Chapter 10 Statistical Foundations: Hypothesis Testing

Within this chapter is discussed the two major classifications of statistical procedures, the logic of hypothesis testing, statistical decision-making, statistical power, effect size estimation, and estimating a population parameter using confidence intervals.

I. Introduction

A. The concepts and procedural models, presented in this unit, are fundamental to the application of statistics to inform data analysis and interpretation.

1. Descriptive statistics summarize population or sample characteristic(s). Inferential statistics are used to estimate a population parameter (i.e., characteristic) or draw a conclusion about a population from sample data.
   a. Population parameters (or characteristics) are denoted by Greek letters, such as μ (mean), δ (standard deviation), δ^2 (variance), or ρ (proportion).
   b. Sample characteristics are denoted by symbols or lower case Arabic letters, such as $\bar{X}$ (mean), *s* or $\hat{s}$ (standard deviation), s^2 (variance), or *r* (proportion).

2. Sample descriptive statistics are routinely employed in inferential statistical procedures. Inferential procedures are classified into parametric and non-parametric classifications. The authors have keyed formulas to Spatz (2011).

B. Statistical Test Classifications: Parametric & Non-parametric Procedures

1. Parametric Statistics
   a. Parametric statistical tests are used to estimate (i.e., make an inference about) a population parameter, such as, μ, δ, or ρ. Parametric statistics are applied to interval and ratio data.
   b. Primarily, this classification of statistical tests includes those based on the t-distribution and f-distribution as well as those employed in correlation and regression analysis.
   c. Parametric procedures rely on fairly rigid assumptions which when materially violated adversely affect inference accuracy. An assumption example is that sample data are drawn from a normally distributed population.
   d. Parametric methods tend to make use of more information within a data set than non-parametric methods; for example, reducing interval data to ordinal data causes the researcher to be unable to measure distances between points.

2. Non-parametric or "distribution-free"
   a. In social science research, it is often not possible for sample data to meet all the rigid parametric assumptions. However, such data still need to be analyzed. In these instances, non-parametric statistical procedures are applied.
   b. Nominal and ordinal data are analyzed with non-parametric procedures.

c. Non-parametric procedures don't require that samples be drawn from normally distributed populations, nor are most concerned with estimating a population parameter.
d. As assumptions are minimal, errors in application tend to be minimal. For many non-parametric tests, computations are simple and can be quickly done with a calculator or by hand if convenient; however, such calculations can be tedious and labor intensive.
e. Apply non-parametric procedures when
   (1) The hypothesis being tested does not estimate a population parameter, e.g., Chi-Square Goodness of Fit Test;
   (2) Parametric test assumption(s) are materially violated;
   (3) Data are on the nominal or ordinal scales (i.e., ranks); or
   (4) Calculations are needed quickly to enable analysis and interpretation.

C. Relationship between Data Type and Statistical Procedure

1. The type of data drives the type of statistical procedure used in data analysis (Table 10.1a); for example, nominal data require the use of a nonparametric statistical test such as the Chi-Square. Also, there are equivalent parametric and non-parametric tests (Table 10.1b); thus, if a parametric test's assumptions are materially violated there is a nearly equally powerful nonparametric alternative.

Table 10.1a
Type of Data and Sample Statistical Test

Type of Data	**Statistic Type**	**Sample Statistical Test**
Nominal	Nonparametric	X^2 Chi-Square
Ordinal	Nonparametric	Wilcoxon Matched Pairs Test
Interval	Parametric	t-tests
Ratio	Parametric	Analysis of Variance (ANOVA)

Table 10.1b
Parametric/Nonparametric Test by Application

Application	**Parametric Test**	**Non-Parametric Test**
Dependent Samples, 2	z- or t-test	Wilcoxon Matched Pairs Test
Independent Samples, 2	z- or t-test	Mann-Whitney U Test
k- Independent Samples (3 or more samples)	F-test	Kruskal-Wallis Test
k- Dependent Samples (3 or more samples)	Repeated Measures ANOVA (F-test)	The Friedman Two-Way ANOVA by Ranks
Correlation	Pearson Product Moment	Spearman Rank Order Correlation Coefficient (r_s)

2. From the above, one should conclude that the selection of a statistical test is dependent upon the inference to be made, type of data presenting, application circumstance, and degree of assumption compliance.

II. The Logic of Hypothesis Testing

A. Introduction

1. When one wants to estimate a population parameter, a confidence interval is constructed. If one seeks to draw a conclusion (e.g., ascertain differences or associations) about a population from sample data, often using variables, a tentative conclusion (i.e., hypothesis) is posited and then tested; this process is called hypothesis testing.

2. Hypothesis testing follows a strict routine employing standard terms and relationships, decision rules, the traditional or p-value procedures. One first learns the standard term definitions and relationships, followed by the decision rules, and finally, the procedures.

B. Standard Term Definitions and Relationships

1. **Null Hypothesis (H_o):** This is a statement about a population parameter (e.g., mean, standard deviation, proportion, etc.) and must include the condition of equality using the "=, ≥, or ≤" symbols. The null hypothesis is what is actually tested and is sometimes referred to as the statistical hypothesis.
   a. The most commonly tested population parameters are μ, δ, or ρ. Mu (μ) is the population parameter for the mean or average. Sigma (δ) is the symbol for a population standard deviation. Rho (ρ) is the symbol for a population proportion.
   b. It is usual when testing a single sample for the null to be written H_o: $\mu = 0$ or H_o: μ = some specified value so standard as to be considered a population parameter (μ, δ, or ρ). When testing two means, it is common for the null to be H_o: $\mu_1 = \mu_2$. When testing three means, the null is written H_o: $\mu_1 = \mu_2 = \mu_3$.

