Biostatistics: A Refresher

Kevin M. Sowinski, Pharm.D., FCCP

Purdue University College of Pharmacy
Indiana University School of Medicine
West Lafayette and Indianapolis, Indiana

Learning Objectives

1. Describe differences between descriptive and inferential statistics.
2. Identify different types of data (nominal, ordinal, continuous [ratio and interval]) to determine an appropriate type of statistical test (parametric vs. nonparametric).
3. Describe strengths and limitations of different types of measures of central tendency (mean, median, and mode) and data spread (standard deviation, standard error of the mean, range, and interquartile range).
4. Describe the concepts of normal distribution and the associated parameters that describe the distribution.
5. State the types of decision errors that can occur when one is using statistical tests and the conditions under which they can occur.
6. Describe hypothesis testing and state the meaning of and distinguish between p-values and confidence intervals.
7. Describe areas of misuse or misrepresentation that are associated with various statistical methods.
8. Select appropriate statistical tests on the basis of the sample distribution, data type, and study design.
9. Interpret statistical significance for results from commonly used statistical tests.
10. Describe the similarities and differences between statistical tests; state how to apply them appropriately.
11. Identify the use of survival analysis and different ways to perform and report it.

Self-Assessment Questions

Answers and explanations to these questions can be found at the end of the chapter.

1. A randomized controlled trial assesses the effects of heart failure treatment on global functioning in three groups of adults after 6 months of treatment. Investigators wanted to assess global functioning with the New York Heart Association (NYHA) functional classification, an ordered scale from I to IV, and to compare the patient classification after 6 months of treatment. Which statistical test is most appropriate to assess differences in functional classification between the groups?
   A. Kruskal-Wallis.
   B. Wilcoxon signed rank test.
   C. Analysis of variance (ANOVA).
   D. Analysis of covariance (ANCOVA).

2. You are evaluating a randomized, double-blind, parallel-group controlled trial that compares four antihypertensive drugs for their effect on blood pressure. The authors conclude that hydrochlorothiazide is better than atenolol ($p<0.05$) and that enalapril is better than hydrochlorothiazide ($p<0.01$), but no difference is observed between any other drugs. The investigators used an unpaired (independent samples) t-test to test the hypothesis that each drug was equal to the other. Which statement is most appropriate?
   A. Investigators used the appropriate statistical test to analyze their data.
   B. Enalapril is the most effective of these drugs.
   C. ANOVA would have been a more appropriate test.
   D. A paired t-test is a more appropriate test.

3. In the results of a randomized, double-blind, controlled clinical trial, it is reported that the difference in hospital readmission rates between the intervention group and the control group is 6% ($p=0.01$), and it is concluded that there is a statistically significant difference between the groups. Which statement is most consistent with this finding and conclusions?
   A. The chance of making a type I error is 5 in 100.
   B. The trial does not have enough power.
   C. There is a high likelihood of having made a type II error.
   D. The chance of making an alpha error is 1 in 100.

4. You are reading a manuscript that evaluates the impact of obesity on enoxaparin pharmacokinetics. The authors used an unpaired t-test to compare the baseline values of body mass index (BMI) in normal subjects and obese subjects. You are evaluating the use of an unpaired t-test to compare the BMI between the two groups. Which choice

represents the most appropriate criteria to be met to use this parametric test?

A. The sample sizes in the normal and obese subjects should be equal to allow the use of a t-test.
B. A t-test is not appropriate because BMI data are ordinal.
C. The variance of the BMI data needs must be similar in each group.
D. The pre-study power should be at least 90%.

5. You are evaluating the results and discussion of a journal club article to present to the pharmacy residents at your institution. The randomized, prospective, controlled trial evaluated the efficacy of a new controller drug for asthma. The primary end point was the morning forced expiratory volume in 1 second (FEV_1) in two groups of subjects (men and women). The difference in FEV_1 between the two groups was 15% (95% confidence interval [CI], 10%–21%). Which statement is most appropriate, given the results?

A. Without the reporting of a p-value, it is not possible to conclude whether these results were statistically significant.
B. There is a statistically significant difference between the men and women ($p<0.05$).
C. There is a statistically significant difference between the men and women ($p<0.01$).
D. There is no statistically significant difference between the men and women.

6. An early-phase clinical trial of 40 subjects evaluated a new drug known to increase high-density lipoprotein cholesterol (HDL-C) concentrations. The objective of the trial was to compare the new drug's ability to increase HDL-C with that of lifestyle modifications (active control group). At the beginning of the study, the mean baseline HDL-C was 37 mg/dL in the active control group and 38 mg/dL in the new drug group. At the end of the 3-month trial, the mean HDL-C for the control group was 44 mg/dL and for the new drug group, 49 mg/dL. The p-value for the comparison at 3 months was 0.08. Which statement provides the best interpretation of these results?

A. An a priori α of less than 0.10 would have made the study more clinically useful.
B. The new drug and active control appear to be equally efficacious in increasing HDL-C.
C. The new drug is better than lifestyle modifications because it increases HDL-C to a greater extent.
D. This study is potentially underpowered.

7. Researchers planned a study to evaluate the percentage of subjects who achieved less than a target blood pressure (less than 140/90 mm Hg) when initiated on two different doses of amlodipine. In the study of 100 subjects, the amlodipine 5-mg group (n=50) and the amlodipine 10-mg group (n=50) were compared. The investigators used a blood pressure goal as their primary end point, defined as the percentage of subjects who successfully achieved the blood pressure goal at 3 months. Which is the most appropriate statistical test to answer such a question?

A. Independent samples t-test.
B. Chi-square or Fisher exact test.
C. Wilcoxon signed rank test.
D. One-sample t-test.

8. An investigational drug is being compared with an existing drug for the treatment of anemia in patients with chronic kidney disease. The study is designed to detect a minimum 20% difference in response rates between the groups, if one exists, with an a priori α of 0.05 or less. The investigators are unclear whether the 20% difference between response rates is too large and think a smaller difference might be more clinically meaningful. In revising their study, they decide they want to be able to detect a minimum 10% difference in response. Which change in the study parameters is most appropriate?

