READING

11

Hypothesis Testing

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Trilogy Global Advisors (USA).

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
□	**a.** define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative hypotheses;
□	**b.** distinguish between one-tailed and two-tailed tests of hypotheses;
□	**c.** explain a test statistic, Type I and Type II errors, a significance level, and how significance levels are used in hypothesis testing;
□	**d.** explain a decision rule, the power of a test, and the relation between confidence intervals and hypothesis tests;
□	**e.** distinguish between a statistical result and an economically meaningful result;
□	**f.** explain and interpret the *p*-value as it relates to hypothesis testing;
□	**g.** identify the appropriate test statistic and interpret the results for a hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed and the variance is 1) known or 2) unknown;
□	**h.** identify the appropriate test statistic and interpret the results for a hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations, based on independent random samples with 1) equal or 2) unequal assumed variances;
□	**i.** identify the appropriate test statistic and interpret the results for a hypothesis test concerning the mean difference of two normally distributed populations;

(continued)

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
□	**j.** identify the appropriate test statistic and interpret the results for a hypothesis test concerning 1) the variance of a normally distributed population, and 2) the equality of the variances of two normally distributed populations based on two independent random samples;
□	**k.** formulate a test of the hypothesis that the population correlation coefficient equals zero and determine whether the hypothesis is rejected at a given level of significance;
□	**l.** distinguish between parametric and nonparametric tests and describe situations in which the use of nonparametric tests may be appropriate.

1 INTRODUCTION

Analysts often confront competing ideas about how financial markets work. Some of these ideas develop through personal research or experience with markets; others come from interactions with colleagues; and many others appear in the professional literature on finance and investments. In general, how can an analyst decide whether statements about the financial world are probably true or probably false?

When we can reduce an idea or assertion to a definite statement about the value of a quantity, such as an underlying or population mean, the idea becomes a statistically testable statement or hypothesis. The analyst may want to explore questions such as the following:

- Is the underlying mean return on this mutual fund different from the underlying mean return on its benchmark?
- Did the volatility of returns on this stock change after the stock was added to a stock market index?
- Are a security's bid-ask spreads related to the number of dealers making a market in the security?
- Do data from a national bond market support a prediction of an economic theory about the term structure of interest rates (the relationship between yield and maturity)?

To address these questions, we use the concepts and tools of hypothesis testing. Hypothesis testing is part of statistical inference, the process of making judgments about a larger group (a population) on the basis of a smaller group actually observed (a sample). The concepts and tools of hypothesis testing provide an objective means to gauge whether the available evidence supports the hypothesis. After a statistical test of a hypothesis we should have a clearer idea of the probability that a hypothesis is true or not, although our conclusion always stops short of certainty. Hypothesis testing has been a powerful tool in the advancement of investment knowledge and science. As Robert L. Kahn of the Institute for Social Research (Ann Arbor, Michigan) has written, "The mill of science grinds only when hypothesis and data are in continuous and abrasive contact."

The main emphases of this reading are the framework of hypothesis testing and tests concerning mean, variance, and correlation, three quantities frequently used in investments. We give an overview of the procedure of hypothesis testing in the next section. We then address testing hypotheses about the mean and hypotheses about the differences between means. In the fourth section of this reading, we address testing hypotheses about a single variance, the differences between variances, and a correlation coefficient. We end the reading with an overview of some other important issues and techniques in statistical inference.

HYPOTHESIS TESTING

2

Hypothesis testing, as we have mentioned, is part of the branch of statistics known as statistical inference. Traditionally, the field of statistical inference has two subdivisions: **estimation** and **hypothesis testing**. Estimation addresses the question "What is this parameter's (e.g., the population mean's) value?" The answer is in the form of a confidence interval built around a point estimate. Take the case of the mean: We build a confidence interval for the population mean around the sample mean as a point estimate. For the sake of specificity, suppose the sample mean is 50 and a 95 percent confidence interval for the population mean is 50 ± 10 (the confidence interval runs from 40 to 60). If this confidence interval has been properly constructed, there is a 95 percent probability that the interval from 40 to 60 contains the population mean's value.[1] The second branch of statistical inference, hypothesis testing, has a somewhat different focus. A hypothesis testing question is "Is the value of the parameter (say, the population mean) 45 (or some other specific value)?" The assertion "the population mean is 45" is a hypothesis. A **hypothesis** is defined as a statement about one or more populations.

This section focuses on the concepts of hypothesis testing. The process of hypothesis testing is part of a rigorous approach to acquiring knowledge known as the scientific method. The scientific method starts with observation and the formulation of a theory to organize and explain observations. We judge the correctness of the theory by its ability to make accurate predictions—for example, to predict the results of new observations.[2] If the predictions are correct, we continue to maintain the theory as a possibly correct explanation of our observations. When risk plays a role in the outcomes of observations, as in finance, we can only try to make unbiased, probability-based judgments about whether the new data support the predictions. Statistical hypothesis testing fills that key role of testing hypotheses when chance plays a role. In an analyst's day-to-day work, he may address questions to which he might give answers of varying quality. When an analyst correctly formulates the question into a testable hypothesis and carries out and reports on a hypothesis test, he has provided an element of support to his answer consistent with the standards of the scientific method. Of course, the analyst's logic, economic reasoning, information sources, and perhaps other factors also play a role in our assessment of the answer's quality.[3]

We organize this introduction to hypothesis testing around the following list of seven steps.

- **Steps in Hypothesis Testing.** The steps in testing a hypothesis are as follows:[4]

[1] We discussed the construction and interpretation of confidence intervals in the reading on sampling and estimation.

[2] To be testable, a theory must be capable of making predictions that can be shown to be wrong.

[3] See Freeley and Steinberg (2013) for a discussion of critical thinking applied to reasoned decision making.

[4] This list is based on one in Daniel and Terrell (1995).

1 Stating the hypotheses.
2 Identifying the appropriate test statistic and its probability distribution.
3 Specifying the significance level.
4 Stating the decision rule.
5 Collecting the data and calculating the test statistic.
6 Making the statistical decision.
7 Making the economic or investment decision.

We will explain each of these steps using as illustration a hypothesis test concerning the sign of the risk premium on US stocks. The steps above constitute a traditional approach to hypothesis testing. We will end the section with a frequently used alternative to those steps, the *p*-value approach.

The first step in hypothesis testing is stating the hypotheses. We always state two hypotheses: the null hypothesis (or null), designated H_0, and the alternative hypothesis, designated H_a.

- **Definition of Null Hypothesis.** The null hypothesis is the hypothesis to be tested. For example, we could hypothesize that the population mean risk premium for US equities is less than or equal to zero.

The null hypothesis is a proposition that is considered true unless the sample we use to conduct the hypothesis test gives convincing evidence that the null hypothesis is false. When such evidence is present, we are led to the alternative hypothesis.

- **Definition of Alternative Hypothesis.** The alternative hypothesis is the hypothesis accepted when the null hypothesis is rejected. Our alternative hypothesis is that the population mean risk premium for US equities is greater than zero.

Suppose our question concerns the value of a population parameter, θ, in relation to one possible value of the parameter, θ_0 (these are read, respectively, “theta” and “theta sub zero”).[5] Examples of a population parameter include the population mean, μ, and the population variance, σ^2. We can formulate three different sets of hypotheses, which we label according to the assertion made by the alternative hypothesis.

- **Formulations of Hypotheses.** In the following discussion we formulate the null and alternative hypotheses in three different ways:

  1 H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$ (a “not equal to” alternative hypothesis)
  2 H_0: $\theta \leq \theta_0$ versus H_a: $\theta > \theta_0$ (a “greater than” alternative hypothesis)
  3 H_0: $\theta \geq \theta_0$ versus H_a: $\theta < \theta_0$ (a “less than” alternative hypothesis)

In our US example, $\theta = \mu_{RP}$ and represents the population mean risk premium on US equities. Also, $\theta_0 = 0$ and we are using the second of the above three formulations.

The first formulation is a **two-sided hypothesis test** (or **two-tailed hypothesis test**): We reject the null in favor of the alternative if the evidence indicates that the population parameter is either smaller or larger than θ_0. In contrast, Formulations 2 and 3 are each a **one-sided hypothesis test** (or **one-tailed hypothesis test**). For Formulations 2 and 3, we reject the null only if the evidence indicates that the population parameter is respectively greater than or less than θ_0. The alternative hypothesis has one side.

[5] Greek letters, such as σ, are reserved for population parameters; Roman letters in italics, such as *s*, are used for sample statistics.

Notice that in each case above, we state the null and alternative hypotheses such that they account for all possible values of the parameter. With Formulation 1, for example, the parameter is either equal to the hypothesized value θ_0 (under the null hypothesis) or not equal to the hypothesized value θ_0 (under the alternative hypothesis). Those two statements logically exhaust all possible values of the parameter.

Despite the different ways to formulate hypotheses, we always conduct a test of the null hypothesis at the point of equality, $\theta = \theta_0$. Whether the null is H_0: $\theta = \theta_0$, H_0: $\theta \leq \theta_0$, or H_0: $\theta \geq \theta_0$, we actually test $\theta = \theta_0$. The reasoning is straightforward. Suppose the hypothesized value of the parameter is 5. Consider H_0: $\theta \leq 5$, with a "greater than" alternative hypothesis, H_a: $\theta > 5$. If we have enough evidence to reject H_0: $\theta = 5$ in favor of H_a: $\theta > 5$, we definitely also have enough evidence to reject the hypothesis that the parameter, θ, is some smaller value, such as 4.5 or 4. To review, the calculation to test the null hypothesis is the same for all three formulations. What is different for the three formulations, we will see shortly, is how the calculation is evaluated to decide whether or not to reject the null.

How do we choose the null and alternative hypotheses? Probably most common are "not equal to" alternative hypotheses. We reject the null because the evidence indicates that the parameter is either larger or smaller than θ_0. Sometimes, however, we may have a "suspected" or "hoped for" condition for which we want to find supportive evidence.[6] In that case, we can formulate the alternative hypothesis as the statement that this condition is true; the null hypothesis that we test is the statement that this condition is not true. If the evidence supports rejecting the null and accepting the alternative, we have statistically confirmed what we thought was true. For example, economic theory suggests that investors require a positive risk premium on stocks (the **risk premium** is defined as the expected return on stocks minus the risk-free rate). Following the principle of stating the alternative as the "hoped for" condition, we formulate the following hypotheses:

H_0: The population mean risk premium on US stocks is less than or equal to 0.

H_a: The population mean risk premium on US stocks is positive.

Note that "greater than" and "less than" alternative hypotheses reflect the beliefs of the researcher more strongly than a "not equal to" alternative hypothesis. To emphasize an attitude of neutrality, the researcher may sometimes select a "not equal to" alternative hypothesis when a one-sided alternative hypothesis is also reasonable.

The second step in hypothesis testing is identifying the appropriate test statistic and its probability distribution.

- **Definition of Test Statistic.** A test statistic is a quantity, calculated based on a sample, whose value is the basis for deciding whether or not to reject the null hypothesis.

The focal point of our statistical decision is the value of the test statistic. Frequently, the test statistic has the form[7]

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Value of the population parameter under } H_0}{\text{Standard error of the sample statistic}} \tag{1}$$

[6] Part of this discussion of the selection of hypotheses follows Bowerman, O'Connell, and Murphree (2016).

[7] In some cases, the test statistic may have a different form. For example, as we discuss in Section 4.3, the form of the test statistic for correlation coefficient is different.

For our risk premium example, the population parameter of interest is the population mean risk premium, μ_{RP}. We label the hypothesized value of the population mean under H_0 as μ_0. Restating the hypotheses using symbols, we test H_0: $\mu_{RP} \leq \mu_0$ versus H_a: $\mu_{RP} > \mu_0$. However, because under the null we are testing $\mu_0 = 0$, we write H_0: $\mu_{RP} \leq 0$ versus H_a: $\mu_{RP} > 0$.

The sample mean provides an estimate of the population mean. Therefore, we can use the sample mean risk premium calculated from historical data, $\overline{X}_{RP}$, as the sample statistic in Equation 1. The standard deviation of the sample statistic, known as the "standard error" of the statistic, is the denominator in Equation 1. For this example, the sample statistic is a sample mean. For a sample mean, $\overline{X}$, calculated from a sample generated by a population with standard deviation σ, the standard error is given by one of two expressions:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{2}$$

when we know σ (the population standard deviation), or

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} \tag{3}$$

when we do not know the population standard deviation and need to use the sample standard deviation s to estimate it. For this example, because we do not know the population standard deviation of the process generating the return, we use Equation 3. The test statistic is thus

$$\frac{\overline{X}_{RP} - \mu_0}{s_{\overline{X}}} = \frac{\overline{X}_{RP} - 0}{s/\sqrt{n}}$$

In making the substitution of 0 for μ_0, we use the fact already highlighted that we test any null hypothesis at the point of equality, as well as the fact that $\mu_0 = 0$ here.

We have identified a test statistic to test the null hypothesis. What probability distribution does it follow? We will encounter four distributions for test statistics in this reading:

- the t-distribution (for a t-test);
- the standard normal or z-distribution (for a z-test);
- the chi-square (χ^2) distribution (for a chi-square test); and
- the F-distribution (for an F-test).

We will discuss the details later, but assume we can conduct a z-test based on the central limit theorem because our US sample has many observations.[8] To summarize, the test statistic for the hypothesis test concerning the mean risk premium is $\overline{X}_{RP}/s_{\overline{X}}$. We can conduct a z-test because we can plausibly assume that the test statistic follows a standard normal distribution.

The third step in hypothesis testing is specifying the significance level. When the test statistic has been calculated, two actions are possible: 1) We reject the null hypothesis or 2) we do not reject the null hypothesis. The action we take is based on comparing the calculated test statistic to a specified possible value or values. The comparison values we choose are based on the level of significance selected. The level of significance reflects how much sample evidence we require to reject the null. Analogous to

[8] The central limit theorem says that the sampling distribution of the sample mean will be approximately normal with mean μ and variance σ^2/n when the sample size is large. The sample we will use for this example has 118 observations.

its counterpart in a court of law, the required standard of proof can change according to the nature of the hypotheses and the seriousness of the consequences of making a mistake. There are four possible outcomes when we test a null hypothesis:

1 We reject a false null hypothesis. This is a correct decision.
2 We reject a true null hypothesis. This is called a **Type I error**.
3 We do not reject a false null hypothesis. This is called a **Type II error**.
4 We do not reject a true null hypothesis. This is a correct decision.

We illustrate these outcomes in Exhibit 1.

Exhibit 1 Type I and Type II Errors in Hypothesis Testing

	True Situation	
Decision	**H_0 True**	**H_0 False**
Do not reject H_0	Correct Decision	Type II Error
Reject H_0 (accept H_a)	Type I Error	Correct Decision

When we make a decision in a hypothesis test, we run the risk of making either a Type I or a Type II error. These are mutually exclusive errors: If we mistakenly reject the null, we can only be making a Type I error; if we mistakenly fail to reject the null, we can only be making a Type II error.

The probability of a Type I error in testing a hypothesis is denoted by the Greek letter alpha, α. This probability is also known as the **level of significance** of the test. For example, a level of significance of 0.05 for a test means that there is a 5 percent probability of rejecting a true null hypothesis. The probability of a Type II error is denoted by the Greek letter beta, β.

Controlling the probabilities of the two types of errors involves a trade-off. All else equal, if we decrease the probability of a Type I error by specifying a smaller significance level (say 0.01 rather than 0.05), we increase the probability of making a Type II error because we will reject the null less frequently, including when it is false. The only way to reduce the probabilities of both types of errors simultaneously is to increase the sample size, n.

Quantifying the trade-off between the two types of error in practice is usually impossible because the probability of a Type II error is itself hard to quantify. Consider H_0: $\theta \leq 5$ versus H_a: $\theta > 5$. Because every true value of θ greater than 5 makes the null hypothesis false, each value of θ greater than 5 has a different β (Type II error probability). In contrast, it is sufficient to state a Type I error probability for $\theta = 5$, the point at which we conduct the test of the null hypothesis. Thus, in general, we specify only α, the probability of a Type I error, when we conduct a hypothesis test. Whereas the significance level of a test is the probability of incorrectly rejecting the null, the **power of a test** is the probability of *correctly* rejecting the null—that is, the probability of rejecting the null when it is false.[9] When more than one test statistic is available to conduct a hypothesis test, we should prefer the most powerful, all else equal.[10]

[9] The power of a test is, in fact, 1 minus the probability of a Type II error.

[10] We do not always have information on the relative power of the test for competing test statistics, however.

To summarize, the standard approach to hypothesis testing involves specifying a level of significance (probability of Type I error) only. It is most appropriate to specify this significance level prior to calculating the test statistic. If we specify it after calculating the test statistic, we may be influenced by the result of the calculation, which detracts from the objectivity of the test.

We can use three conventional significance levels to conduct hypothesis tests: 0.10, 0.05, and 0.01. Qualitatively, if we can reject a null hypothesis at the 0.10 level of significance, we have *some evidence* that the null hypothesis is false. If we can reject a null hypothesis at the 0.05 level, we have *strong evidence* that the null hypothesis is false. And if we can reject a null hypothesis at the 0.01 level, we have *very strong evidence* that the null hypothesis is false. For the risk premium example, we will specify a 0.05 significance level.

The fourth step in hypothesis testing is stating the decision rule. The general principle is simply stated. When we test the null hypothesis, if we find that the calculated value of the test statistic is more extreme than a given value or values determined by the specified level of significance, α, we reject the null hypothesis. We say the result is **statistically significant**. Otherwise, we do not reject the null hypothesis and we say the result is not statistically significant. The value or values with which we compare the calculated test statistic to make our decision are the rejection points (critical values) for the test.[11]

- **Definition of a Rejection Point (Critical Value) for the Test Statistic.** A rejection point (critical value) for a test statistic is a value with which the computed test statistic is compared to decide whether to reject or not reject the null hypothesis.