2. **Alternative Hypothesis (H_1 or H_A):** The alternative hypothesis is the exact opposite of the Null hypothesis. The signs are: ≠, <, or >. For example, if H_o: $\mu_1 = \mu_2$, then the alternative is written as H_1: $\mu_1 \neq \mu_2$. The same conventions apply to either Sigma (δ) or variance and Rho (ρ) or relationship.

3. In hypothesis testing, researchers are very concerned about two types of errors when applying the decision rules which determine whether or the null is true (i.e., retained) or false (rejected). These are called Type I or Type II error.
   a. T**ype I Error** is the mistake of rejecting the null when it's really true. We could think of Type I Error as a false positive. Alpha (α) is the probability of a Type I error.
   b. **Type II Error** is retaining the null when it's actually false. We could think of Type II Error as a false negative. Beta (β) is the probability of a Type II error.

c. Considerable effort is expended to control these errors. Strategies are:
   (1) For any fixed α, increasing the sample size will reduce β.
   (2) For any fixed sample size *n*, a decrease in α will cause an increase in β. An increase in α, will cause a decrease β.
   (3) To decrease α and β, increase sample size.

4. Standard Normal Curve (SNC): Remember the standard normal curve from Chapter 9. Much of the follow discussion is based on it. Please see Figure 10.1 to refresh your memory.

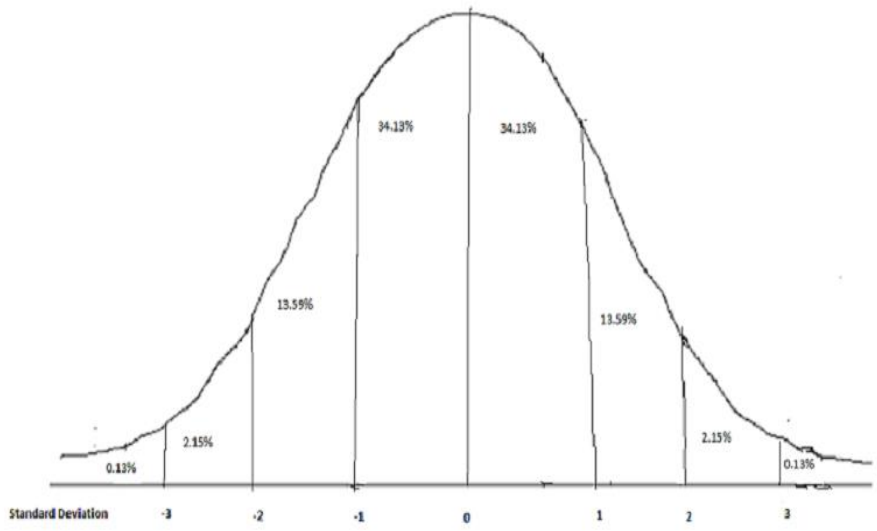

Figure 10.1 Areas under the Standard Normal Curve

4. **Test Statistic (TS)**: The test statistic is the computed value produced through application of a statistical test.

5. **Critical Value (CV)**: The critical value is taken from a critical value table which is specific to the statistical test being applied. A critical value defines (or separates) the standard normal curve into the critical or rejection region and acceptance region.

6. **The Rejection and Acceptance Regions**
   a. **Critical (Rejection) Region**: The area under the curve, defined by the critical value(s), beyond which the test statistic falls which leads to null hypothesis rejection.
   b. **Acceptance Region**: The area under the curve, defined by the critical value(s), either below which or between which, the test statistic falls leading to the acceptance of the null hypothesis. Spatz (2011, p. 163) provides a graphic example of portioning the SNC into Critical (Rejection) Region (shaded) and Acceptance Regions (un-shaded).

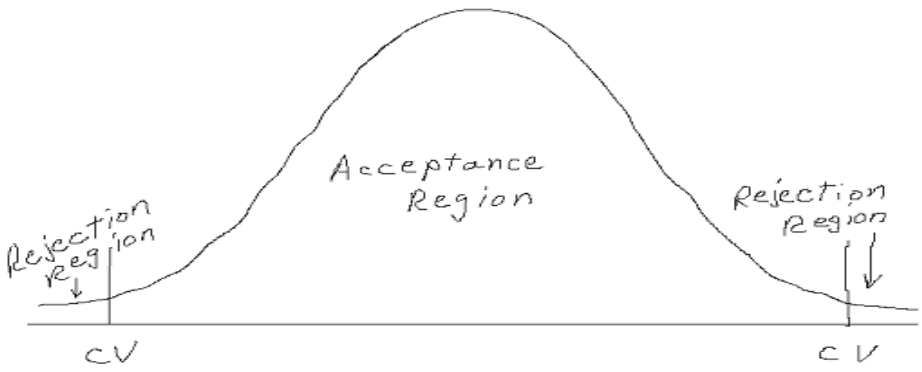

Figure 10.2 Acceptance & Rejection Regions

c. If $\partial = 0.05$, 0.025% of the SNC becomes the rejection regions (located in the SNC tails) for a two (2) tail test as in Figure 10.2. If the Test statistic falls in either rejection region (TS > CV), then the null (H_o) is rejected. If the test statistic ≤ critical value, the null (H_o) is retained, i.e., the test statistic falls into the acceptance region.

7. **Decision Rules and Permissible Conclusions**
   a. Decision Rules
      (1) If TS >CV, reject the null hypothesis
      (2) If TS ≤ CV, retain the null hypothesis
   b. Permissible Conclusions
      (1) Accept the null
      (2) Reject the null
      (3) Suspend judgment

8. **Effect Size**
   a. Pedhazur and Schmelkin (1991, pp. 204) describe effect size as, "the magnitude of findings (e.g., a correlation between two variables, the difference between two means)." In other words, the effect the independent or predictor variable had on the dependent variable. See Chapter 2 for a discussion of the various types of variables.
   b. Effect size is related to statistical social validity in that it helps answer the "So what?" question.
   c. Effect size can be used as an indicator of a practical significance when considered within the context of subject matter expertise. Practical significance considers the question, "Is the impact of the independent variable on the dependent variable, of such magnitude that program changes should be made, regardless of potential disruption?"