A. Increase the sample size.
B. Select an α of 0.001 as a cutoff for statistical significance.
C. Select an α of 0.10 as a cutoff for statistical significance.
D. Decrease the sample size.

9. You are designing a new computer alert system to investigate the impact of several factors on the risk of QTc prolongation. You want to develop a model to predict which patients are most likely to experience QTc prolongation after the administration of certain drugs or the presence of certain conditions. You plan to assess the presence or absence of several different variables. Which technique will be most useful in completing such an analysis?

   A. Correlation.

   B. Kaplan-Meier curve.

   C. Regression.

   D. Confidence intervals.

I. INTRODUCTION TO STATISTICS

A. Method for Collecting, Classifying, Summarizing, and Analyzing Data

B. Useful Tool for Quantifying Clinical and Laboratory Data in a Meaningful Way

C. Assists in Determining Whether and by How Much a Treatment or Procedure Affects a Group of Patients

D. Why Pharmacists Need to Know Statistics

E. As Statistics Pertains to Most of You:
   1. *Pharmacotherapy Specialty Examination content outline*
      - Domain 2: Retrieval, Generation, Interpretation, and Dissemination of Knowledge in Pharmacotherapy (25%)
      - Interpret biomedical literature with respect to study design and methodology, statistical analysis, and significance of reported data and conclusions.
      - Knowledge of biostatistical methods, clinical and statistical significance, research hypothesis generation, research design and methodology, and protocol and proposal development

F. Several articles have investigated the various types of statistical tests used in the biomedical literature; the data from one of these articles are illustrated below.

Table 1. Statistical Content of Original Articles in *The New England Journal of Medicine,* 2004–2005

Statistical Procedure	% of Articles Containing Methods	Statistical Procedure	% of Articles Containing Methods
No statistics or descriptive statistics	13	Adjustment and standardization	1
t-tests	26	Multiway tables	13
Contingency tables	53	Power analyses	39
Nonparametric tests	27	Cost-benefit analysis	<1
Epidemiologic statistics	35	Sensitivity analysis	6
Pearson correlation	3	Repeated-measures analysis	12
Simple linear regression	6	Missing data methods	8
Analysis of variance	16	Noninferiority trials	4
Transformation	10	Receiver operating characteristics	2
Nonparametric correlation	5	Resampling	2
Survival methods	61	Principal component and cluster analyses	2
Multiple regression	51	Other methods	4
Multiple comparisons	23		

Table 2. Statistical Content of Original Articles from Six Major Medical Journals from January to March 2005 (n=239 articles).

Statistical Test	No. (%)	Statistical Test	No. (%)
Descriptive statistics (mean, median, frequency, SD, and IQR)	219 (91.6)	Others	
Simple statistics	120 (50.2)	Intention-to-treat analysis	42 (17.6)
Chi-square analysis	70 (29.3)	Incidence or prevalence	39 (16.3)
t-test	48 (20.1)	Relative risk or risk ratio	29 (12.2)
Kaplan-Meier analysis	48 (20.1)	Sensitivity analysis	21 (8.8)
Wilcoxon rank sum test	38 (15.9)	Sensitivity or specificity	15 (6.3)
Fisher exact test	33 (13.8)		
Analysis of variance	21 (8.8)		
Correlation	16 (6.7)		
Multivariate analysis	164 (68.6)		
Cox proportional hazards	64 (26.8)		
Multiple logistic regression	54 (22.6)		
Multiple linear regression	7 (2.9)		
Other regression analysis	38 (15.9)		
None	5 (2.1)		

IQR = interquartile range; SD = standard deviation.

Articles published in American Journal of Medicine, Annals of Internal Medicine, BMJ, JAMA, Lancet, and The New England Journal of Medicine. Table modified from JAMA 2007;298:1010-22.

II. TYPES OF VARIABLES AND DATA

A. Definition: Random variables—A variable with observed values that may be considered outcomes of an experiment and whose values cannot be anticipated with certainty before the experiment is conducted

B. Two Types of Random Variables
   1. Discrete variables (e.g., dichotomous, categorical)
   2. Continuous variables

C. Discrete Variables
   1. Can take only a limited number of values within a given range
   2. Nominal: Classified into groups in an unordered manner and with no indication of relative severity (e.g., male/female sex, mortality [dead or alive], disease presence [yes or no], race, marital status). These data are often expressed as a frequency or proportion.
   3. Ordinal: Ranked in a specific order but with no consistent level of magnitude of difference between ranks (e.g., New York Heart Association [NYHA] functional class describes the functional status of patients with heart failure, and subjects are classified in increasing order of symptoms: I, II, III, IV; Likert-type scales)
   4. Common error: Measure of central tendency—In most cases, means and standard deviations (SDs) should not be reported with ordinal data. What is a common incorrect use of means and SDs to show ordinal data?

D. Continuous Variables, Sometimes Called Counting Variables
   1. Continuous variables can take on any value within a given range.
   2. Interval: Data are ranked in a specific order with a consistent change in magnitude between units; the zero point is arbitrary (e.g., degrees Fahrenheit).
   3. Ratio: Like interval but with an absolute zero (e.g., degrees Kelvin, heart rate, blood pressure, time, distance)

III. TYPES OF STATISTICS

A. Descriptive Statistics: Used to summarize and describe data that are collected or generated in research studies. This is done both visually and numerically.
   1. Visual methods of describing data
      a. Frequency distribution
      b. Histogram
      c. Scatterplot
   2. Numerical methods of describing data: Measures of central tendency
      a. Arithmetic mean (i.e., average)
         i. Sum of all values divided by the total number of values
         ii. Should generally be used only for continuous and normally distributed data
         iii. Very sensitive to outliers and tend toward the tail, which has the outliers
         iv. Most commonly used and most understood measure of central tendency
         v. Geometric mean
      b. Median
         i. Midpoint of the values when placed in order from highest to lowest. Half of the observations are above and below. When there are an even number of observations, it is the mean of the two middle values.
         ii. Also called 50th percentile
         iii. Can be used for ordinal or continuous data (especially good for skewed populations)
         iv. Insensitive to outliers
      c. Mode
         i. Most common value that occurs in a distribution
         ii. Can be used for nominal, ordinal, or continuous data
         iii. Sometimes, there may be more than one mode (e.g., bimodal, trimodal).
         iv. Does not help describe meaningful distributions with a large range of values, each of which occurs infrequently
   3. Numerical methods of describing data: Measures of data spread or variability
      a. Standard deviation
         i. Measure of the variability around the mean; most common measure used to describe the spread of data
         ii. Square root of the variance (average squared difference of each observation from the mean), so the SD is reported in the original units (nonsquared).
         iii. Appropriately applied only to continuous data that are normally or near-normally distributed or that can be transformed to be normally distributed
         iv. By the empirical rule, 68% of the sample values are found within ±1 SD, 95% are found within ±2 SD, and 99% are found within ±3 SD.
         v. The coefficient of variation relates the mean and the SD (SD/mean × 100%).

