
- Interval – variables with a constant interval but no true zero (ex. Temperature in F or C)
- Ratio – variables with a constant interval 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 = PP (Positive Test | Has Disease)
- **Specificity** = True Negative Fraction = PN (Negative Test | No Disease)

**Descriptive Statistics**

- **Frequency distribution**
  - shows the percentage of all the elements that fall within each class interval

**Distribution**

- 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.

**Central tendency**

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

**Dispersion**

- The spread or variability of the data around the central tendency.
- Measures of dispersion are range and standard deviation.
- **Range** = the highest value - the lowest value
- **Standard deviation** = 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.

**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.
- Irrelevant, different types of bias can be introduced when selecting a sample group being tested, which is known as **sampling error**.

**Central limit theorem:**
- 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.
  - The random sampling distribution of means will become closer to normal as the size of the sample increases

**Hypothesis testing**

- To derive statistical conclusions from medical research, one must first construct two mutually exclusive hypotheses, usually termed the **null hypothesis** (H₀), which states that there is no difference or effect, for example, of a given treatment or drug, versus the **alternative hypothesis** (H₁), which posits that there is a difference/effect.

**Type I Error a**
- 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.

**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

**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.

**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.

**References**