9. **Two-tailed, Right-tailed, and Left-tailed Tests**
   a. Picture a bell curve, specifically the Standard Normal Curve (SNC), with its left and right extending tails. Spatz (2011, pp. 186-187) discusses.
      (1) Within one or both of these tails, is the entire rejection region, as defined by a critical value or values.
      (2) The size of the rejection region is determined by the specified alpha level, or the probability of a Type I Error. For example, when $\partial = .05$, the rejection region comprises 5% of the area under the curve.
      (3) Location of the Rejection Region
         (a) A rejection region may be placed in the left tail (Figure 10.3) or the right tail (Figure 10.4). When a rejection region is so placed, it is referred to as either a left or right tailed test. If $\partial = .05$, then the rejection region comprises 5% of the area under the curve in that tail. The one-tail test is statistically more powerful than a two tail-test as it's easier to reject the null hypothesis. A one tail test is often referred to as a directional test because the null hypothesis specifies a relationship.
         (b) When a rejection region is placed in each tail, a two tailed test is conducted. One-half of the alpha (∂) is placed in each tail (Figure 10.2). The null hypothesis specifies no relationship. The two-tail test is the more routine.
         (c) One-tail tests vs. Two-tail tests
            (1) Using a one-tail test, it easier to reject a null hypothesis when the results actually "run" as the directional hypothesis posits.
            (2) By loading the rejection region into one tail, an extreme score in the other tail is ignored, potentially losing important data.
            (3) A two-tail test is more stringent and preserves all the data.

   b. Clues to Directionality
      (1) If the alternative hypothesis is written as $H_1: \mu \neq 0$, or $H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 \neq \mu_2 \neq \mu_3$, then we would know that a two-tail test is being conducted. Remember H_1 or H_A mean the same.
      (2) If "<" or ">" is used in the alternative hypotheses, then a left-tailed (<) (Figure 10.3) or right-tailed (>) (Figure 10.4) test is being conducted.

Figure 10.3 Left Tail Test

Figure 10.4 Right Tail Test

C. The Classical and p-value Hypothesis Testing Procedures

1. Classical Method of Hypothesis Testing
   a. State the hypothesized relationship in plain English.
   b. Cast the English language hypothesized relationship in symbolic form.
   c. State the exact opposite of the hypothesized relationship in symbolic form as well.
   d. The symbolic statement which contains the condition of equality (=, ≥, or ≤) becomes the null hypothesis (H_o) which is to be tested.
   e. The exact opposite symbolic statement becomes the alternative hypothesis H_1.
   f. Given the consequences of a Type I Error, specify the alpha (α) level.
   g. Considering the hypotheses, type of data, identify the statistical procedure that is relevant to this application.
   h. Compute the test statistic and select the relevant critical value(s). When using a critical value table,
      (1) Select the degrees of freedom (*df*) associated with the specific test.
      (2) Next, identify the alpha level (usually 0.05 or 0.01) for either a one-tail (Figure 10.3 or 10.4) or a two-tail test (Figure 10.2) test.
      (3) Finally, locate the critical value at the intersection of the correct *df* and alpha level.

Table 10.1c
Sample Critical Value Table

df	**$\partial = 0.05$**	**$\partial = 0.01$**
1	Critical Value	Critical Value
2	Critical Value	Critical Value
3	Critical Value	Critical Value
4	Critical Value	Critical Value
5	Critical Value	Critical Value

   i. Apply the standard decision rules and draw one of the three permissible conclusions.
      (a) If the test statistic is greater than the critical value, reject the null.
      (b) If the test statistic is less than the critical value, retain the null.

j. Restate the statistical conclusion in simple non-technical English.
k. Effect size estimation is made when statistical significance is established.

2. The p-value Method of Testing Hypotheses
   a. Virtually all statistical computer programs print out exact probabilities for the null. Spatz (2011, p. 185-186) discusses.
   b. Use the following guidelines to interpret:
      (1) $p <$ specified α (e.g., 0.05 or 0.01) reject the null hypothesis.
      (2) $p >$ specified α (e.g., 0.05 or 0.01) retain the null hypothesis.
   c. This link <http://graphpad.com/quickcalcs/PValue1.cfm> (GraphPad, n.d.) takes the interested reader to a calculator, which will compute the exact probability of committing a Type 1 error given the degrees of freedom (*df*) and the test statistic applicable to the situation. The calculator produces an exact probability for the z-test, t-test, F-test (ANOVA), and the chi-square test. If the exact probability is less than the specified alpha (∂) level, reject the null hypothesis. If the exact probability is equal to or greater than the specified alpha level, retain the null hypothesis.

D. Statistical Power

1. Statistical power is the probability of rejecting the null hypothesis when it is really false or stated otherwise, the ability of a statistical procedure to detect statistically significant differences or associations when they in fact exist.

2. Statistical power is dependent on a combination of sample size (n), alpha level (α), and anticipated effect size (ES). Power is computed as ($1-\beta$). Thus, when $\beta = .35$, then power = .65 or there is a 65% chance of rejecting the null hypothesis when it's false (Stevens, 1999, p. 122).