b. Range
   i. Difference between the smallest and largest value in a data set; does not give a tremendous amount of information by itself
   ii. Easy to compute (simple subtraction)
   iii. Size of range is very sensitive to outliers.
   iv. Often reported as the actual values rather than the difference between the two extreme values
c. Percentiles
   i. The point (value) in a distribution in which a value is larger than some percentage of the other values in the sample; can be calculated by ranking all data in a data set
   ii. The 75th percentile lies at a point at which 75% of the other values are smaller.
   iii. Does not assume the population has a normal distribution (or any other distribution)
   iv. The interquartile range (IQR) is an example of the use of percentiles to describe the middle 50% values. The IQR encompasses the 25th–75th percentile.

4. Presenting data using only measures of central tendency can be misleading without some idea of data spread. Studies that report only medians or means without their accompanying measures of data spread should be closely scrutinized. What are the measures of spread that should be used with means and medians?
5. Example data set

Table 3. Twenty Baseline HDL-C Concentrations from an Experiment Evaluating the Impact of Green Tea on HDL-C

64	60	59	65	64	62	54
54	68	67	79	55	48	65
59	65	87	49	46	46	

HDL-C = high-density lipoprotein cholesterol.

a. Calculate the mean, median, and mode of the above data set.
b. Calculate the range, SD (will not have to do this by hand), and standard error of the mean (SEM) of the above data set (we will describe this later).
c. Evaluate the visual presentation of the data.

B. Inferential Statistics
1. Conclusions or generalizations made about a population (large group) from the study of a sample of that population
2. Choosing and evaluating statistical methods depend, in part, on the type of data used.
3. An educated statement about an unknown population is commonly referred to in statistics as an inference.
4. Statistical inference can be made by estimation or hypothesis testing.

IV. POPULATION DISTRIBUTIONS

A. Discrete Distributions
   1. Binomial distribution
   2. Poisson distribution

B. Normal (Gaussian) Distribution
   1. Most common model for population distributions
   2. Symmetric or bell-shaped frequency distribution
   3. Landmarks for continuous, normally distributed data
      a. μ: Population mean
      b. σ: Population SD
      c. *x* and *s* represent the sample mean and SD.
   4. When measuring a random variable in a large-enough sample of any population, some values will occur more often than will others.
   5. A visual check of a distribution can help determine whether it is normally distributed (whether it appears symmetric and bell shaped). Need the data to perform these checks:
      a. Frequency distribution and histograms (visually look at the data; you should do this anyway)
      b. Median and mean will be about equal for normally distributed data (most practical and easiest to use).
      c. Formal test: Kolmogorov-Smirnov test
      d. More challenging to evaluate this when we do not have access to the data (when we are reading an article), because most articles do not present all data or both the mean and median
   6. The parameters mean and SD completely define a normally distributed population.
   7. Probability: The likelihood that any one event will occur given all the possible outcomes
   8. Estimation and sampling variability
      a. One method that can be used to make an inference about a population parameter
      b. Separate samples (even of the same size) from a single population will give slightly different estimates.
      c. The distribution of means from these separate random samples approximates a normal distribution.
         i. The mean of this "distribution of means" = the unknown population mean, μ.
         ii. The SD of the means is estimated by the SEM, which conceptually represents the variability of the distribution of means.
         iii. As in any normal distribution, 95% of the sample means lie within ±2 SEM of the population mean.
      d. The distribution of means from these random samples is about normal regardless of the underlying population distribution (central limit theorem). You will get slightly different mean and SD values each time you repeat this experiment.
      e. The SEM is estimated for a single sample by dividing the SD by the square root of the sample size (n). The SEM quantifies uncertainty in the estimate of the mean, not variability in the sample. Important for hypothesis testing and 95% CI estimation
      f. Why is all of this information about the difference between the SEM and SD worth knowing?
         i. Calculation of CIs (95% CI is about mean ± 2 times the SEM)
         ii. Hypothesis testing
         iii. Deception (e.g., makes results look less "variable," especially when used in graphic format)
   9. Recall the previous example about HDL-C and green tea. From the calculated values in section III, do these data appear to be normally distributed?

V. CONFIDENCE INTERVALS

A. Commonly Reported as a Way to Estimate a Population Parameter
   1. In the medical literature, 95% CIs are the most commonly reported CIs. In repeated samples, 95% of all CIs include true population value (i.e., the likelihood or confidence [or probability] that the population value is contained within the interval). In some cases, 90% or 99% CIs are reported.
   2. Why are 95% CIs most often reported?
      a. Assume a baseline birth weight in a group n = 13 with a mean ± SD of 1.18 ± 0.4 kg.
      b. 95% CI is about equal to the mean ± 1.96 × SEM (or 2 × SEM). In reality, it depends on the distribution being used and is a bit more complicated.
      c. What is the 95% CI? It is (1.07–1.29), meaning there is 95% certainty that the true mean of the entire population studied is between 1.07 and 1.29 kg.
      d. What is the 90% CI? The 90% CI is calculated to be (1.09–1.27). Of note, the 95% CI will always be wider than the 90% CI for any given sample. Therefore, the wider the CI, the more likely it is to encompass the true population mean.
   3. The differences between the SD, SEM, and CIs should be noted when interpreting the literature because they are often used interchangeably. Although it is common for CIs to be confused with SDs, the information each provides is quite different and must be assessed correctly.
   4. Recall the previous example about HDL-C and green tea. What is the 95% CI of the data set, and what does that mean?

