The following is a review of the Quantitative Methods principles designed to address the learning outcome statements set forth by CFA Institute®. This topic is also covered in:

COMMON PROBABILITY DISTRIBUTIONS

Study Session 3

EXAM FOCUS

This topic review contains a lot of very testable material. Learn the difference between discrete and continuous probability distributions. The binomial and normal distributions are the most important here. You must learn the properties of both distributions and memorize the formulas for the mean and variance of the binomial distribution and for the probability of a particular value when given a binomial probability distribution. Learn what shortfall risk is and how to calculate and use Roy's safety-first criterion. Know how to standardize a normally distributed random variable, use a *z*-table, and construct confidence intervals. These skills will be used repeatedly in the topic reviews that follow. Additionally, understand the basic features of the lognormal distribution, Monte Carlo simulation, and historical simulation. Finally, it would be a good idea to know how to get continuously compounded rates of return from holding period returns. Other than that, no problem.

LOS 9.a: Define and explain a probability distribution and distinguish between discrete and continuous random variables.

LOS 9.b: Describe the set of possible outcomes of a specified discrete random variable.

A **probability distribution** describes the probabilities of all the possible outcomes for a random variable. The probabilities of all possible outcomes must sum to 1. A simple probability distribution is that for the roll of one fair die; there are six possible outcomes and each one has a probability of 1/6, so they sum to 1. The probability distribution of all the possible returns on the S&P 500 index for the next year is a more complex version of the same idea.

A **discrete random variable** is one for which the number of possible outcomes can be counted, and for each possible outcome, there is a measurable and positive probability. An example of a discrete random variable is the number of days it rains in a given month, because there is a finite number of possible outcomes—the number of days it can rain in a month is defined by the number of days in the month.

A **continuous random variable** is one for which the number of possible outcomes is infinite, even if lower and upper bounds exist. The actual amount of daily rainfall between zero and 100 inches is an example of a continuous random variable because the actual amount of rainfall can take on an infinite number of values. Daily rainfall can be measured in inches, half inches, quarter inches, thousandths of inches, or in even smaller increments. Thus, the number of possible daily rainfall amounts between zero and 100 inches is essentially infinite.

The assignment of probabilities to the possible outcomes for discrete and continuous random variables provides us with discrete probability distributions and continuous probability distributions. The difference between these types of distributions is most apparent for the following properties:

- For a *discrete distribution,* $p(x) = 0$ when x cannot occur, or $p(x) > 0$ if it can. Recall that $p(x)$ is read: "the probability that random variable $X = x$." For example, the probability of it raining on 33 days in June is zero because this cannot occur, but the probability of it raining 25 days in June has some positive value.

- For a *continuous distribution,* $p(x) = 0$ even though x can occur. We can only consider $P(x_1 \leq X \leq x_2)$ where x_1 and x_2 are actual numbers. For example, the probability of receiving two inches of rain in June is zero because two inches is a single point in an infinite range of possible values. On the other hand, the probability of the amount of rain being between 1.99999999 and 2.00000001 inches has some positive value. In the case of continuous distributions, it is interesting to note that $P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2)$ because $p(x_1) = p(x_2) = 0$.

In finance, some discrete distributions are treated as though they are continuous because the number of possible outcomes is very large. For example, the increase or decrease in the price of a stock traded on an American exchange is recorded in dollars and cents. Yet, the probability of a change of exactly \$1.33 or \$1.34 or any other specific change is almost zero. It is customary, therefore, to speak in terms of the probability of a range of possible price change, say between \$1.00 and \$2.00. In other words p(price change = 1.33) is essentially zero, but p(\$1 < price change < \$2) > 0.

LOS 9.c: Define and interpret a probability function, a probability density function, and a cumulative distribution function, and calculate and interpret probabilities for a random variable, given its cumulative distribution function.

A **probability function**, denoted p(x), specifies the probability that a random variable is equal to a specific value. More formally, p(x) is the probability that random variable X takes on the value x, or $p(x) = P(X = x)$.

The **two key properties of a probability function** are:

- $0 \leq p(x) \leq 1$.
- $\sum p(x) = 1$, the sum of the probabilities for *all* possible outcomes, x, for a random variable, X, equals 1.

Example: Evaluating a probability function

Consider the following function: $X = \{1, 2, 3, 4\}$, $p(x) = \frac{x}{10}$, else $p(x) = 0$

Determine whether this function satisfies the conditions for a probability function.

Answer:

Note that all of the probabilities are between zero and one, and the sum of all probabilities equals one:

$$\sum p(x) = \frac{1}{10} + \frac{2}{10} + \frac{3}{10} + \frac{4}{10} = 0.1 + 0.2 + 0.3 + 0.4 = 1$$

Both conditions for a probability function are satisfied.

A **probability density function** (pdf) is a function, denoted f(x), that can be used to generate the probability that outcomes of a continuous distribution lie within a particular range of outcomes. For a continuous distribution, it is the equivalent of a *probability function* for a discrete distribution. Remember, for a continuous distribution the probability of any one particular outcome (of the infinite possible outcomes) is zero. A pdf is used to calculate the probability of an outcome between two values (i.e., the probability of the outcome falling within a specified range). How that is actually done (it involves using calculus to take the integral of the function) is, thankfully, beyond the scope of the material required for the exam.

A **cumulative distribution function** (cdf), or simply *distribution function*, defines the probability that a random variable, *X*, takes on a value equal to or less than a specific value, *x*. It represents the sum, or *cumulative value*, of the probabilities for the outcomes up to and including a specified outcome. The cumulative distribution function for random variable, *X*, may be expressed as $F(x) = P(X \leq x)$. For example, consider the probability function defined earlier for $X = \{1, 2, 3, 4\}$, $p(x) = x / 10$. For this distribution, $F(3) = 0.6 = 0.1 + 0.2 + 0.3$, and $F(4) = 1 = 0.1 + 0.2 + 0.3 + 0.4$. This means that F(3) is the cumulative probability that outcomes 1, 2, or 3 occur, and F(4) is the cumulative probability that one of the possible outcomes occurs.

LOS 9.d: Define a discrete uniform random variable and a binomial random variable, calculate and interpret probabilities given the discrete uniform and the binomial distribution functions, and constuct a binomial tree to describe stock price movement.

A **discrete uniform random variable** is one for which the probabilities for all possible outcomes for a discrete random variable are equal. For example, consider the *discrete uniform probability distribution* defined as $X = \{1, 2, 3, 4, 5\}$, $p(x) = 0.2$. Here, the probability for each outcome is equal to 0.2 [i.e., $p(1) = p(2) = p(3) = p(4) = p(5) = 0.2$]. Also, the cumulative distribution function for the *n*th outcome, $F(x_n) = np(x)$, and the probability for a range of outcomes is p(x)k, where *k* is the number of possible outcomes in the range.

Example: Discrete uniform distribution

Determine p(6), F(6), and $P(2 \leq X \leq 8)$ for the discrete uniform distribution function defined as:

$X \{2, 4, 6, 8, 10\}$, $p(x) = 0.2$

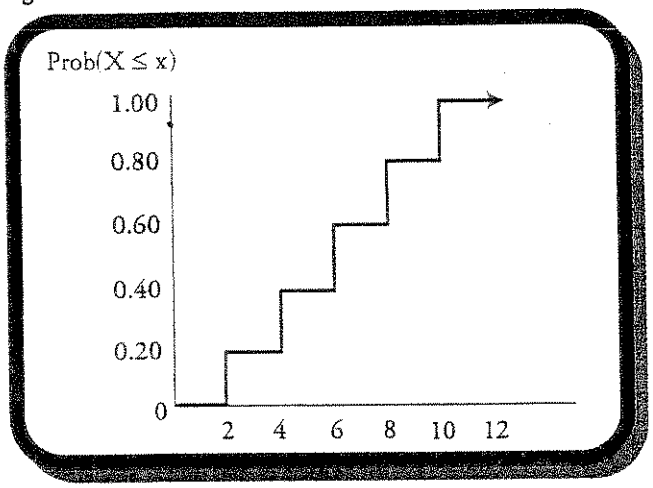
Answer:

$p(6) = 0.2$, since $p(x) = 0.2$ for all *x*. $F(6) = P(X \leq 6) = np(x) = 3(0.2) = 0.6$. Note that $n = 3$ since 6 is the third outcome in the range of possible outcomes. $P(2 \leq X \leq 8) = 4(0.2) = 0.8$. Note that $k = 4$, since there are four outcomes in the range $2 \leq X \leq 8$. Figures 1 and 2 illustrate the concepts of a probability function and cumulative distribution function for this distribution.

Figure 1: Probability and Cumulative Distribution Functions

X = x	Probability of x Prob (X = x)	Cumulative Distribution Function Prob ($X \leq x$)
2	0.20	0.20
4	0.20	0.40
6	0.20	0.60
8	0.20	0.80

Figure 2: Cumulative Distribution Function for X ~ Uniform {2, 4, 6, 8, 10}



The Binomial Distribution

A **binomial random variable** may be defined as the number of "successes" in a given number of trials, whereby the outcome can be either "success" or "failure." The probability of success, *p*, is constant for each trial, and the trials are independent. Think of a trial as a mini-experiment. The final outcome is the number of successes in a series of *n* trials. Under these conditions, the binomial probability function defines the probability of *x* successes in *n* trials. It can be expressed using the following formula:

$p(x) = P(X = x) = (\text{number of ways to choose } x \text{ from } n)p^x(1-p)^{n-x}$

where:

$(\text{number of ways to choose } x \text{ from } n) = \frac{n!}{(n-x)!x!}$ which may also be denoted as $\binom{n}{x}$ or stated as "*n* choose *x*"

p = the probability of "success" on each trial (don't confuse it with p(x))

So the probability of exactly *x* successes in *n* trials is:

$$p(x) = \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}$$

Example: Binomial probability

Assuming a binomial distribution, compute the probability of drawing three black beans from a bowl of black and white beans if the probability of selecting a black bean in any given attempt is 0.6. You will draw five beans from the bowl.

Answer:

$$P(X = 3) = p(3) = \frac{5!}{2!3!}(0.6)^3(0.4)^2 = (120 / 12)(0.216)(0.160) = 0.3456$$

Some intuition about these results may help you remember the calculations. Consider that a (very large) bowl of black and white beans has 60% black beans and that each time you select a bean, you replace it in the bowl before drawing again. We want to know the probability of selecting exactly three black beans in five draws, as in the above problem.

One way this might happen is BBBWW. Since the draws are independent, the probability of this is easy to calculate. The probability of drawing a black bean is 60% and the probability of drawing a white bean is 1 – 60% = 40%. Therefore, the probability of selecting BBBWW, in order is, 0.6 × 0.6 × 0.6 × 0.4 × 0.4 = 3.456%. This is the $p^3(1-p)^2$ from the formula and p is 60%, the probability of selecting a black bean on any single draw from the bowl.

BBBWW is not, however, the only way to choose exactly three black beans in five trials. Another possibility is BBWWB, and a third is BWWBB. Each of these will have exactly the same probability of occurring as our initial outcome, BBBWW. That's why we need to answer the question of how many ways (different orders) there are for us to choose three black beans in five draws. Using the formula, there are $\frac{5!}{3!(5-3)!} = 10$ ways; 10 × 3.456% = 34.56%, the answer we computed above.