3. Stevens (1999, pp. 136-137) suggests four ways to improve statistical power:
   a. Select a more lenient alpha, i.e., chose .10 instead of .05.
   b. Where the literature or well documented theory supports a directional hypothesis, use a one-tail test.
   c. Reduce within group variability. A more homogeneous sample will differ less variance on the dependent variable. One could use a factorial or covariance research design (beyond the scope of this primer).
   d. Ensure that there is a probable strong linkage between the treatment (independent variable) and dependent variable and that the treatment extends for a sufficient period of time to produce a larger effect size.
   e. Remember that while statistical power is important, we are more concerned with reducing Type I Errors.

III. Single and Two Sample Cases

A. Single Sample Tests

1. Single sample tests are often applied to survey research data. When tested for statistical significance, survey data (i.e., responses) are being tested against

either the SNC (z-test) or the t-distribution (t-test) to test for significant departures. Assumptions:

a. There is a single sample drawn randomly from a population. If the sample size is 30 or greater, use the z-test. If the sample size is less than 30, use the t-test (for small samples).
b. We are (1) actually testing the single sample against the SNC for the population from which the sample was drawn, so randomization is very important or (2) we are testing the sample value against an acknowledged standard which is viewed as the corresponding population parameter.
c. For the z-value in a two tail test, when $\alpha = .05$, $z = \pm 1.96$; & when $\alpha = .01$, $z = \pm 2.575$.

2. Testing a Single Large Population Mean (z-test)
   a. z-test Assumptions
      (1) The sample was drawn from a normal population.
      (2) σ is known. It is rare that σ is known; if not known, use the t-test for a single sample.
      (3) The sample was randomly drawn.
      (4) Data are on at least an interval scale.

   b. Formula 10.1 z-test for a Single Large Sample (Triola, 1998, p. 359)

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\delta}{\sqrt{n}}}$$

   where: Z = the "z" value; $\bar{x}$ = group mean; μ = estimated population mean; δ = population standard deviation; n = sample size. Spatz (2011, p. 157) offers a similar formula.

   c. **Case 10.1**: The management of a financial services company, "We are Rich and You are Not", has set a performance standard for loan application processors. You have been asked to assist the local branch manager to determine whether or not her unit meets the mandated standard of an average of 13 applications processed over a 4 hour period per processor. She has 34 loan application processors. The mean number of applications processed is 11.9 with a standard deviation of 1.2. You are going to test the claim that the average number of applications processed meets the mandated standard.

   d. Applying the Hypothesis Testing Model
      (1) Is the performance standard for loan processors at this manager's office being met?
      (2) $\mu = 13$
      (3) $\mu \neq 13$
      (4) H_o: $\mu = 13$
      (5) H_1: $\mu \neq 13$ (two tail test)
      (6) $\alpha = 0.05$

(7) z-test for a large sample
(8) Compute test statistic
  (a) Construct Data Table: See case.
  (b) Compute Degrees of Freedom: Doesn't apply
  (c) Substitute into Formula 10.1

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\delta}{\sqrt{n}}} = \frac{11.9 - 13}{\frac{1.2}{\sqrt{34}}} = \frac{-1.1}{\frac{1.2}{\sqrt{34}}} = \frac{-1.1}{.2058} = -5.34$$

  (d) Critical Value: ±1.96 (for ∂ = .05 in a two tail test)

(9) Apply Decision Rule: Since the test statistic of z = -5.34 is greater than the critical value of z = ±1.96, reject (i.e., don't accept) the null, H_o: μ = 13, as p < .05. The *GraphPad Software QuickCalcs* (n.d.) calculator confirms the traditional hypothesis testing method at the 0.0001 level.
(10) The performance standard for loan processors at this manager's office is not being met.
(11) Effect Size estimation: Doesn't apply.

3. Testing a Single Small Population Mean (t-test)

a. The t-distribution varies for each sample size. However, as the sample size reaches 30, the standard normal curve is approximated.

b. Formula 10.2 t-test for a Small Population (Triola, 1998, p. 379)

$$t = \frac{\bar{X} - \mu_{\bar{x}}}{\frac{\delta}{\sqrt{n}}}$$

where: t = "t" value; $\bar{x}$ = group mean; μ = population mean; δ = population standard deviation (use s if δ is unknown); n = sample size. (Spatz, 2011, pp. 181-183) offers an equivalent formula.

c. **Case 10.2**: The credit card company you work for as a manager in Tampa has just launched a third shift to meet service needs for its Asian customers. You have spent the last three weeks training your 300 bilingual staff members in the company's service policies and procedures. The Company requires that a performance standard of 135 points be earned before trainees are allowed to service "live" customers. As time is of the essence, you have randomly selected nine trainees and tested them. You are asserting that the nine trainees' scores are drawn from a population (the 300) with a mean score greater than 135 points.

d. Applying the Hypothesis Testing Model
(1) Are the nine trainees' scores representative of the population?
(2) Question in symbolic form: μ > 135

(3) Opposite of question in symbolic form: $\mu \leq 135$
(4) H_o: $\mu \leq 135$
(5) H_1: $\mu > 135$ (right tail test)
(6) $\alpha = 0.05$
(7) t-test for a small sample
(8) Compute the Test Statistic
  (a) Construct Data Table: See data box.
  (b) Compute Degrees of Freedom: 8 (n-1)
  (c) Substitute into Formula 10.2

$\bar{x}$ = 146.778
s = 24.299
Scores:
140, 157, 119, 127, 190, 121, 136, 160, 171

$$t = \frac{\bar{X} - \mu_{\bar{x}}}{\frac{\delta}{\sqrt{n}}} = \frac{146.778 - 135}{\frac{24.299}{\sqrt{9}}} = \frac{11.778}{\frac{24.299}{3}} = \frac{11.778}{8.0997} = 1.45$$

  (d) Critical Value: 1.860 at $\partial = .05$ and $df = 8$ in a right tail (Spatz, 2011, p. 392; Triola, 1998, p. 715). *GraphPad Software QuickCalcs* (n.d.) isn't applicable, as it computes an exact probability statement for a two tail test. The test performed immediately was a right tail test. So, we have to use a critical value table for one tail tests located at http://faculty.vassar.edu/lowry/PDF/t_tables.pdf . Here, we can confirm that the critical value is 1.8595 or 1.860.