For a one-tailed test, we indicate a rejection point using the symbol for the test statistic with a subscript indicating the specified probability of a Type I error, α; for example, z_α. For a two-tailed test, we indicate $z_{\alpha/2}$. To illustrate the use of rejection points, suppose we are using a z-test and have chosen a 0.05 level of significance.

- For a test of H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$, two rejection points exist, one negative and one positive. For a two-sided test at the 0.05 level, the total probability of a Type I error must sum to 0.05. Thus, 0.05/2 = 0.025 of the probability should be in each tail of the distribution of the test statistic under the null. Consequently, the two rejection points are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$. Let z represent the calculated value of the test statistic. We reject the null if we find that $z < -1.96$ or $z > 1.96$. We do not reject if $-1.96 \leq z \leq 1.96$.
- For a test of H_0: $\theta \leq \theta_0$ versus H_a: $\theta > \theta_0$ at the 0.05 level of significance, the rejection point is $z_{0.05} = 1.645$. We reject the null hypothesis if $z > 1.645$. The value of the standard normal distribution such that 5 percent of the outcomes lie to the right is $z_{0.05} = 1.645$.
- For a test of H_0: $\theta \geq \theta_0$ versus H_a: $\theta < \theta_0$, the rejection point is $-z_{0.05} = -1.645$. We reject the null hypothesis if $z < -1.645$.

Exhibit 2 illustrates a test H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$ at the 0.05 significance level using a z-test. The "acceptance region" is the traditional name for the set of values of the test statistic for which we do not reject the null hypothesis. (The traditional name, however, is inaccurate. We should avoid using phrases such as "accept the null hypothesis" because such a statement implies a greater degree of conviction about the

[11] "Rejection point" is a descriptive synonym for the more traditional term "critical value."

null than is warranted when we fail to reject it.)[12] On either side of the acceptance region is a rejection region (or critical region). If the null hypothesis that $\mu = \mu_0$ is true, the test statistic has a 2.5 percent chance of falling in the left rejection region and a 2.5 percent chance of falling in the right rejection region. Any calculated value of the test statistic that falls in either of these two regions causes us to reject the null hypothesis at the 0.05 significance level. The rejection points of 1.96 and −1.96 are seen to be the dividing lines between the acceptance and rejection regions.

Exhibit 2 Rejection Points (Critical Values), 0.05 Significance Level, Two-Sided Test of the Population Mean Using a *z*-Test

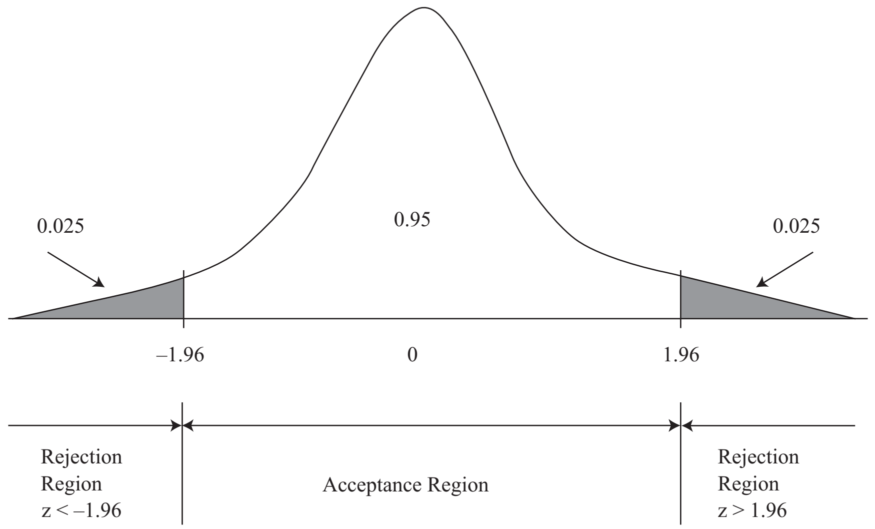


Exhibit 2 affords a good opportunity to highlight the relationship between confidence intervals and hypothesis tests. A 95 percent confidence interval for the population mean, μ, based on sample mean, $\overline{X}$, is given by $\overline{X} - 1.96s_{\overline{X}}$ to $\overline{X} + 1.96s_{\overline{X}}$, where $s_{\overline{X}}$ is the standard error of the sample mean (Equation 3).[13]

Now consider one of the conditions for rejecting the null hypothesis:

$$\frac{\overline{X} - \mu_0}{s_{\overline{X}}} > 1.96$$

Here, μ_0 is the hypothesized value of the population mean. The condition states that rejection is warranted if the test statistic exceeds 1.96. Multiplying both sides by $s_{\overline{X}}$, we have $\overline{X} - \mu_0 > 1.96s_{\overline{X}}$, or after rearranging, $\overline{X} - 1.96s_{\overline{X}} > \mu_0$, which we can also write as $\mu_0 < \overline{X} - 1.96s_{\overline{X}}$. This expression says that if the hypothesized population mean, μ_0, is less than the lower limit of the 95 percent confidence interval based on the sample mean, we must reject the null hypothesis at the 5 percent significance level (the test statistic falls in the rejection region to the right).

[12] The analogy in some courts of law (for example, in the United States) is that if a jury does not return a verdict of guilty (the alternative hypothesis), it is most accurate to say that the jury has failed to reject the null hypothesis, namely, that the defendant is innocent.

[13] Just as with the hypothesis test, we can use this confidence interval, based on the standard normal distribution, when we have large samples. An alternative hypothesis test and confidence interval uses the *t*-distribution, which requires concepts that we introduce in the next section.

Now, we can take the other condition for rejecting the null hypothesis:

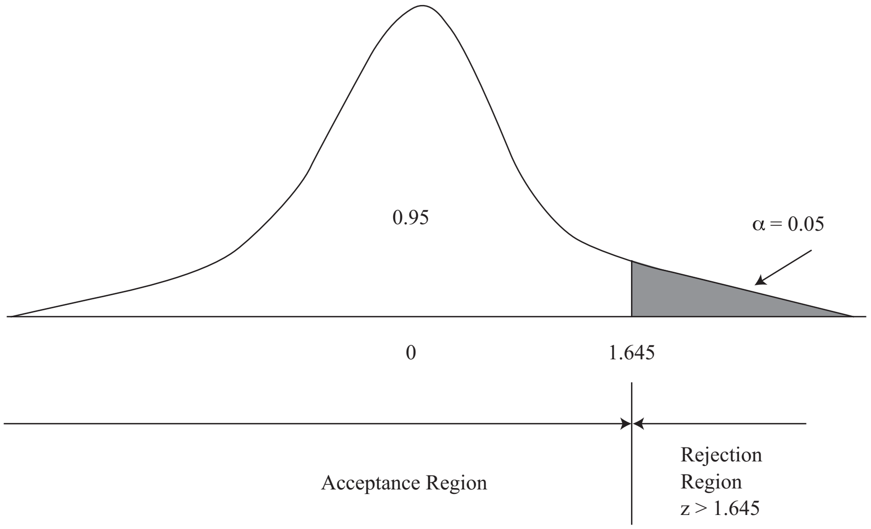
$$\frac{\overline{X} - \mu_0}{s_{\overline{X}}} < -1.96$$

and, using algebra as before, rewrite it as $\mu_0 > \overline{X} + 1.96 s_{\overline{X}}$. If the hypothesized population mean is larger than the upper limit of the 95 percent confidence interval, we reject the null hypothesis at the 5 percent level (the test statistic falls in the rejection region to the left). Thus, an α significance level in a two-sided hypothesis test can be interpreted in exactly the same way as a (1 − α) confidence interval.

In summary, when the hypothesized value of the population parameter under the null is outside the corresponding confidence interval, the null hypothesis is rejected. We could use confidence intervals to test hypotheses; practitioners, however, usually do not. Computing a test statistic (one number, versus two numbers for the usual confidence interval) is more efficient. Also, analysts encounter actual cases of one-sided confidence intervals only rarely. Furthermore, only when we compute a test statistic can we obtain a *p*-value, a useful quantity relating to the significance of our results (we will discuss *p*-values shortly).

To return to our risk premium test, we stated hypotheses H_0: $\mu_{RP} \le 0$ versus H_a: $\mu_{RP} > 0$. We identified the test statistic as $\overline{X}_{RP} / s_{\overline{X}}$ and stated that it follows a standard normal distribution. We are, therefore, conducting a one-sided *z*-test. We specified a 0.05 significance level. For this one-sided *z*-test, the rejection point at the 0.05 level of significance is 1.645. We will reject the null if the calculated *z*-statistic is larger than 1.645. Exhibit 3 illustrates this test.

Exhibit 3 Rejection Point (Critical Value), 0.05 Significance Level, One-Sided Test of the Population Mean Using a *z*-Test



The fifth step in hypothesis testing is collecting the data and calculating the test statistic. The quality of our conclusions depends not only on the appropriateness of the statistical model but also on the quality of the data we use in conducting the test. We first need to check for measurement errors in the recorded data. Some other issues to be aware of include sample selection bias and time-period bias. Sample selection bias refers to bias introduced by systematically excluding some members of the population according to a particular attribute. One type of sample selection bias is survivorship bias. For example, if we define our sample as US bond mutual funds currently operating and we collect returns for just these funds, we will systematically exclude funds that have not survived to the present date. Nonsurviving funds are likely to have underperformed surviving funds, on average; as a result the performance reflected in the sample may be biased upward. Time-period bias refers to the possibility that when we use a time-series sample, our statistical conclusion may be sensitive to the starting and ending dates of the sample.[14]

To continue with the risk premium hypothesis, we focus on US equities. According to Dimson, Marsh, and Staunton (2018) for the period 1900 to 2017 inclusive (118 annual observations), the arithmetic mean equity risk premium for US stocks relative to bill returns, $\overline{X}_{\text{RP}}$, was 7.5 percent per year. The sample standard deviation of the annual risk premiums was 19.5 percent. Using Equation 3, the standard error of the sample mean is $s_{\overline{X}} = s/\sqrt{n} = 19.5\%/\sqrt{118} = 1.795\%$. The test statistic is $z = \overline{X}_{\text{RP}}/s_{\overline{X}} = 7.5\%/1.795\% = 4.18$.

The sixth step in hypothesis testing is making the statistical decision. For our example, because the test statistic $z = 4.18$ is larger than the rejection point of 1.645, we reject the null hypothesis in favor of the alternative hypothesis that the risk premium on US stocks is positive. The first six steps are the statistical steps. The final decision concerns our use of the statistical decision.

The seventh and final step in hypothesis testing is making the economic or investment decision. The economic or investment decision takes into consideration not only the statistical decision but also all pertinent economic issues. In the sixth step, we found strong statistical evidence that the US risk premium is positive. The magnitude of the estimated risk premium, 7.5 percent a year, is economically very meaningful as well. Based on these considerations, an investor might decide to commit funds to US equities. A range of nonstatistical considerations, such as the investor's tolerance for risk and financial position, might also enter the decision-making process.

The preceding discussion raises an issue that often arises in this decision-making step. We frequently find that slight differences between a variable and its hypothesized value are statistically significant but not economically meaningful. For example, we may be testing an investment strategy and reject a null hypothesis that the mean return to the strategy is zero based on a large sample. Equation 1 shows that the smaller the standard error of the sample statistic (the divisor in the formula), the larger the value of the test statistic and the greater the chance the null will be rejected, all else equal. The standard error decreases as the sample size, n, increases, so that for very large samples, we can reject the null for small departures from it. We may find that although a strategy provides a statistically significant positive mean return, the results are not economically significant when we account for transaction costs, taxes, and risk. Even if we conclude that a strategy's results are economically meaningful, we should explore the logic of why the strategy might work in the future before actually implementing it. Such considerations cannot be incorporated into a hypothesis test.

Before leaving the subject of the process of hypothesis testing, we should discuss an important alternative approach called the p-value approach to hypothesis testing. Analysts and researchers often report the p-value (also called the marginal significance level) associated with hypothesis tests.

- **Definition of p-Value.** The p-value is the smallest level of significance at which the null hypothesis can be rejected.

For the value of the test statistic of 4.18 in the risk premium hypothesis test, using a spreadsheet function for the standard normal distribution, we calculate a p-value of 0.000015. We can reject the null hypothesis at that level of significance. The smaller

[14] These issues are discussed further in the reading on sampling.

the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis. The p-value for a two-sided test that a parameter equals zero is frequently generated automatically by statistical and econometric software programs.[15]

We can use p-values in the hypothesis testing framework presented above as an alternative to using rejection points. If the p-value is less than our specified level of significance, we reject the null hypothesis. Otherwise, we do not reject the null hypothesis. Using the p-value in this fashion, we reach the same conclusion as we do using rejection points. For example, because 0.000015 is less than 0.05, we would reject the null hypothesis in the risk premium test. The p-value, however, provides more precise information on the strength of the evidence than does the rejection points approach. The p-value of 0.000015 indicates that the null is rejected at a far smaller level of significance than 0.05.

If one researcher examines a question using a 0.05 significance level and another researcher uses a 0.01 significance level, the reader may have trouble comparing the findings. This concern has given rise to an approach to presenting the results of hypothesis tests that features p-values and omits specification of the significance level (Step 3). The interpretation of the statistical results is left to the consumer of the research. This has sometimes been called the p-value approach to hypothesis testing.[16]

3 HYPOTHESIS TESTS CONCERNING THE MEAN

Hypothesis tests concerning the mean are among the most common in practice. In this section we discuss such tests for several distinct types of problems. In one type (discussed in Section 3.1), we test whether the population mean of a single population is equal to (or greater or less than) some hypothesized value. Then, in Sections 3.2 and 3.3, we address inference on means based on two samples. Is an observed difference between two sample means due to chance or different underlying (population) means? When we have two random samples that are independent of each other—no relationship exists between the measurements in one sample and the measurements in the other—the techniques of Section 3.2 apply. When the samples are dependent, the methods of Section 3.3 are appropriate.[17]

3.1 Tests Concerning a Single Mean

An analyst who wants to test a hypothesis concerning the value of an underlying or population mean will conduct a t-test in the great majority of cases. A ***t*-test** is a hypothesis test using a statistic (t-statistic) that follows a t-distribution. The t-distribution is a probability distribution defined by a single parameter known as degrees of freedom (df). Each value of degrees of freedom defines one distribution in this family of

[15] We can use spreadsheets to calculate p-values as well. In Microsoft Excel, for example, we may use the worksheet functions TDIST, NORMSDIST, CHIDIST, and FDIST to calculate p-values for t-tests, z-tests, chi-square tests, and F-tests, respectively.

[16] Davidson and MacKinnon (1993) argued the merits of this approach: "The P value approach does not necessarily force us to make a decision about the null hypothesis. If we obtain a P value of, say, 0.000001, we will almost certainly want to reject the null. But if we obtain a P value of, say, 0.04, or even 0.004, we are not *obliged* to reject it. We may simply file the result away as information that casts some doubt on the null hypothesis, but that is not, by itself, conclusive. We believe that this somewhat agnostic attitude toward test statistics, in which they are merely regarded as pieces of information that we may or may not want to act upon, is usually the most sensible one to take." (p. 80)

[17] When we want to test whether the population means of more than two populations are equal, we use analysis of variance (ANOVA). We introduce ANOVA in its most common application, regression analysis, in the reading on linear regression.

distributions. The *t*-distribution is closely related to the standard normal distribution. Like the standard normal distribution, a *t*-distribution is symmetrical with a mean of zero. However, the *t*-distribution is more spread out: It has a standard deviation greater than 1 (compared to 1 for the standard normal)[18] and more probability for outcomes distant from the mean (it has fatter tails than the standard normal distribution). As the number of degrees of freedom increases with sample size, the spread decreases and the *t*-distribution approaches the standard normal distribution as a limit.

Why is the *t*-distribution the focus for the hypothesis tests of this section? In practice, investment analysts need to estimate the population standard deviation by calculating a sample standard deviation. That is, the population variance (or standard deviation) is unknown. For hypothesis tests concerning the population mean of a normally distributed population with unknown variance, the theoretically correct test statistic is the *t*-statistic. What if a normal distribution does not describe the population? The *t*-test is **robust** to moderate departures from normality, except for outliers and strong skewness.[19] When we have large samples, departures of the underlying distribution from the normal are of increasingly less concern. The sample mean is approximately normally distributed in large samples according to the central limit theorem, whatever the distribution describing the population. In general, a sample size of 30 or more usually can be treated as a large sample and a sample size of 29 or less is treated as a small sample.[20]

- **Test Statistic for Hypothesis Tests of the Population Mean (Practical Case—Population Variance Unknown).** If the population sampled has unknown variance and either of the conditions below holds:

  1 the sample is large, or

  2 the sample is small but the population sampled is normally distributed, or approximately normally distributed,

  then the test statistic for hypothesis tests concerning a single population mean, μ, is

$$t_{n-1} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \tag{4}$$

  where

  t_{n-1} = *t*-statistic with $n - 1$ degrees of freedom (n is the sample size)

  $\overline{X}$ = the sample mean

  μ_0 = the hypothesized value of the population mean

  s = the sample standard deviation

[18] The formula for the variance of a *t*-distribution is df/(df − 2).

[19] See Moore, McCabe, and Craig (2016). A statistic is robust if the required probability calculations are insensitive to violations of the assumptions.

[20] Although this generalization is useful, we caution that the sample size needed to obtain an approximately normal sampling distribution for the sample mean depends on how non-normal the original population is. For some populations, "large" may be a sample size well in excess of 30.

The denominator of the t-statistic is an estimate of the sample mean standard error, $s_{\overline{X}} = s/\sqrt{n}$.[21]

In Example 1, because the sample size is small, the test is called a small sample test concerning the population mean.

EXAMPLE 1

Risk and Return Characteristics of an Equity Mutual Fund (1)

You are analyzing Sendar Equity Fund, a midcap growth fund that has been in existence for 24 months. During this period, it has achieved a mean monthly return of 1.50 percent with a sample standard deviation of monthly returns of 3.60 percent. Given its level of systematic (market) risk and according to a pricing model, this mutual fund was expected to have earned a 1.10 percent mean monthly return during that time period. Assuming returns are normally distributed, are the actual results consistent with an underlying or population mean monthly return of 1.10 percent?