B. CIs can also be used for any sample estimate. Estimates derived from categorical data such as risk, risk differences, and risk ratios are often presented with the CI and will be discussed later.

C. CIs Instead of Hypothesis Testing
   1. Hypothesis testing and calculation of p-values tell us (ideally) whether there is or is not a statistically significant difference between groups, but they do not tell us anything about the magnitude of the difference.
   2. CIs help us determine the importance of a finding or findings, which we can apply to a situation.
   3. CIs give us an idea of the magnitude of the difference between groups and the statistical significance.
   4. CIs are a "range" of data, together with a point estimate of the difference.
   5. Wide CIs
      a. Many results are possible, either larger or smaller than the point estimate provided by the study.
      b. All values contained in the CI are statistically plausible.
   6. If the estimate is the difference between two continuous variables, a CI that includes zero (no difference between two variables) can be interpreted as not statistically significant (a p-value of 0.05 or greater). There is no need to show both the 95% CI and the p-value.
   7. The interpretation of CIs for odds ratios and relative risks is somewhat different. In that case, a value of 1 indicates no difference in risk, and if the CI includes 1, there is no statistical difference. (See the discussions of case-control and cohort in other sections for how to interpret CIs for odds ratios and relative risks.)

VI. HYPOTHESIS TESTING

A. Null and Alternative Hypotheses (See Table 4 for other types of examples.)
   1. Null hypothesis (H_0): Example: No difference between groups being compared (treatment A = treatment B)
   2. Alternative hypothesis (Ha): Example: Opposite of null hypothesis; states that there is a difference (treatment A ≠ treatment B)
   3. The structure or the manner in which the hypothesis is written dictates which statistical test is used. Two-sample t-test: H_0: mean 1 = mean 2
   4. Used to assist in determining whether any observed differences between groups can be explained by chance
   5. Tests for statistical significance (hypothesis testing) determine whether the data are consistent with H_0 (no difference).
   6. The results of the hypothesis testing will indicate whether enough evidence exists for H_0 to be rejected.
      a. If H_0 is rejected = statistically significant difference between groups (unlikely attributable to chance)
      b. If H_0 is not rejected = no statistically significant difference between groups (any "apparent" differences may be attributable to chance). Note that we are not concluding that the treatments are equal.
   7. Types of hypothesis testing. These are situations in which two groups are being compared. There are numerous other examples of situations these procedures could be applied to.

Table 4. Types of Hypothesis Testing

	Question	**Hypothesis**	**Method**
Non-directional			
Difference	Are the means different?	H_0: $Mean_1 = Mean_2$ H_A: $Mean_1 \neq Mean_2$ OR H_0: $Mean_1 - Mean_2 = 0$ H_A: $Mean_1 - Mean_2 \neq 0$	Traditional 2-sided t-test Confidence intervals
Equivalence	Are the means practically equivalent?	H_0: $Mean_1 - Mean_2 \geq \Delta$ H_A: $Mean_1 - Mean_2 < \Delta$	Two 1-sided t-test procedures Confidence intervals
Directional			
Superiority	Is mean 1 > mean 2? (or some other similarly worded question)	H_0: $Mean_1 \leq Mean_2$ H_A: $Mean_1 > Mean_2$ or H_0: $Mean_1 - Mean_2 \leq 0$ H_A: $Mean_1 - Mean_2 > 0$	Traditional 1-sided t-test Confidence intervals
Noninferiority	Is mean 1 no more than a certain amount lower than mean 2?	H_0: $Mean_1 - Mean_2 \geq \Delta$ H_A: $Mean_1 - Mean_2 < \Delta$	Confidence intervals

B. To Determine What Is Sufficient Evidence to Reject H_0: Set the a priori significance level (α) and generate the decision rule.
   1. Developed after the research question has been stated in hypothesis form
   2. Used to determine the level of acceptable error caused by a false positive (also known as level of significance)
      a. Convention: A priori α is usually 0.05.
      b. Critical value is calculated, capturing how extreme the sample data must be to reject H_0.

C. Perform the Experiment and Estimate the Test Statistic.
   1. A test statistic is calculated from the observed data in the study, which is compared with the critical value.
   2. Depending on this test statistic's value, H_0 is not rejected (often called fail to reject) or rejected.
   3. In general, the test statistic and critical value are not presented in the literature; instead, p-values are generally reported and compared with a priori α values to assess statistical significance. p-Value: Probability of obtaining a test statistic and critical value as extreme, or more extreme, than the one actually obtained
   4. Because computers are used in these tests, this step is often transparent; the p-value estimated in the statistical test is compared with the a priori α (usually 0.05), and the decision is made.

VII. STATISTICAL TESTS AND CHOOSING A STATISTICAL TEST

A. Which Tests Do You Need to Know?

B. Choosing the Appropriate Statistical Test Depends on
   1. Type of data (nominal, ordinal, or continuous)
   2. Distribution of data (e.g., normal)
   3. Number of groups
   4. Study design (e.g., parallel, crossover)
   5. Presence of confounding variables
   6. One-tailed versus two-tailed
   7. Parametric versus nonparametric tests
      a. Parametric tests assume:
         i. Data being investigated have an underlying distribution that is normal or close to normal or, more correctly, randomly drawn from a parent population with a normal distribution.
         ii. Data measured are continuous data, measured on either an interval or a ratio scale.
         iii. Parametric tests assume that the data being investigated have variances that are homogeneous between the groups investigated. This is often called homoscedasticity.
      b. Nonparametric tests are used when data are not normally distributed or do not meet other criteria for parametric tests (e.g., discrete data).d

C. Parametric Tests
   1. Student t-test: Several different types
      a. One-sample test: Compares the mean of the study sample with the population mean

Group 1	Known population mean

b. Two-sample, independent samples, or unpaired test: Compares the means of two independent samples. This is an independent samples test.

Group 1	Group 2

i. Equal variance test
   (a) Rule for variances: If the ratio of larger variance to smaller variance is greater than 2, we generally conclude the variances are different.
   (b) Formal test for differences in variances: F test
   (c) Adjustments can be made for cases of unequal variance.

ii. Unequal variance

c. Paired test: Compares the mean difference of paired or matched samples. This is a related samples test.