(9) Apply Decision Rule: Since the test statistic of t = 1.45 is less than the critical value of t = 1.860, retain (i.e., don't reject) the null, H_o: $\mu \leq 135$, as $p > .05$. We would note this as $t_{(8)} = 1.45$, $p > 0.05$.
(10) The 9 trainees' scores are not representative of the population and don't meet the standard of 135 points. The training manager must test all 300 trainees. Those trainees that don't meet the standard will either re-test or re-train.

(11) Effect Size estimation: Not possible in this instance as the null hypothesis was retained. When the null is rejected use the formula presented below. In theory, one should use "σ", but it is rarely, if ever known; use $\hat{s}$ or the standard deviation of the sample as the acceptable parameter estimate. Formula 10.3a (Spatz, 2011, pp. 189) computes *d.*

$$d = \frac{\bar{X} - \mu_0}{\delta or \hat{s}}$$

where: d = effect size index
$\bar{X}$ = the sample mean
μ_0 = the mean specified by the null hypothesis

$\hat{s}$ = estimated σ (population standard deviation)
can be computed by using the N-1 standard deviation key on your calculator or use Formula 10.3b (Spatz, 2011, p. 65):

$$\hat{s} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

B. t-tests for Independent & Dependent Samples

1. Assumptions
   a. The t-distribution varies for different sample sizes
   b. Has the same general symmetric bell shape as the SNC, but it reflects greater variability that is expected for small samples. There are those that argue that the t-distribution can be used with any sample size, small ($n \leq 30$) or large ($n \geq 31$)
   c. The t-distribution has a mean, t = 0.
   d. The t-distribution has a standard deviation which varies with sample size, but is greater than one.
   e. As sample size increases, the t-distribution approximates the SNC. When sample size reaches n = 31 or so, the t-distribution is similar to the SNC.

2. t-test for Independent Samples

a. Formula 10.4a t-test for Independent Sample, when $N_1 = N_2$ (Spatz, 2011, p. 206)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

where: t = "t" value
$\bar{X}_1$ = group one mean
$\bar{X}_2$ = group two mean
s^2 = variance for groups one and two
n = sample sizes for groups one and two
$s_{\bar{X}_1 - \bar{X}_2}$ = standard error of the difference

b. Formula 10.4b Standard Error of the Difference ($s_{\bar{X}_1 - \bar{X}_2}$), when $N_1 = N_2$ (Spatz, 2011, p. 206)

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{N_1} + \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{N_2}}{N_1(N_2 - 1)}}$$

c. Formula 10.4c Standard Error of the Difference ($s_{\bar{X}_1-\bar{X}_2}$), when $N_1 \neq N_2$ (Spatz, 2011, p. 206)

$$s_{\bar{X}_1-\bar{X}_2}=\sqrt{\left(\frac{\sum X_1^2-\frac{(\sum X_1)^2}{N_1}+\sum X_2^2-\frac{(\sum X_2)^2}{N_2}}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$$

d. **Case 10.3**: Your home healthcare organization is considering adopting a new home health service protocol (treatment program) which is designed to improve the functional ability of home-bound elderly patients to prevent placement into a nursing home. You have decided to test the effectiveness of the new protocol, as your company might buy it to replace your current protocol.

Accordingly, you have randomly selected and assigned 13 patients to a treatment (i.e., the new protocol or Experimental group) and 13 more to a control group (i.e., the company's current treatment protocol). Further, you have randomly selected and assigned home health care service providers, who have been trained in one of the service protocols, to each group. At the end of your trial, a test was administered by independent raters to assess the functional ability (i.e., the ability to take care of one-self, with minimal help). The scores range from zero to 30 (highest); the higher the score, the better a patient's functional ability. Alpha = .05. The order of the scores does not imply that they are matched.

E-Group: 27, 29, 30, 26, 23, 18, 21, 18, 19, 23, 13, 26, & 24.
C-Group: 21, 20, 14, 17, 19, 22, 26, 9, 16, 17, 19, 18, & 18.

e. Applying the Hypothesis Testing Model
(1) Are patients treated with the newer protocol more functional?
(2) $\mu_1 = \mu_2$
(3) $\mu_1 \neq \mu_2$
(4) H_o: $\mu_1 = \mu_2$
(5) H_1: $\mu_1 \neq \mu_2$
(6) $\alpha = 0.05$ (two-tail test)
(7) t-distribution for independent samples
(8) Compute the Test Statistic
(a) Construct Data Table: See data box.
(b) Compute Degrees of Freedom: $N_1 + N_2$ - 2 or 13 + 13 – 2 = 24
(c) Substitute into Formula 10.4b (Spatz, 2011, p. 206)

E- Group	C-Group
$\bar{x}_1$ = 22.85	$\bar{x}_2$ = 18.15
$\sum x_1 = 307$	$\sum x_2 = 236$
$\sum(X_1)^2 = 7655$	$\sum(X_2)^2 = 4419$
n = 13	n = 13

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\Sigma X_1^2-\frac{(\Sigma X_1)^2}{N_1}+\Sigma X_2^2-\frac{(\Sigma X_2)^2}{N_2}}{N_1(N_2-1)}}=\sqrt{\frac{7655-\frac{(307)^2}{13}+4419-\frac{(236)^2}{13}}{13(12)}}=$$

$\sqrt{3.4601}=1.860$ (Then compute Formula 10.4a)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{22.85 - 18.15}{1.860} = \frac{4.70}{1.860} = 2.5269$$

(d) Critical Value: = ±2.064 at δ = .05 and *df* = 24 in a two tail test (Spatz, 2011, p. 392; Triola 1998, p. 715).