1 Formulate null and alternative hypotheses consistent with the description of the research goal.

2 Identify the test statistic for conducting a test of the hypotheses in Part 1.

3 Identify the rejection point or points for the hypothesis tested in Part 1 at the 0.10 level of significance.

4 Determine whether the null hypothesis is rejected or not rejected at the 0.10 level of significance. (Use the tables in the back of this book.)

Solution to 1:

We have a "not equal to" alternative hypothesis, where μ is the underlying mean return on Sendar Equity Fund—H_0: μ = 1.10 versus H_a: μ ≠ 1.10.

Solution to 2:

Because the population variance is not known, we use a t-test with 24 − 1 = 23 degrees of freedom.

Solution to 3:

Because this is a two-tailed test, we have the rejection point $t_{\alpha/2,n-1} = t_{0.05,23}$. In the table for the t-distribution, we look across the row for 23 degrees of freedom to the 0.05 column, to find 1.714. The two rejection points for this two-sided test are −1.714 and 1.714. We will reject the null if we find that $t < -1.714$ or $t > 1.714$.

Solution to 4:

$$t_{23} = \frac{1.50 - 1.10}{3.60/\sqrt{24}} = \frac{0.40}{0.734847} = 0.544331 \text{ or } 0.544$$

[21] A technical note, for reference, is required. When the sample comes from a finite population, estimates of the standard error of the mean, whether from Equation 2 or Equation 3, overestimate the true standard error. To address this, the computed standard error is multiplied by a shrinkage factor called the finite population correction factor (fpc), equal to $\sqrt{(N-n)/(N-1)}$, where N is the population size and n is the sample size. When the sample size is small relative to the population size (less than 5 percent of the population size), the fpc is usually ignored. The overestimation problem arises only in the usual situation of sampling without replacement (after an item is selected, it cannot be picked again) as opposed to sampling with replacement.

Because 0.544 does not satisfy either $t > 1.714$ or $t < -1.714$, we do not reject the null hypothesis.

The confidence interval approach provides another perspective on this hypothesis test. The theoretically correct 100(1 − α)% confidence interval for the population mean of a normal distribution with unknown variance, based on a sample of size n, is

$$\overline{X} - t_{\alpha/2}s_{\overline{X}} \text{ to } \overline{X} + t_{\alpha/2}s_{\overline{X}}$$

where $t_{\alpha/2}$ is the value of t such that α/2 of the probability remains in the right tail and where $-t_{\alpha/2}$ is the value of t such that α/2 of the probability remains in the left tail, for $n - 1$ degrees of freedom. Here, the 90 percent confidence interval runs from 1.5 − (1.714)(0.734847) = 0.240 to 1.5 + (1.714)(0.734847) = 2.760, compactly [0.240, 2.760]. The hypothesized value of mean return, 1.10, falls within this confidence interval, and we see from this perspective also that the null hypothesis is not rejected. At a 10 percent level of significance, we conclude that a population mean monthly return of 1.10 percent is consistent with the 24-month observed data series. Note that 10 percent is a relatively high probability of rejecting the hypothesis of a 1.10 percent population mean monthly return when it is true.

EXAMPLE 2

A Slowdown in Payments of Receivables

FashionDesigns, a supplier of casual clothing to retail chains, is concerned about a possible slowdown in payments from its customers. The controller's office measures the rate of payment by the average number of days in receivables.[22] FashionDesigns has generally maintained an average of 45 days in receivables. Because it would be too costly to analyze all of the company's receivables frequently, the controller's office uses sampling to track customers' payment rates. A random sample of 50 accounts shows a mean number of days in receivables of 49 with a standard deviation of 8 days.

1 Formulate null and alternative hypotheses consistent with determining whether the evidence supports the suspected condition that customer payments have slowed.

2 Identify the test statistic for conducting a test of the hypotheses in Part 1.

3 Identify the rejection point or points for the hypothesis tested in Part 1 at the 0.05 and 0.01 levels of significance.

4 Determine whether the null hypothesis is rejected or not rejected at the 0.05 and 0.01 levels of significance.

Solution to 1:

The suspected condition is that the number of days in receivables has increased relative to the historical rate of 45 days, which suggests a "greater than" alternative hypothesis. With μ as the population mean number of days in receivables, the hypotheses are H_0: $\mu \leq 45$ versus H_a: $\mu > 45$.

[22] This measure represents the average length of time that the business must wait after making a sale before receiving payment. The calculation is (Accounts receivable)/(Average sales per day).

Solution to 2:

Because the population variance is not known, we use a *t*-test with 50 − 1 = 49 degrees of freedom.

Solution to 3:

The rejection point is found across the row for degrees of freedom of 49. To find the one-tailed rejection point for a 0.05 significance level, we use the 0.05 column: The value is 1.677. To find the one-tailed rejection point for a 0.01 level of significance, we use the 0.01 column: The value is 2.405. To summarize, at a 0.05 significance level, we reject the null if we find that $t > 1.677$; at a 0.01 significance level, we reject the null if we find that $t > 2.405$.

Solution to 4:

$$t_{49} = \frac{49 - 45}{8/\sqrt{50}} = \frac{4}{1.131371} = 3.536$$

Because 3.536 > 1.677, the null hypothesis is rejected at the 0.05 level. Because 3.536 > 2.405, the null hypothesis is also rejected at the 0.01 level. We can say with a high level of confidence that FashionDesigns has experienced a slowdown in customer payments. The level of significance, 0.01, is a relatively low probability of rejecting the hypothesized mean of 45 days or less. Rejection gives us confidence that the mean has increased above 45 days.

We stated above that when population variance is not known, we use a *t*-test for tests concerning a single population mean. Given at least approximate normality, the *t*-test is always called for when we deal with small samples and do not know the population variance. For large samples, the central limit theorem states that the sample mean is approximately normally distributed, whatever the distribution of the population. So the *t*-test is still appropriate, but an alternative test may be more useful when sample size is large.

For large samples, practitioners sometimes use a *z*-test in place of a *t*-test for tests concerning a mean.[23] The justification for using the *z*-test in this context is twofold. First, in large samples, the sample mean should follow the normal distribution at least approximately, as we have already stated, fulfilling the normality assumption of the *z*-test. Second, the difference between the rejection points for the *t*-test and *z*-test becomes quite small when sample size is large. For a two-sided test at the 0.05 level of significance, the rejection points for a *z*-test are 1.96 and −1.96. For a *t*-test, the rejection points are 2.045 and −2.045 for df = 29 (about a 4 percent difference between the *z* and *t* rejection points) and 2.009 and −2.009 for df = 50 (about a 2.5 percent difference between the *z* and *t* rejection points). Because the *t*-test is readily available as statistical program output and theoretically correct for unknown population variance, we present it as the test of choice.

In a very limited number of cases, we may know the population variance; in such cases, the *z*-test is theoretically correct.[24]

- **The *z*-Test Alternative.**

[23] These practitioners choose between *t*-tests and *z*-tests based on sample size. For small samples ($n < 30$), they use a *t*-test, and for large samples, a *z*-test.

[24] For example, in Monte Carlo simulation, we prespecify the probability distributions for the risk factors. If we use a normal distribution, we know the true values of mean and variance. Monte Carlo simulation involves the use of a computer to represent the operation of a system subject to risk; we discuss Monte Carlo simulation in the reading on common probability distributions.

1 If the population sampled is normally distributed with known variance σ^2, then the test statistic for a hypothesis test concerning a single population mean, μ, is

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \qquad (5)$$

2 If the population sampled has unknown variance and the sample is large, in place of a *t*-test, an alternative test statistic (relying on the central limit theorem) is

$$z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \qquad (6)$$

In the above equations,

σ = the known population standard deviation
s = the sample standard deviation
μ_0 = the hypothesized value of the population mean

When we use a *z*-test, we most frequently refer to a rejection point in the list below.

- **Rejection Points for a *z*-Test.**

A Significance level of $\alpha = 0.10$.

1 H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$. The rejection points are $z_{0.05} = 1.645$ and –$z_{0.05} = -1.645$.

Reject the null hypothesis if $z > 1.645$ or if $z < -1.645$.

2 H_0: $\theta \leq \theta_0$ versus H_a: $\theta > \theta_0$. The rejection point is $z_{0.10} = 1.28$.

Reject the null hypothesis if $z > 1.28$.

3 H_0: $\theta \geq \theta_0$ versus H_a: $\theta < \theta_0$. The rejection point is $-z_{0.10} = -1.28$.

Reject the null hypothesis if $z < -1.28$.

B Significance level of $\alpha = 0.05$.

1 H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$. The rejection points are $z_{0.025} = 1.96$ and –$z_{0.025} = -1.96$.

Reject the null hypothesis if $z > 1.96$ or if $z < -1.96$.

2 H_0: $\theta \leq \theta_0$ versus H_a: $\theta > \theta_0$. The rejection point is $z_{0.05} = 1.645$.

Reject the null hypothesis if $z > 1.645$.

3 H_0: $\theta \geq \theta_0$ versus H_a: $\theta < \theta_0$. The rejection point is $-z_{0.05} = -1.645$.

Reject the null hypothesis if $z < -1.645$.

C Significance level of $\alpha = 0.01$.

1 H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$. The rejection points are $z_{0.005} = 2.575$ and $-z_{0.005} = -2.575$.

Reject the null hypothesis if $z > 2.575$ or if $z < -2.575$.

2 H_0: $\theta \leq \theta_0$ versus H_a: $\theta > \theta_0$. The rejection point is $z_{0.01} = 2.33$.

Reject the null hypothesis if $z > 2.33$.

3 H_0: $\theta \geq \theta_0$ versus H_a: $\theta < \theta_0$. The rejection point is $-z_{0.01} = -2.33$.

Reject the null hypothesis if $z < -2.33$.

Next, we present a historical example of conducting a hypothesis test on the potential impact of negative internal control disclosure by a company on its stock price.

EXAMPLE 3

The Effect of Control Deficiency Disclosures under the Sarbanes–Oxley Act on Share Prices

The Sarbanes–Oxley Act came into effect in 2002 and introduced major changes to the regulation of corporate governance and financial practice in the United States. One of the requirements of this Act is for firms to periodically assess and report certain types of internal control deficiencies to the audit committee, external auditors, and to the Securities and Exchange Commission (SEC). When a company makes an internal control weakness disclosure, does it convey information that affects the market value of the firm's stock?

Gupta and Nayar (2007) addressed this question by studying a number of voluntary disclosures made in the very early days of Sarbanes–Oxley implementation. Their final sample for this study consisted of 90 firms that had made control deficiency disclosures to the SEC from March 2003 to July 2004. This 90-firm sample was termed the "full sample". These firms were further examined to see if there were any other contemporaneous announcements, such as earnings announcements, associated with the control deficiency disclosures. Of the 90 firms, 45 did not have any such confounding announcements, and the sample of these firms was termed the "clean sample".

The announcement day of the internal control weakness was designated $t = 0$. If these announcements provide *new* information useful for equity valuation, the information should cause a change in stock prices and returns once it is available. Only one component of stock returns is of interest: the return in excess of that predicted given a stock's market risk or beta, called the abnormal return. Significant negative (positive) abnormal returns indicate that investors perceive unfavorable (favorable) corporate news in the internal control weakness announcement. Although Gupta and Nayar examined abnormal returns for various time horizons or event windows, we report a selection of their findings for the window [0, +1], which includes a two-day period of the day of and the day after the announcement. The researchers chose to use z-tests for statistical significance.

Full sample (90 firms). The null hypothesis that the average abnormal stock return during [0, +1] was 0 would be true if stock investors did not find either positive or negative information in the announcement.

Mean abnormal return = –3.07 percent.
z-statistic for abnormal return = –5.938.

Clean sample (45 firms). The null hypothesis that the average abnormal stock return during [0, +1] was 0 would be true if stock investors did not find either positive or negative information in the announcement.

Mean abnormal return = –1.87 percent.
z-statistic for abnormal return = –3.359.

1 With respect to both of the cases, suppose that the null hypothesis reflects the belief that investors do not, on average, perceive either positive or negative information in control deficiency disclosures. State one set of hypotheses (a null hypothesis and an alternative hypothesis) that covers both cases.

2 Determine whether the null hypothesis formulated in Part 1 is rejected or not rejected at the 0.05 and 0.01 levels of significance for the *full sample* case. Interpret the results.

3 Determine whether the null hypothesis formulated in Part 1 is rejected or not rejected at the 0.05 and 0.01 levels of significance for the *clean sample* case. Interpret the results.

Solution to 1:

A set of hypotheses consistent with no information in control deficiency disclosures relevant to stock investors is

H_0: The population mean abnormal return during [0, +1] equals 0.

H_a: The population mean abnormal return during [0, +1] does not equal 0.

Solution to 2:

From the information on rejection points for z-tests, we know that we reject the null hypothesis at the 0.05 significance level if $z > 1.96$ or if $z < -1.96$, and at the 0.01 significance level if $z > 2.575$ or if $z < -2.575$. The z-statistic reported by the researchers is −5.938, which is significant at the 0.05 and 0.01 levels. The null is rejected. The control deficiency disclosures appear to contain valuation-relevant information.

Because it is possible that significant results could be due to outliers, the researchers also reported the number of cases of positive and negative abnormal returns. The ratio of cases of positive to negative abnormal returns was 32:58, which tends to support the conclusion from the z-test of statistically significant negative abnormal returns.

Solution to 3:

The z-statistic reported by the researchers for the clean sample is −3.359, which is significant at the 0.05 and 0.01 levels. Although both the mean abnormal return and the z-statistic are smaller in magnitude for the clean sample than for the full sample, the results continue to be statistically significant.

The ratio of cases of positive to negative abnormal returns was 16:29, which tends to support the conclusion from the z-test of statistically significant negative abnormal returns.

Nearly all practical situations involve an unknown population variance. Exhibit 4 summarizes our discussion for tests concerning the population mean when the population variance is unknown.

Exhibit 4 Test Concerning the Population Mean (Population Variance Unknown)

	Large Sample ($n \geq 30$)	Small Sample ($n < 30$)
Population normal	t-Test (z-Test alternative)	t-Test
Population non-normal	t-Test (z-Test alternative)	Not Available

3.2 Tests Concerning Differences between Means

We often want to know whether a mean value—for example, a mean return—differs between two groups. Is an observed difference due to chance or to different underlying values for the mean? We have two samples, one for each group. When it is reasonable to believe that the samples are from populations at least approximately normally distributed and that the samples are also independent of each other, the techniques of this section apply. We discuss two *t*-tests for a test concerning differences between the means of two populations. In one case, the population variances, although unknown, can be assumed to be equal. Then, we efficiently combine the observations from both samples to obtain a pooled estimate of the common but unknown population variance. A pooled estimate is an estimate drawn from the combination of two different samples. In the second case, we do not assume that the unknown population variances are equal, and an approximate *t*-test is then available. Letting μ_1 and μ_2 stand, respectively, for the population means of the first and second populations, we most often want to test whether the population means, although unknown, are equal or whether one is larger than the other. Thus we usually formulate the following hypotheses:

1. H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$ (the alternative is that $\mu_1 \neq \mu_2$)
2. H_0: $\mu_1 - \mu_2 \leq 0$ versus H_a: $\mu_1 - \mu_2 > 0$ (the alternative is that $\mu_1 > \mu_2$)
3. H_0: $\mu_1 - \mu_2 \geq 0$ versus H_a: $\mu_1 - \mu_2 < 0$ (the alternative is that $\mu_1 < \mu_2$)

We can, however, formulate other hypotheses, such as H_0: $\mu_1 - \mu_2 = 2$ versus H_a: $\mu_1 - \mu_2 \neq 2$. The procedure is the same.

The definition of the *t*-test follows.

- **Test Statistic for a Test of the Difference between Two Population Means (Normally Distributed Populations, Population Variances Unknown but Assumed Equal).** When we can assume that the two populations are normally distributed and that the unknown population variances are equal, a *t*-test based on independent random samples is given by

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\left(\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}\right)^{1/2}} \tag{7}$$

where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is a pooled estimator of the common variance.

The number of degrees of freedom is $n_1 + n_2 - 2$.

EXAMPLE 4

Mean Returns on the S&P BSE SENSEX: A Test of Equality across Two Time Periods

The S&P BSE SENSEX is an index designed to measure the performance of the Indian stock market. The realized mean monthly return on this index in years 2012–2014 appears to have been substantially different than the mean return in years 2015–2017. Was the difference statistically significant? The data, shown in Exhibit 5, indicate that the difference in standard deviations during these two periods is small. Therefore, assuming equal population variances for returns in the two periods is not unreasonable.

Exhibit 5 S&P BSE SENSEX Monthly Return and Standard Deviation for Two Time Periods

Time Period	Number of Months (n)	Mean Monthly Return (%)	Standard Deviation
2012 through 2014	36	1.694	4.115
2015 through 2017	36	0.665	3.779

Source of data returns: https://www.asiaindex.co.in/indices/equity/sp-bse-sensex accessed 18 August 2018.

1 Formulate null and alternative hypotheses consistent with a two-sided hypothesis test.

2 Identify the test statistic for conducting a test of the hypotheses in Part 1.

3 Identify the rejection point or points for the hypothesis tested in Part 1 at the 0.10, 0.05, and 0.01 levels of significance.

4 Determine whether the null hypothesis is rejected or not rejected at the 0.10, 0.05, and 0.01 levels of significance.

Solution to 1:

Letting μ_1 represent the population mean return for the 2012 through 2014 and μ_2 represent the population mean return for the 2015 through 2017, we formulate the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0 \text{ versus } H_a: \mu_1 - \mu_2 \neq 0$$

Solution to 2:

Because the two samples are drawn from two different time periods, they are independent samples. The population variances are not known but can be assumed to be equal. Given all these considerations, the t-test given in Equation 7 has 36 + 36 − 2 = 70 degrees of freedom.