The Expected Value of a Binomial Random Variable

For a given series of n trials, the expected number of successes or E(X) and the variance of X or Var(X) are given by the following formulas:

expected value of X = E(X) = np

variance of X = Var(X) = np(1 – p)

Example: Expected value of a binomial random variable

Based on empirical data, the probability that the Dow Jones Industrial Average (DJIA) will increase on any given day has been determined to equal 0.67. Assuming that the only other outcome is that it decreases, we can state p(UP) = 0.67 and p(DOWN) = 0.33. Further, assume that movements in the DJIA are independent (i.e., an increase in one day is independent of what happened on another day).

Using the information provided, compute the expected value of the number of up days in a 5-day period.

Answer:

Using binomial terminology, we define success as UP, so p = 0.67. Note that the definition of success is critical to any binomial problem.

E(X | n = 5, p = 0.67) = (5)(0.67) = 3.35

Recall that the "|" symbol means "*given.*" Hence, the preceding statement is read as: the expected value of X given that n = 5 and the probability of success = 67% is 3.35.

We should note that since the binomial distribution is a discrete distribution, the result X = 3.35 is not possible. However, if we were to record the results of many 5-day periods, the average number of up days (successes) would converge to 3.35.

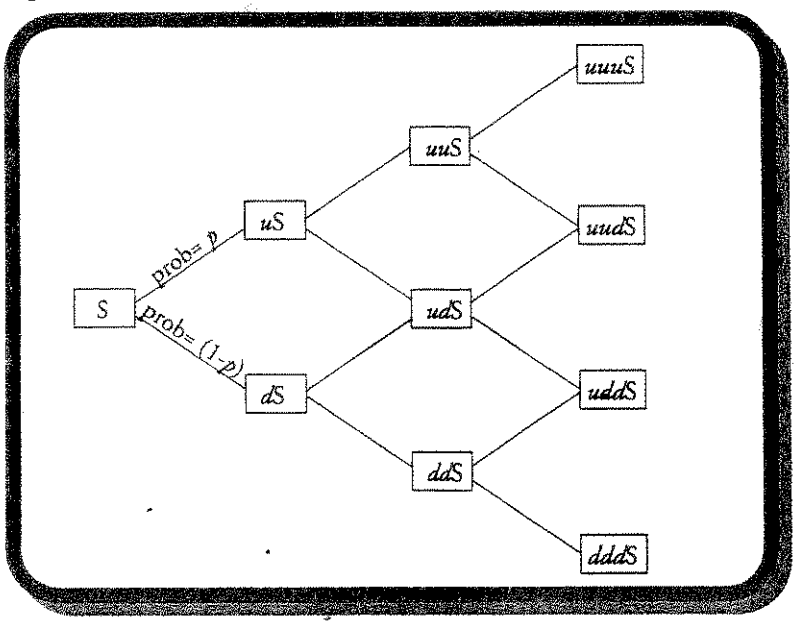
A Binomial Tree to Describe Stock Price Movement

A binomial model can be applied to stock price movements. We just need to define the two possible outcomes and the probability that each outcome will occur. Consider a stock with current price S that will, over the next period, either increase in value by 1% or decrease in value by 1% (the only two possible outcomes). The

probability of an up-move (u) is *p* and the probability of a down-move (d) is (1 – p). For our example, the up-move factor (U) is 1.01 and the down-move factor (D) is 1/1.01. So there is a probability *p* that the stock price will move to S(1.01) over the next period and a probability (1 – p) that the stock price will move to S/1.01.

A binomial tree is constructed by showing all the possible combinations of up-moves and down-moves over a number of successive periods. For two periods, these combinations are UU, UD, DU, and DD. Importantly, UD and DU result in the same stock price S after two periods since S (1.01)(1/1.01) = S and the order of the moves does not change the result. Figure 3 illustrates a binomial tree for three periods.

Figure 3: A Binomial Tree



With an initial stock price S = 50, U = 1.01, D = $1/1.01$, and prob(u) = 0.6, we can calculate the possible stock prices after two periods as:

$uuS = 1.01^2 \times 50 = 51.01$ with probability $(0.6)^2 = 0.36$

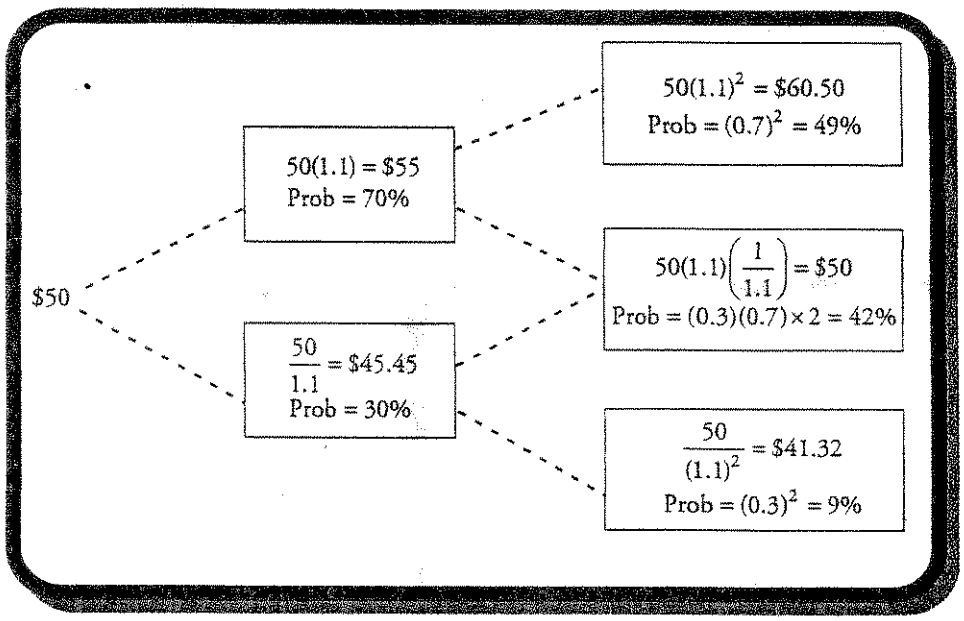
$udS = 1.01\left(1/1.01\right) \times 50 = 50$ with probability $(0.6)(0.4) = 0.24$

$duS = \left(1/1.01\right)(1.01) \times 50 = 50$ with probability $(0.4)(0.6) = 0.24$

$ddS = \left(1/1.01\right)^2 \times 50 = 49.01$ with probability $(0.4)^2 = 0.16$

Since a stock price of 50 can result from either *ud* or *du* moves, the probability of a stock price of 50 after two periods (the middle value) is 2 × (0.6)(0.4) = 48%. A binomial tree with S = 50, U = 1.1, and Prob(0) = 0.7 is illustrated in Figure 4.

Figure 4: A Two-Period Binomial Tree
S = $50, U = 1.10, Prob(U) = 0.7



One of the important applications of a binomial stock price model is in pricing options. We can make a binomial tree for asset prices more realistic by shortening the length of the periods and increasing the number of periods and possible outcomes.

LOS 9.e: Describe the continuous uniform distribution, and calculate and interpret probabilities, given a continuous uniform probability distribution.

The **continuous uniform distribution** is defined over a range that spans between some lower limit, *a*, and some upper limit, *b*, which serve as the parameters of the distribution. Outcomes can only occur between *a* and *b*, and since we are dealing with a continuous distribution, even if $a < x < b$, $P(X = x) = 0$. Formally, the properties of a continuous uniform distribution may be described as follows:

For all $a \leq x_1 < x_2 \leq b$, (i.e., for all x_1 and x_2 between the boundaries *a* and *b*)

$P(X < a \text{ or } X > b) = 0$, (i.e., the probability of *X* outside the boundaries is zero)
and
$P(x_1 \leq X \leq x_2) = (x_2 - x_1)/(b - a)$ (this defines the probability between x_1 and x_2)

Don't miss how simple this is just because the notation is so mathematical. For a continuous uniform distribution, the probability of outcomes in a range that is one-half the whole range is 50%. The probability of outcomes in a range that is one-quarter as large as the whole possible range is 25%.

Example: Continuous uniform distribution

X is uniformly distributed between 2 and 12. Calculate the probability that *X* will be between 4 and 8.

Answer:

$$\frac{8-4}{12-2} = \frac{4}{10} = 40\%$$

LOS 9.f: Explain the key properties of the normal distribution, distinguish between a univariate and a multivariate distribution, and explain the role of correlation in the multivariate normal distribution.

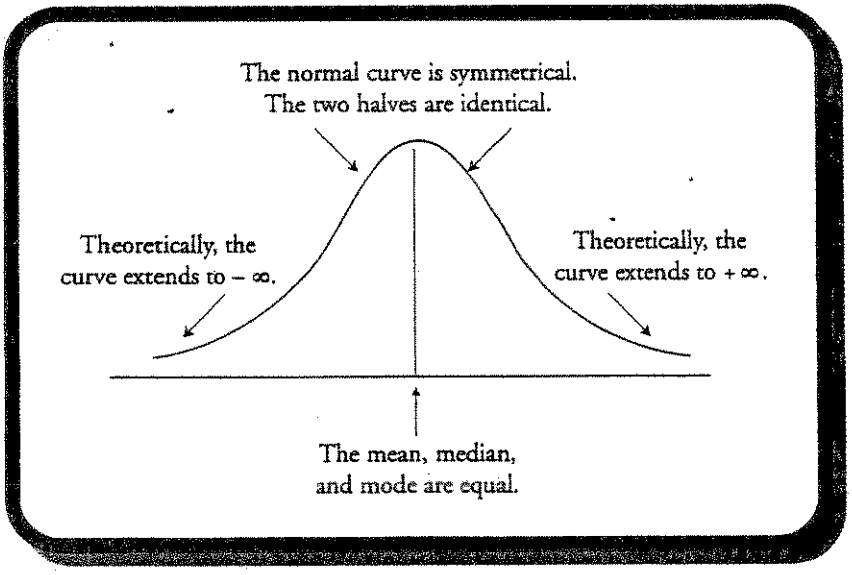
The normal distribution is important for many reasons. Besides the high probability that it will be covered on the exam, many of the random variables that are relevant to finance and other professional disciplines follow a normal distribution. In the area of investment and portfolio management, the normal distribution plays a central role in portfolio theory.

The **normal distribution** has the following key properties:

- It is completely described by its mean, μ, and variance, σ^2, stated as $X \sim N(\mu, \sigma^2)$. In words, this says that "X is normally distributed with mean μ and variance σ^2."
- Skewness = 0, meaning that the normal distribution is symmetric about its mean, so that $P(X \leq \mu) = P(\mu \leq X) = 0.5$, and mean = median = mode.
- Kurtosis = 3; this is a measure of how flat the distribution is. Recall that excess kurtosis is measured relative to 3, the kurtosis of the normal distribution.
- A linear combination of normally distributed random variables is also normally distributed.
- The probabilities of outcomes further above and below the mean get smaller and smaller but do not go to zero (the tails get very thin but extend infinitely).

Many of these properties are evident from examining the graph of a normal distribution's probability density function as illustrated in Figure 5.

Figure 5: Normal Distribution Probability Density Function



Univariate and Multivariate Distributions

Up to this point, our discussion has been strictly focused on **univariate distributions**, (i.e., the distribution of a single random variable). In practice, however, the relationships between two or more random variables are often relevant. For instance, investors and investment managers are frequently interested in the interrelationship among the returns of one or more assets. In fact, as you will see in your study of asset pricing models and modern portfolio theory, the return on a given stock and the return on the S&P 500 or some other market index will have special significance. Regardless of the specific variables, the simultaneous analysis of two or more random variables requires an understanding of multivariate distributions.