(9) Apply Decision Rule: Since the test statistic, t = 2.5269, is greater than the critical value of t = 2.064, reject (i.e., don't retain) the null, Ho: $\mu_1 = \mu_2$, p < .05. *GraphPad Software QuickCalcs* (n.d.) calculator, we see that the exact probability is 0.0185; we reject the null hypothesis as the exact probability is less than the specified alpha level. We would note this as $t_{(24)} = 2.5269$, p < 0.05.

(10) It appears that the new protocol is more effective in preventing the institutionalization of elderly patients than the current one. Patients treated under the new protocol were more functional than those treated under the current protocol.

(11) Effect Size Estimate: An effect size of 0.9909 indicates a large effect.

(a) Formula when $N_1 = N_2$ (Spatz, 2011, p. 216):

Formula 10.5a $d = \frac{|\bar{X}_1 - \bar{X}_2|}{\hat{s}}$ Formula 10.5b $\hat{s} = \sqrt{\frac{N_1}{2}}(s_{\bar{X}_1 - \bar{X}_2})$

where: d = effect size index
$\bar{X}_1$ = the mean of group one
$\bar{X}_2$ = the mean of group two
$\hat{s}$ = estimated δ (population standard deviation)

(b) Formula (10.6) when $N_1 \neq N_2$ (Spatz, 2011, p. 216):

$$\hat{s} = \sqrt{\frac{s_1^2(df_1) + s_2^2(df_2)}{df_1 + df_2}}$$

where: $\hat{s}$ = estimated σ (population standard deviation)
$\hat{s}_1^2$ = variance for group one
$\hat{s}_2^2$ = variance for group two
df_1 = degrees of freedom (n-1) for group one
df_2 = degrees of freedom (n-1) for group two

(c) Substitute into Formula 10.5b and 10.5a (Spatz, 2011, p. 216)

$$\hat{s} = \sqrt{\frac{N_1}{2}}(s_{\bar{x}_1 - \bar{x}_2}) = \sqrt{\frac{13}{2}}(1.860) = (2.55)(1.860) = 4.743$$

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\hat{s}} = \frac{4.7}{4.743} = 0.9909$$

(d) Interpretation: The independent variable (i.e., the new program had a large effect on the dependent variable (i.e., functionality) based on Cohen's (1988, pp. 25-26) criterions: d = 0.2 (small effect), d = 0.5 (medium effect) & d = 0.8 (large effect).

3. t-test for Dependent Samples (Pre-test/Post-test)

a. The t-test for dependent samples is applied to instances where
   (1) Natural pairs are based on some logical pairing of the scores, e.g., fathers' and sons' height, IQ, religious belief, etc.
   (2) Matched pairs are those constructed by the researcher, e.g., pairs matched based on a pre-test score or criteria which would ensure equivalence at the start of an experiment.
   (3) Repeated measures are taken from the same individual (e.g., pretest /posttest).

b. The Formula (Spatz, 2011, pp. 210-2011)

   (1) Formula (10.6a) (Spatz, 2011, p, 211):

$$\hat{s}_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N-1}}$$

   (2) Formula 10.6b (Spatz, 2011, p. 211): $s_{\bar{D}} = \frac{\hat{s}_D}{\sqrt{N}}$

   (3) Formula 10.6c (Spatz, 2011, p. 210): $t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}} = \frac{\bar{D}}{s_{\bar{D}}}$

   where: $\hat{s}_D$ = difference score standard deviation (also noted as "s_d")
   D or d = the difference scores or X-Y
   $s_{\bar{D}}$ = standard error of the difference scores
   N = number of score pairs
   $\bar{X}$ = mean of the X scores
   $\bar{Y}$ = mean of the Y scores

c. **Case 10.4**: The foundation which funds your community service organization has asked it to demonstrate that its six week, residential program actually increases a client's level of civic responsibility. Research has shown that youth with high levels of civic responsibility are less likely to violate the law and get into trouble. A civic responsibility

index was given to clients at program entry and program exit. Scores range from zero to 20 (highest). A higher score indicates a greater level of civic responsibility. Eight clients were tested. Alpha = .05.

d. Applying the Hypothesis Testing Model
   (1) Does the program improve levels of civic responsibility?
   (2) $\mu_d = 0$
   (3) $\mu_d \neq 0$
   (4) H_o: $\mu_d = 0$
   (5) H_1: $\mu_d \neq 0$ (two-tailed test)
   (6) $\alpha = 0.05$
   (7) t-distribution for dependent samples
   (8) Compute the Test Statistic
      (a) Construct Data Table:

Table 10.2
Dependent Samples t-test Case Data

Subject	**Pretest (X)**	**Posttest (Y)**	**X-Y = D**	**D^2**
A	10	14	-4	16
B	11	19	-8	64
C	9	13	-4	16
D	11	12	-1	1
E	14	19	-5	25
F	6	10	-4	16
G	13	17	-4	16
H	12	13	-1	1
	$\bar{X} = 10.75$	$\bar{Y} = 14.63$	$\Sigma D = -31$	$\Sigma D^2 = 155$

(b) Compute Degrees of Freedom: N (#of pairs)-1 or 8-1 = 7

(c) Substitute into Formulae 10.6a, 10.6b, and 10.6c.

$$\hat{s}_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N-1}} = \sqrt{\frac{155 - \frac{(-31)^2}{8}}{7}} = \sqrt{\frac{155 - 120.13}{7}} = 2.23$$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{N}} = \frac{2.23}{\sqrt{8}} = \frac{2.23}{2.83} = .788$$

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{D}}} = \frac{10.75 - 14.63}{.788} = \frac{-3.88}{.788} = -4.92$$

(d) Critical Value: = ±2.365 at $\partial = .05$ and *df* = 7 in a two tail test (Spatz, 2011, p. 392; Triola, 1998, p. 715).