Solution to 3:

In the tables (Appendix B), for a two-sided test, the rejection points are ±1.667, ±1.994, and ±2.648 for, respectively, the 0.10, 0.05, and 0.01 levels for df = 70. To summarize, at the 0.10 level, we will reject the null if $t < -1.667$ or $t > 1.667$; at the 0.05 level, we will reject the null if $t < -1.994$ or $t > 1.994$; and at the 0.01 level, we will reject the null if $t < -2.648$ or $t > 2.648$.

Solution to 4:

In calculating the test statistic, the first step is to calculate the pooled estimate of variance:

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(36-1)(4.115)^2 + (36-1)(3.779)^2}{36 + 36 - 2} \\ &= \frac{1{,}092.4923}{70} \\ &= 15.6070 \end{aligned}$$

$$
\begin{aligned}
t &= \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\left(\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}\right)^{1/2}} \\
&= \frac{(1.694 - 0.665) - 0}{\left(\dfrac{15.6070}{36} + \dfrac{15.6070}{36}\right)^{1/2}} \\
&= \frac{1.029}{0.9312} \\
&= 1.11
\end{aligned}
$$

The calculated t statistic of 1.11 is not significant at the 0.10 level, so it is also not significant at the 0.05 and 0.01 levels. Therefore, we do not reject the null hypothesis at any of the three levels.

In many cases of practical interest, we cannot assume that population variances are equal. The following test statistic is often used in the investment literature in such cases:

- **Test Statistic for a Test of the Difference between Two Population Means (Normally Distributed Populations, Unequal and Unknown Population Variances).** When we can assume that the two populations are normally distributed but do not know the population variances and cannot assume that they are equal, an approximate t-test based on independent random samples is given by

$$
t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^{1/2}} \tag{8}
$$

where we use tables of the t-distribution using "modified" degrees of freedom computed with the formula

$$
\text{df} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2}} \tag{9}
$$

A practical tip is to compute the t-statistic before computing the degrees of freedom. Whether or not the t-statistic is significant will sometimes be obvious.

EXAMPLE 5

Recovery Rates on Defaulted Bonds: A Hypothesis Test

How are the required yields on risky corporate bonds determined? Two key factors are the expected probability of default and the expected amount that will be recovered in the event of default, or the recovery rate. Jankowitsch, Nagler, and Subrahmanyam (2014) examine the recovery rates of defaulted bonds in the US corporate bond market based on an extensive set of traded prices and volumes around various types of default events. For their study period, 2002 to

2012, Jankowitsch et al. confirm that the type of default event (e.g., distressed exchanges and formal bankruptcy filings), the seniority of the bond, and the industry of the firm are important in explaining the recovery rate. In one of their analyses, they focus on non-financial firms, and find that electricity firms recover more in default than firms in the retail industry. We want to test if the difference in recovery rates between those two types of firms is statistically significant. With μ_1 denoting the population mean recovery rate for the bonds of electricity firms and μ_2 denoting the population mean recovery rate for the bonds of retail firms, the hypotheses are H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$.

Exhibit 6 excerpts from their findings.

Exhibit 6 Recovery Rates by Industry of Firm

Electricity			Retail		
Number of Observations	Average Price[a]	Standard Deviation	Number of Observations	Average Price[a]	Standard Deviation
39	\$48.03	\$22.67	33	\$33.40	\$34.19

[a] This is the average traded price over the default day and the following 30 days after default; the average price provides an indication of the amount of money that can be recovered.

Source: Jankowitsch, Nagler, and Subrahmanyam (2013), Table 2.

We assume that the populations (recovery rates) are normally distributed and that the samples are independent. Based on the data in the table, address the following:

1 Discuss whether we should choose a test based on Equation 8 or Equation 7.

2 Calculate the test statistic to test the null hypothesis given above.

3 What is the value of the test's modified degrees of freedom?

4 Determine whether to reject the null hypothesis at the 0.10 level.

Solution to 1:

The sample standard deviation for the recovery rate on the bonds of electricity firms (\$22.67) appears much smaller than the sample standard deviation of the bonds for retail firms (\$34.19). Therefore, we should not assume equal variances, and accordingly, we should employ the approximate *t*-test given in Equation 8.

Solution to 2:

The test statistic is

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^{1/2}}$$

where

$\overline{X}_1$= sample mean recovery rate for electricity firms = 48.03

$\overline{X}_2$ = sample mean recovery rate for retail firms = 33.40

s_1^2 = sample variance for electricity firms = 22.67^2 = 513.9289

s_2^2 = sample variance for retail firms = 34.19^2 = 1,168.9561

n_1 = sample size of the electricity firms sample = 39
n_2 = sample size of the retail firms sample = 33

Thus, $t = (48.03 - 33.40)/[(513.9289/39) + (1{,}168.9561/33)]^{1/2} = 14.63/(13.177664 + 35.422912)^{1/2} = 14.63/6.971411 = 2.099$. The calculated t-statistic is thus 2.099.

Solution to 3:

$$\text{df} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2}} = \frac{\left(\dfrac{513.9289}{39} + \dfrac{1{,}168.9561}{33}\right)^2}{\dfrac{(513.9289/39)^2}{39} + \dfrac{(1{,}168.9561/33)^2}{33}}$$

$$= \frac{2362.016009}{42.476304} = 55.61 \text{ or } 56 \text{ degrees of freedom}$$

Solution to 4:

The closest entry to df = 56 in the tables for the t-distribution is df = 60. For $\alpha = 0.10$, we find $t_{\alpha/2} = 1.671$. Thus, we reject the null if $t < -1.671$ or $t > 1.671$. Based on the computed value of 2.099, we reject the null hypothesis at the 0.10 level. Some evidence exists that recovery rates differ between electricity and retail industries. Why? Studies on recovery rates suggest that the higher recovery rates of electricity firms may be explained by their higher levels of tangible assets.

3.3 Tests Concerning Mean Differences

In the previous section, we presented two t-tests for discerning differences between population means. The tests were based on two samples. An assumption for those tests' validity was that the samples were independent—i.e., unrelated to each other. When we want to conduct tests on two means based on samples that we believe are dependent, the methods of this section apply.

The t-test in this section is based on data arranged in **paired observations**, and the test itself is sometimes called a **paired comparisons test**. Paired observations are observations that are dependent because they have something in common. A paired comparisons test is a statistical test for differences in dependent items. For example, we may be concerned with the dividend policy of companies before and after a change in the tax law affecting the taxation of dividends. We then have pairs of "before" and "after" observations for the same companies. We may test a hypothesis about the mean of the differences (mean differences) that we observe across companies. In other cases, the paired observations are not on the same units. For example, we may be testing whether the mean returns earned by two investment strategies were equal over a study period. The observations here are dependent in the sense that there is one observation for each strategy in each month, and both observations depend on underlying market risk factors. Because the returns to both strategies are likely to be related to some common risk factors, such as the market return, the samples are dependent. By calculating a standard error based on differences, the t-test presented below takes account of correlation between the observations.

Letting A represent "after" and B "before," suppose we have observations for the random variables X_{A} and X_{B} and that the samples are dependent. We arrange the observations in pairs. Let d_i denote the difference between two paired observations. We can use the notation $d_i = x_{\text{A}i} - x_{\text{B}i}$, where $x_{\text{A}i}$ and $x_{\text{B}i}$ are the ith pair of observations, $i = 1, 2, \ldots, n$ on the two variables. Let μ_d stand for the population mean difference. We can formulate the following hypotheses, where μ_{d0} is a hypothesized value for the population mean difference:

1 H_0: $\mu_d = \mu_{d0}$ versus H_a: $\mu_d \neq \mu_{d0}$

2 H_0: $\mu_d \leq \mu_{d0}$ versus H_a: $\mu_d > \mu_{d0}$

3 H_0: $\mu_d \geq \mu_{d0}$ versus H_a: $\mu_d < \mu_{d0}$

In practice, the most commonly used value for μ_{d0} is 0.

As usual, we are concerned with the case of normally distributed populations with unknown population variances, and we will formulate a t-test. To calculate the t-statistic, we first need to find the sample mean difference:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{10}$$

where n is the number of pairs of observations. The sample variance, denoted by s_d^2, is

$$s_d^2 = \frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1} \tag{11}$$

Taking the square root of this quantity, we have the sample standard deviation, s_d, which then allows us to calculate the standard error of the mean difference as follows:[25]

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} \tag{12}$$

- **Test Statistic for a Test of Mean Differences (Normally Distributed Populations, Unknown Population Variances).** When we have data consisting of paired observations from samples generated by normally distributed populations with unknown variances, a t-test is based on

$$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}} \tag{13}$$

with $n - 1$ degrees of freedom, where n is the number of paired observations, $\bar{d}$ is the sample mean difference (as given by Equation 10), and $s_{\bar{d}}$ is the standard error of $\bar{d}$ (as given by Equation 12).

Exhibit 7 reports the quarterly returns for a six-year period for two managed portfolios specializing in precious metals. The two portfolios were closely similar in risk (as measured by standard deviation of return and other measures) and had nearly identical expense ratios. A major investment services company rated Portfolio B more

[25] We can also use the following equivalent expression, which makes use of the correlation between the two variables: $s_{\bar{d}} = \sqrt{s_{\text{A}}^2 + s_{\text{B}}^2 - 2r(X_{\text{A}}, X_{\text{B}})s_{\text{A}}s_{\text{B}}} \Big/ \sqrt{n}$ where s_{A}^2 is the sample variance of X_{A}, s_{B}^2 is the sample variance of X_{B}, and $r(X_{\text{A}}, X_{\text{B}})$ is the sample correlation between X_{A} and X_{B}.

highly than Portfolio A. In investigating the portfolios' relative performance, suppose we want to test the hypothesis that the mean quarterly return on Portfolio A equaled the mean quarterly return on Portfolio B during the six-year period. Because the two portfolios shared essentially the same set of risk factors, their returns were not independent, so a paired comparisons test is appropriate. Let μ_d stand for the population mean value of difference between the returns on the two portfolios during this period. We test H_0: $\mu_d = 0$ versus H_a: $\mu_d \neq 0$ at a 0.05 significance level.

Exhibit 7 Quarterly Returns on Two Managed Portfolios

Quarter	Portfolio A (%)	Portfolio B (%)	Difference (Portfolio A – Portfolio B)
4Q:Year 6	11.40	14.64	−3.24
3Q:Year 6	−2.17	0.44	−2.61
2Q:Year 6	10.72	19.51	−8.79
1Q:Year 6	38.91	50.40	−11.49
4Q:Year 5	4.36	1.01	3.35
3Q:Year 5	5.13	10.18	−5.05
2Q:Year 5	26.36	17.77	8.59
1Q:Year 5	−5.53	4.76	−10.29
4Q:Year 4	5.27	−5.36	10.63
3Q:Year 4	−7.82	−1.54	−6.28
2Q:Year 4	2.34	0.19	2.15
1Q:Year 4	−14.38	−12.07	−2.31
4Q:Year 3	−9.80	−9.98	0.18
3Q:Year 3	19.03	26.18	−7.15
2Q:Year 3	4.11	−2.39	6.50
1Q:Year 3	−4.12	−2.51	−1.61
4Q:Year 2	−0.53	−11.32	10.79
3Q:Year 2	5.06	0.46	4.60
2Q:Year 2	−14.01	−11.56	−2.45
1Q:Year 2	12.50	3.52	8.98
4Q:Year 1	−29.05	−22.45	−6.60
3Q:Year 1	3.60	0.10	3.50
2Q:Year 1	−7.97	−8.96	0.99
1Q:Year 1	−8.62	−0.66	−7.96
Mean	1.87	2.52	−0.65

Sample standard deviation of differences = 6.71

The sample mean difference, $\overline{d}$, between Portfolio A and Portfolio B is −0.65 percent per quarter. The standard error of the sample mean difference is $s_{\overline{d}} = 6.71/\sqrt{24} = 1.369673$. The calculated test statistic is $t = (-0.65 - 0)/1.369673 = -0.475$ with $n - 1 = 24 - 1 = 23$ degrees of freedom. At the 0.05 significance level, we reject the null if $t > 2.069$ or if $t < -2.069$. Because −0.475 is not less than −2.069, we fail to reject the null. At the 0.10 significance level, we reject the null if $t > 1.714$ or if $t < -1.714$. Thus, the difference in mean quarterly returns is not significant at any conventional significance level.

The following example illustrates the application of this test to evaluate two competing investment strategies.

EXAMPLE 6

A Comparison of Two Portfolios

You are investigating whether the performance of a portfolio of stocks from the entire world differs from the performance of a portfolio of only US stocks. For the worldwide portfolio, you choose to focus on Vanguard Total World Stock Index ETF. This ETF seeks to track the performance of the FTSE Global All Cap Index, which is a market-capitalization-weighted index designed to measure the market performance of stock of companies from both developed and emerging markets. For the US portfolio, you choose to focus on SPDR S&P 500, an ETF that seeks to track the performance of the S&P 500 Index. You analyze the monthly returns on both ETFs from August 2013 to July 2018 and prepare the following summary table.

Exhibit 8 Montly Return Summary for Vanguard Total World Stock Index ETF and SPDR S&P 500 ETF: August 2013 to July 2018 (*n* = 60)

Strategy	Mean Return	Standard Deviation
Worldwide	0.79%	2.93%
US	1.06	2.81
Difference	–0.27	1.00[a]

[a] Sample standard deviation of differences.
Source of data returns: finance.yahoo.com accessed 18 August 2018.

From Exhibit 8 we have $\overline{d}$ = –0.27% and s_d = 1.00%.

1. Formulate null and alternative hypotheses consistent with a two-sided test that the mean difference between the worldwide and only US strategies equals 0.
2. Identify the test statistic for conducting a test of the hypotheses in Part 1.
3. Identify the rejection point or points for the hypothesis tested in Part 1 at the 0.01 level of significance.
4. Determine whether the null hypothesis is rejected or not rejected at the 0.01 level of significance. (Use the tables in the back of this volume.)
5. Discuss the choice of a paired comparisons test.

Solution to 1:

With μ_d as the underlying mean difference between the worldwide and US strategies, we have H_0: $\mu_d = 0$ versus H_a: $\mu_d \neq 0$.

Solution to 2:

Because the population variance is unknown, the test statistic is a *t*-test with 60 – 1 = 59 degrees of freedom.

Solution to 3:

In the table for the t-distribution, the closest entry to df = 59 is df = 60. We look across the row for 60 degrees of freedom to the 0.005 column, to find 2.66. We will reject the null if we find that $t > 2.66$ or $t < -2.66$.

Solution to 4:

$$t_{59} = \frac{-0.27}{1.00/\sqrt{60}} = \frac{-0.27}{0.129099} = -2.09$$

Because −2.09 > −2.66, we cannot reject the null hypothesis. Accordingly, we conclude that the difference in mean returns for the two strategies is not statistically significant.

Solution to 5:

Several US stocks that are part of the S&P 500 index are also included in the Vanguard Total World Stock Index ETF. The profile of the World ETF indicates that nine of the top ten holdings in the ETF are US stocks. As a result, they are not independent samples; in general, the correlation of returns on the Vanguard Total World Stock Index ETF and SPDR S&P 500 ETF should be positive. Because the samples are dependent, a paired comparisons test was appropriate.

4 HYPOTHESIS TESTS CONCERNING VARIANCE AND CORRELATION

Because variance and standard deviation are widely used quantitative measures of risk in investments, analysts should be familiar with hypothesis tests concerning variance. The correlation between two variables is also widely used in investments. For example, investment managers often need to understand the correlations among returns on different assets. Therefore, analysts should also be familiar with hypothesis tests concerning correlation. The tests of variance and correlation discussed in this section make regular appearances in investment literature. Next, we examine two types of tests concerning variance: tests concerning the value of a single population variance and tests concerning the differences between two population variances. We then examine how to test the significance of a correlation coefficient.

4.1 Tests Concerning a Single Variance

In this section, we discuss testing hypotheses about the value of the variance, σ^2, of a single population. We use σ_0^2 to denote the hypothesized value of σ^2. We can formulate hypotheses as follows:

1 $H_0{:}\sigma^2 = \sigma_0^2$ versus $H_a{:}\ \sigma^2 \neq \sigma_0^2$ (a "not equal to" alternative hypothesis)

2 $H_0{:}\ \sigma^2 \leq \sigma_0^2$ versus $H_a{:}\ \sigma^2 > \sigma_0^2$ (a "greater than" alternative hypothesis)

3 $H_0{:}\ \sigma^2 \geq \sigma_0^2$ versus $H_a{:}\ \sigma^2 < \sigma_0^2$ (a "less than" alternative hypothesis)

In tests concerning the variance of a single normally distributed population, we make use of a chi-square test statistic, denoted χ^2. The chi-square distribution, unlike the normal and t-distributions, is asymmetrical. Like the t-distribution, the chi-square distribution is a family of distributions. A different distribution exists for each possible value of degrees of freedom, $n - 1$ (n is sample size). Unlike the t-distribution, the chi-square distribution is bounded below by 0; χ^2 does not take on negative values.

- **Test Statistic for Tests Concerning the Value of a Population Variance (Normal Population).** If we have n independent observations from a normally distributed population, the appropriate test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{14}$$

with $n - 1$ degrees of freedom. In the numerator of the expression is the sample variance, calculated as

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} \tag{15}$$

In contrast to the t-test, for example, the chi-square test is sensitive to violations of its assumptions. If the sample is not actually random or if it does not come from a normally distributed population, inferences based on a chi-square test are likely to be faulty.

If we choose a level of significance, α, the rejection points for the three kinds of hypotheses are as follows:

- **Rejection Points for Hypothesis Tests on the Population Variance.**
  1. "Not equal to" H_a: Reject the null hypothesis if the test statistic is greater than the upper α/2 point (denoted $\chi^2_{\alpha/2}$) or less than the lower α/2 point (denoted $\chi^2_{1-\alpha/2}$) of the chi-square distribution with df = $n - 1$.[26]
  2. "Greater than" H_a: Reject the null hypothesis if the test statistic is greater than the upper α point of the chi-square distribution with df = $n - 1$.
  3. "Less than" H_a: Reject the null hypothesis if the test statistic is less than the lower α point of the chi-square distribution with df = $n - 1$.

