

Introduction to Biostatistics

Richard P. DiCarlo, MD
Assistant Dean for Undergraduate
Medical Education
LSU School of Medicine New Orleans

Normal Distribution (Normal Curve, Bell-shaped Curve, or Gaussian Distribution)

- A mathematical abstraction that describes the distribution of observations in a population. Quantitative measurements are along the x-axis and relative frequency (or density function) is along the y-axis. The shape and location of the curve are determined by the mean (or other measures of central tendency) and the standard deviation (or other measures of variance).

Measures of Central Tendency

- **Mean (numeric average):** the sum of all values divided by their number.
- **Median (50th percentile):** the value that divides a distribution into two equal parts such that the number of values $>$ the median equals the number of values $<$ the median.
- **Mode:** the value that occurs most frequently among a group of observation
- *Note: in a true normal distribution, the mean, median, and mode are equivalent and values.*

Skewed Distributions

- **Positively skewed distribution:** a distribution of values that is influenced by some extremely high values such that the mean is higher than the median and mode.
- **Negatively skewed distribution:** a distribution of values that is influenced by some extremely low values such that the mean is lower than the median and mode.
- *Note: when the distribution is skewed it is often advisable to use the median or mode as the measure of central tendency.*

Probability Curves

- The sum of the probabilities of all possible observations in a population equals 1.0. Therefore, we can assign the area under the normal curve a value of 1.0. The area under a particular segment of the curve is a percentage of the total area and represents the probability that a random member of the population will have a value that falls between the two points on the x-axis. This can be determined using z-scores.

Standard Deviation

- is a standardized measure of the variance of the data. If perpendiculars are constructed at a distance of plus or minus 1-standard deviation from the mean, then 68% of the total area (68% of the sample population) is captured; at a distance of plus or minus 2-standard deviations from the mean, 95% of the total area is captured (95% of the sample population); at a distance of three standard deviations from the mean, 99% of the total area is captured.
- *Note: when we say that a normal curve is determined by the mean and the standard deviation, the mean determines the location along the x-axis and the standard deviation determines the width of the curve. The standard deviation is a measure of the variance of the data*

Confidence Interval

- We can easily determine the mean value of a sample population, but we are usually interested in extrapolating our data to the true population mean. How can we use samples to estimate the corresponding true population parameter? How can we be certain the sample is representative? We use 'standard error' calculations and derive the confidence interval.

Confidence Interval

- a range of values within which a population parameter is likely to occur with a specified probability. The most commonly used confidence interval (CI) is the 95% CI, in which the limits of the interval are 2 standard errors from the mean.

Confidence Interval

- There are two interpretations of the 95% confidence interval:
 - Practical interpretation: we are 95% confident that the true population mean is contained within the specified interval. (*This is not the statistically correct interpretation, but it is relatively easy to understand as such, and this interpretation is commonly used.*)
 - Probabilistic interpretation: in repeated random samples of the population, 95% of all such intervals will, in the long run, contain the true population mean.
- Confidence intervals help us determine the degree of imprecision in a sample mean, odds ratio, relative risk, etc. A wide confidence interval indicates that the estimate is imprecise.

Confidence Interval

- 3 things that determine width of CI:
 - *****Sample size:***** in general, the larger the sample size, the narrower the confidence interval will be for any given measurement.
 - **Variance of the data:** less variance (smaller standard deviation) generally means a narrower the confidence interval.
 - **Degree of confidence required:** a 90% confidence interval will be narrower than a 95% confidence interval (for the most part, 95% confidence intervals are reported).

Hypothesis Testing and p-values

- **Scientific hypothesis:** a testable statement of anticipated findings.
- **Statistical hypothesis:** a formal statement of the scientific hypothesis that includes the parameters being measured (the test statistic). It is usually stated as two hypotheses: the null hypothesis and the alternate hypothesis.
 - **Null hypothesis:** a formal statement that the anticipated effect will not be observed.
 - **Alternate hypothesis:** the formal statement of the scientific hypothesis.