(9) Apply Decision Rule: Since the test statistic of t = |-4.92|is greater than the critical value of t = 2.365, reject (i.e., don't retain) the null, H_o: $\mu_d = 0$. Using the *GraphPad Software QuickCalcs* (n.d.) calculator, we see that the exact probability is 0.0017; we reject the null hypothesis as the exact probability is less than the specified alpha level. We would note this as $t_{(7)} = 4.92$, $p < 0.05$.

(10) It appears that the organization's services increase civic responsibility.

(11) Effect Size Estimation: The independent variable (i.e., the residential program) had a large effect on the dependent variable (i.e., civic responsibility) as seen by using Formula 10.8d and then 10.8e. Cohen's (1988, pp. 25-26) criterions of d = .2 (small effect), d = 5 (medium effect) & d = .8 (large effect).

[1] Formula 10.6d (Spatz, 2011, p. 217): $\hat{s} = \sqrt{N}(s_{\bar{D}})$

$$\hat{s} = \sqrt{N}(s_{\bar{D}}) = \sqrt{8}(.788) = (2.828)(.788) = 2.2285$$

[2] Formula 10.6e (Spatz, 2011, p. 207):

$$d = \frac{\left|\bar{X} - \bar{Y}\right|}{\hat{s}} = \frac{\left|-3.88\right|}{2.2285} = 1.7411$$

IV. Estimating a Population Parameter using Confidence Intervals

A. Introduction

1. Confidence intervals are used to judge how accurate an estimate of a population parameter or point estimate is, usually a mean or standard deviation. Confidence intervals (CI) are used primarily in the health sciences and in statistical process control (SPC). They provide a basis for comparing individual medical test or SPC score values to an acceptable range of similar values on the same measurement scale to ascertain whether or not the individual value (e.g., test result or score) falls into the acceptable range.

2. **Confidence intervals are an alternative to one sample tests (e.g., z-test or t-test) for either rejecting or retaining a null hypothesis.**
   a. If the value of μ, assumed under the null, doesn't fall into the interval, the null is rejected.
   b. If the value of μ, assumed under the null, falls into the interval, the null is retained.
   c. The sample mean used to construct the interval will always fall into that interval.
   d. The narrower the CI, the more precise is the estimate. To increase precision, increase the sample size and/or decrease the standard error.

e. The discussion to follow is based on Triola (1998). Spatz (2011, pp. 164-167, 218-219) offers an alternative explanation.

3. Definition of Terms
   a. Point Estimate: A single sample statistic (or point) which is used to estimate a population parameter is called a point estimate.
   b. Confidence Interval: A confidence interval is an interval of values within which the true population parameter value is thought to reside.
   c. Degree of Confidence: The degree of confidence is the probability, 1 - α, that the confidence interval actually contains the true value of the point parameter. For example, if $\alpha = .05$, then we say we are 95% confident or sure that the population parameter falls between "a…b".
   d. Critical Value: A confidence interval is a specific distribution value (e.g., z-distribution or t-distribution), just like those for the one or two sample tests discussed above.
      (1) Critical *z* values,
         (a) Depending on the degree of confidence, are fixed, and in confidence interval construction, employed with large sample sizes (n > 30). See Table 10.3.

Table 10.3
Critical z Values

1- α	**α**	**Critical z Value**
99%	.01	± 2.575
95%	.05	± 1.96
90%	.10	± 1.645

         (b) For example, picture a bell shaped curve, specifically the Standard Normal Curve (SNC) (Figure 9.8) facing you with its left (to your left) and right (to your right) extending tails. Let's assume a 95% confidence interval with a critical value of z = ±1.96. The z = -1.96 (in the left tail) demarcates the 2.5% of the area under the curve to its left from other SNC area. The z = +1.96 in the right-tail does the same for SNC area to its right. The critical values are ±1.96.

         (c) The SNC area to the left and right of the critical z-values represent those other z-values considered highly unlikely to occur (i.e., the rejection region) given a specified 1- α. The process is the same for small samples using the *t*-distribution.

      (2) Critical *t* values, used in constructing small sample (n≤ 25-30) confidence intervals, are selected from a *t* table where the t-value magnitude depends on α and the degrees of freedom, using the "two tail" column. Degrees of freedom (*df*) are computed using n-1, where n = sample size.

Table 10.4
Critical t Values

1- α	α	Critical *t* Value
99%	.01	Varies with *df*
95%	.05	Varies with *df*
90%	.10	Varies with *df*

e. Margin of error: When constructing a confidence interval for a population mean, μ, the margin of error or standard error of the mean is the largest likely difference between $\bar{X}$ (an observed sample mean) and μ, for a specified degree of confidence (1 – α).

4. It is possible to compute the margin of error as long as δ is known or a sample standard deviation, *s*, is available and the sample size is known.

B. Estimating a Population Mean using a Large Sample (n > 30)

1. To estimate a population mean from a large sample, we rely on z-values to build the confidence interval (CI). The critical z-value for a 95% CI is ± 1.96 and the critical z-value for a 99% CI is ± 2.575.

$$E = Z_{\alpha/2} \bullet \frac{\delta}{\sqrt{n}}$$

2. Formula 10.7 The Margin of Error, E (Triola, 1998, p. 293) Where: E = margin of error; $Z_{\alpha/2}$ = critical z-value; σ = standard deviation; n = sample size. (Note: s is substituted for δ, when δ is unknown.)

