Introduction to Biostatistics

Objectives

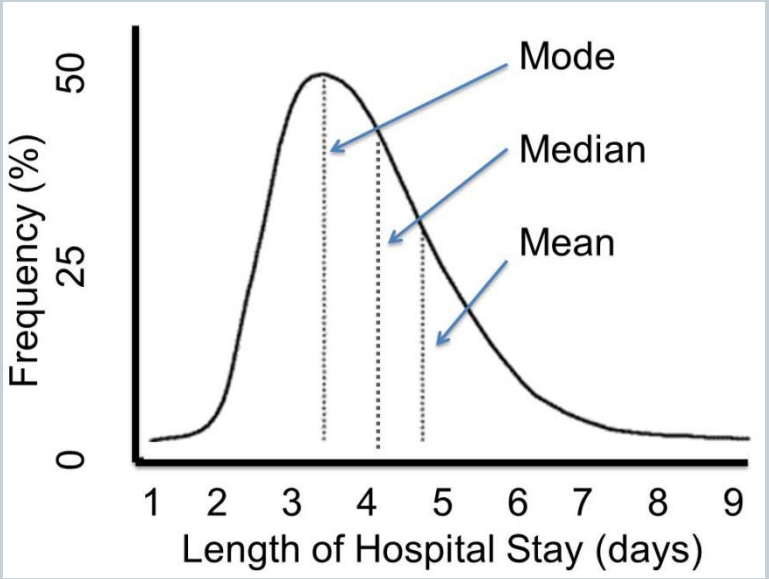
- Review factors that may influence statistical distribution.
- Discuss confidence intervals and their significance.
- Review different types of hypotheses and their role in medical research.

Normal Distribution

- Mathematical abstraction that describes the distribution of observations in a population.
- AKA Normal curve, bell-shaped curve, or Gaussian distribution.
- Quantitative measurements are along the x-axis, and relative frequency (or density function) is along the y-axis.
- 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

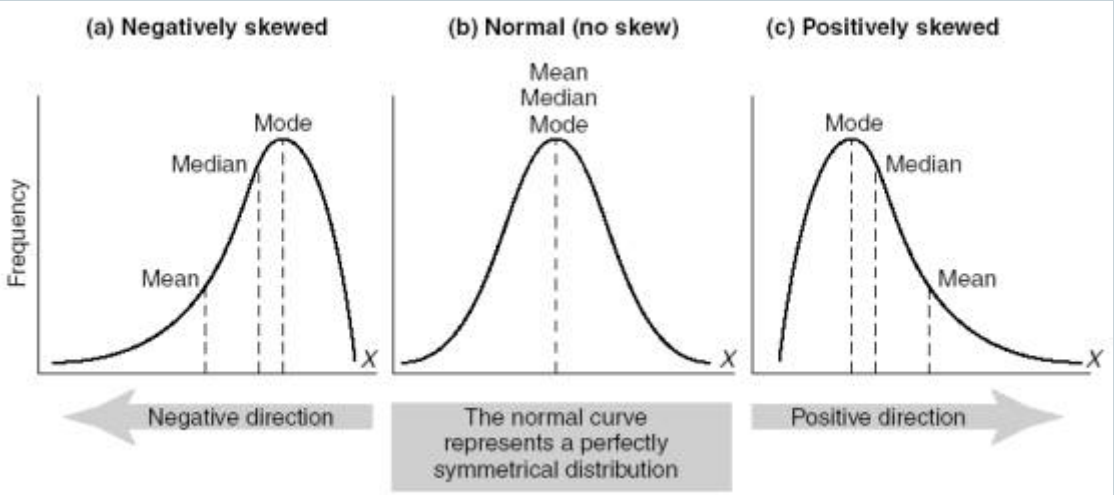
- Mean (numeric average): sum of all values divided by their number.
- Median (50th percentile): value that divides a distribution into two equal parts such that the number of values > the median equals the number of values < the median.
- Mode: value that occurs most frequently among a group of observation.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



Sullivan L, The Role of Probability, http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability8.html.

Skewed Distributions

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



Descriptive statistics lecture, Dr. R. Burke Johnson, Univ of S Alabama.