EXAMPLE 7

Risk and Return Characteristics of an Equity Mutual Fund (2)

You continue with your analysis of Sendar Equity Fund, a midcap growth fund that has been in existence for only 24 months. Recall that during this period, Sendar Equity achieved a sample standard deviation of monthly returns of 3.60 percent. You now want to test a claim that the specific investment approach followed by Sendar result in a standard deviation of monthly returns of less than 4 percent.

1. Formulate null and alternative hypotheses consistent with the verbal description of the research goal.

[26] Just as with other hypothesis tests, the chi-square test can be given a confidence interval interpretation. Unlike confidence intervals based on z- or t-statistics, however, chi-square confidence intervals for variance are asymmetric. A two-sided confidence interval for population variance, based on a sample of size n, has a lower limit $\text{L} = (n-1)s^2 \big/ \chi^2_{\alpha/2}$ and an upper limit $\text{U} = (n-1)s^2 \big/ \chi^2_{1-\alpha/2}$. Under the null hypothesis, the hypothesized value of the population variance should fall within these two limits.

2 Identify the test statistic for conducting a test of the hypotheses in Part 1.

3 Identify the rejection point or points for the hypothesis tested in Part 1 at the 0.05 level of significance.

4 Determine whether the null hypothesis is rejected or not rejected at the 0.05 level of significance. (Use the tables in the back of this volume.)

Solution to 1:

We have a "less than" alternative hypothesis, where σ is the underlying standard deviation of return on Sendar Equity Fund. Being careful to square standard deviation to obtain a test in terms of variance, the hypotheses are H_0: $\sigma^2 \geq 16.0$ versus H_a: $\sigma^2 < 16.0$.

Solution to 2:

The test statistic is χ^2 with 24 − 1 = 23 degrees of freedom.

Solution to 3:

The lower 0.05 rejection point is found on the line for df = 23, under the 0.95 column (95 percent probability in the right tail, to give 0.95 probability of getting a test statistic this large or larger). The rejection point is 13.091. We will reject the null if we find that χ^2 is less than 13.091.

Solution to 4:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{23 \times 3.60^2}{4^2} = \frac{298.08}{16} = 18.63$$

Because 18.63 (the calculated value of the test statistic) is not less than 13.091, we do not reject the null hypothesis. We cannot conclude that Sendar's investment disciplines result in a standard deviation of monthly returns of less than 4 percent.

4.2 Tests Concerning the Equality (Inequality) of Two Variances

Suppose we have a hypothesis about the relative values of the variances of two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We can formulate all hypotheses as one of the choices below:

1 H_0: $\sigma_1^2 = \sigma_2^2$ versus H_a: $\sigma_1^2 \neq \sigma_2^2$

2 H_0: $\sigma_1^2 \leq \sigma_2^2$ versus H_a: $\sigma_1^2 > \sigma_2^2$

3 H_0: $\sigma_1^2 \geq \sigma_2^2$ versus H_a: $\sigma_1^2 < \sigma_2^2$

Note that at the point of equality, the null hypothesis $\sigma_1^2 = \sigma_2^2$ implies that the ratio of population variances equals 1: $\sigma_1^2/\sigma_2^2 = 1$. Given independent random samples from these populations, tests related to these hypotheses are based on an *F*-test, which is the ratio of sample variances. Suppose we use n_1 observations in calculating the sample variance s_1^2 and n_2 observations in calculating the sample variance s_2^2. Tests concerning the difference between the variances of two populations make use of the *F*-distribution. Like the chi-square distribution, the *F*-distribution is a family of asymmetrical distributions bounded from below by 0. Each *F*-distribution is defined by two values of degrees of freedom, called the numerator and denominator degrees of freedom.[27] The *F*-test, like the chi-square test, is not robust to violations of its assumptions.

- **Test Statistic for Tests Concerning Differences between the Variances of Two Populations (Normally Distributed Populations).** Suppose we have two samples, the first with n_1 observations and sample variance s_1^2, the second with n_2 observations and sample variance s_2^2. The samples are random, independent of each other, and generated by normally distributed populations. A test concerning differences between the variances of the two populations is based on the ratio of sample variances

$$F = \frac{s_1^2}{s_2^2} \tag{16}$$

with $\text{df}_1 = n_1 - 1$ numerator degrees of freedom and $\text{df}_2 = n_2 - 1$ denominator degrees of freedom. Note that df_1 and df_2 are the divisors used in calculating s_1^2 and s_2^2, respectively.

A convention, or usual practice, is to use the larger of the two ratios s_1^2/s_2^2 or s_2^2/s_1^2 as the actual test statistic. When we follow this convention, the value of the test statistic is always greater than or equal to 1; tables of critical values of *F* then need include only values greater than or equal to 1. Under this convention, the rejection point for any formulation of hypotheses is a single value in the right-hand side of the relevant *F*-distribution. Note that the labeling of populations as "1" or "2" is arbitrary in any case.

- **Rejection Points for Hypothesis Tests on the Relative Values of Two Population Variances.** Follow the convention of using the larger of the two ratios s_1^2/s_2^2 and s_2^2/s_1^2 and consider two cases:

  1 A "not equal to" alternative hypothesis: Reject the null hypothesis at the α significance level if the test statistic is greater than the upper α/2 point of the *F*-distribution with the specified numerator and denominator degrees of freedom.

  2 A "greater than" or "less than" alternative hypothesis: Reject the null hypothesis at the α significance level if the test statistic is greater than the upper α point of the *F*-distribution with the specified number of numerator and denominator degrees of freedom.

Thus, if we conduct a two-sided test at the $\alpha = 0.01$ level of significance, we need to find the rejection point in *F*-tables at the $\alpha/2 = 0.01/2 = 0.005$ significance level for a one-sided test (Case 1). But a one-sided test at 0.01 uses rejection points in *F*-tables for $\alpha = 0.01$ (Case 2). As an example, suppose we are conducting a two-sided test at the 0.05 significance level. We calculate a value of *F* of 2.77 with 12 numerator and 19 denominator degrees of freedom. Using the *F*-tables for 0.05/2 = 0.025 in the back of the volume, we find that the rejection point is 2.72. Because the value 2.77 is greater than 2.72, we reject the null hypothesis at the 0.05 significance level.

[27] The relationship between the chi-square and *F*-distributions is as follows: If χ_1^2 is one chi-square random variable with *m* degrees of freedom and χ_2^2 is another chi-square random variable with *n* degrees of freedom, then $F = \left(\chi_1^2/m\right)/\left(\chi_2^2/n\right)$ follows an *F*-distribution with *m* numerator and *n* denominator degrees of freedom.

If the convention stated above is not followed and we are given a calculated value of F less than 1, can we still use F-tables? The answer is yes; using a reciprocal property of F-statistics, we can calculate the needed value. The easiest way to present this property is to show a calculation. Suppose our chosen level of significance is 0.05 for a two-tailed test and we have a value of F of 0.11, with 7 numerator degrees of freedom and 9 denominator degrees of freedom. We take the reciprocal, 1/0.11 = 9.09. Then we look up this value in the F-tables for 0.025 (because it is a two-tailed test) with degrees of freedom reversed: F for 9 numerator and 7 denominator degrees of freedom. In other words, $F_{9,7} = 1/F_{7,9}$ and 9.09 exceeds the critical value of 4.82, so $F_{7,9} = 0.11$ is significant at the 0.05 level.

EXAMPLE 8

Volatility and the Global Financial Crisis of the Late 2000s

You are investigating whether the population variance of returns on the KOSPI Index of the South Korean stock market changed subsequent to the global financial crisis that peaked in 2008. For this investigation, you are considering 1999 to 2006 as the pre-crisis period and 2010 to 2017 as the post-crisis period. You gather the data in Exhibit 9 for 418 weeks of returns during 1999 to 2006 and 418 weeks of returns during 2010 to 2017. You have specified a 0.01 level of significance.

Exhibit 9 KOSPI Index Returns and Variance before and after the Global Financial Crisis of the Late 2000s

	n	Mean Weekly Return (%)	Variance of Returns
Before crisis: 1999 to 2006	418	0.307	18.203
After crisis: 2010 to 2017	418	0.114	3.919

Source of data for returns: finance.yahoo.com accessed 19 August 2018.

1 Formulate null and alternative hypotheses consistent with the verbal description of the research goal.

2 Identify the test statistic for conducting a test of the hypotheses in Part 1.

3 Determine whether or not to reject the null hypothesis at the 0.01 level of significance. (Use the F-tables in the back of this volume.)

Solution to 1:

We have a "not equal to" alternative hypothesis:

$$H_0\colon \sigma^2_{\text{Before}} = \sigma^2_{\text{After}} \text{ versus } H_a\colon \sigma^2_{\text{Before}} \neq \sigma^2_{\text{After}}$$

Solution to 2:

To test a null hypothesis of the equality of two variances, we use $F = s_1^2 \big/ s_2^2$ with 418 − 1 = 417 numerator and denominator degrees of freedom.

Solution to 3:

The "before" sample variance is larger, so following a convention for calculating *F*-statistics, the "before" sample variance goes in the numerator: *F* = 18.203/3.919 = 4.645. Because this is a two-tailed test, we use *F*-tables for the 0.005 level (= 0.01/2) to give a 0.01 significance level. In the tables in the back of the volume, the closest value to 417 degrees of freedom is 120 degrees of freedom. At the 0.01 level, the rejection point is 1.61. Because 4.645 is greater than the critical value 1.61, we reject the null hypothesis that the population variance of returns is the same in the pre- and post-global financial crisis periods.[28] It seems that the South Korean market was more volatile before the financial crisis.

EXAMPLE 9

The Volatility of Derivatives Expiration Days

Since 2001, the financial markets in the United States have seen the quadruple occurrence of stock option, index option, index futures, and single stock futures expirations on the same day during four months of the year. Such days are known as "quadruple witching days." You are interested in investigating whether quadruple witching days exhibit greater volatility than normal days. Exhibit 10 presents the daily standard deviation of return for normal days and options/futures expiration days during a four-year period. The tabled data refer to options and futures on the 30 stocks that constitute the Dow Jones Industrial Average.

Exhibit 10 Standard Deviation of Return: Normal Trading Days and Derivatives Expiration Days

Type of Day	*n*	Standard Deviation (%)
Normal trading	138	0.821
Options/futures expiration	16	1.217

1 Formulate null and alternative hypotheses consistent with the belief that quadruple witching days display above-normal volatility.

2 Identify the test statistic for conducting a test of the hypotheses in Part 1.

3 Determine whether to reject the null hypothesis at the 0.05 level of significance. (Use the *F*-tables in the back of this volume.)

Solution to 1:

We have a "greater than" alternative hypothesis:

$$H_0\colon \sigma^2_{\text{Expirations}} \leq \sigma^2_{\text{Normal}} \text{ versus } H_a\colon \sigma^2_{\text{Expirations}} > \sigma^2_{\text{Normal}}$$

[28] The critical value decreases as the degrees of freedom increase. Therefore, the critical value for 417 degrees of freedom is even smaller than 1.61, and we can reject the null hypothesis.

Solution to 2:

Let σ_1^2 represent the variance of quadruple witching days, and σ_2^2 represent the variance of normal days, following the convention for the selection of the numerator and the denominator stated earlier. To test the null hypothesis, we use $F = s_1^2 / s_2^2$ with 16 – 1 = 15 numerator and 138 – 1 = 137 denominator degrees of freedom.

Solution to 3:

$F = (1.217)^2/(0.821)^2 = 1.481/0.674 = 2.20$. Because this is a one-tailed test at the 0.05 significance level, we use *F*-tables for the 0.05 level directly. In the tables in the back of the volume, the closest value to 137 degrees of freedom is 120 degrees of freedom. At the 0.05 level, the rejection point is 1.75. Because 2.20 is greater than 1.75, we reject the null hypothesis. It appears that quadruple witching days have above-normal volatility.

4.3 Tests Concerning Correlation

In many contexts in investments, we want to assess the strength of the linear relationship between two variables—the correlation between them. A common approach is to use the correlation coefficient. A significance test of a correlation coefficient allows us to assess whether the relationship between two random variables is the result of chance. If we decide that the relationship does not result from chance, we are inclined to use this information in predictions because a good prediction of one variable will help us predict the other variable.

If the correlation coefficient between two variables is zero, we would conclude that there is no linear relation between the two variables. We use a test of significance to assess whether the correlation is different from zero. After we estimate a correlation coefficient, we need to ask whether the estimated correlation is significantly different from 0. Before we can answer this question, we must know some details about the distribution of the underlying variables themselves. For purposes of simplicity, assume that both of the variables are normally distributed.[29]

We propose two hypotheses: the null hypothesis, H_0, that the correlation in the population is 0 ($\rho = 0$); and the alternative hypothesis, H_a, that the correlation in the population is different from 0 ($\rho \neq 0$). The alternative hypothesis is a test that the correlation is not equal to 0; therefore, a two-tailed test is appropriate. As long as the two variables are distributed normally, we can test to determine whether the null hypothesis should be rejected using the sample correlation, *r*. The formula for the *t*-test is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{17}$$

This test statistic has a *t*-distribution with $n - 2$ degrees of freedom if the null hypothesis is true. One practical observation concerning Equation 17 is that the magnitude of *r* needed to reject the null hypothesis H_0: $\rho = 0$ decreases as sample size *n* increases, for two reasons. First, as *n* increases, the number of degrees of freedom increases and the absolute value of the critical value t_c decreases. Second, the absolute value of the numerator increases with larger *n*, resulting in larger-magnitude *t*-values. For

[29] Actually, we must assume that each observation (x, y) on the two variables (X, Y) is a random observation from a bivariate normal distribution. Informally, in a bivariate or two-variable normal distribution, each individual variable is normally distributed and their joint relationship is completely described by the correlation, ρ, between them. For more details, see, for example, Daniel and Terrell (1995) and Greene (2018).

example, with sample size n = 12, r = 0.58 results in a t-statistic of 2.252 that is just significant at the 0.05 level (t_c = 2.228). With a sample size n = 32, a smaller sample correlation r = 0.35 yields a t-statistic of 2.046 that is just significant at the 0.05 level (t_c = 2.042); the r = 0.35 would not be significant with a sample size of 12 even at the 0.10 significance level. Another way to make this point is that sampling from the same population, a false null hypothesis H_0: ρ = 0 is more likely to be rejected as we increase sample size, all else equal, because a higher number of observations increases the numerator of the test statistic.

EXAMPLE 10

Testing the Yen–Canadian Dollar Return Correlation

The sample correlation between the GBP monthly returns to Japanese yen and Canadian dollar is 0.5132 for the period from January 2011 through December 2017 (*Source of exchange rate data*: http://fx.sauder.ubc.ca/).

Can we reject a null hypothesis that the underlying or population correlation equals 0 at the 0.05 level of significance?

Solution:

With 84 months from January 2011 through December 2017, we use the following statistic to test the null hypothesis, H_0, that the true correlation in the population is 0, against the alternative hypothesis, H_a, that the correlation in the population is different from 0:

$$t = \frac{0.5132\sqrt{84-2}}{\sqrt{1-0.5132^2}} = 5.4146$$

In the tables at the back of this volume, at the 0.05 significance level, the critical level for this test statistic is 1.99 (n = 84, degrees of freedom = 82). When the test statistic is either larger than 1.99 or smaller than −1.99, we can reject the hypothesis that the correlation in the population is 0. The test statistic is 5.4146, so we can reject the null hypothesis.

OTHER ISSUES: NONPARAMETRIC INFERENCE

5

The hypothesis-testing procedures we have discussed to this point have two characteristics in common. First, they are concerned with parameters, and second, their validity depends on a definite set of assumptions. Mean and variance, for example, are two parameters, or defining quantities, of a normal distribution. The tests also make specific assumptions—in particular, assumptions about the distribution of the population producing the sample. Any test or procedure with either of the above two characteristics is a **parametric test** or procedure. In some cases, however, we are concerned about quantities other than parameters of distributions. In other cases, we may believe that the assumptions of parametric tests do not hold for the particular data we have. In such cases, a nonparametric test or procedure can be useful. A **nonparametric test** is a test that is not concerned with a parameter, or a test that makes minimal assumptions about the population from which the sample comes.[30]

We primarily use nonparametric procedures in three situations: when the data we use do not meet distributional assumptions, when the data are given in ranks, or when the hypothesis we are addressing does not concern a parameter.

The first situation occurs when the data available for analysis suggest that the distributional assumptions of the parametric test are not satisfied. For example, we may want to test a hypothesis concerning the mean of a population but believe that neither a *t*-test nor a *z*-test is appropriate because the sample is small and may come from a markedly non-normally distributed population. In that case, we may use a nonparametric test. The nonparametric test will frequently involve the conversion of observations (or a function of observations) into ranks according to magnitude, and sometimes it will involve working with only "greater than" or "less than" relationships (using the signs + and – to denote those relationships). Characteristically, one must refer to specialized statistical tables to determine the rejection points of the test statistic, at least for small samples.[31] Such tests, then, typically interpret the null hypothesis as a thesis about ranks or signs. In Exhibit 11, we give examples of nonparametric alternatives to the parametric tests concerning means we have discussed in this reading.[32] The reader should consult a comprehensive business statistics textbook for an introduction to such tests, and a specialist textbook for details.[33]

Exhibit 11 Nonparametric Alternatives to Parametric Tests Concerning Means

	Parametric	Nonparametric
Tests concerning a single mean	*t*-test *z*-test	Wilcoxon signed-rank test
Tests concerning differences between means	*t*-test Approximate *t*-test	Mann–Whitney U test
Tests concerning mean differences (paired comparisons tests)	*t*-test	Wilcoxon signed-rank test Sign test

We pointed out that when we use nonparametric tests, we often convert the original data into ranks. In some cases, the original data are already ranked. In those cases, we also use nonparametric tests because parametric tests generally require a stronger measurement scale than ranks. For example, if our data were the rankings of investment managers, hypotheses concerning those rankings would be tested using nonparametric procedures. Ranked data also appear in many other finance contexts. For example, Heaney, Koga, Oliver, and Tran (1999) studied the relationship between the size of Japanese companies (as measured by revenue) and their use of derivatives. The companies studied used derivatives to hedge one or more of five types of risk exposure: interest rate risk, foreign exchange risk, commodity price risk, marketable

[30] Some writers make a distinction between "nonparametric" and "distribution-free" tests. They refer to procedures that do not concern the parameters of a distribution as nonparametric and to procedures that make minimal assumptions about the underlying distribution as distribution free. We follow a commonly accepted, inclusive usage of the term nonparametric.