3. Formula 10.8 The Confidence Interval (Triola, 1998, p. 293)

$$\bar{X} - E < \mu < \bar{X} + E \text{ or } \mu = \pm E$$

4. **Case 10.5**: Your organization provides training services to small and medium size companies to improve support staff clerical skills. The sales manger has informed you that she must respond to an inquiry from a potentially significant customer about what score they might expect their employees to earn on your competency test which is given at the end of an eight hour workshop.

   Accordingly, you have computed the following descriptive statistics for the past year which are $\bar{X}$ = 93, s = 2.1, n = 612. Test scores can range from zero to 100.

To add further context to the point estimate, you have decided to construct a 95% confidence interval. You computed E = .166; therefore the 95% CI: Thus, your marketing manager can tell the prospective client that 95 out of 100 times a trainee can expect his or her competency test score to between 92.83-93.166 points.

C. Estimating the Mean Using a Small Sample

1. For small samples (n ≤ 25-30), the computational process is essentially the same as for large samples, except that the t-distribution is used to identify the critical values. To use t-values, we must assume that the distribution of a population is essentially normal. To locate the appropriate t- critical value in the t-table, we need to know the degree of freedom (n-1 for this application) and the α level of the CI (a 95% CI has an α = .05 while a 99% CI has an α = .01).

2. Formula 10.9 The Modified Margin of Error (Triola, 1998, p. 309)

$$E = t_{\partial/2} \bullet \frac{s}{\sqrt{n}}$$

Where: E = margin of error; $t_{\alpha/2}$ = critical t-value, with n-1 degrees of freedom; *s* = standard deviation; n = sample size. (Note: s is substituted for σ when δ is unknown which is almost always the case.)

3. The confidence interval formula remains unchanged. The basic interpretation is as above in the example for a specified degree of confidence and measurement scale. Formula 10.8 remains unchanged.

$$\bar{X} - E < \mu < \bar{X} + E \text{ or } \mu = \pm E$$

4. **Case 10.6**: Your manufacturing company builds a key machine used for constructing an oasis. While well constructed, maintenance costs are high. A prospective customer, Sultan Charles, has inquired as to expected first year maintenance costs and wants a 99% CI because he took a statistics course in Sultan School. Your research turned up the following first year maintenance cost information for the last 12 machines produced: $\bar{X}$ = $105,000 and s = $35,750. E = $9,252.77 (t = 3.106 @ df = 11 @ α = .01). We are 99% confident that first year maintenance costs will be between $95,747.23 and $114,252.77.

Review Questions

Directions. Read each item carefully; either fill-in-the-blank or circle letter associated with the term that best answers the item.

1. The degree of confidence is noted by:
   a. 1 + α
   b. 1 - α
   c. 1/ α
   d. 1 * α

2. Concerning the Null Hypothesis, which one of the following statements is inaccurate?
   a. We test assuming the null is true.
   b. Must contain the equality.
   c. Symbolized as H_o
   d. Symbolized as H_1

3. When we fail to reject a false null hypothesis, we are committing what type of error?
   a. Type I
   b. Type II
   c. Alpha error
   d. Critical error

4. What are the two most widely used permissible options, when drawing conclusions in hypothesis testing?
   a._____
   b._____

5. With respect to controlling error in hypothesis testing, which one of the following statements is untrue?
   a. For any fixed alpha, an increase in *n*, will decrease β.
   b. For any fixed sample size, an increase in alpha will increase β.
   c. To decrease both alpha and β, increase *n*.
   d. For Type I errors, with serious consequences, select a smaller alpha.

6. Of the following, which is usually the computed value?
   a. Test statistic
   b. Critical region
   c. Critical Value
   d. Beta Value

7. An alternative hypothesis of the form H_1 < some value is said to be a:
   a. A right tailed test
   b. A left tailed test
   c. A Two tailed test
   d. A Null hypothesis

8. What is the value of $\alpha = .05$ in each tail for a two tail test?
   a. .05
   b. .25
   c. .025
   d. .50

9. Which hypothesis testing method, is also referred to as the classical approach?
   a. P-value
   b. Computerized testing
   c. Confidence interval
   d. Traditional approach

10. When applying the p-value approach to hypothesis testing, which one of the following statements is untrue?
    a. Retain the null if the p-value is less than or equal to the significance level.
    b. Retain the null if the p-value is greater than the significance level.
    c. P-values measure how confident we are in rejecting the null.
    d. In a two tailed test, the p-value is no different as in a one-tail test.

11. You have been asked to analyze data from a training program where trainees were pre-tested and then post-tested after the training classes. Which type of test will you use?
    a. Confidence intervals
    b. Dependent samples
    c. Independent samples
    d. Chi-square test

12. You have been asked to determine which one of two training programs produces better trained graduates. Thirty-nine students completed Course A and 61 completed Course B. Which type of test do you use?
    a. Confidence intervals
    b. Dependent samples
    c. Independent samples
    d. Test for Two Proportion

13. Which one of the following statistical tests is most often used with nominal data?
    a. Dependent samples t-test
    b. Independent samples t-test
    c. Chi-square test
    d. Correlation

Answers: 1. b, 2. d, 3. b, 4. Reject Null & Retain Null, 5. b, 6. a, 7. b, 8. c, 9. d, 10. a, 11. b, 12. c, 13. c.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

GraphPad Software QuickCalcs. (n.d.). Retrieved from http://graphpad.com/quickcalcs/Pvalue2.cfm

Pedhazur, E. J. & Schmelkin, L. P. (1991) *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Spatz, C. (2011). *Basic statistics* (10th ed.). Belmont, CA: Wadsworth.

Stevens, J. P. (1999) *Intermediate statistics: A modern approach* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Triola, M. F. (1998). *Elementary statistics* (7th ed.). Reading, MA: Addison-Wesley.