Group 1	
Measurement 1	Measurement 2

d. Common error: Use of multiple t-tests with more than two groups.

2. Analysis of variance (ANOVA): A more generalized version of the t-test that can apply to more than two groups

a. One-way ANOVA: Compares the means of three or more groups in a study. Also known as single-factor ANOVA. This is an independent samples test.

Group 1	Group 2	Group 3

b. Two-way ANOVA: Additional factor (e.g., age) added

Young groups	Group 1	Group 2	Group 3
Old groups	Group 1	Group 2	Group 3

c. Repeated-measures ANOVA: This is a related samples test.

	Related Measurements		
Group 1	Measurement 1	Measurement 2	Measurement 3

d. Several more complex factorial ANOVAs can be used.

e. Many comparison procedures are used to determine which groups actually differ from each other. Post hoc tests: Tukey HSD (honestly significant difference), Bonferroni, Scheffé, Newman-Keuls

3. Analysis of covariance (ANCOVA): Provides a method to explain the influence of a categorical variable (independent variable) on a continuous variable (dependent variable) while statistically controlling for other variables (confounding)

D. Nonparametric Tests

1. These tests may also be used for continuous data that do not meet the assumptions of the t-test or ANOVA.
2. Tests for independent samples

a. Wilcoxon rank sum, Mann-Whitney *U* test, or Wilcoxon-Mann-Whitney Test: Compare two independent samples (related to a t-test)

b. Kruskal-Wallis one-way ANOVA by ranks
   i. Compares three or more independent groups (related to one-way ANOVA)
   ii. Post hoc testing

3. Tests for related or paired samples
   a. Sign test and Wilcoxon signed rank test: Compares two matched or paired samples (related to a paired t-test)
   b. Friedman ANOVA by ranks: Compares three or more matched or paired groups

E. Nominal Data
1. Chi-square (χ^2) test: Compares expected and observed proportions between two or more groups
   a. Test of independence
   b. Test of goodness of fit
2. Fisher exact test: Specialized version of the chi-square test for small groups (cells) containing less than five predicted observations
3. McNemar: Paired samples
4. Mantel-Haenszel: Controls for the influence of confounders

F. Correlation and Regression (see section IX)

G. Choosing the Most Appropriate Statistical Test: Example 1
1. A trial was conducted to determine whether rosuvastatin was better than simvastatin at lowering low-density lipoprotein cholesterol (LDL-C) concentrations. The trial was designed such that the subjects' baseline characteristics were as comparable as possible with each other. *The intended primary end point for this 3-month trial was the difference in LDL-C between the two drugs.* The results of the trial are reported as follows:

Table 5. Rosuvastatin and Simvastatin Effect on LDL-C

	Rosuvastatin (n=25)	**Simvastatin (n=25)**
Men/women	12/13	10/15
Smokers	10	13
Baseline LDL-C (mg/dL)	152 ± 5	151 ± 4
Final LDL-C (mg/dL)	138 ± 7	135 ± 5

LDL-C = low-density lipoprotein cholesterol.

2. Which is the appropriate statistical test to determine baseline differences in:
   a. Sex distribution?
   b. LDL-C?
   c. Percentage of smokers and nonsmokers?
3. Which is the appropriate statistical test to determine:
   a. The effect of rosuvastatin on LDL-C?
   b. The primary end point?
4. The authors concluded that rosuvastatin is similar to simvastatin. What else would you like to know in evaluating this study?

VIII. DECISION ERRORS

Table 6. Summary of Decision Errors

	Underlying "Truth" or Reality	
Test Result	**H_0 *Is True* (No difference)**	**H_0 *Is False* (Difference)**
Accept H_0 (no difference)	No error (correct decision)	Type II error (beta error)
Reject H_0 (difference)	Type I error (alpha error)	No error (correct decision)

H_0 = null hypothesis.

A. Type I Error: The probability of making this error is defined as the significance level α.
   1. Convention is to set the α to 0.05, effectively meaning that, 1 in 20 times, a type I error will occur when the H_0 is rejected. So, 5.0% of the time, a researcher will conclude that there is a statistically significant difference when one does not actually exist.
   2. The calculated chance that a type I error has occurred is called the p-value.
   3. The p-value tells us the likelihood of obtaining a given (or a more extreme) test result if the H_0 is true. When the α level is set a priori, H_0 is rejected when p is less than α. In other words, the p-value tells us the probability of being wrong when we conclude that a true difference exists (false positive).
   4. A lower p-value does not mean the result is more important or more meaningful but only that it is statistically significant and not likely to be attributable to chance.

B. Type II Error: The probability of making this error is called β.
   1. Concluding that no difference exists when one truly does (not rejecting H_0 when it should be rejected)
   2. It has become a convention to set β to between 0.20 and 0.10.

C. Power (1 − β)
   1. The probability of making a correct decision when H_0 is false; the ability to detect differences between groups if one actually exists
   2. Dependent on the following factors:
      a. Predetermined α
      b. Sample size
      c. The size of the difference between the outcomes you want to detect, called the effect size. Often not known before the experiment is conducted, so to estimate the power of your test, you will have to specify how large a change is worth detecting.
      d. The variability of the outcomes that are being measured
      e. Items c and d are generally determined from previous data or the literature.
   3. Power is decreased by (in addition to the above criteria):
      a. Poor study design
      b. Incorrect statistical tests (use of nonparametric tests when parametric tests are appropriate)
   4. Statistical power analysis and sample size calculation
      a. Related to above discussion of power and sample size
      b. Sample size estimates should be performed in all studies a priori.
      c. Necessary components for estimating appropriate sample size
         i. Acceptable type II error rate (usually 0.10–0.20)
         ii. Observed difference in predicted study outcomes that is clinically significant

iii. The expected variability in item ii
iv. Acceptable type I error rate (usually 0.05)
v. Statistical test that will be used for primary end point

5. Statistical significance versus clinical significance
   a. As stated earlier, the size of the p-value is not necessarily related to the clinical importance of the result. Smaller values mean only that chance is less likely to explain observed differences.
   b. Statistically significant does not necessarily mean clinically significant.
   c. Lack of statistical significance does not mean that results are not clinically important.
   d. When considering nonsignificant findings, consider sample size, estimated power, difference study was powered to detect, and observed variability and observed variability.