A **multivariate distribution** specifies the probabilities associated with a group of random variables and is meaningful only when the behavior of each random variable in the group is in some way dependent upon the behavior of the others. Both discrete and continuous random variables can have multivariate distributions. Multivariate distributions between two discrete random variables are described using joint probability tables. For continuous random variables, a multivariate *normal* distribution may be used to describe them if all of the individual variables follow a normal distribution. As previously mentioned, one of the characteristics of a normal distribution is that a linear combination of normally distributed random variables is normally distributed as well. For example, if the return of each stock in a portfolio is normally distributed, the return on the portfolio will also be normally distributed.

The Role of Correlation in the Multivariate Normal Distribution

Similar to a univariate normal distribution, a multivariate normal distribution can be described by the mean and variance of the individual random variables. Additionally, it is necessary to specify the correlation between the individual pairs of variables when describing a multivariate distribution. Correlation is the feature that distinguishes a multivariate distribution from a univariate normal distribution. *Correlation indicates the strength of the linear relationship between a pair of random variables.*

Using asset returns as our random variables, the multivariate normal distribution for the returns on *n* assets can be completely defined by the following three sets of parameters:

- *n* means of the *n* series of returns ($\mu_1, \mu_2, \ldots, \mu_n$).
- *n* variances of the *n* series of returns ($\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_n$).
- $0.5n(n-1)$ pair-wise correlations.

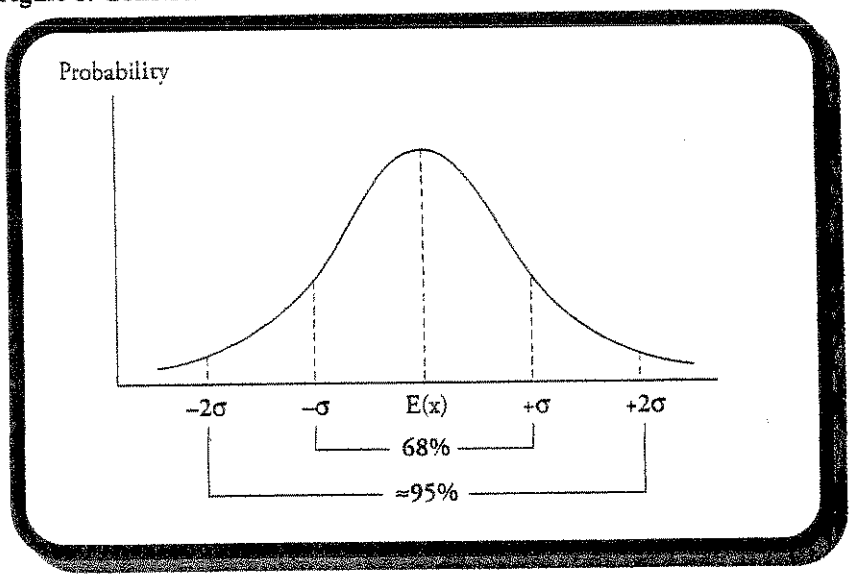
For example, if there are two assets, n = 2, then the multivariate returns distribution can be described with two means, two variances, and one correlation [0.5(2)(2 – 1) = 1]. If there are four assets, n = 4, the multivariate distribution can be described with four means, four variances, and six correlations [0.5(4)(4 – 1) = 6]. When building a portfolio of assets, all other things being equal, it is desirable to combine assets having low returns correlation because this will result in a portfolio with a lower variance than one composed of assets with higher correlations.

LOS 9.g: Construct and explain confidence intervals for a normally distributed random variable, and interpret the probability that a normally distributed random variable takes its value inside the constructed confidence interval.

A **confidence interval** is a range of values around the expected outcome within which we expect the actual outcome to be some specified percentage of the time. A 95% confidence interval is a range that we expect the random variable to be in 95% of the time. For a normal distribution, this interval is based on the expected value (sometimes called a point estimate) of the random variable and on its variability, which we measure with standard deviation.

Confidence intervals for a normal distribution are illustrated in Figure 6. For any normally distributed random variable, 68% of the outcomes are within one standard deviation of the expected value (mean) and approximately 95% of the outcomes are within two standard deviations of the expected value.

Figure 6: Confidence Intervals for a Normal Distribution



In practice we will not know the actual values for the mean and standard deviation of the distribution, but will have estimated them as $\bar{X}$ and s. The three confidence intervals of most interest are given by:

- The *90% confidence interval* for X is $\bar{X} - 1.65s$ to $\bar{X} + 1.65s$.
- The *95% confidence interval* for X is $\bar{X} - 1.96s$ to $\bar{X} + 1.96s$.
- The *99% confidence interval* for X is $\bar{X} - 2.58s$ to $\bar{X} + 2.58s$.

Example: Confidence intervals

The average return of a mutual fund is 10.5% per year and the standard deviation of annual returns is 18%. If returns are approximately normal, what is the 95% confidence interval for the mutual fund return next year?

Answer:

Here μ and σ are 10.5% and 18%, respectively. Thus, the 95% confidence interval for the return, R, is:

$10.5 \pm 1.96(18) = -24.78\%$ to 45.78%

Symbolically, this result can be expressed as:

$P(-24.78 < R < 45.78) = 0.95$ or 95%

The interpretation is that the annual return is expected to be within this interval 95% of the time or 95 out of 100 years.

LOS 9.h: Define the standard normal distribution, explain how to standardize a random variable, and calculate and interpret probabilities using the standard normal distribution.

The **standard normal distribution** is a normal distribution that has been standardized so that it has a mean of zero and a standard deviation of 1 [i.e., N~(0,1)]. To standardize an observation from a given normal distribution, the *z-value* of the observation must be calculated. The z-value represents the number of standard

deviations a given observation is from the population mean. *Standardization* is the process of converting an observed value for a random variable to its *z*-value. The following formula is used to *standardize a random variable*:

$$z = \frac{\text{observation} - \text{population mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

Professor's Note: The term z-value *will be used for a standardized observation in this document. The terms* z-score *and* z-statistic *are also commonly used.*

Example: Standardizing a random variable (*z*-values)

Assume that the annual earnings per share (EPS) for a large sample of firms are normally distributed with a mean of $6.00 and a standard deviation of $2.00.

What is the approximate probability of an observed EPS value falling between $2.00 and $8.00?

Answer:

If EPS = x = $8, then $z = (x - \mu)/\sigma = (\$8 - \$6) / \$2 = +1$
If EPS = x = $2, then $z = (x - \mu)/\sigma = (\$2 - \$6) / \$2 = -2$

Here, z = +1 indicates that an EPS of $8 is one standard deviation above the mean, and z = –2 means that an EPS of $2 is two standard deviations below the mean.

Owing to the symmetry property of a normal distribution, we know that 68% of all observations will fall within ± one standard deviation of the mean and that 95% will fall within ± two standard deviations of the mean. Since the mean of the standard normal distribution is zero, we can approximate the probability of the EPS falling between $2 and $8 as 0.68 / 2 + 0.95 / 2 = 0.815, or $P(2 \leq EPS \leq 8) = P(-2 \leq Z \leq 1) = 81.5\%$.

Calculating Standard Normal Probabilities

In the preceding example, we approximated the probability of a range of values for a random variable. Now we will show how to use standardized values and a table of probabilities for *Z* to determine the exact probability of a normally distributed random variable falling between any two values. A portion of a table of the cumulative distribution function for *Z* is shown in Figure 7. We will refer to this table as the *z*-table, as it contains values generated using the cumulative density function for *Z*, denoted by F(Z). Thus, the values in the *z*-table are the probabilities of observing a *z*-value that is less than a given value, *z* (i.e., $P(Z < z)$). The numbers in the first column are *z*-values that have only one decimal place. The columns to the right supply probabilities for *z*-values with two decimal places.

Note that the *z*-table in Figure 7 only provides probabilities for positive *z*-values. This is not a problem because we know from the symmetry of the standard normal distribution that $F(-Z) = 1 - F(Z)$. The tables in the back of many texts actually provide probabilities for negative *z*-values, but we will work with only the positive portion of the table because this may be all you get on the exam. In Figure 4 we can find the probability that a standard normal random variable will be less than 1.66, for example. The table value is 95.15%.

Figure 7: Cumulative Probabilities for a Standard Normal Distribution

Cdf Values for the Standard Normal Distribution: The z-table										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.5	.6915	Please note that several of the rows have been deleted to save space.*								
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

* A complete normal table is shown at the back of this book.

Example: Using a standard normal probability table (z-table)

What is the probability of a normally distributed random variable taking on a z-value that is between +2 and –2? That is, what is $P(-2 \leq Z \leq +2)$?

Answer:

From the z-table, we see that $F(2) = 0.9772$, meaning that the cumulative probability of an observed value falling below +2 is 0.9772, or 97.72%. Because of the distribution's symmetry, we also know that $F(-2) = 1 - 0.9772 = 0.0228$. This means that the probability of an observed value falling below –2 is 0.0228. These results indicate that 2.28% of all observations fall below $z = -2$ and an equal amount falls above $z = +2$. Thus:

$$P(-2 \leq Z \leq 2) = (1 - 0.0228) - 0.0228 = 0.9544$$

Another way to determine this probability is as follows:

$$P(-2 \leq Z \leq 2) = F(2) - F(-2) = 0.9772 - 0.0228 = 0.9544$$

Note how close this probability comes to our approximation rule that states that approximately 95% of all observations fall in the interval $\mu \pm 2\sigma$.

We can also use the *z*-table to measure confidence intervals for normally distributed random variables. For example, the 95% confidence interval for Z is the range of *z*-values such that 2.5% of the *z*-distribution falls above the upper value and 2.5% falls below the lower limit. That is, $P(z_1 \leq Z \leq z_2) = 0.95$. This corresponds to the *z*-value in the *z*-table for which the probability (the area under the curve to the left of the value) is 0.975, or $F(Z) = 0.975$. Why 0.975? Because we want a *z*-value such that $F(Z) - F(-Z) = 0.95$, or $0.975 - (1 - 0.975) = 0.95$ (i.e., we want to split the extra 5% evenly at each end of the curve). The probability 0.9750 in the table in Figure 7 corresponds to $z = 1.96$. Thus, $F(1.96) - F(-1.96) = 0.975 - (1 - 0.975) = 0.95$ or 95%. This confirms our previously stated confidence interval where we said that $P(X \text{ will be within } \bar{X} \pm 1.96s) = 0.95$ or 95%.

Example: Using the *z*-Table

Using the distribution of EPS ($\mu = \$6.00$, $\sigma = \$2.00$) again, what is the area under the standard normal distribution curve between \$3.50 and \$9.34? That is, what is $P(3.5 \leq EPS \leq 9.34)$?