[31] For large samples, there is often a transformation of the test statistic that permits the use of tables for the standard normal or *t*-distribution.

[32] In some cases, there are several nonparametric alternatives to a parametric test.

[33] See, for example, Hettmansperger and McKean (2010) or Siegel and Castellan (1988).

security price risk, and credit risk. The researchers gave a "perceived scope of risk exposure" score to each company that was equal to the number of types of risk exposure that the company reported hedging. Although revenue is measured on a strong scale (a ratio scale), scope of risk exposure is measured on only an ordinal scale.[34] The researchers thus employed nonparametric statistics to explore the relationship between derivatives usage and size.

A third situation in which we use nonparametric procedures occurs when our question does not concern a parameter. For example, if the question concerns whether a sample is random or not, we use the appropriate nonparametric test (a so-called "runs test"). Another type of question nonparametrics can address is whether a sample came from a population following a particular probability distribution (using the Kolmogorov–Smirnov test, for example).

We end this reading by describing in some detail a nonparametric statistic that has often been used in investment research, the Spearman rank correlation.

5.1 Nonparametric Tests Concerning Correlation: The Spearman Rank Correlation Coefficient

Earlier in this reading, we examined the *t*-test of the hypothesis that two variables are uncorrelated, based on the correlation coefficient. As we pointed out there, this test relies on fairly stringent assumptions. When we believe that the population under consideration meaningfully departs from those assumptions, we can employ a test based on the **Spearman rank correlation coefficient**, r_S. The Spearman rank correlation coefficient is essentially equivalent to the usual correlation coefficient calculated on the *ranks* of the two variables (say X and Y) within their respective samples. Thus it is a number between −1 and +1, where −1 (+1) denotes a perfect inverse (positive) straight-line relationship between the variables and 0 represents the absence of any straight-line relationship (no correlation). The calculation of r_S requires the following steps:

1. Rank the observations on X from largest to smallest. Assign the number 1 to the observation with the largest value, the number 2 to the observation with second-largest value, and so on. In case of ties, we assign to each tied observation the average of the ranks that they jointly occupy. For example, if the third- and fourth-largest values are tied, we assign both observations the rank of 3.5 (the average of 3 and 4). Perform the same procedure for the observations on Y.
2. Calculate the difference, d_i, between the ranks of each pair of observations on X and Y.
3. Then, with n the sample size, the Spearman rank correlation is given by[35]

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \qquad (18)$$

Suppose an investor wants to invest in a diversified emerging markets mutual fund. He has narrowed the field to 10 such funds, which are rated as 5-star funds by Morningstar. In examining the funds, a question arises as to whether the funds' most recent reported Sharpe ratios and expense ratios as of mid-2018 are related. Because the assumptions of the *t*-test on the correlation coefficient may not be met, it is appropriate to conduct

[34] We discussed scales of measurement in the reading on statistical concepts and market returns.

[35] Calculating the usual correlation coefficient on the ranks would yield approximately the same result as Equation 18.

a test on the rank correlation coefficient.[36] Exhibit 12 presents the calculation of r_S. The first two rows contain the original data. The row of X ranks converts the Sharpe ratios to ranks; the row of Y ranks converts the expense ratios to ranks. We want to test H_0: $\rho = 0$ versus H_a: $\rho \neq 0$, where ρ is defined in this context as the population correlation of X and Y after ranking. For small samples, the rejection points for the test based on r_S must be looked up in Exhibit 13. For large samples (say $n > 30$), we can conduct a t-test using

$$t = \frac{(n-2)^{1/2} r_s}{\left(1 - r_s^2\right)^{1/2}} \qquad (19)$$

based on $n - 2$ degrees of freedom.

Exhibit 12 The Spearman Rank Correlation: An Example

	Mutual Fund									
	1	2	3	4	5	6	7	8	9	10
Sharpe Ratio (X)	0.65	0.80	0.68	0.72	0.64	0.54	0.71	0.76	0.62	0.64
Expense Ratio (Y)	1.04	1.05	1.79	1.26	1.33	1.64	1.01	3.20	6.81	1.07
X Rank	5.5	1	5.5	3	7.5	10	4	2	9	7.5
Y Rank	9	8	3	6	5	4	10	2	1	7
d_i	−3.5	−7	2.5	−3	2.5	6	−6	0	8	0.5
d_i^2	12.25	49	6.25	9	6.25	36	36	0	64	0.25

$$r_s = 1 - \frac{6\sum d_i^2}{10(100-1)} = 1 - \frac{6(219)}{10(100-1)} = -0.3273$$

Source of Sharpe and Expense Ratios: http://markets.on.nytimes.com/research/screener/mutual_funds/mutual_funds.asp accessed 19 August 2018.

In the example at hand, a two-tailed test with a 0.05 significance level, Exhibit 13 gives the upper-tail rejection point for $n = 10$ as 0.6364 (we use the 0.025 column for a two-tailed test at a 0.05 significance level). Accordingly, we reject the null hypothesis if r_S is less than −0.6364 or greater than 0.6364. With r_S equal to−0.3273, we do not reject the null hypothesis.

Exhibit 13 Spearman Rank Correlation Distribution Approximate Upper-Tail Rejection Points

Sample Size: n	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
5	0.8000	0.9000	0.9000
6	0.7714	0.8286	0.8857

[36] The expense ratio (the ratio of a fund's operating expenses to average net assets) is bounded both from below (by zero) and from above. The Sharpe ratio is also observed within a limited range, in practice. Thus, neither variable can be normally distributed, and hence jointly they cannot follow a bivariate normal distribution. In short, the assumptions of a t-test are not met.

Exhibit 13 (Continued)

Sample Size: *n*	α = 0.05	α = 0.025	α = 0.01
7	0.6786	0.7450	0.8571
8	0.6190	0.7143	0.8095
9	0.5833	0.6833	0.7667
10	0.5515	0.6364	0.7333
11	0.5273	0.6091	0.7000
12	0.4965	0.5804	0.6713
13	0.4780	0.5549	0.6429
14	0.4593	0.5341	0.6220
15	0.4429	0.5179	0.6000
16	0.4265	0.5000	0.5824
17	0.4118	0.4853	0.5637
18	0.3994	0.4716	0.5480
19	0.3895	0.4579	0.5333
20	0.3789	0.4451	0.5203
21	0.3688	0.4351	0.5078
22	0.3597	0.4241	0.4963
23	0.3518	0.4150	0.4852
24	0.3435	0.4061	0.4748
25	0.3362	0.3977	0.4654
26	0.3299	0.3894	0.4564
27	0.3236	0.3822	0.4481
28	0.3175	0.3749	0.4401
29	0.3113	0.3685	0.4320
30	0.3059	0.3620	0.4251

Note: The corresponding lower tail critical value is obtained by changing the sign of the upper-tail critical value.

In the mutual fund example, we converted observations on two variables into ranks. If one or both of the original variables were in the form of ranks, we would need to use r_S to investigate correlation.

5.2 Nonparametric Inference: Summary

Nonparametric statistical procedures extend the reach of inference because they make few assumptions, can be used on ranked data, and may address questions unrelated to parameters. Quite frequently, nonparametric tests are reported alongside parametric tests. The reader can then assess how sensitive the statistical conclusion is to the assumptions underlying the parametric test. However, if the assumptions of the parametric test are met, the parametric test (where available) is generally preferred to the nonparametric test because the parametric test usually permits us to draw sharper conclusions.[37] For complete coverage of all the nonparametric procedures that may be encountered in the finance and investment literature, it is best to consult a specialist textbook.[38]

SUMMARY

In this reading, we have presented the concepts and methods of statistical inference and hypothesis testing.

- A hypothesis is a statement about one or more populations.
- The steps in testing a hypothesis are as follows:
  1. Stating the hypotheses.
  2. Identifying the appropriate test statistic and its probability distribution.
  3. Specifying the significance level.
  4. Stating the decision rule.
  5. Collecting the data and calculating the test statistic.
  6. Making the statistical decision.
  7. Making the economic or investment decision.
- We state two hypotheses: The null hypothesis is the hypothesis to be tested; the alternative hypothesis is the hypothesis accepted when the null hypothesis is rejected.
- There are three ways to formulate hypotheses:
  1. H_0: $\theta = \theta_0$ versus H_a: $\theta \neq \theta_0$
  2. H_0: $\theta \leq \theta_0$ versus H_a: $\theta > \theta_0$
  3. H_0: $\theta \geq \theta_0$ versus H_a: $\theta < \theta_0$

  where θ_0 is a hypothesized value of the population parameter and θ is the true value of the population parameter. In the above, Formulation 1 is a two-sided test and Formulations 2 and 3 are one-sided tests.
- When we have a "suspected" or "hoped for" condition for which we want to find supportive evidence, we frequently set up that condition as the alternative hypothesis and use a one-sided test. To emphasize a neutral attitude, however, the researcher may select a "not equal to" alternative hypothesis and conduct a two-sided test.
- A test statistic is a quantity, calculated on the basis of a sample, whose value is the basis for deciding whether to reject or not reject the null hypothesis. To decide whether to reject, or not to reject, the null hypothesis, we compare the computed value of the test statistic to a critical value (rejection point) for the same test statistic.
- In reaching a statistical decision, we can make two possible errors: We may reject a true null hypothesis (a Type I error), or we may fail to reject a false null hypothesis (a Type II error).

[37] To use a concept introduced in an earlier section, the parametric test is often more powerful.
[38] See, for example, Hettmansperger and McKean (2010) or Siegel and Castellan (1988).

- The level of significance of a test is the probability of a Type I error that we accept in conducting a hypothesis test. The probability of a Type I error is denoted by the Greek letter alpha, α. The standard approach to hypothesis testing involves specifying a level of significance (probability of Type I error) only.
- The power of a test is the probability of correctly rejecting the null (rejecting the null when it is false).
- A decision rule consists of determining the rejection points (critical values) with which to compare the test statistic to decide whether to reject or not to reject the null hypothesis. When we reject the null hypothesis, the result is said to be statistically significant.
- The (1 − α) confidence interval represents the range of values of the test statistic for which the null hypothesis will not be rejected at an α significance level.
- The statistical decision consists of rejecting or not rejecting the null hypothesis. The economic decision takes into consideration all economic issues pertinent to the decision.
- The *p*-value is the smallest level of significance at which the null hypothesis can be rejected. The smaller the *p*-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis. The *p*-value approach to hypothesis testing does not involve setting a significance level; rather it involves computing a *p*-value for the test statistic and allowing the consumer of the research to interpret its significance.
- For hypothesis tests concerning the population mean of a normally distributed population with unknown (known) variance, the theoretically correct test statistic is the *t*-statistic (*z*-statistic). In the unknown variance case, given large samples (generally, samples of 30 or more observations), the *z*-statistic may be used in place of the *t*-statistic because of the force of the central limit theorem.
- The *t*-distribution is a symmetrical distribution defined by a single parameter: degrees of freedom. Compared to the standard normal distribution, the *t*-distribution has fatter tails.
- When we want to test whether the observed difference between two means is statistically significant, we must first decide whether the samples are independent or dependent (related). If the samples are independent, we conduct tests concerning differences between means. If the samples are dependent, we conduct tests of mean differences (paired comparisons tests).
- When we conduct a test of the difference between two population means from normally distributed populations with unknown variances, if we can assume the variances are equal, we use a *t*-test based on pooling the observations of the two samples to estimate the common (but unknown) variance. This test is based on an assumption of independent samples.
- When we conduct a test of the difference between two population means from normally distributed populations with unknown variances, if we cannot assume that the variances are equal, we use an approximate *t*-test using modified degrees of freedom given by a formula. This test is based on an assumption of independent samples.
- In tests concerning two means based on two samples that are not independent, we often can arrange the data in paired observations and conduct a test of mean differences (a paired comparisons test). When the samples are from normally distributed populations with unknown variances, the appropriate test statistic is a *t*-statistic. The denominator of the *t*-statistic, the standard error of the mean differences, takes account of correlation between the samples.

- In tests concerning the variance of a single, normally distributed population, the test statistic is chi-square (χ^2) with $n - 1$ degrees of freedom, where n is sample size.
- For tests concerning differences between the variances of two normally distributed populations based on two random, independent samples, the appropriate test statistic is based on an F-test (the ratio of the sample variances).
- The F-statistic is defined by the numerator and denominator degrees of freedom. The numerator degrees of freedom (number of observations in the sample minus 1) is the divisor used in calculating the sample variance in the numerator. The denominator degrees of freedom (number of observations in the sample minus 1) is the divisor used in calculating the sample variance in the denominator. In forming an F-test, a convention is to use the larger of the two ratios, s_1^2/s_2^2 or s_2^2/s_1^2, as the actual test statistic.
- In tests concerning correlation, we use a t-statistic to test whether a population correlation coefficient is significantly different from 0. If we have n observations for two variables, this test statistic has a t-distribution with $n - 2$ degrees of freedom.
- A parametric test is a hypothesis test concerning a parameter or a hypothesis test based on specific distributional assumptions. In contrast, a nonparametric test either is not concerned with a parameter or makes minimal assumptions about the population from which the sample comes.
- A nonparametric test is primarily used in three situations: when data do not meet distributional assumptions, when data are given in ranks, or when the hypothesis we are addressing does not concern a parameter.
- The Spearman rank correlation coefficient is calculated on the ranks of two variables within their respective samples.

REFERENCES

Bowerman, Bruce L., Richard T. O'Connell, and Emily S. Murphree. 2016. *Business Statistics in Practice*, 8th edition. New York: McGraw-Hill/Irwin.

Daniel, Wayne W., and James C. Terrell. 1995. *Business Statistics for Management & Economics*, 7th edition. Boston: Houghton-Mifflin.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Dimson, Elroy, Paul Marsh, and Mike Staunton. 2018. "Credit Suisse Global Investment Returns Yearbook 2018 (Summary Edition)". Credit Suisse Research Institute.

Freeley, Austin J., and David L. Steinberg. 2013. *Argumentation and Debate: Critical Thinking for Reasoned Decision Making*, 13th edition. Boston, MA: Wadsworth Cengage Learning.

Gupta, Parveen P, and Nandkumar Nayar. 2007. "Information Content of Control Deficiency Disclosures under the Sarbanes-Oxley Act: An Empirical Investigation." *International Journal of Disclosure and Governance*, vol. 4:3–23.

Heaney, Richard, Chitoshi Koga, Barry Oliver, and Alfred Tran. 1999. "The Size Effect and Derivative Usage in Japan." Working paper: The Australian National University.

Hettmansperger, Thomas P., and Joseph W. McKean. 2010. *Robust Nonparametric Statistical Methods*, 2nd edition. Boca Raton, FL: CRC Press.

Jankowitsch, Rainer, Florian Nagler, and Marti G. Subrahmanyam. 2014. "The Determinants of Recovery Rates in the US Corporate Bond Market." *Journal of Financial Economics*, vol. 114, no. 1:155–177.

Moore, David S., George P. McCabe, and Bruce Craig. 2016. *Introduction to the Practice of Statistics*, 9th edition. New York: W.H. Freeman.

Siegel, Sidney, and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. New York: McGraw-Hill.

PRACTICE PROBLEMS

1 Which of the following statements about hypothesis testing is correct?

A The null hypothesis is the condition a researcher hopes to support.

B The alternative hypothesis is the proposition considered true without conclusive evidence to the contrary.

C The alternative hypothesis exhausts all potential parameter values not accounted for by the null hypothesis.

2 Identify the appropriate test statistic or statistics for conducting the following hypothesis tests. (Clearly identify the test statistic and, if applicable, the number of degrees of freedom. For example, "We conduct the test using an *x*-statistic with *y* degrees of freedom.")

A H_0: $\mu = 0$ versus H_a: $\mu \neq 0$, where μ is the mean of a normally distributed population with unknown variance. The test is based on a sample of 15 observations.

B H_0: $\mu = 0$ versus H_a: $\mu \neq 0$, where μ is the mean of a normally distributed population with unknown variance. The test is based on a sample of 40 observations.

C H_0: $\mu \leq 0$ versus H_a: $\mu > 0$, where μ is the mean of a normally distributed population with known variance σ^2. The sample size is 45.

D H_0: $\sigma^2 = 200$ versus H_a: $\sigma^2 \neq 200$, where σ^2 is the variance of a normally distributed population. The sample size is 50.

E $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a: \sigma_1^2 \neq \sigma_2^2$, where σ_1^2 is the variance of one normally distributed population and σ_2^2 is the variance of a second normally distributed population. The test is based on two independent random samples.

F H_0: (Population mean 1) – (Population mean 2) = 0 versus H_a: (Population mean 1) – (Population mean 2) ≠ 0, where the samples are drawn from normally distributed populations with unknown variances. The observations in the two samples are correlated.

G H_0: (Population mean 1) – (Population mean 2) = 0 versus H_a: (Population mean 1) – (Population mean 2) ≠ 0, where the samples are drawn from normally distributed populations with unknown but assumed equal variances. The observations in the two samples (of size 25 and 30, respectively) are independent.

3 For each of the following hypothesis tests concerning the population mean, μ, state the rejection point condition or conditions for the test statistic (e.g., $t > 1.25$); *n* denotes sample size.