Hypothesis Testing and p-values

- **Hypothesis testing** involves the statistical evaluation of the null hypothesis. If the statistical analysis falls within certain parameters, the null hypothesis is accepted and the scientific hypothesis is rejected (the anticipated effect was not observed). If the statistical analysis falls outside of those parameters, the null hypothesis is rejected and the scientific hypothesis is accepted (the anticipated effect was observed).

P-value

- The **p-value** is a computed statistic that gives us the probability that the measured difference between the two groups would occur if the null hypothesis were true. (e.g. if $p = .05$, it means that if the null hypothesis were true, the measured difference would occur by chance only five times in one-hundred similarly conducted trials, or a 1 in 20 chance; if $p = .01$, then if the null hypothesis were true, the likelihood of the measured difference occurring by chance is 1 in 100; $p = .001$ correlates with a 1 in 1000 chance if the null hypothesis were true; etc.)
- The smaller the p-value, the more surprising the results would be if the null hypothesis were true.

Level of Significance

- The **level of significance** is the degree of difference between two groups that will result in rejection of the null hypothesis. A level of significance is chosen such that the difference is so great that it is not likely to have occurred by chance. $P < .05$ is generally regarded as a statistically significant difference between the two groups (it is usually the level of significance at which point we conclude that the null hypothesis is false).
- If $p > .05$, the null hypothesis is usually accepted (the scientific hypothesis is rejected) and any measured difference is thought to be a chance event.
 - Keep in mind that this is an arbitrary cutoff point. If $p = .05$ there is still a 1 in 20 chance that the null hypothesis is actually true, but that the measured difference occurred by chance. This is called **type I error** (alpha error). If $p = .05$, there is a 5% chance of type I error (5% chance that the null hypothesis is actually true); if $p = .01$ there is a 1% chance of type I error; etc.

Hypothesis testing versus confidence intervals

- Confidence intervals can provide the same information as hypothesis testing with respect to rejection of the null hypothesis. However, confidence intervals provide the reader with more information:
 - - It specifies a range of values within which the parameter of interest is likely to occur.
 - Confidence intervals reflect the variability associated with the parameter of interest. A large difference between two groups becomes more meaningful when the confidence intervals are known. If the confidence intervals are wide, then it may be due to a small sample size or to a large amount of biologic variability in the parameter being measured (variance).
 - Confidence intervals highlight the importance of sample size. As sample size increases, a small difference between a test group and a control group may become statistically significant, without really being clinically significant. Confidence intervals provide information about the actual magnitude of the difference between the two groups.

Prevalence:

- the proportion of individuals in a population who have a specific characteristic or disease at a specific point in time. It is usually expressed as a number per 1000 or 100,000. This is an appropriate measure for chronic conditions, but it may underestimate acute diseases

- $$\frac{\text{Cases of Disease } X \text{ at Time } t}{\text{Population at Time } t}$$

Incidence

- a measure of *new* cases of disease in a defined population over a specified period of time. It is the number of new cases of disease per the number of people at risk in a defined period of time of observation. It is usually expressed as cases per 1000 or 100,000 person-years.
- $$\frac{\text{New Cases of Disease } X}{\text{Number of Persons at risk (over a specified time period)}}$$

Mortality

- the incidence of death in a specified population during a specific time period. It may be *crude* (deaths due to all causes) or *disease-specific*.

Case Fatality

- the number of people who die from a particular disease over the total number of persons with that disease or condition. It is usually expressed as a percent.

Risk

■ **Absolute risk :**

- the fundamental measure of risk is disease incidence

■ **Attributable risk (risk difference):**

- the difference between the cumulative incidence of disease in the exposed and unexposed groups.

■ **Relative risk (risk ratio):**

- the incidence of disease in the exposed group divided by the incidence of disease in the unexposed group.

Odds Ratio

- is derived from a case-control study and represents the odds that a case patient has been exposed divided by the odds that a control patient has been exposed.