Answer:

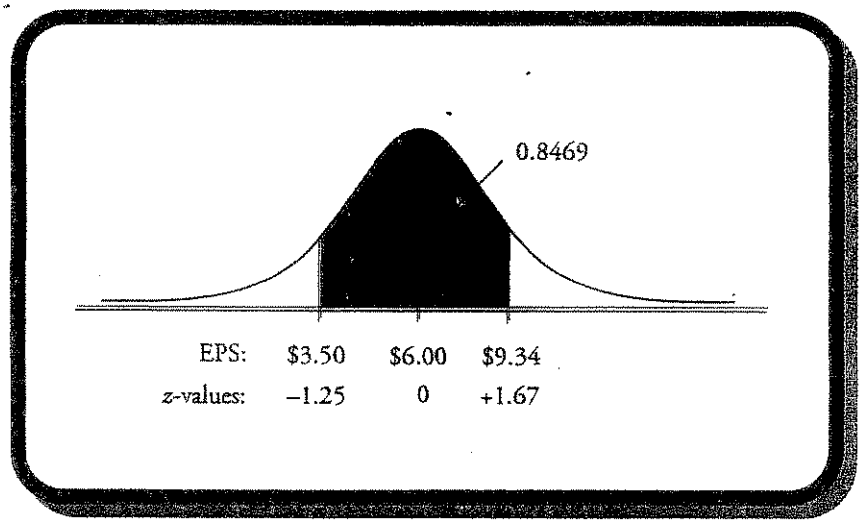
The *z*-values for the corresponding EPS values are:

EPS = \$3.50: $z_1 = (3.50 - 6) / 2 = -1.25$
EPS = \$9.34: $z_2 = (9.34 - 6) / 2 = +1.67$

Using the *z*-table, and referencing the distribution shown in Figure 8, the area under the curve between these *z*-values is:

$$\begin{aligned} P(3.5 \leq EPS \leq 9.34) &= P(-1.25 \leq Z \leq 1.67) = F(1.67) - F(-1.25) \\ &= 0.9525 - [1 - 0.8944] \\ &= 0.8469, \text{ or } 84.69\% \end{aligned}$$

Figure 8: Using the *z*-Table



Example: Using the *z*-table

Returning again to the EPS figures ($\mu = \$6$, $\sigma = \$2$), what percent of the EPS values are \$9.70 or more?

Answer:

Here we want to know $P(EPS > \$9.70)$, which is the area under the curve to the right of the *z*-value corresponding to EPS = \$9.70 (see Figure 9).

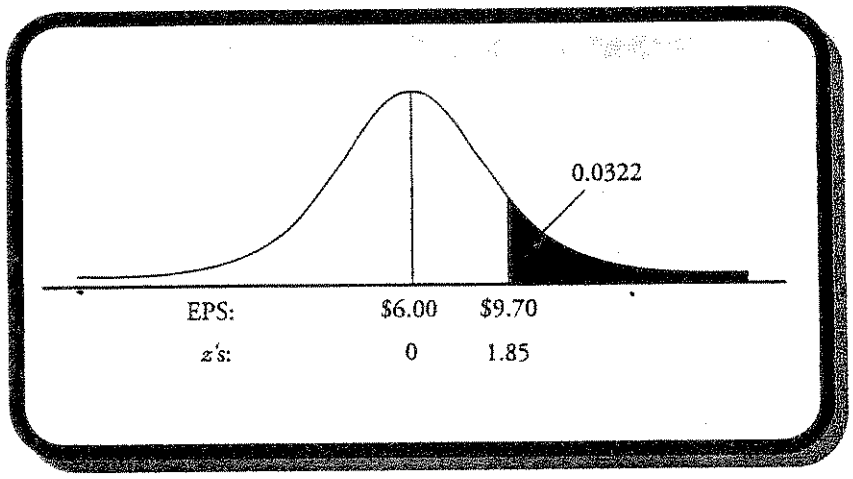
The z-value for EPS = \$9.70 is:

$$z = \frac{(x - \mu)}{\sigma} = \frac{(9.70 - 6)}{2} = 1.85$$

From the table we have F(1.85) = 0.9678, but this is P(EPS ≤ 9.70). We want P(EPS > 9.70), which is determined as:

$$\begin{aligned} P(EPS > 9.70) &= 1 - P(EPS \leq 9.70) \\ &= 1 - P(Z \leq 1.85) = 1 - F(1.85) \\ &= 1 - 0.9678 = 0.0322, \text{ or } 3.2\% \end{aligned}$$

Figure 9: P(EPS>\$9.70)



Example: Using the z-table

Using the distribution of EPS (μ = \$6.00, σ = \$2.00) again, what percent of the observed EPS values are likely to be less than \$4.10?

Answer:

As shown graphically in Figure 10, we want to know P(EPS < \$4.10). This requires a two-step approach like the one taken in the preceding example.

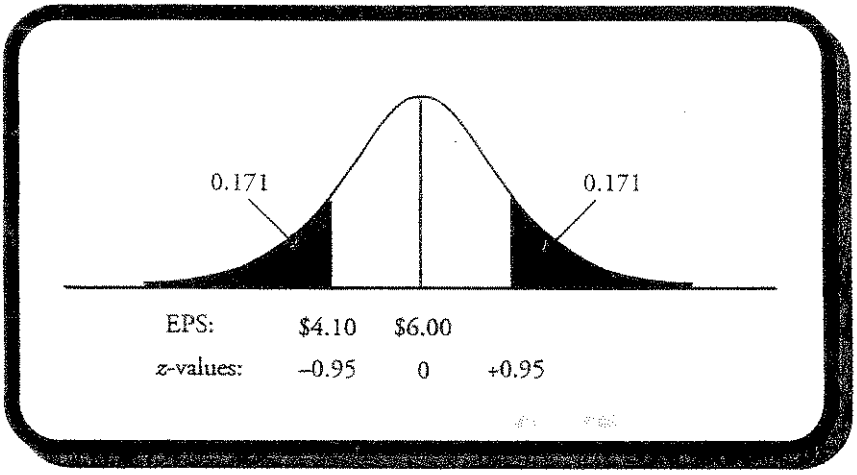
First, the corresponding z-value must be determined as follows:

$$z = \frac{(\$4.10 - \$6)}{2} = -0.95$$

Now, from the z-table *in the back of this book*, we find that F(0.95) = 0.8289, but this is P(Z ≤ +0.95) and we want P(Z ≤ –0.95). The probability that EPS will fall short of \$4.10 is determined as:

$$\begin{aligned} P(EPS < 4.10) &= P(Z < -0.95) = P(Z > 0.95) \\ &= 1 - F(0.95) = 1 - 0.8289 \\ &= 0.1711, \text{ or } 17.11\% \end{aligned}$$

Figure 10: Finding a Left-Tail Probability



Here, we have used the fact that the probability of being more than 0.95 standard deviations above the mean is equal to the probability of being 0.95 standard deviations below the mean. The *z*-table gave us the probability that the outcome will be less than 0.95 standard deviations above the mean. We subtract this probability from 1 to get the probability of being more than 0.95 standard deviations above the mean of $6.00, which is the same as the probability of outcomes more than 0.95 standard deviations *below* the mean.

Example: Using the *z*-table

Continuing with the EPS values, determine the probability that EPS will exceed $3.64.

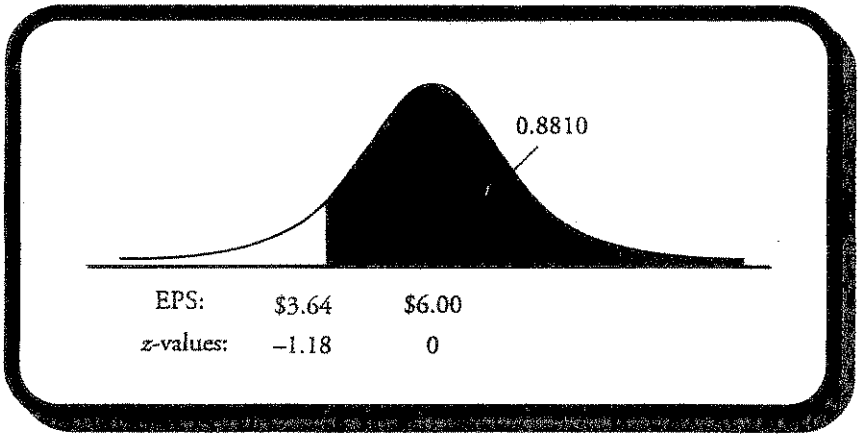
Answer:

The *z*-value associated with EPS = $3.64 is $z = (3.64 - 6) / 2 = -1.18$. $3.64 is 1.18 standard deviations below $6.00.

As shown in Figure 11, we are interested in P(EPS > $3.64) = P(Z > –1.18) which is determined as follows:

$$\begin{aligned} P(EPS > \$3.64) &= P(Z > -1.18) \\ &= 1 - F(-1.18) \\ &= 1 - [1 - F(1.18)] = F(1.18) \\ &= 0.8810 \text{ or } 88.10\% \end{aligned}$$

Figure 11: P(EPS > $3.64)



Note: Refer to the z-table at the back of this book to get F(1.18).

LOS 9.i: Define shortfall risk, calculate the safety-first ratio and select an optimal portfolio using Roy's safety-first criterion.

Shortfall risk is the probability that a portfolio value or return will fall below a particular (target) value or return over a given time period.

Roy's **safety-first criterion** states that the optimal portfolio minimizes the probability that the return of the portfolio falls below some minimum acceptable level. This minimum acceptable level is called the "threshold" level. Symbolically, Roy's safety-first criterion can be stated as:

minimize $P(R_p < R_L)$

where:
R_p = portfolio return
R_L = threshold level return

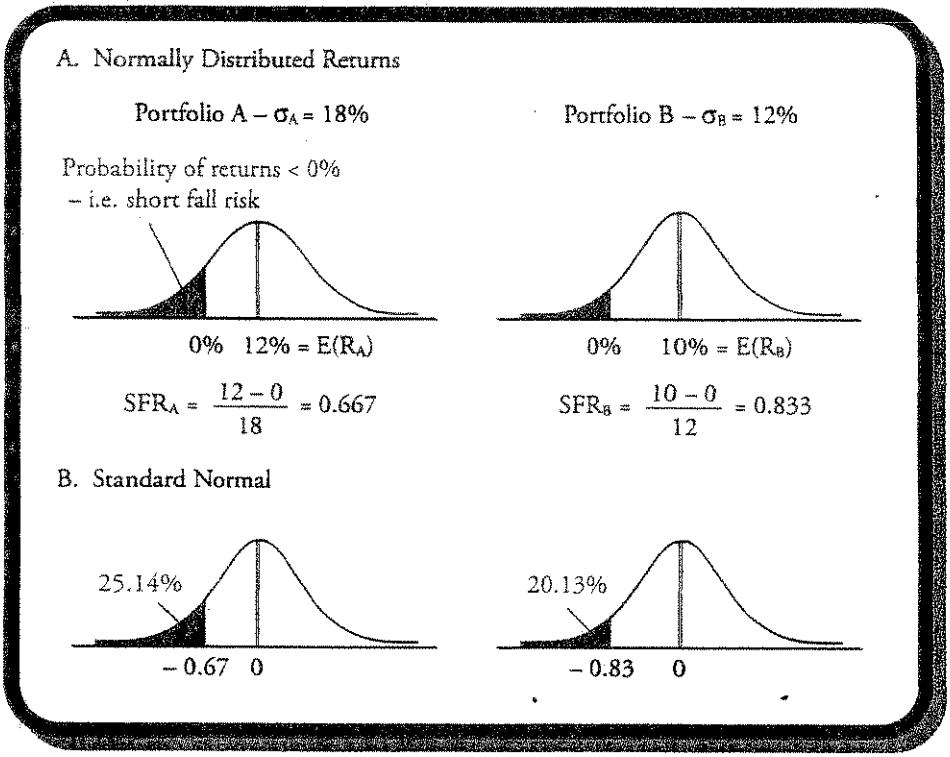
If portfolio returns are normally distributed, then Roy's safety-first criterion can be stated as:

maximize the SFRatio where SFRatio = $\frac{\left[E(R_p) - R_L\right]}{\sigma_p}$

Professors's Note: Notice the similarity to the Sharpe ratio: Sharpe = $\frac{\left[E(R_p) - R_f\right]}{\sigma_p}$. *The only difference is that the SFRatio utilizes the excess return over the threshold return, R_L, where the Sharpe ratio uses the excess return over the risk-free rate, R_f.*

The reasoning behind the safety-first criterion is illustrated in Figure 12. Assume an investor is choosing between two portfolios: Portfolio A with expected return of 12% and standard deviation of returns of 18%, and Portfolio B with expected return of 10% and standard deviation of returns of 12%. The investor has stated that he wants to minimize the probability of losing money (negative returns). Assuming that returns are normally distributed, the portfolio with the larger SFR using 0% as the threshold return (R_L) will be the one with the lower probability of negative returns.