A H_0: $\mu = 10$ versus H_a: $\mu \neq 10$, using a *t*-test with $n = 26$ and $\alpha = 0.05$

B H_0: $\mu = 10$ versus H_a: $\mu \neq 10$, using a *t*-test with $n = 40$ and $\alpha = 0.01$

C H_0: $\mu \leq 10$ versus H_a: $\mu > 10$, using a *t*-test with $n = 40$ and $\alpha = 0.01$

D H_0: $\mu \leq 10$ versus H_a: $\mu > 10$, using a *t*-test with $n = 21$ and $\alpha = 0.05$

E H_0: $\mu \geq 10$ versus H_a: $\mu < 10$, using a *t*-test with $n = 19$ and $\alpha = 0.10$

F H_0: $\mu \geq 10$ versus H_a: $\mu < 10$, using a *t*-test with $n = 50$ and $\alpha = 0.05$

4 For each of the following hypothesis tests concerning the population mean, μ, state the rejection point condition or conditions for the test statistic (e.g., $z > 1.25$); n denotes sample size.

A H_0: μ = 10 versus H_a: μ ≠ 10, using a z-test with n = 50 and α = 0.01

B H_0: μ = 10 versus H_a: μ ≠ 10, using a z-test with n = 50 and α = 0.05

C H_0: μ = 10 versus H_a: μ ≠ 10, using a z-test with n = 50 and α = 0.10

D H_0: μ ≤ 10 versus H_a: μ > 10, using a z-test with n = 50 and α = 0.05

5 Willco is a manufacturer in a mature cyclical industry. During the most recent industry cycle, its net income averaged $30 million per year with a standard deviation of $10 million ($n$ = 6 observations). Management claims that Willco's performance during the most recent cycle results from new approaches and that we can dismiss profitability expectations based on its average or normalized earnings of $24 million per year in prior cycles.

A With μ as the population value of mean annual net income, formulate null and alternative hypotheses consistent with testing Willco management's claim.

B Assuming that Willco's net income is at least approximately normally distributed, identify the appropriate test statistic.

C Identify the rejection point or points at the 0.05 level of significance for the hypothesis tested in Part A.

D Determine whether or not to reject the null hypothesis at the 0.05 significance level.

The following information relates to Questions 6–7

Performance in Forecasting Quarterly Earnings per Share

	Number of Forecasts	Mean Forecast Error (Predicted – Actual)	Standard Deviations of Forecast Errors
Analyst A	101	0.05	0.10
Analyst B	121	0.02	0.09

6 Investment analysts often use earnings per share (EPS) forecasts. One test of forecasting quality is the zero-mean test, which states that optimal forecasts should have a mean forecasting error of 0. (Forecasting error = Predicted value of variable − Actual value of variable.)

You have collected data (shown in the table above) for two analysts who cover two different industries: Analyst A covers the telecom industry; Analyst B covers automotive parts and suppliers.

A With μ as the population mean forecasting error, formulate null and alternative hypotheses for a zero-mean test of forecasting quality.

B For Analyst A, using both a t-test and a z-test, determine whether to reject the null at the 0.05 and 0.01 levels of significance.

C For Analyst B, using both a t-test and a z-test, determine whether to reject the null at the 0.05 and 0.01 levels of significance.

7 Reviewing the EPS forecasting performance data for Analysts A and B, you want to investigate whether the larger average forecast errors of Analyst A are due to chance or to a higher underlying mean value for Analyst A. Assume that the forecast errors of both analysts are normally distributed and that the samples are independent.

A Formulate null and alternative hypotheses consistent with determining whether the population mean value of Analyst A's forecast errors (μ_1) are larger than Analyst B's (μ_2).

B Identify the test statistic for conducting a test of the null hypothesis formulated in Part A.

C Identify the rejection point or points for the hypothesis tested in Part A, at the 0.05 level of significance.

D Determine whether or not to reject the null hypothesis at the 0.05 level of significance.

8 The table below gives data on the monthly returns on the S&P 500 and small-cap stocks for a forty-year period and provides statistics relating to their mean differences. Furthermore, the entire sample period is split into two subperiods of 20 years each and the returns data for these subperiods is also given in the table.

Measure	S&P 500 Return (%)	Small-Cap Stock Return (%)	Differences (S&P 500–Small-Cap Stock)
Entire sample period, 480 months			
Mean	1.0542	1.3117	–0.258
Standard deviation	4.2185	5.9570	3.752
First subperiod, 240 months			
Mean	0.6345	1.2741	–0.640
Standard deviation	4.0807	6.5829	4.096
Second subperiod, 240 months			
Mean	1.4739	1.3492	0.125
Standard deviation	4.3197	5.2709	3.339

Let μ_d stand for the population mean value of difference between S&P 500 returns and small-cap stock returns. Use a significance level of 0.05 and suppose that mean differences are approximately normally distributed.

A Formulate null and alternative hypotheses consistent with testing whether any difference exists between the mean returns on the S&P 500 and small-cap stocks.

B Determine whether or not to reject the null hypothesis at the 0.05 significance level for the entire sample period.

C Determine whether or not to reject the null hypothesis at the 0.05 significance level for the first subperiod.

D Determine whether or not to reject the null hypothesis at the 0.05 significance level for the second subperiod.

9 During a 10-year period, the standard deviation of annual returns on a portfolio you are analyzing was 15 percent a year. You want to see whether this record is sufficient evidence to support the conclusion that the portfolio's underlying variance of return was less than 400, the return variance of the portfolio's benchmark.

A Formulate null and alternative hypotheses consistent with the verbal description of your objective.

B Identify the test statistic for conducting a test of the hypotheses in Part A.

C Identify the rejection point or points at the 0.05 significance level for the hypothesis tested in Part A.

D Determine whether the null hypothesis is rejected or not rejected at the 0.05 level of significance.

10 You are investigating whether the population variance of returns on the S&P 500/BARRA Growth Index changed subsequent to the October 1987 market crash. You gather the following data for 120 months of returns before October 1987 and for 120 months of returns after October 1987. You have specified a 0.05 level of significance.

Time Period	*n*	Mean Monthly Return (%)	Variance of Returns
Before October 1987	120	1.416	22.367
After October 1987	120	1.436	15.795

A Formulate null and alternative hypotheses consistent with the verbal description of the research goal.

B Identify the test statistic for conducting a test of the hypotheses in Part A.

C Determine whether or not to reject the null hypothesis at the 0.05 level of significance. (Use the *F*-tables in the back of this volume.)

11 The following table shows the sample correlations between the monthly returns for four different mutual funds and the S&P 500. The correlations are based on 36 monthly observations. The funds are as follows:

Fund 1	Large-cap fund
Fund 2	Mid-cap fund
Fund 3	Large-cap value fund
Fund 4	Emerging markets fund
S&P 500	US domestic stock index

	Fund 1	Fund 2	Fund 3	Fund 4	S&P 500
Fund 1	1				
Fund 2	0.9231	1			
Fund 3	0.4771	0.4156	1		
Fund 4	0.7111	0.7238	0.3102	1	
S&P 500	0.8277	0.8223	0.5791	0.7515	1

Test the null hypothesis that each of these correlations, individually, is equal to zero against the alternative hypothesis that it is not equal to zero. Use a 5 percent significance level.

12 In the step "stating a decision rule" in testing a hypothesis, which of the following elements must be specified?

A Critical value

B Power of a test

C Value of a test statistic

13 Which of the following statements is correct with respect to the null hypothesis?

A It is considered to be true unless the sample provides evidence showing it is false.

B It can be stated as “not equal to” provided the alternative hypothesis is stated as “equal to.”

C In a two-tailed test, it is rejected when evidence supports equality between the hypothesized value and population parameter.

14 An analyst is examining a large sample with an unknown population variance. To test the hypothesis that the historical average return on an index is less than or equal to 6%, which of the following is the *most* appropriate test?

A One-tailed z-test

B Two-tailed z-test

C One-tailed F-test

15 A hypothesis test for a normally-distributed population at a 0.05 significance level implies a:

A 95% probability of rejecting a true null hypothesis.

B 95% probability of a Type I error for a two-tailed test.

C 5% critical value rejection region in a tail of the distribution for a one-tailed test.

16 Which of the following statements regarding a one-tailed hypothesis test is correct?

A The rejection region increases in size as the level of significance becomes smaller.

B A one-tailed test more strongly reflects the beliefs of the researcher than a two-tailed test.

C The absolute value of the rejection point is larger than that of a two-tailed test at the same level of significance.

17 The value of a test statistic is *best* described as the basis for deciding whether to:

A reject the null hypothesis.

B accept the null hypothesis.

C reject the alternative hypothesis.

18 Which of the following is a Type I error?

A Rejecting a true null hypothesis

B Rejecting a false null hypothesis

C Failing to reject a false null hypothesis

19 A Type II error is *best* described as:

A rejecting a true null hypothesis.

B failing to reject a false null hypothesis.

C failing to reject a false alternative hypothesis.

20 The level of significance of a hypothesis test is *best* used to:

A calculate the test statistic.

B define the test’s rejection points.

C specify the probability of a Type II error.

21 You are interested in whether excess risk-adjusted return (alpha) is correlated with mutual fund expense ratios for US large-cap growth funds. The following table presents the sample.

Mutual Fund	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alpha (*X*)	−0.52	−0.13	−0.60	−1.01	−0.26	−0.89	−0.42	−0.23	−0.60
Expense Ratio (*Y*)	1.34	0.92	1.02	1.45	1.35	0.50	1.00	1.50	1.45

A Formulate null and alternative hypotheses consistent with the verbal description of the research goal.

B Identify the test statistic for conducting a test of the hypotheses in Part A.

C Justify your selection in Part B.

D Determine whether or not to reject the null hypothesis at the 0.05 level of significance.

22 All else equal, is specifying a smaller significance level in a hypothesis test likely to increase the probability of a:

	Type I error?	Type II error?
A	No	No
B	No	Yes
C	Yes	No

23 The probability of correctly rejecting the null hypothesis is the:

A *p*-value.

B power of a test.

C level of significance.

24 The power of a hypothesis test is:

A equivalent to the level of significance.

B the probability of not making a Type II error.

C unchanged by increasing a small sample size.

25 When making a decision in investments involving a statistically significant result, the:

A economic result should be presumed meaningful.

B statistical result should take priority over economic considerations.

C economic logic for the future relevance of the result should be further explored.

26 An analyst tests the profitability of a trading strategy with the null hypothesis being that the average abnormal return before trading costs equals zero. The calculated *t*-statistic is 2.802, with critical values of ± 2.756 at significance level $\alpha = 0.01$. After considering trading costs, the strategy's return is near zero. The results are *most likely*:

A statistically but not economically significant.

B economically but not statistically significant.

C neither statistically nor economically significant.

27 Which of the following statements is correct with respect to the *p*-value?

A It is a less precise measure of test evidence than rejection points.

B It is the largest level of significance at which the null hypothesis is rejected.

C It can be compared directly with the level of significance in reaching test conclusions.

28 Which of the following represents a correct statement about the *p*-value?

A The *p*-value offers less precise information than does the rejection points approach.

B A larger *p*-value provides stronger evidence in support of the alternative hypothesis.

C A *p*-value less than the specified level of significance leads to rejection of the null hypothesis.

29 Which of the following statements on *p*-value is correct?

A The *p*-value is the smallest level of significance at which H_0 can be rejected.

B The *p*-value indicates the probability of making a Type II error.

C The lower the p-value, the weaker the evidence for rejecting the H_0.

30 The following table shows the significance level (α) and the *p*-value for three hypothesis tests.

	α	*p*-value
Test 1	0.05	0.10
Test 2	0.10	0.08
Test 3	0.10	0.05

The evidence for rejecting H_0 is strongest for:

A Test 1.

B Test 2.

C Test 3.

31 Which of the following tests of a hypothesis concerning the population mean is *most* appropriate?

A A *z*-test if the population variance is unknown and the sample is small

B A *z*-test if the population is normally distributed with a known variance

C A *t*-test if the population is non-normally distributed with unknown variance and a small sample

32 For a small sample with unknown variance, which of the following tests of a hypothesis concerning the population mean is most appropriate?

A A *t*-test if the population is normally distributed

B A *t*-test if the population is non-normally distributed

C A *z*-test regardless of the normality of the population distribution

33 For a small sample from a normally distributed population with unknown variance, the *most* appropriate test statistic for the mean is the:

A *z*-statistic.

B *t*-statistic.

C χ^2 statistic.

34 An investment consultant conducts two independent random samples of 5-year performance data for US and European absolute return hedge funds. Noting a 50 basis point return advantage for US managers, the consultant decides to test whether the two means are statistically different from one another at a 0.05 level of significance. The two populations are assumed to be normally distributed with unknown but equal variances. Results of the hypothesis test are contained in the tables below.

	Sample Size	Mean Return %	Standard Deviation
US Managers	50	4.7	5.4
European Managers	50	4.2	4.8

Null and Alternative Hypotheses	H_0: $\mu_{US} - \mu_E = 0$; H_a: $\mu_{US} - \mu_E \neq 0$
Test Statistic	0.4893
Critical Value Rejection Points	±1.984

μ_{US} is the mean return for US funds and μ_E is the mean return for European funds.

The results of the hypothesis test indicate that the:

A null hypothesis is not rejected.

B alternative hypothesis is statistically confirmed.

C difference in mean returns is statistically different from zero.

35 A pooled estimator is used when testing a hypothesis concerning the:

A equality of the variances of two normally distributed populations.

B difference between the means of two at least approximately normally distributed populations with unknown but assumed equal variances.

C difference between the means of two at least approximately normally distributed populations with unknown and assumed unequal variances.

36 When evaluating mean differences between two dependent samples, the *most* appropriate test is a:

A chi-square test.

B paired comparisons test.

C *z*-test.

37 A fund manager reported a 2% mean quarterly return over the past ten years for its entire base of 250 client accounts that all follow the same investment strategy. A consultant employing the manager for 45 client accounts notes that their mean quarterly returns were 0.25% less over the same period. The consultant tests the hypothesis that the return disparity between the returns of his clients and the reported returns of the fund manager's 250 client accounts are significantly different from zero.

Assuming normally distributed populations with unknown population variances, the *most* appropriate test statistic is:

A a paired comparisons *t*-test.

B a *t*-test of the difference between the two population means.

C an approximate *t*-test of mean differences between the two populations.

38 A chi-square test is *most* appropriate for tests concerning:

A a single variance.

B differences between two population means with variances assumed to be equal.

C differences between two population means with variances assumed to not be equal.

39 Which of the following should be used to test the difference between the variances of two normally distributed populations?

A *t*-test

B *F*-test

C Paired comparisons test

40 Jill Batten is analyzing how the returns on the stock of Stellar Energy Corp. are related with the previous month's percent change in the US Consumer Price Index for Energy (CPIENG). Based on 248 observations, she has computed the sample correlation between the Stellar and CPIENG variables to be −0.1452. She also wants to determine whether the sample correlation is statistically significant. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. Batten should conclude that the statistical relationship between Stellar and CPIENG is:

A significant, because the calculated test statistic has a lower absolute value than the critical value for the test statistic.

B significant, because the calculated test statistic has a higher absolute value than the critical value for the test statistic.

C not significant, because the calculated test statistic has a higher absolute value than the critical value for the test statistic.

41 In which of the following situations would a non-parametric test of a hypothesis *most likely* be used?

A The sample data are ranked according to magnitude.

B The sample data come from a normally distributed population.

C The test validity depends on many assumptions about the nature of the population.

42 An analyst is examining the monthly returns for two funds over one year. Both funds' returns are non-normally distributed. To test whether the mean return of one fund is greater than the mean return of the other fund, the analyst can use:

A a parametric test only.

B a nonparametric test only.

C both parametric and nonparametric tests.

SOLUTIONS

1 C is correct. Together, the null and alternative hypotheses account for all possible values of the parameter. Any possible values of the parameter not covered by the null must be covered by the alternative hypothesis (e.g., H_0: $\theta \leq 5$ versus H_a: $\theta > 5$).

2 A The appropriate test statistic is a *t*-statistic with $n - 1 = 15 - 1 = 14$ degrees of freedom. A *t*-statistic is theoretically correct when the sample comes from a normally distributed population with unknown variance. When the sample size is also small, there is no practical alternative.

B The appropriate test statistic is a *t*-statistic with $40 - 1 = 39$ degrees of freedom. A *t*-statistic is theoretically correct when the sample comes from a normally distributed population with unknown variance. When the sample size is large (generally, 30 or more is a "large" sample), it is also possible to use a *z*-statistic, whether the population is normally distributed or not. A test based on a *t*-statistic is more conservative than a *z*-statistic test.

C The appropriate test statistic is a *z*-statistic because the sample comes from a normally distributed population with known variance. (The known population standard deviation is used to compute the standard error of the mean using Equation 2 in the text.)

D The appropriate test statistic is chi-square (χ^2) with $50 - 1 = 49$ degrees of freedom.

E The appropriate test statistic is the *F*-statistic (the ratio of the sample variances).

F The appropriate test statistic is a *t*-statistic for a paired observations test (a paired comparisons test), because the samples are correlated.

G The appropriate test statistic is a *t* –statistic using a pooled estimate of the population variance. The *t*-statistic has $25 + 30 - 2 = 53$ degrees of freedom. This statistic is appropriate because the populations are normally distributed with unknown variances; because the variances are assumed equal, the observations can be pooled to estimate the common variance. The requirement of independent samples for using this statistic has been met.

3 A With degrees of freedom (df) $n - 1 = 26 - 1 = 25$, the rejection point conditions for this two-sided test are $t > 2.060$ and $t < -2.060$. Because the significance level is 0.05, 0.05/2 = 0.025 of the probability is in each tail. The tables give one-sided (one-tailed) probabilities, so we used the 0.025 column. Read across df = 25 to the $\alpha = 0.025$ column to find 2.060, the rejection point for the right tail. By symmetry, −2.060 is the rejection point for the left tail.