IX. CORRELATION AND REGRESSION

A. Introduction: Correlation Versus Regression
   1. Correlation examines the strength of the association between two variables. It does not necessarily assume that one variable is useful in predicting the other.
   2. Regression examines the ability of one or more variables to predict another variable.

B. Pearson Correlation
   1. The strength of the relationship between two variables that are normally distributed, ratio or interval scaled, and linearly related is measured with a correlation coefficient.
   2. Often referred to as the degree of association between the two variables
   3. Does not necessarily imply that one variable is dependent on the other (regression analysis will do that)
   4. Pearson correlation (r) ranges from −1 to +1 and can take any value in between:

−1	0	+1
Perfect negative linear relationship	No linear relationship	Perfect positive linear relationship

   5. Hypothesis testing is performed to determine whether the correlation coefficient is different from zero. This test is highly influenced by sample size.

C. Pearls About Correlation
   1. The closer the magnitude of r to 1 (either + or −), the more highly correlated the two variables. The weaker the relationship between the two variables, the closer r is to 0.
   2. There is no agreed-on or consistent interpretation of the value of the correlation coefficient. It is dependent on the environment of the investigation (laboratory vs. clinical experiment).
   3. Pay more attention to the magnitude of the correlation than to the p-value because it is influenced by sample size.
   4. Crucial to the proper use of correlation analysis is interpretation of the graphic representation of the two variables. Before using correlation analysis, it is essential to generate a scatterplot of the two variables to visually examine the relationship.

D. Spearman Rank Correlation: Nonparametric test that quantifies the strength of an association between two variables but does not assume a normal distribution of continuous data. Can be used for ordinal data or nonnormally distributed continuous data

E. Regression
   1. A statistical technique related to correlation. There are many different types; for simple linear regression, one continuous outcome (dependent) variable and one continuous independent (causative) variable
   2. Two main purposes of regression: (1) development of prediction model and (2) accuracy of prediction
   3. Prediction model: Making predictions of the dependent variable from the independent variable; $Y = mx + b$ (dependent variable = slope × independent variable + intercept)
   4. Accuracy of prediction: How well the independent variable predicts the dependent variable. Regression analysis determines the extent of variability in the dependent variable that can be explained by the independent variable.
      a. Coefficient of determination (r^2) measured describing this relationship. Values of r^2 can range from 0 to 1.
      b. An r^2 of 0.80 could be interpreted as saying that 80% of the variability in *Y* is explained by the variability in *X*.
      c. This does not provide a mechanistic understanding of the relationship between *X* and *Y* but rather a description of how clearly such a model (linear or otherwise) describes the relationship between the two variables.
      d. Like the interpretation of r, the interpretation of r^2 depends on the scientific arena (e.g., clinical research, basic research, social science research) to which it is applied.
   5. For simple linear regression, two statistical tests can be used.
      a. To test the hypothesis that the *y*-intercept differs from zero
      b. To test the hypothesis that the slope of the line is different from zero
   6. Regression is useful in constructing predictive models. The literature is full of examples of predictions. The process involves developing a formula for a regression line that best fits the observed data.
   7. There are many different types of regression analysis.
      a. Multiple linear regression: One continuous independent variable and two or more continuous dependent variables
      b. Simple logistic regression: One categorical response variable and one continuous or categorical explanatory variable
      c. Multiple logistic regression: One categorical response variable and two or more continuous or categorical explanatory variables
      d. Nonlinear regression: Variables are not linearly related (or cannot be transformed into a linear relationship). This is where our pharmacokinetic equations come from.
      e. Polynomial regression: Any number of response and continuous variables with a curvilinear relationship (e.g., cubed, squared)
   8. Example of regression
      a. The following data are taken from a study evaluating enoxaparin use. The authors were interested in predicting patient response (measured as antifactor Xa concentrations) from the enoxaparin dose in the 75 subjects who were studied.

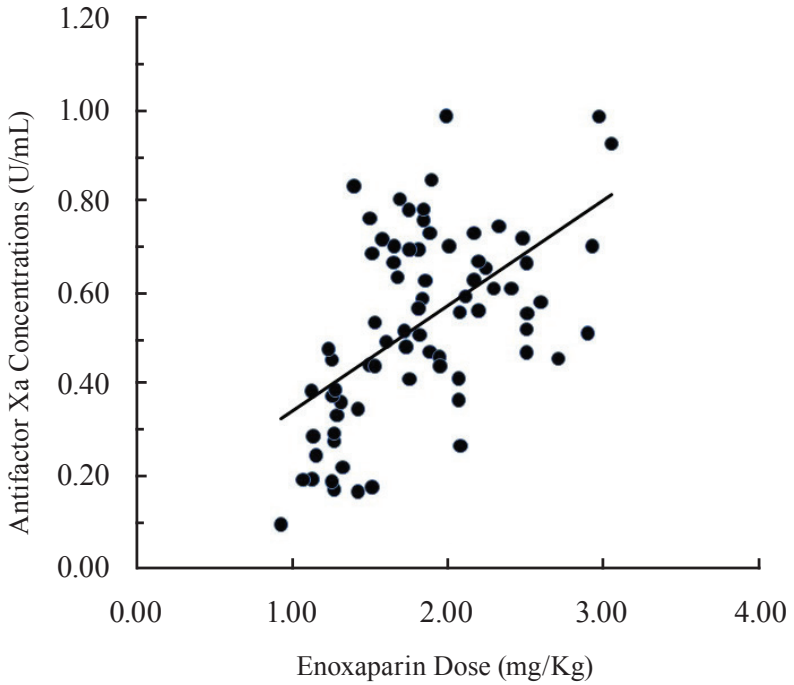

Figure 1. Relationship between antifactor Xa concentrations and enoxaparin dose.

b. The authors performed regression analysis and reported the following: slope, 0.227; *y*-intercept, 0.097; $p<0.05$; $r^2 = 0.31$.
c. Answer the following questions:
   i. What are the necessary assumptions to use regression analysis?
   ii. Provide an interpretation of the coefficient of determination.
   iii. Predict antifactor Xa concentrations at enoxaparin doses of 2 and 3.75 mg/kg.
   iv. What does the $p<0.05$ value indicate?