Figure 12: The Safety-First Criterion and Shortfall Risk



Panel B of Figure 12 relates the SFRatio to the standard normal distribution. Note that the SFR is the number of standard deviations *below* the mean. Thus, the portfolio with the larger SFR has the lower probability of returns below the threshold return, zero in our example. Using the standard normal distribution tables, we can find the probabilities in the left-hand tails as indicated. These probabilities (25% for Portfolio A and 20% for Portfolio B) are also the shortfall risk for a target return of 0%. Portfolio B has the higher SFR which means it has the lower probability of negative returns.

In summary, when choosing among portfolios with normally distributed returns using Roy's safety-first criterion, there are two steps:

Step 1: Calculate the SFRatio = $\frac{\left[E(R_P) - R_L\right]}{\sigma_P}$.

Step 2: Choose the portfolio that has the *largest* SFRatio.

Example: Roy's safety-first criterion

For the next year, the managers of a $120 million college endowment plan, have set a minimum acceptable end-of-year portfolio value of $123.6 million. Three portfolios are being considered which have the expected returns and standard deviation shown in the first two rows of Figure 13. Determine which of these portfolios is the most desirable using Roy's safety-first criterion.

Answer:

The threshold return is R_L = (123.6 – 120) / 120 = 0.030 = 3%. The SFRs are shown in Figure 13. As indicated, the best choice is Portfolio A because it has the largest SFR.

Figure 13: Roy's Safety-First Ratios

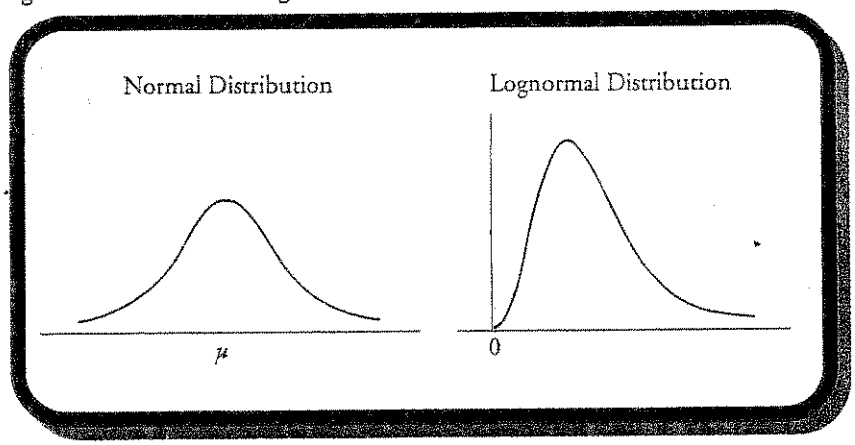
Portfolio	*Portfolio A*	*Portfolio B*	*Portfolio C*
$E(R_p)$	9%	11%	6.6%
σ_p	12%	20%	8.2%
SFRatio	$0.5 = \frac{(9-3)}{12}$	$0.4 = \frac{(11-3)}{20}$	$0.44 = \frac{(6.6-3)}{8.2}$

LOS 9.j: Explain the relationship between the lognormal and normal distributions and explain and interpret the use of the lognormal distribution in modeling asset prices.

The **lognormal distribution** is generated by the function e^x, where x is normally distributed. Since the natural logarithm, ln, of e^x is x, the logarithms of lognormally distributed random variables are normally distributed, thus the name.

Figure 14 illustrates the differences between a normal distribution and a lognormal distribution.

Figure 14: Normal vs. Lognormal Distributions



In Figure 14, we can see that:

- The lognormal distribution is skewed to the right.
- The lognormal distribution is bounded from below by zero so that it is useful for modeling asset prices which never take negative values.

If we used a normal distribution of returns to model asset prices over time, we would admit the possibility of returns less than –100%, which would admit the possibility of asset prices less than zero. Using a lognormal distribution to model *price relatives* avoids this problem. A price relative is just the end-of-period price of the asset over the beginning price (S_1/S_0) and is equal to (1 + the holding period return). To get the end-of-period asset price, we can simply multiply the price relative times the beginning-of-period asset price. Since a lognormal distribution takes a minimum value of zero, end-of-period asset prices cannot be less than zero. A price relative of

zero corresponds to a holding period return of –100% (i.e., the asset price has gone to zero). Recall that we used price relatives as the up-move and down-move (multiplier) terms in constructing a binomial tree for stock price changes over a number of periods.

LOS 9.k: Distinguish between discretely and continuously compounded rates of return; and calculate and interpret the continuously compounded rate of return, given a specific holding period return.

Discretely compounded returns are just the compound returns we are familiar with, given some discrete compounding period, such as semiannual or quarterly. Recall that the more frequent the compounding period, the greater the effective annual return. For a stated rate of 10%, semiannual compounding results in an effective yield of $\left(1+\frac{0.10}{2}\right)^2-1=10.25\%$ and monthly compounding results in an effective yield of $\left(1+\frac{0.10}{12}\right)^{12}-1=10.47\%$. Daily or even hourly compounding will produce still larger effective yields. The limit of this exercise, as the compounding period gets shorter and shorter, is called **continuous compounding**. The effective annual rate, based on continuous compounding for a stated annual rate of *i*, can be calculated from the formula:

$$\text{effective annual rate} = e^{i} - 1$$

Based on a stated rate of 10%, the effective rate with continuous compounding is $e^{0.10} - 1 = 10.5171\%$. Please verify this by entering 0.1 in your calculator and finding the e^x function.

Since the natural log, ln, of e^x is *x*, we can get the continuously compounded rate from an effective annual rate by using the ln calculator function. Using our previous example, ln (1 + 10.517%) = ln 1.105171 = 10%. Verify this by entering 1.105171 in your calculator and then entering the ln key. (Using the HP calculator, the keystrokes are 1.1 [ENTER] [g] [ln].)

We can use this method to find the continuously compounded rate that will generate a particular holding period return. If we are given a holding period return of 12.5% for the year, the equivalent continuously compounded rate is ln 1.125 = 11.778%. Since the calculation is based on 1 plus the holding period return, we can also do the calculation directly from the *price relative*. The price relative is just the end-of-period value divided by the beginning of period value. The continuously compounded rate of return is:

$$\ln\left(\frac{S_1}{S_0}\right)=\ln(1+\text{HPR})$$

Example: Calculating continuously compounded returns

A stock was purchased for $100 and sold one year later for $120. Calculate the investor's annual rate of return on a continuously compounded basis.

Answer:

$$\ln\left(\frac{120}{100}\right)=18.232\%$$

If we had been given the return (20%) instead, the calculation is:

$$\ln(1+0.20)=18.232\%$$

LOS 9.l: Explain Monte Carlo simulation and historical simulation and describe their major applications and limitations.

Monte Carlo simulation is a technique based on the repeated generation of one or more risk factors that affect security values, in order to generate a distribution of security values. For each of the risk factors, the analyst must specify the parameters of the probability distribution that the risk factor is assumed to follow. A computer is then used to generate random values for each risk factor based on its assumed probability distributions. Each set of randomly generated risk factors is used with a pricing model to value the security. This procedure is repeated many times (100s, 1,000s, or 10,000s) and the distribution of simulated asset values is used to draw inferences about the expected (mean) value of the security and possibly the variance of security values about the mean as well.

As an example, consider the valuation of stock options that can only be exercised on a particular date. The main risk factor is the value of the stock itself, but interest rates could affect the valuation as well. The simulation procedure would be to:

1. Specify the probability distributions of stock prices and of the relevant interest rate, as well as the parameters (mean, variance, possibly skewness) of the distributions.
2. Randomly generate values for both stock prices and interest rates.
3. Value the options for each pair of risk factor values.
4. After many iterations, calculate the mean option value and use that as your estimate of the option's value.

Monte Carlo simulation is used to:

- Value complex securities.
- Simulate the profits/losses from a trading strategy.
- Calculate estimates of value at risk (VAR) to determine the riskiness of a portfolio of assets and liabilities.
- Simulate pension fund assets and liabilities over time to examine the variability of the difference between the two.
- Value portfolios of assets that have non-normal returns distributions.

The **limitations of Monte Carlo simulation** are that it is fairly complex and will provide answers that are no better than the assumptions about the distributions of the risk factors and the pricing/valuation model that is used. Also, simulation is not an analytic method but a statistical one, and cannot provide the insights that analytic methods can.

Historical simulation is based on actual changes in value or actual changes in risk factors over some prior period. Rather than model the distribution of risk factors, as in Monte Carlo simulation, the set of all changes in the relevant risk factors over some prior period is used. Each iteration of the simulation involves randomly selecting one of these past changes for each risk factor and calculating the value of the asset or portfolio in question, based on those changes in risk factors.

Historical simulation has the advantage of using the actual distribution of risk factors so that the distribution of changes in the risk factors does not have to be estimated. It suffers from the fact that past changes in risk factors may not be a good indication of future changes. Events that occur infrequently may not be reflected in historical simulation results unless the events occurred during the period from which the values for risk factors are drawn.

An additional limitation of historical simulation is that it cannot address the sort of "what if" questions that Monte Carlo simulation can. With Monte Carlo simulation we can investigate the effect on the distribution of security/portfolio values of increasing the variance of one of the risk factors by 20%; with historical simulation we cannot do this.