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

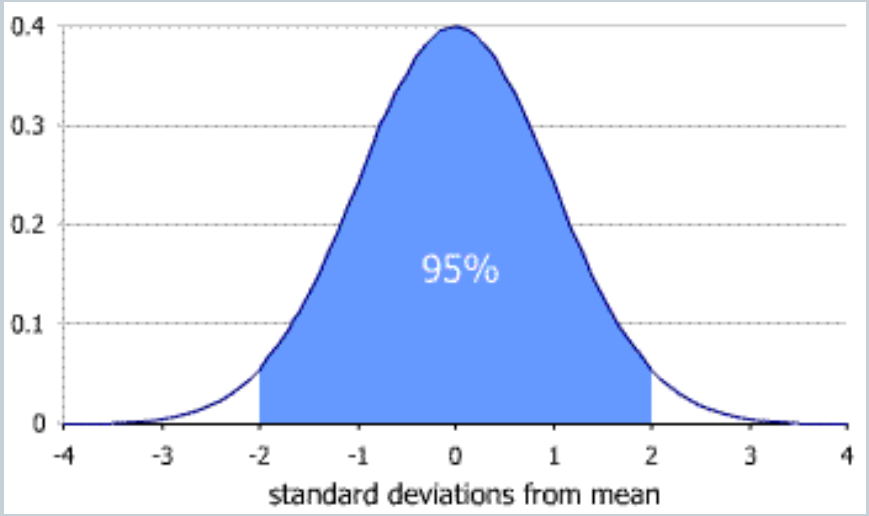
- Standardized measure of the variance of data.
  - If perpendiculars are constructed at a distance of plus or minus 1-standard deviation from the mean, then 68% of the total area (68% of the sample population) is captured; at a distance of plus or minus 2-standard deviations from the mean, 95% of the 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: The mean determines the location along the x-axis, and the standard deviation determines the width of the curve.

Confidence Interval (CI)

- In extrapolating data to 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 (CI).
- Range of values within which a population parameter is likely to occur with a specified probability.

Confidence Interval (CI)

- Help us determine the degree of imprecision in a sample mean, odds ratio, relative risk, etc.
- Wide CI indicates that estimate is imprecise.
- Most commonly used: 95% CI, in which the limits of the interval are 2 standard errors from the mean.



"Reference Range," Wikimedia Commons.

Confidence Interval (CI)

- Two interpretations of the 95% CI:
  - 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 relatively easy to understand as such, and this interpretation is commonly used.)
  - Probabilistic interpretation: in repeated random sample of the population, 95% of all such intervals will, in the long run, contain the true population mean.

Confidence Interval (CI)

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

http://pixabay.com/en/test-tube-lab-medical-research-218181/

Hypotheses

- Scientific hypothesis: testable statement of anticipated findings.
- Statistical hypothesis: formal statement of the scientific hypothesis that includes the parameters being measured (the test statistic). Usually stated as two hypotheses:
  - Null hypothesis: formal statement that the anticipated effect will not be observed.
  - Alternate hypothesis: formal statement of the scientific hypothesis.

Hypothesis Testing

- Hypothesis testing: statistical evaluation of 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).

Hypothesis Testing vs Confidence Intervals

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

P-Value

- Computed statistic that gives the probability that the measured difference between the two groups would occur if the null hypothesis were true.
  - 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

- Degree of difference between two groups that will result in rejection of the null hypothesis.
  - 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).

Level of Significance

- 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.
  - 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.

Prevalence

- Proportion of individuals in a population who have a specific characteristic or disease at a specific point in time.
- Usually expressed as a number per 1,000 or 100,000.
- Appropriate measure for chronic conditions but may underestimate acute diseases.

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

Incidence

- 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 1,000 or 100,000 person-years.

$$\frac{\text{New Cases of Disease X}}{\text{Number of Persons at Risk}}$$
(over a specified time period)

Mortality and Case Fatality

- Mortality: Incidence of death in a specified population during a specific time period.
  - May be crude (deaths due to all causes) or disease-specific.
- Case fatality: number of people who die from a particular disease over the total number of persons with that disease or condition.
  - Usually expressed as a percent.

Risk

- Absolute risk: the fundamental measure of risk is disease incidence.
- Attributable risk (risk difference): difference between the cumulative incidence of disease in the exposed and unexposed groups.
- Relative risk (risk ratio): incidence of disease in the exposed group divided by the incidence of disease in the unexposed group.

Odds Ratio

- Derived from a case-control study.
- Represents the odds that a case patient has been exposed divided by the odds that a control patient has been exposed.

Acknowledgements

- Richard DiCarlo, MD
- Rebecca Frey, PhD
- Stacey Holman, MD
- Richard Tejedor, MD
- Murtuza Ali, MD