X. SURVIVAL ANALYSIS

A. Studies the Time Between Entry in a Study and Some Event (e.g., death, myocardial infarction)
   1. Censoring makes survival methods unique; considers that some subjects leave the study for reasons other than the event (e.g., lost to follow-up, end of study period).
   2. Considers that all subjects do not enter the study at the same time
   3. Standard methods of statistical analysis such as t-tests and linear or logistic regression may not be appropriately applied to survival data because of censoring.

B. Estimating the Survival Function
   1. Kaplan-Meier method
      a. Uses survival times (or censored survival times) to estimate the proportion of people who would survive a given length of time under the same circumstances
      b. Allows the production of a table ("life table") and a graph ("survival curve")
      c. We can visually evaluate the curves, but we need a test to evaluate them formally.

2. Log-rank test: Compare the survival distributions between two or more groups.
   a. This test precludes an analysis of the effects of several variables or the magnitude of difference between groups or the CI (see below for Cox proportional hazards model).
   b. H_0: No difference in survival between the two populations
   c. Log-rank test uses several assumptions:
      i. Random sampling and subjects chosen independently
      ii. Consistent criteria for entry or end point
      iii. Baseline survival rate does not change as time progresses.
      iv. Censored subjects have the same average survival time as uncensored subjects.
3. Cox proportional hazards model
   a. Most popular method to evaluate the impact of covariates; reported (graphically) like Kaplan-Meier
   b. Investigates several variables at a time
   c. Actual method of construction and calculation is complex.
   d. Compares survival in two or more groups after other variables are adjusted for
   e. Allows calculation of a hazard ratio (and CI)

XI. SELECTED REPRESENTATIVE STATISTICAL TESTS

Table 7. Representative Statistical Tests

Type of Variable	2 Samples (independent)	2 Samples (related)	>2 Samples (independent)	>2 Samples (related)
Nominal	χ^2 or Fisher exact test	McNemar test	χ^2	Cochran *Q*
Ordinal	Wilcoxon rank sum Mann-Whitney *U* test Wilcoxon-Mann-Whitney	Wilcoxon signed rank Sign test	Kruskal-Wallis (MCP)	Friedman ANOVA
Continuous No factors	Equal variance t-test Unequal variance t-test	Paired t-test	1-way ANOVA (MCP)	Repeated-measures ANOVA
1 factor	ANCOVA	2-way repeated-measures ANOVA	2-way ANOVA (MCP)	2-way repeated-measures ANOVA

ANCOVA = analysis of covariance; ANOVA = analysis of variance; MCP = multiple comparisons procedure.

REFERENCES

1. Crawford SL. Correlation and regression. Circulation 2006;114:2083-8.
2. Davis RB, Mukamal KJ. Hypothesis testing: means. Circulation 2006;114:1078-82.
3. DeYoung GR. Understanding biostatistics: an approach for the clinician. In: Zarowitz B, Shumock G, Dunsworth T, et al., eds. Pharmacotherapy Self-Assessment Program, 5th ed. Kansas City, MO: ACCP, 2005:1-20.
4. DiCenzo R, ed. Clinical Pharmacist's Guide to Biostatistics and Literature Evaluation. Lenexa, KS: ACCP, 2010.
5. Gaddis ML, Gaddis GM. Introduction to biostatistics. Part 1, basic concepts. Ann Emerg Med 1990;19:86-9.
6. Gaddis ML, Gaddis GM. Introduction to biostatistics. Part 2, descriptive statistics. Ann Emerg Med 1990;19:309-15.
7. Gaddis ML, Gaddis GM. Introduction to biostatistics. Part 3, sensitivity, specificity, predictive value, and hypothesis testing. Ann Emerg Med 1990;19:591-7.
8. Gaddis ML, Gaddis GM. Introduction to biostatistics. Part 4, statistical inference techniques in hypothesis testing. Ann Emerg Med 1990;19:820-5.
9. Gaddis ML, Gaddis GM. Introduction to biostatistics. Part 5, statistical inference techniques for hypothesis testing with nonparametric data. Ann Emerg Med 1990;19:1054-9.
10. Gaddis ML, Gaddis GM. Introduction to biostatistics. Part 6, correlation and regression. Ann Emerg Med 1990;19:1462-8.
11. Harper ML. Biostatistics for the clinician. In: Zarowitz B, Shumock G, Dunsworth T, et al., eds. Pharmacotherapy Self-Assessment Program, 4th ed. Kansas City, MO: ACCP, 2002:183-200.
12. Hayney MS, Meek PD. Essential clinical concepts of biostatistics. In: Carter BL, Lake KD, Raebel MA, et al., eds. Pharmacotherapy Self-Assessment Program, 3rd ed. Kansas City, MO: ACCP, 1999:19-46.
13. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20:453-8.
14. Kier KL. Biostatistical methods in epidemiology. Pharmacotherapy 2011;31:9-22.
15. Kusuoka H, Hoffman JIE. Advice on statistical analysis for circulation research. Circ Res 2002;91:662-71.
16. Larson MG. Descriptive statistics and graphical displays. Circulation 2006;114:76-81.
17. Larson MG. Analysis of variance. Circulation 2008;117:115-21.
18. Overholser BR, Sowinski KM. Biostatistics primer. Part 1. Nutr Clin Pract 2007;22:629-35.
19. Overholser BR, Sowinski KM. Biostatistics primer. Part 2. Nutr Clin Pract 2008;23:76-84.
20. Rao SR, Schoenfeld DA. Survival methods. Circulation 2007;115:109-13.
21. Rector TS, Hatton RC. Statistical concepts and methods used to evaluate pharmacotherapy. In: Zarowitz B, Shumock G, Dunsworth T, et al., eds. Pharmacotherapy Self-Assessment Program, 2nd ed. Kansas City, MO: ACCP, 1997:130-61.
22. Strassels SA. Biostatistics. In: Dunsworth TS, Richardson MM, Chant C, et al., eds. Pharmacotherapy Self-Assessment Program, 6th ed. Lenexa, KS: ACCP, 2007:1-16.
23. Sullivan LM. Estimation from samples. Circulation 2006;114:445-9.
24. Tsuyuki RT, Garg S. Interpreting data in cardiovascular disease clinical trials: a biostatistical toolbox. In: Richardson MM, Chant C, Cheng JWM, et al., eds. Pharmacotherapy Self-Assessment Program, 7th ed. Lenexa, KS: ACCP, 2010:241-55.
25. Windish DM, Huot SJ, Green ML. Medicine resident's understanding of the biostatistics and results in the medical literature. JAMA 2007;298:1010-22.
26. Horton NJ, Switzer SS. Statistical methods in the journal. N Engl J Med. 2005;353:1977-79.