KEY CONCEPTS

1. A probability distribution lists all the possible outcomes of an experiment along with their associated probabilities.
2. A probability function specifies the probability that a random variable is equal to a specific value; $P(X = x) = p(x)$.
3. A probability density function (pdf) is the expression for probability function for a continuous random variable.
4. The two key properties of a probability function are: (i) $0 \leq p(x) \leq 1$, and (ii) $\sum p(x) = 1$.
5. A cumulative distribution function (cdf) gives the probability of the random variable being equal to or less than each specific value. It is the area under the probability distribution to the left of a specified value.
6. A discrete random variable has positive probabilities associated with specific single number outcomes.
7. A continuous random variable has positive probabilities associated with a range of outcome values—the probability of it equaling any single value is zero.
8. The binomial distribution is a probability distribution for a binomial (discrete) random variable, X, that has one of two possible outcomes: success or failure, where the probability of success is p. The probability of a specific number of successes in n independent trials is:

   $$p(x) = P(X = x) = \frac{n!}{(n-x)!x!} p^x(1-p)^{n-x}$$

   $E(X) = np$ = expected value of X
9. A discrete uniform distribution is one where there are n discrete, equally likely outcomes, so that for each outcome $p(x) = 1/n$.
10. A continuous uniform distribution is one where the probability of X occurring in a possible range is the length of the range relative to the total of all possible values. Letting a and b be the lower and upper limit of the uniform distribution, respectively, then for $a \leq x_1 < x_2 \leq b$, $P(x_1 \leq X \leq x_2) = \frac{(x_2 - x_1)}{(b-a)}$.
11. The normal probability distribution and normal curve have the following characteristics:
    - The normal curve is symmetrical and bell-shaped with a single peak at the exact center of the distribution.
    - Mean = median = mode, and all are in the exact center of the distribution.
    - The normal distribution can be completely defined by its mean and standard deviation.
12. A confidence interval is a range within which we have a given level of confidence of finding a point estimate (e.g., the 90% confidence interval for X is $\overline{X} - 1.65s$ to $\overline{X} + 1.65s$).
13. The standard normal probability distribution has a mean of zero and a standard deviation of 1.

    A normally distributed random variable X can be normalized by $Z = \frac{(x-\mu)}{\sigma}$.

    A table that gives the cumulative probabilities for Z, the z-table, is used to find the probability that X falls within certain regions.

    $P(X < x) = F(x) = F\left[\frac{(x-\mu)}{\sigma}\right] = F(z)$, which is found in the standard normal probability table.

    $P(X > x) = 1 - P(X < x) = 1 - F(z)$

14. Multivariate distributions describe groups of random variables. A normal multivariate distribution with n individual random variables is completely described by the n means, n variances, and the $n(n-1)/2$ correlations (or covariances).
15. Correlation is a measure of the strength of the linear relationship between two random variables.
16. Safety-first rules focus on the probability of a portfolio return, R_p, falling below a certain lower-threshold return, R_L. The goal is to minimize $P(R_p < R_L)$ or equivalently, maximize:

$$\text{SFRatio} = \frac{\left[E(R_p) - R_L\right]}{\sigma_p}$$

17. Shortfall risk is the probability that a portfolio's value will fall below a specified minimum value over a specified period of time.
18. A lognormal distribution exists for random variable Y, when $Y = e^X$, and X is normally distributed.
19. As we decrease the length of discrete compounding periods (e.g., from quarterly to monthly) the EAY increases. As the length of the compounding period in discrete compounding gets shorter and shorter, the compounding becomes continuous where the effective annual rate = $e^i - 1$.
20. For a holding period return (HPR) over any period t, the equivalent continuously compounded rate over the period = ln(1 + HPR).
21. Monte Carlo simulation uses randomly generated values for risk factors, based on their assumed distributions, to produce a distribution of security values. Historical simulation uses randomly selected past changes in these risk factors to generate a distribution of security values.

CONCEPT CHECKERS: COMMON PROBABILITY DISTRIBUTIONS

1. Which of the following is NOT an example of a discrete random variable?
   A. The number of stocks a person owns.
   B. The time spent by a portfolio manager with a client.
   C. The number of days it rains in a month in Iowa City.
   D. The number of people holding Microsoft in their portfolios.

2. For a continuous random variable *X*, the probability of any single value of *X* is:
   A. one.
   B. zero.
   C. determined by the cdf.
   D. determined by the pdf.

Use the following table to answer Questions 3 through 7.

Probability distribution of a discrete random variable X								
X	0	1	2	3	4	5	6	7
P(X)	0.04	0.11	0.18	0.24	0.14	0.17	0.09	0.03

3. The probability that X = 3 is:
   A. 0.18.
   B. 0.24.
   C. 0.43.
   D. 0.70.

4. The cdf of 5, or F(5) is:
   A. 0.14.
   B. 0.17.
   C. 0.71.
   D. 0.88.

5. The probability that *X* is *greater* than 3 is:
   A. 0.24.
   B. 0.43.
   C. 0.57.
   D. 0.67.

6. What is $P(2 \leq X \leq 5)$?
   A. 0.12.
   B. 0.17.
   C. 0.38.
   D. 0.73.

7. The expected value of the random variable *X* is:
   A. 1.89.
   B. 3.35.
   C. 3.70.
   D. 5.47.

8. Which of the following is NOT a condition of a binomial experiment?
   A. There are only two trials.
   B. The trials are independent.
   C. Each trial has two and only two possible outcomes.
   D. If p is the probability of success, and q is the probability of failure, then $p + q = 1$.

9. A recent study indicated that 60% of all businesses have a fax machine. From the binomial probability distribution table, the probability that exactly four businesses will have a fax machine in a random selection of six businesses is:
   A. 0.138.
   B. 0.276.
   C. 0.311.
   D. 0.324.

10. Ten percent of all college graduates hired stay with the same company for more than five years. In a random sample of six recently hired college graduates, the probability that exactly two will stay with the same company for more than five years is *closest* to:
   A. 0.015.
   B. 0.098.
   C. 0.114.
   D. 0.185.

11. Assume that 40% of candidates who sit for the CFA® examination pass it the first time. Of a random sample of 15 candidates who are sitting for the exam for the first time, what is the expected number of candidates that will pass?
   A. 0.375.
   B. 4.000.
   C. 6.000.
   D. 6.667.

12. For the standard normal distribution, the z-value gives the distance between the mean and a point in terms of the:
   A. mean.
   B. variance.
   C. standard deviation.
   D. center of the curve.

13. For a standard normal distribution, F(0) is:
   A. 0.00.
   B. 0.10.
   C. 0.50.
   D. 1.00.

14. For the standard normal distribution, $P(0 \leq Z \leq 1.96)$ is:
   A. 0.4713.
   B. 0.4761.
   C. 0.4745.
   D. 0.4750.

15. For the standard normal distribution, $P(-2.05 \leq Z \leq 0.00)$ is:
A. 0.4798.
B. 0.4803.
C. 0.4938.
D. 0.9586.

Use the following data to answer Questions 16 through 19.

A study of hedge fund investors found that their annual household incomes are normally distributed with a mean of $175,000 and a standard deviation of $25,000.

16. What percent of hedge fund investors have incomes *less* than $100,000?
A. 0.05%.
B. 0.10%.
C. 0.13%.
D. 0.25%.

17. Approximately what percent of hedge fund investors have incomes between $150,000 and $200,000?
A. 50%.
B. 68%.
C. 75%.
D. 95%.

18. What percent of hedge fund investors have incomes *greater* than $225,000?
A. 0.50%.
B. 1.10%.
C. 2.28%.
D. 3.46%.

19. What percent of hedge fund investors have incomes *greater* than $150,000?
A. 15.87%.
B. 34.13%.
C. 68.26%.
D. 84.13%.

Use the following table to answer Questions 20 and 21.

Portfolio	Portfolio A	Portfolio B	Portfolio C	Portfolio D
$E(R_p)$	5%	11%	14%	18%
σ_p	8%	21%	34%	40%

20. Given a threshold level of return of 4%, use Roy's safety-first criterion to choose the optimal portfolio. Portfolio:
A. A.
B. B.
C. C.
D. D.

29. A stock doubled in value last year. Its continuously compounded return over the period was *closest* to:
A. 18.2%.
B. 69.3%.
C. 100.0%.
D. 200.0%.

30. Portfolio A has a safety-first ratio of 1.3 with a threshold return of 2%. What is the shortfall risk for a target return of 2%?
A. 90.30%.
B. 40.30%.
C. 9.68%.
D. 49.68%.

31. Of all the bonds currently rated B, 20% will default over an investor's horizon. The expected number of defaults and the variance of the number of defaults in a randomly selected 40-bond portfolio over the investor's horizon is:

	Expected Defaults	Variance
A.	8	32.0
B.	32	6.4
C.	8	6.4
D.	32	32.0

COMPREHENSIVE PROBLEMS: COMMON PROBABILITY DISTRIBUTIONS

1. A stock's price is $8.50 today. You decide to model the stock price over time using a binomial model (as a Bernoulli random variable) with a probability of an up-move of 60%. The up-move factor is 1.05.

A. How many different prices are possible for the stock at the end of two periods?

B. What are the possible prices after two periods?

C. What is the probability that the stock price will be $8.50 after three periods?

D. What is the probability that the stock price will be $8.925 after three periods?

E. What is the probability that the stock price will be unchanged after six periods?

2. An analyst has developed a model of option prices as a function of a short-term interest rate and the price of the underlying stock. She decides to test the model with a Monte Carlo simulation.

A. What steps does she need to perform to run the simulation?

B. What limitations of Monte Carlo simulation does she need to keep in mind when she interprets the results?

C. What would be the advantages of using historical simulation instead of Monte Carlo simulation? What would be the drawbacks?

21. Given a threshold level of return of 0%, use Roy's safety-first criterion to choose the optimal portfolio.
Portfolio:
A. A.
B. B.
C. C.
D. D.

22. If a stock's initial price is $20 and its year-end price is $23, then its continuously compounded annual rate of return is:
A. 9.86%.
B. 13.64%.
C. 13.98%.
D. 15.00%.

23. For a lognormal distribution, the:
A. mean equals the median.
B. standard deviation equals 1.
C. probability of a negative outcome is zero.
D. probability of a positive outcome is 50%.

24. Using hypothesized parameter values and a random number generator to study the behavior of certain asset returns is part of:
A. historical analysis.
B. normalizing a random variable.
C. Monte Carlo simulation.
D. standardizing a random variable.

25. A continuous uniform distribution has the parameters a = 4 and b = 10. The F(20) is:
A. 0.25.
B. 0.50.
C. 1.00.
D. 2.00.

26. All of the following statements accurately describe the binomial distribution EXCEPT:
A. the trials are independent.
B. it is a discrete distribution.
C. the probability of an outcome of zero is zero.
D. the combination formula is used in computing probabilities.

27. Approximately 50% of all observations for a normally distributed random variable fall in the interval:
A. $\mu \pm 0.67\sigma$.
B. $\mu \pm \sigma$.
C. $\mu \pm 2\sigma$.
D. $\mu \pm 3\sigma$.

28. The probability that a normally distributed random variable will be more than two standard deviations above its mean is:
A 0.0217.
B. 0.0228.
C. 0.4772.
D. 0.9772.

3. The monthly returns on an index of investment-grade corporate bonds for the last ten years have averaged 0.7% with a standard deviation of 2.0%.

   A. Assuming the returns are approximately normally distributed, what are the 90%, 95%, and 99% confidence intervals for the monthly return on this index?

   B. You are considering whether to use a lognormal distribution to model the value of one of the bonds in the index. In what ways is the lognormal distribution different from the normal distribution? What property of the lognormal distribution makes it useful for modeling asset prices?