B With df = 39, the rejection point conditions for this two-sided test are $t > 2.708$ and $t < -2.708$. This is a two-sided test, so we use the 0.01/2 = 0.005 column. Read across df = 39 to the $\alpha = 0.005$ column to find 2.708, the rejection point for the right tail. By symmetry, −2.708 is the rejection point for the left tail.

C With df = 39, the rejection point condition for this one-sided test is $t > 2.426$. Read across df = 39 to the $\alpha = 0.01$ column to find 2.426, the rejection point for the right tail. Because we have a "greater than" alternative, we are concerned with only the right tail.

D With df = 20, the rejection point condition for this one-sided test is $t > 1.725$. Read across df = 20 to the $\alpha = 0.05$ column to find 1.725, the rejection point for the right tail. Because we have a "greater than" alternative, we are concerned with only the right tail.

E With df = 18, the rejection point condition for this one-sided test is $t < -1.330$. Read across df = 18 to the $\alpha = 0.10$ column to find 1.330, the rejection point for the right tail. By symmetry, the rejection point for the left tail is −1.330.

F With df = 49, the rejection point condition for this one-sided test is $t < -1.677$. Read across df = 49 to the $\alpha = 0.05$ column to find 1.677, the rejection point for the right tail. By symmetry, the rejection point for the left tail is −1.677.

4 Recall that with a *z*-test (in contrast to the *t*-test), we do not employ degrees of freedom. The standard normal distribution is a single distribution applicable to all *z*-tests. You should refer to "Rejection Points for a *z*-Test" in Section 3.1 to answer these questions.

A This is a two-sided test at a 0.01 significance level. In Part C of "Rejection Points for a *z*-Test," we find that the rejection point conditions are $z > 2.575$ and $z < -2.575$.

B This is a two-sided test at a 0.05 significance level. In Part B of "Rejection Points for a *z*-Test," we find that the rejection point conditions are $z > 1.96$ and $z < -1.96$.

C This is a two-sided test at a 0.10 significance level. In Part A of "Rejection Points for a *z*-Test," we find that the rejection point conditions are $z > 1.645$ and $z < -1.645$.

D This is a one-sided test at a 0.05 significance level. In Part B of "Rejection Points for a *z*-Test," we find that the rejection point condition for a test with a "greater than" alternative hypothesis is $z > 1.645$.

5 **A** As stated in the text, we often set up the "hoped for" or "suspected" condition as the alternative hypothesis. Here, that condition is that the population value of Willco's mean annual net income exceeds $24 million. Thus we have H_0: $\mu \le 24$ versus H_a: $\mu > 24$.

B Given that net income is normally distributed with unknown variance, the appropriate test statistic is t with $n - 1 = 6 - 1 = 5$ degrees of freedom.

C In the *t*-distribution table in the back of the book, in the row for df = 5 under $\alpha = 0.05$, we read the rejection point (critical value) of 2.015. We will reject the null if $t > 2.015$.

D The *t*-test is given by Equation 4:

$$t_5 = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{30 - 24}{10/\sqrt{6}} = \frac{6}{4.082483} = 1.469694$$

or 1.47. Because 1.47 does not exceed 2.015, we do not reject the null hypothesis. The difference between the sample mean of $30 million and the hypothesized value of $24 million under the null is not statistically significant.

6 **A** H_0: $\mu = 0$ versus H_a: $\mu \ne 0$.

B The *t*-test is based on $t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$ with $n - 1 = 101 - 1 = 100$ degrees of freedom. At the 0.05 significance level, we reject the null if $t > 1.984$ or if $t < -1.984$. At the 0.01 significance level, we reject the null if $t > 2.626$ or if $t <$

−2.626. For Analyst A, we have $t = (0.05 - 0)\Big/\left(0.10/\sqrt{101}\right)$ = 0.05/0.00995 = 5.024938 or 5.025. We clearly reject the null hypothesis at both the 0.05 and 0.01 levels.

The calculation of the *z*-statistic with unknown variance, as in this case, is the same as the calculation of the *t*-statistic. The rejection point conditions for a two-tailed test are as follows: $z > 1.96$ and $z < -1.96$ at the 0.05 level; and $z > 2.575$ and $z < -2.575$ at the 0.01 level. Note that the *z*-test is a less conservative test than the *t*-test, so when the *z*-test is used, the null is easier to reject. Because $z = 5.025$ is greater than 2.575, we reject the null at the 0.01 level; we also reject the null at the 0.05 level.

In summary, Analyst A's EPS forecasts appear to be biased upward—they tend to be too high.

C For Analyst B, the *t*-test is based on *t* with 121 − 1 = 120 degrees of freedom. At the 0.05 significance level, we reject the null if $t > 1.980$ or if $t < -1.980$. At the 0.01 significance level, we reject the null if $t > 2.617$ or if $t < -2.617$. We calculate $t = (0.02 - 0)\Big/\left(0.09/\sqrt{121}\right)$ = 0.02/0.008182 = 2.444444 or 2.44. Because 2.44 > 1.98, we reject the null at the 0.05 level. However, 2.44 is not larger than 2.617, so we do not reject the null at the 0.01 level.

For a *z*-test, the rejection point conditions are the same as given in Part B, and we come to the same conclusions as with the *t*-test. Because 2.44 > 1.96, we reject the null at the 0.05 significance level; however, because 2.44 is not greater than 2.575, we do not reject the null at the 0.01 level.

The mean forecast error of Analyst B is only $0.02; but because the test is based on a large number of observations, it is sufficient evidence to reject the null of mean zero forecast errors at the 0.05 level.

7 **A** Stating the suspected condition as the alternative hypothesis, we have

$$H_0: \mu_1 - \mu_2 \le 0 \text{ versus } H_a: \mu_1 - \mu_2 > 0$$

where

μ_1 = the population mean value of Analyst A's forecast errors
μ_2 = the population mean value of Analyst B's forecast errors

B We have two normally distributed populations with unknown variances. Based on the samples, it is reasonable to assume that the population variances are equal. The samples are assumed to be independent; this assumption is reasonable because the analysts cover quite different industries. The appropriate test statistic is *t* using a pooled estimate of the common variance. The number of degrees of freedom is

$$n_1 + n_2 - 2 = 101 + 121 - 2 = 222 - 2 = 220.$$

C For df = 200 (the closest value to 220), the rejection point for a one-sided test at the 0.05 significance level is 1.653.

D We first calculate the pooled estimate of variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(101 - 1)(0.10)^2 + (121 - 1)(0.09)^2}{101 + 121 - 2}$$

$$= \frac{1.972}{220} = 0.008964$$

Then

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\left(\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}\right)^{1/2}} = \frac{\left(0.05 - 0.02\right) - 0}{\left(\dfrac{0.008964}{101} + \dfrac{0.008964}{121}\right)^{1/2}}$$

$$= \frac{0.03}{0.01276} = 2.351018$$

or 2.35. Because 2.35 > 1.653, we reject the null hypothesis in favor of the alternative hypothesis that the population mean forecast error of Analyst A is greater than that of Analyst B.

8 **A** We test H_0: $\mu_d = 0$ versus H_a: $\mu_d \neq 0$.

B This is a paired comparisons t-test with $n - 1 = 480 - 1 = 479$ degrees of freedom. At the 0.05 significance level, we reject the null hypothesis if either $t > 1.96$ or $t < -1.96$. We use df = ∞ in the t-distribution table under $\alpha = 0.025$ because we have a very large sample and a two-sided test.

$$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}} = \frac{-0.258 - 0}{3.752 / \sqrt{480}} = \frac{-0.258}{0.171255} = -1.506529 \text{ or } -1.51$$

At the 0.05 significance level, because neither rejection point condition is met, we do not reject the null hypothesis that the mean difference between the returns on the S&P 500 and small-cap stocks during the entire sample period was 0.

C This t-test now has $n - 1 = 240 - 1 = 239$ degrees of freedom. At the 0.05 significance level, we reject the null hypothesis if either $t > 1.972$ or $t < -1.972$, using df = 200 in the t-distribution tables.

$$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}} = \frac{-0.640 - 0}{4.096 / \sqrt{240}} = \frac{-0.640}{0.264396} = -2.420615 \text{ or } -2.42$$

Because $-2.42 < -1.972$, we reject the null hypothesis at the 0.05 significance level. During this subperiod, small-cap stocks significantly outperformed the S&P 500.

D This t-test has $n - 1 = 240 - 1 = 239$ degrees of freedom. At the 0.05 significance level, we reject the null hypothesis if either $t > 1.972$ or $t < -1.972$, using df = 200 in the t-distribution tables.

$$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}} = \frac{0.125 - 0}{3.339 / \sqrt{240}} = \frac{0.125}{0.215532} = 0.579962 \text{ or } 0.58$$

At the 0.05 significance level, because neither rejection point condition is met, we do not reject the null hypothesis that for the second subperiod, the mean difference between the returns on the S&P 500 and small-cap stocks was zero.

9 **A** We have a "less than" alternative hypothesis, where σ^2 is the variance of return on the portfolio. The hypotheses are H_0: $\sigma^2 \geq 400$ versus H_a: $\sigma^2 < 400$, where 400 is the hypothesized value of variance, σ_0^2.

B The test statistic is chi-square with $10 - 1 = 9$ degrees of freedom.

C The rejection point is found across degrees of freedom of 9, under the 0.95 column (95 percent of probability above the value). It is 3.325. We will reject the null hypothesis if we find that $\chi^2 < 3.325$.

D The test statistic is calculated as

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9 \times 15^2}{400} = \frac{2{,}025}{400} = 5.0625 \text{ or } 5.06$$

Because 5.06 is not less than 3.325, we do not reject the null hypothesis.

10 A We have a "not equal to" alternative hypothesis:

$$H_0\colon \sigma^2_{\text{Before}} = \sigma^2_{\text{After}} \text{ versus } H_a\colon \sigma^2_{\text{Before}} \neq \sigma^2_{\text{After}}$$

B To test a null hypothesis of the equality of two variances, we use an *F*-test:

$$F = \frac{s_1^2}{s_2^2}$$

C The "before" sample variance is larger, so following a convention for calculating *F*-statistics, the "before" sample variance goes in the numerator. *F* = 22.367/15.795 = 1.416, with 120 − 1 = 119 numerator and denominator degrees of freedom. Because this is a two-tailed test, we use *F*-tables for the 0.025 level (df = 0.05/2). Using the tables in the back of the volume, the closest value to 119 is 120 degrees of freedom. At the 0.05 level, the rejection point is 1.43. (Using the Insert/Function/Statistical feature on a Microsoft Excel spreadsheet, we would find FINV(0.025, 119, 119) = 1.434859 as the critical *F*-value.) Because 1.416 is not greater than 1.43, we do not reject the null hypothesis that the "before" and "after" variances are equal.

11 The critical *t*-value for $n - 2 = 34$ df, using a 5 percent significance level and a two-tailed test, is 2.032. First, take the smallest correlation in the table, the correlation between Fund 3 and Fund 4, and see if it is significantly different from zero. Its calculated *t*-value is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.3102\sqrt{36-2}}{\sqrt{1-0.3102^2}} = 1.903$$

This correlation is not significantly different from zero. If we take the next lowest correlation, between Fund 2 and Fund 3, this correlation of 0.4156 has a calculated *t*-value of 2.664. So this correlation is significantly different from zero at the 5 percent level of significance. All of the other correlations in the table (besides the 0.3102) are greater than 0.4156, so they too are significantly different from zero.

12 A is correct. The critical value in a decision rule is the rejection point for the test. It is the point with which the test statistic is compared to determine whether to reject the null hypothesis, which is part of the fourth step in hypothesis testing.

13 A is correct. The null hypothesis is the hypothesis to be tested. The null hypothesis is considered to be true unless the evidence indicates that it is false, in which case the alternative hypothesis is accepted.

14 A is correct. If the population sampled has unknown variance and the sample is large, a *z*-test may be used. Hypotheses involving "greater than" or "less than" postulations are one-sided (one-tailed). In this situation, the null and alternative hypotheses are stated as H_0: μ ≤ 6% and H_a: μ > 6%, respectively. A one-tailed *t*-test is also acceptable in this case.

15 C is correct. For a one-tailed hypothesis test, there is a 5% critical value rejection region in one tail of the distribution.

16 B is correct. One-tailed tests in which the alternative is "greater than" or "less than" represent the beliefs of the researcher more firmly than a "not equal to" alternative hypothesis.

17 A is correct. Calculated using a sample, a test statistic is a quantity whose value is the basis for deciding whether to reject the null hypothesis.

18 A is correct. The definition of a Type I error is when a true null hypothesis is rejected.

19 B is correct. A Type II error occurs when a false null hypothesis is not rejected.

20 B is correct. The level of significance is used to establish the rejection points of the hypothesis test.

21 **A** We have a "not equal to" alternative hypothesis:

$$H_0: \rho = 0 \text{ versus } H_a: \rho \neq 0$$

B We would use the nonparametric Spearman rank correlation coefficient to conduct the test.

C Mutual fund expense ratios are bounded from above and below, and in practice there is at least a lower bound on alpha (as any return cannot be less than −100 percent). These variables are markedly non-normally distributed, and the assumptions of a parametric test are not likely to be fulfilled. Thus a nonparametric test appears to be appropriate.

D The calculation of the Spearman rank correlation coefficient is given in the following table.

Mutual Fund	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Alpha (X)	−0.52	−0.13	−0.60	−1.01	−0.26	−0.89	−0.42	−0.23	−0.60
Expense Ratio (Y)	1.34	0.92	1.02	1.45	1.35	0.50	1.00	1.50	1.45
X Rank	5	1	6.5	9	3	8	4	2	6.5
Y Rank	5	8	6	2.5	4	9	7	1	2.5
d_i	0	−7	0.5	6.5	−1	−1	−3	1	4
d_i^2	0	49	0.25	42.25	1	1	9	1	16

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(119.50)}{9(81 - 1)} = 0.0042$$

We use Table 11 to tabulate the rejection points for a test on the Spearman rank correlation. Given a sample size of 9 in a two-tailed test at a 0.05 significance level, the upper-tail rejection point is 0.6833 (we use the 0.025 column). Thus we reject the null hypothesis if the Spearman rank correlation coefficient is less than −0.6833 or greater than 0.6833. Because r_S is equal to 0.0042, we do not reject the null hypothesis.

22 B is correct. Specifying a smaller significance level decreases the probability of a Type I error (rejecting a true null hypothesis), but increases the probability of a Type II error (not rejecting a false null hypothesis). As the level of significance decreases, the null hypothesis is less frequently rejected.

23 B is correct. The power of a test is the probability of rejecting the null hypothesis when it is false.

24 B is correct. The power of a hypothesis test is the probability of correctly rejecting the null when it is false. Failing to reject the null when it is false is a Type II error. Thus, the power of a hypothesis test is the probability of not committing a Type II error.

25 C is correct. When a statistically significant result is also economically meaningful, one should further explore the logic of why the result might work in the future.

26 A is correct. The hypothesis is a two-tailed formulation. The *t*-statistic of 2.802 falls outside the critical rejection points of less than –2.756 and greater than 2.756, therefore the null hypothesis is rejected; the result is statistically significant. However, despite the statistical results, trying to profit on the strategy is not likely to be economically meaningful because the return is near zero after transaction costs

27 C is correct. When directly comparing the *p*-value with the level of significance, it can be used as an alternative to using rejection points to reach conclusions on hypothesis tests. If the *p*-value is smaller than the specified level of significance, the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected.

28 C is correct. The *p*-value is the smallest level of significance at which the null hypothesis can be rejected for a given value of the test statistic. The null hypothesis is rejected when the *p*-value is less than the specified significance level.

29 A is correct. The *p*-value is the smallest level of significance (α) at which the null hypothesis can be rejected.

30 C is correct. The *p*-value is the smallest level of significance (α) at which the null hypothesis can be rejected. If the *p*-value is less than α, the null can be rejected. The smaller the *p*-value, the stronger the evidence is against the null hypothesis and in favor of the alternative hypothesis. Thus, the evidence for rejecting the null is strongest for Test 3.

31 B is correct. The *z*-test is theoretically the correct test to use in those limited cases when testing the population mean of a normally distributed population with known variance.

32 A is correct. A *t*-test is used if the sample is small and drawn from a normally or approximately normally distributed population.

33 B is correct. A *t*-statistic is the most appropriate for hypothesis tests of the population mean when the variance is unknown and the sample is small but the population is normally distributed.

34 A is correct. The *t*-statistic value of 0.4893 does not fall into the critical value rejection regions (≤ –1.984 or > 1.984). Instead it falls well within the acceptance region. Thus, H_0 cannot be rejected; the result is not statistically significant at the 0.05 level.

35 B is correct. The assumption that the variances are equal allows for the combining of both samples to obtain a pooled estimate of the common variance.

36 B is correct. A paired comparisons test is appropriate to test the mean differences of two samples believed to be dependent.

37 A is correct. The sample sizes for both the fund manager and the consultant's accounts consists of forty quarterly periods of returns. However, the consultant's client accounts are a subset of the fund manager's entire account base. As such, they are not independent samples. When samples are dependent, a paired comparisons test is appropriate to conduct tests of the differences in dependent items.

38 A is correct. A chi-square test is used for tests concerning the variance of a single normally distributed population.

39 B is correct. An *F*-test is used to conduct tests concerning the difference between the variances of two normally distributed populations with random independent samples.

40 B is correct. The calculated test statistic is

$$\begin{aligned} t &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ &= \frac{-0.1452\sqrt{248-2}}{\sqrt{1-\left(-0.1452\right)^2}} = -2.30177 \end{aligned}$$

Because the absolute value of $t = -2.30177$ is greater than 1.96, the correlation coefficient is statistically significant.

41 A is correct. A non-parametric test is used when the data are given in ranks.

42 B is correct. There are only 12 (monthly) observations over the one year of the sample and thus the samples are small. Additionally, the funds' returns are non-normally distributed. Therefore, the samples do not meet the distributional assumptions for a parametric test. The Mann–Whitney U test (a nonparametric test) could be used to test the differences between population means.