ANSWERS AND EXPLANATIONS TO SELF-ASSESSMENT QUESTIONS

1. Answer: A

The NYHA functional class is an ordinal scale from I (no symptoms) to IV (severe symptoms). Neither ANOVA nor ANCOVA is appropriate for ordinal or noncontinuous data (Answer C and Answer D are incorrect). The Wilcoxon signed rank test is an appropriate nonparametric test to use for paired ordinal data, such as the change in NYHA functional class over time on the same person (Answer B is incorrect). The Kruskal-Wallis test is the nonparametric analog of a one-way ANOVA and is appropriate for this analysis (Answer A is correct).

2. Answer: C

You cannot determine which finding is more important (in this case, the best drug) on the basis of the p-value (i.e., a lower p-value does not mean more important) (Answer B is incorrect). All statistically significant results are interpreted as significant without respect to the size of the p-value. This trial had four independent samples, and use of the unpaired (independent samples) t-test is not appropriate because it requires several unnecessary tests and increases the chances of making a type I error (Answer A is incorrect). In this setting, ANOVA is the correct test (Answer C is correct), followed by a multiple comparisons procedure to determine where the actual differences between groups lie. A paired t-test is inappropriate because this is a parallel-group trial (Answer D is incorrect). The use of ANOVA in this case assumes a normal distribution and equal variance in each of the four groups.

3. Answer: D

The typical a priori alpha error (type I rate) rate is 5% (i.e., when the study was designed, the error rate was designed to be 5% or less) (Answer D is correct). The actual type I error rate is reported in the question as 0.01 (1%) (Answer A is incorrect). Answer B and Answer C are related; the study did have enough power because a statistically significant difference was observed. Similarly, a type II error was not made because this error has to do with not finding a difference when one truly exists. In this question, the type I error rate is 1%, the value of the p-value.

4. Answer: C

Sample sizes need not be equal for a t-test to be appropriate (Answer A is incorrect). Body mass index data are not ordinal but rather continuous; thus, a t-test is appropriate (Answer B is incorrect). The assumption of equal variances is necessary to use any parametric test (Answer C is correct). A specific value for power is not necessary to use a test (Answer D is incorrect).

5. Answer: B

Many think reporting the mean difference and CI is a superior means of presenting the results from a clinical trial because it describes both precision and statistical significance versus a p-value, which distills everything into one value, making Answer A incorrect. The presentation of the data in this manner clearly shows all the necessary information for making the appropriate conclusion. To assess statistical significance by use of CIs, the 95% CI (corresponding to the 5% type I error rate used in most studies) may not contain zero (signifying no difference between men and women) for the mean difference, making Answer D incorrect. Answer B is correct because the p-value of less than 0.05 corresponds to the 95% CI in that item. To evaluate Answer C, we would need to know the 99% CI.

6. Answer: D

Answer A is incorrect because it uses unconventional approaches to determine statistical significance. Although this can be done, it is unlikely to be accepted by other readers and investigators. This study observed a nonsignificant increase in HDL-C between the two groups. With a small sample size, such as the one used in this study, there is always concern about adequate power to observe a difference between the two treatments. A difference may exist between these two drugs, but the number of subjects studied may be too small to detect it statistically. Answer D is correct because, given the lack of information provided in this narrative, it is not possible to estimate power; thus, more information is needed. Answer B may be correct, but without first addressing the question of adequate power, it would be an inappropriate conclusion to draw. Answer C is incorrect because even though the new drug increased HDL-C more than the other treatment, it is inappropriate to conclude that it is better because, statistically, it is not.

7. Answer: B

The primary end point in this study, the percentage of subjects at or below the target blood pressure, is nominal data. Subjects at target blood pressure (less than 140/90 mm Hg) are defined as having reached the target. This type of data requires either a chi-square test or a Fisher exact test (depending on the sample size or, more accurately, the number of counts in the individual contingency table cells) (Answer B is correct). An independent samples t-test is not appropriate because actual blood pressure values are not being compared (at least not in this question or this end point) (Answer A is incorrect). If we were comparing the actual blood pressure between the two groups, the test might be appropriate, if parametric assumptions were met. The Wilcoxon signed rank test is the appropriate nonparametric test when paired samples are compared (usually in a crossover trial) (Answer C is incorrect). Finally, a one-sample t-test is used to compare the mean of a single group with the mean of a reference group. This is also incorrect in this situation because two groups are being compared (Answer D is incorrect).

8. Answer: A

Detecting the smaller difference between the treatments requires more power. Power can be increased in several different ways. Answer A is correct because the most common approach is to increase the sample size, which is expensive for the researchers. Answer D is incorrect because smaller sample sizes decrease a study's ability to detect differences between groups. Power can also be increased by increasing α, but doing so increases the chances of a type I error. Answer B decreases α, thus making it more difficult to detect differences between groups. Answer C certainly makes it easier to detect a difference between the two groups; however, it uses an unconventional α value and is thus not the most appropriate technique.

9. Answer: C

Regression analysis is the most effective way to develop models to predict outcomes or variables (Answer C is correct). There are many different types of regression, but all share the ability to evaluate the impact of multiple variables simultaneously on an outcome variable. Correlation analysis is used to assess the association between two (or more) variables, not to make predictions (Answer A is incorrect). Kaplan-Meier curves are used to graphically depict survival curves or time to an event (Answer B is incorrect). Confidence intervals are not used to make predictions (Answer D is incorrect).