ANSWERS – CONCEPT CHECKERS: COMMON PROBABILITY DISTRIBUTIONS

1. B Time is usually a continuous random variable; all the others are discrete.

2. B For a continuous distribution p(x) = 0 for all X; only ranges of value of X have positive probabilities.

3. B From the table.

4. D (0.04 + 0.11 + 0.18 + 0.24 + 0.14 + 0.17) = 0.88

5. B (0.14 + 0.17 + 0.09 + 0.03) = 0.43

6. D (0.18 + 0.24 + 0.14 + 0.17) = 0.73

7. B 0 + 1(0.11) + 2(0.18) + 3(0.24) + 4(0.14) + 5(0.17) + 6(0.09) + 7(0.03) = 3.35

8. A There may be any number of independent trials, each with only two possible outcomes.

9. C Success = having a fax machine. $[6! / 4!(6 - 4)!](0.6)^4(0.4)^{6-4} = 15(0.1296)(0.16) = 0.311$.

10. B Success = staying for five years. $[6! / 2!(6 - 2)!](0.10)^2(0.90)^{6-2} = 15(0.01)(0.656) = 0.0984$.

11. C Success = passing the exam. Then, E(success) = np = 15 × 0.4 = 6.

12. C This is true by the formula for z.

13. C By the symmetry of the z-distribution and F(0) = 0.5. Half the distribution lies on each side of the mean.

14. D From the table F(1.96) = 0.9750, thus the answer is 0.9750 – 0.5 = 0.4750. Knowing that 95% lie between –1.96 and +1.96, and that 0 is the midpoint, we can say that $\frac{95\%}{2} = 47.5\%$ lie between 0 and +1.96.

15. A From the table, and via symmetry, F(2.05) = 0.9798, thus the answer is 0.9798 – 0.5 = 0.4798.

16. C z = –3 = (100 – 175) / 25, F(–3) = 1 – 0.9987 = 0.0013

17. B This is $\pm 1\sigma$ from the mean. For a normal distribution, 68% of observations are between $\pm 1\sigma$ from the mean.

18. C 1 – F(2), where F(2) equals 0.9772. Hence, 1 – 0.9772 = 0.0228.

19. D 1 – F(–1) = F(1) = 0.8413

20. D SFR = (18 – 4) / 40 is the largest value.

21. A SFR = (5 – 0) / 8 is the largest value.

22. C ln(23 / 20) = 0.1398

23. C A lognormally distributed variable is never negative.

24. C Monte Carlo simulation.

25. C F(x) = 1 for all x > b. Remember F(x) is the cumulative probability, P(x < 20) here.

26. C With only two possible outcomes, there must be some positive probability for each. It does not matter if one of the possible outcomes happens to be zero. If this were not the case, the variable in question would not be a random variable, and a probability distribution would be meaningless.

27. **A** $\mu \pm 0.67\sigma$

28. **B** 1 – F(2) = 1 – 0.9772 = 0.0228.

29. **B** ln(2) = 0.6931.

30. C Using the tables, the cdf for –1.3 is 9.68%, which is the probability of returns less than 2%.

31. C Treating the number of defaults as a binomial random variable, *x*, E(x) = 0.2(40) = 8 and the Var(x) = (0.2)(0.8)40 = 6.4.

ANSWERS – COMPREHENSIVE PROBLEMS: COMMON PROBABILITY DISTRIBUTIONS

1. A. Using u for an up move and d for a down move, there are four possible outcomes (price paths) over two periods: uu, ud, du, and dd. Since ud and du result in the same price at the end of two periods ($8.50), there are three possible prices after two periods.

   B. An up-move factor of 1.05 means the down-move factor is 1/1.05. Therefore the possible prices for each path are as follows:
   - uu: Price = $8.50(1.05)^2$ = $9.37
   - ud: Price = 8.50(1.05)(1/1.05) = $8.50
   - du: Price = 8.50(1/1.05)(1.05) = $8.50
   - dd: Price = $8.50(1/1.05)^2$ = $7.71

   C. For a 3-period model, there is no possible way that the up moves and down moves can exactly offset since the possibilities are: uuu, uud, udu, duu, ddu, dud, udd, ddd. The probability of a price of $8.50 is zero.

   D. $8.925 is the result of a single up move, 8.50 (1.05) = 8.925. With three periods this price could only result from two up moves and one down move (in any order). The probabilities of prices after n periods follow a binomial distribution. Define an up move as a success so that the probability of a success (p) is 0.60, the probability of an up move. The probability of two successes in three trials is $_3C_2\ (0.6)^2\ (1 - 0.6)$ = 43.2%. This calculation takes account of all three possible price paths that include two up moves and one down move: uud, udu, and duu.

   E. The price will be unchanged after six periods only if the price path includes three up moves and three down moves. The probability of three successes (up moves) in six trials is: $_6C_3\ (0.6)^3\ (1 - 0.6)^3$ = 27.65%.

2. A. To construct a Monte Carlo simulation, the analyst would need to:

      1. Identify the distributions that the input variables follow and their means, variances and any other relevant parameters, such as skewness.

      2. Generate random values from these distributions for both input variables.

      3. Price the option using the randomly generated inputs.

      4. Repeat steps 2 and 3 for many trials. Calculate the mean option price from all the trials. This value is the simulation's estimate of the option value.

   B. Whether the results from a Monte Carlo simulation are useful depends on how well the analyst has specified the distributions of the interest rate and the stock price (the old, garbage in—garbage out problem). Also, the simulation results contain no information about whether the valuation model itself is valid.

C. The main advantages of historical simulation are that it uses actual historical values for the model inputs, so that the analyst does not need to make assumptions about their probability distributions, and it is a less computer-intensive procedure. A disadvantage of historical simulation is that it assumes the past behavior of the variables is a reliable indicator of their future behavior, which might not be the case. Historical simulation also lacks the Monte Carlo approach's ability to model "what if" questions by changing the assumed probability distributions of the model inputs.

3. A. The 90% confidence interval is the mean ± 1.65 standard deviations.
0.7 + 1.65(2.0) = 4.0%
0.7 – 1.65(2.0) = –2.6%

The 95% confidence interval is the mean ± 1.96 standard deviations.
0.7 + 1.96(2.0) = 4.62%
0.7 – 1.96(2.0) = –3.22%

The 99% confidence interval is the mean ± 2.58 standard deviations.
0.7 + 2.58(2.0) = 5.86%
0.7 – 2.58(2.0) = –4.46%

B. The lognormal distribution is skewed to the right, whereas the normal distribution is symmetrical. The lognormal distribution can only have positive values; whereas the normal distribution includes both positive and negative values. This property makes the lognormal distribution useful for modeling asset prices.

The following is a review of the Quantitative Methods principles designed to address the learning outcome statements set forth by CFA Institute®. This topic is also covered in:

SAMPLING AND ESTIMATION

Study Session 3

EXAM FOCUS

This topic review covers random samples and inferences about population means from sample data. It is essential that you know the central limit theorem, for it allows us to use sampling statistics to construct confidence intervals for point estimates of population means. Make sure you can calculate confidence intervals for population means given sample parameter estimates and a level of significance, and know when it is appropriate to use the *z*-statistic versus the *t*-statistic. You should also understand the basic procedures for creating random samples, and recognize the warning signs of various sampling biases from nonrandom samples.

APPLIED STATISTICS

In many real-world statistics applications, it is impractical (or impossible) to study an entire population. When this is the case, a subgroup of the population, called a sample, can be evaluated. Based upon this sample, the parameters of the underlying population can be estimated.

For example, rather than attempting to measure the performance of the U.S. stock market by observing the performance of all 10,000 or so stocks trading in the United States at any one time, the performance of the subgroup of 500 stocks in the S&P 500 can be measured. The results of the statistical analysis of this sample can then be used to draw conclusions about the entire population of U.S. stocks.

LOS 10.a: Define simple random sampling, sampling error, and a sampling distribution, and interpret sampling error.

Simple random sampling is a method of selecting a sample in such a way that each item or person in the population being studied has the same likelihood of being included in the sample. As an example of simple random sampling, assume that you want to draw a sample of five items out of a group of 50 items. This can be accomplished by numbering each of the 50 items, placing them in a hat, and shaking the hat. Next, one number can be drawn randomly from the hat. Repeating this process (experiment) four more times results in a set of five numbers. The five drawn numbers (items) comprise a simple random sample from the population. In applications like this one, a random-number table or a computer random-number generator is often used to create the sample.

Sampling error is the difference between a sample statistic (the mean, variance, or standard deviation of the sample) and its corresponding population parameter (the true mean, variance, or standard deviation of the population). For example, the sampling error for the mean is as follows:

$$\text{sampling error of the mean} = \text{sample mean} - \text{population mean} = \bar{x} - \mu$$

A Sampling Distribution

It is important to recognize that the sample statistic itself is a random variable and, therefore, has a probability distribution. The **sampling distribution** of the sample statistic is a probability distribution of all possible sample

statistics computed from a set of equal-size samples that were randomly drawn from the same population. Think of it as the probability distribution of a statistic from many samples.

For example, suppose a random sample of 100 bonds is selected from a population of a major municipal bond index consisting of 1,000 bonds, and then the mean return of the 100-bond sample is calculated. Repeating this process many times will result in many different estimates of the population mean return (i.e., one for each sample). The distribution of these estimates of the mean is the *sampling distribution of the mean*.

It is important to note that this sampling distribution is distinct from the distribution of the actual prices of the 1,000 bonds in the underlying population and will have different parameters.

LOS 10.b: Distinguish between simple random and stratified random sampling.

Stratified random sampling uses a classification system to separate the population into smaller groups based on one or more distinguishing characteristics. From each subgroup, or stratum, a random sample is taken and the results are pooled. The size of the samples from each stratum is based on the size of the stratum relative to the population.

Stratified sampling is often used in bond indexing because of the difficulty and cost of completely replicating the entire population of bonds. In this case, bonds in a population are categorized (stratified) according to major bond risk factors such as duration, maturity, coupon rate, and the like. Then samples are drawn from each separate category and combined to form a final sample.

To see how this works, suppose you want to construct a bond portfolio that is indexed to the major municipal bond index using a stratified random sampling approach. First, the entire population of 1,000 municipal bonds in the index can be classified on the basis of maturity and coupon rate. Then, cells (stratum) can be created for different maturity/coupon combinations, and random samples can be drawn from each of the maturity/coupon cells. To sample from a cell containing 50 bonds with 2- to 4-year maturities and coupon rates less than 5%, we would select 5 bonds. The number of bonds drawn from a given cell corresponds to the cell's weight relative to the population (index), or (50/1000) × (100) = 5 bonds. This process is repeated for all of the maturity/coupon cells, and the individual samples are combined to form the portfolio.

By using stratified sampling, we guarantee that we sample five bonds from this cell. If we had used simple random sampling, there would be no guarantee that we would sample any of the bonds in the cell. Or, we may have selected more than five bonds.

LOS 10.c: Distinguish between time-series and cross-sectional data.

Time-series data consist of observations taken *over a period of time* at specific and equally spaced time intervals. The set of monthly returns on Microsoft stock from January 1994 to January 2004 is an example of a time-series data sample.

Cross-sectional data are a sample of observations taken *at a single point in time.* The sample of reported earnings per share of all NASDAQ companies as of December 31, 2004, is an example of a cross-sectional data sample.

LOS 10.d: Interpret the central limit theorem and describe its importance.

The **central limit theorem** states that for simple random samples of size n from a *population* with a mean μ and a finite variance σ^2, the sampling distribution of the sample mean $\bar{x}$ approaches a normal probability distribution with mean μ and a variance equal to $\frac{\sigma^2}{n}$ as the sample size becomes large.

The central limit theorem is extremely useful because the normal distribution is relatively easy to apply to hypothesis testing and to the construction of confidence intervals. Specific inferences about the population mean can be made from the sample mean, *regardless of the population's distribution*, as long as the sample size is "sufficiently large," which usually means $n \geq 30$.

Important properties of the central limit theorem include the following:

- If the sample size *n* is sufficiently large ($n \geq 30$), the sampling distribution of the sample means will be approximately normal. Remember what's going on here, random samples of size *n* are repeatedly being taken from an overall larger population. Each of these random samples has its own mean, which is itself a random variable, and this set of sample means has a distribution that is approximately normal.
- The mean of the population, μ, and the mean of the distribution of all possible sample means are equal.
- The variance of the distribution of sample means is $\frac{\sigma^2}{n}$, the population variance divided by the sample size.

LOS 10.e: Calculate and interpret the standard error of the sample mean.

The **standard error of the sample mean** is the standard deviation of the distribution of the sample means.

When the standard deviation of the population, σ, is *known*, the standard error of the sample mean is calculated as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where:
$\sigma_{\bar{x}}$ = standard error of the sample mean
σ = standard deviation of the population
n = size of the sample

Example: Standard error of sample mean (known population variance)

The mean hourly wage for Iowa farm workers is \$13.50 with a *population standard deviation* of \$2.90. Calculate and interpret the standard error of the sample mean for a sample size of 30.

Answer:

Because the population standard deviation, σ, *is known*, the standard error of the sample mean is expressed as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$2.90}{\sqrt{30}} = \$0.53$$

Professor's Calculator Tip: On the TI BAII Plus, the use of the square root key is obvious. On the HP 12C, the square root of 30 is computed as: [30] [ENTER] [g] [$\sqrt{x}$].

This means that if we were to take all possible samples of size 30 from the Iowa farm worker population and prepare a sampling distribution of the sample means, we will obtain a mean of \$13.50 and standard error of \$0.53.

Practically speaking, the *population's standard deviation is almost never known.* Instead, the standard error of the sample mean must be estimated by dividing the standard deviation of *the sample* mean by $\sqrt{n}$:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Note: Use this when the population variance is unknown.

where:

$s_{\bar{x}}$ = standard error of the sample mean

s = standard deviation of the sample = $\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

n = size of the sample

Example: Standard error of sample mean (unknown population variance)

Suppose a sample contains the past 30 monthly returns for McCreary, Inc. The mean return is 2% and the *sample* standard deviation is 20%. Calculate and interpret the standard error of the sample mean.

Answer:

Since σ is unknown, the standard error of the sample mean is:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{30}} = 3.6\%$$

This implies that if we took all possible samples of size 30 from McCreary's monthly returns and prepared a sampling distribution of the sample means, the mean would be 2% with a standard error of 3.6%.

Example: Standard error of sample mean (unknown population variance)

Continuing with our example, suppose that instead of a sample size of 30, we take a sample of the past 200 monthly returns for McCreary, Inc. In order to highlight the effect of sample size on the sample standard error, let's assume that the mean return and standard deviation of this larger sample remain at 2% and 20%, respectively. Now, calculate the standard error of the sample mean for the 200-return sample.

Answer:

The standard error of the sample mean is computed as:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{200}} = 1.4\%$$

The result of the preceding two examples illustrates an important property of sampling distributions. Notice that the value of the standard error of the sample mean decreased from 3.6% to 1.4% as the sample size increased from 30 to 200. This is because as the sample size increases it gets closer to the size of the population, and the distribution of the sample means about the population mean gets smaller and smaller, which causes the standard error of the sample mean to decrease.

LOS 10.f: Distinguish between a point estimate and a confidence interval estimate of a population parameter.

LOS 10.h: Explain the construction of confidence intervals.

Point estimates are single (sample) values used to estimate population parameters. The formula used to compute the point estimate is called the estimator. For example, the sample mean, $\overline{x}$, is an estimator of the population mean μ and is computed using the familiar formula:

$$\overline{x} = \frac{\sum x}{n}$$

The value generated with this calculation for a given sample is called the *point estimate* of the mean.

Confidence interval estimates result in a range of values within which the actual value of a parameter will lie, given the probability of $1 - \alpha$. Here, alpha, α, is called the *level of significance* for the confidence interval, and the probability $1 - \alpha$ is referred to as the *degree of confidence*. For example, we might estimate that the population mean of random variables will range from 15 to 25 with a 95% degree of confidence, or at the 5% level of significance.

Confidence intervals are usually constructed by adding or subtracting an appropriate value from the point estimate. In general, confidence intervals take on the following form:

point estimate ± (reliability factor × standard error)

where:
point estimate = value of a sample statistic of the population parameter
reliability factor = number that depends on the sampling distribution of the point estimate and the probability that the point estimate falls in the confidence interval, $(1 - \alpha)$
standard error = standard error of the point estimate

LOS 10.g: Identify and describe the desirable properties of an estimator.

Regardless of whether we are concerned with point estimates or confidence intervals, there are certain statistical properties that make some estimates more desirable than others. These **desirable properties of an estimator** are **unbiasedness**, **efficiency**, and **consistency**.

- An *unbiased* estimator is one for which the expected value of the estimator is equal to the parameter you are trying to estimate. For example, because the expected value of the sample mean is equal to the population mean $\left[E(\overline{x}) = \mu\right]$, the sample mean is an unbiased estimator of the population mean.
- An unbiased estimator is also *efficient* if the variance of its sampling distribution is smaller than all the other unbiased estimators of the parameter you are trying to estimate. The sample mean, for example, is an unbiased and efficient estimator of the population mean.
- A *consistent* estimator is one for which the accuracy of the parameter estimate increases as the sample size increases. As the sample size increases, the standard error of the sample mean falls, and the sampling distribution bunches more closely around the population mean. In fact, as the sample size approaches infinity, the standard error approaches zero.

LOS 10.i: Describe the properties of Student's *t*-distribution and calculate and interpret its degrees of freedom.

Student's ***t*-distribution**, or simply the *t*-distribution, is a bell-shaped probability distribution that is symmetrical about its mean. It is the appropriate distribution to use when constructing confidence intervals based on *small samples* (n < 30) from populations with *unknown variance* and a normal, or approximately normal, distribution. It may also be appropriate to use the *t*-distribution when the population variance is unknown and the sample size is large enough that the central limit theorem will assure that the sampling distribution is approximately normal.

Student's *t*-distribution has the following properties:

- It is symmetrical.
- It is defined by a single parameter, the **degrees of freedom** (df), where the degrees of freedom are equal to the number of sample observations minus 1, n – 1, for sample means.
- It is less peaked than a normal distribution, with more probability in the tails ("fatter tails").
- As the degrees of freedom (the sample size) gets larger, the shape of the *t*-distribution more closely approaches a standard normal distribution.

When *compared to the normal distribution*, the *t*-distribution is flatter with more area under the tails (i.e., it has fatter tails). As the degrees of freedom for the *t*-distribution increase, however, its shape approaches that of the normal distribution.

The degrees of freedom for tests based on sample means are n – 1 because, given the mean, only n – 1 observations can be unique.

LOS 10.j: Calculate and interpret a confidence interval for a population mean when sampling from a normal distribution with 1) a known population variance, 2) an unknown population variance, or 3) with an unknown variance and the sample size is large.

If the population has a *normal distribution with a known variance*, a **confidence interval for the population mean** can be calculated as:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x}$ = *point estimate* of the population mean (sample mean).

$z_{\alpha/2}$ = *reliability factor*, a standard normal random variable for which the probability in the right-hand tail of the distribution is $\alpha/2$. In other words, this is the *z*-score that leaves $\alpha/2$ of probability in the upper tail.

$\frac{\sigma}{\sqrt{n}}$ = the *standard error* of the sample mean where σ is the known standard deviation of the population, and *n* is the sample size.

The most commonly used standard normal distribution reliability factors are:

$z_{\alpha/2}$ = 1.645 for 90% confidence intervals (the significance level is 10%, 5% in each tail)

$z_{\alpha/2}$ = 1.960 for 95% confidence intervals (the significance level is 5%, 2.5% in each tail)

$z_{\alpha/2}$ = 2.575 for 99% confidence intervals (the significance level is 1%, 0.5% in each tail)

Do these numbers look familiar? They should! In our review of common probability distributions, we found the probability under the standard normal curve between z = –1.96 and z = +1.96 to be 0.95, or 95%. Owing to symmetry, this leaves a probability of 0.025 under each tail of the curve beyond z = –1.96 of z = +1.96, for a total of 0.05, or 5%—just what we need for a significance level of 0.05, or 5%.

Example: Confidence interval

Consider a practice exam that was administered to 100 Level 1 candidates. The mean score on this practice exam was 80 for all 36 of the candidates in the sample who studied at least 10 hours a week in preparation for the exam. Assuming a population standard deviation equal to 15, **construct** and **interpret** a 99% confidence interval for the mean score on the practice exam for 36 candidates who study at least 10 hours a week. *Note that in this example the population standard deviation is known, so we don't have to estimate it.*

Answer:

At a confidence level of 99%, $z_{\alpha/2} = z_{0.005} = 2.575$. So, the 99% confidence interval is calculated as follows:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 80 \pm 2.575\frac{15}{\sqrt{36}} = 80 \pm 6.4$$

Thus, the 99% confidence interval ranges from 73.6 to 86.4.

Confidence intervals can be interpreted from a probabilistic perspective or a practical perspective. With regard to the outcome of the CFA® practice exam example, these two perspectives can be described as follows:

- *Probabilistic interpretation.* After repeatedly taking samples of CFA candidates who studied 10 hours or more per week, administering the practice exam, and constructing confidence intervals for each sample's mean, 99% of the resulting confidence intervals will, in the long run, include the population mean.
- *Practical interpretation.* We are 99% confident that the population mean score is between 73.6 and 86.4 for candidates who study more than 10 hours per week.

Confidence Intervals for the Population Mean: Normal With Unknown Variance

If the distribution of the *population is normal with unknown variance*, we can use the *t*-distribution to construct a confidence interval:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where:

$\bar{x}$ = the point estimate of the population mean

$t_{\alpha/2}$ = the *t*-reliability factor (a.k.a., *t*-statistic or critical *t*-value) corresponding to a *t*-distributed random variable with n – 1 degrees of freedom, where *n* is the sample size. The area under the tail of the *t*-distribution to the right of $t_{\alpha/2}$ is $\alpha/2$.

$\frac{s}{\sqrt{n}}$ = standard error of the sample mean

s = sample standard deviation

Unlike the standard normal distribution, the reliability factors for the *t*-distribution depend on the sample size, so we can't rely on a commonly used set of reliability factors. Instead, reliability factors for the *t*-distribution have to be looked up in a table of Student's *t*-distribution, like the one at the back of this book.

Owing to the relatively fatter tails of the *t*-distribution, confidence intervals constructed using *t*-reliability factors ($t_{\alpha/2}$) will be more conservative (wider) than those constructed using *z* reliability factors ($z_{\alpha/2}$).

Example: Confidence intervals

Let's return to the McCreary, Inc. example. Recall that we took a sample of the past 30 monthly stock returns for McCreary, Inc. and determined that the mean return was 2% and the sample standard deviation was 20%. Since the population variance is unknown, the standard error of the sample was estimated to be:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{30}} = 3.6\%$$

Now, let's construct a 95% confidence interval for the mean monthly return.

Answer:

Here we will use the *t*-reliability factor because the population variance is unknown. Since there are 30 observations, the degrees of freedom are 29 = 30 – 1. Remember, because this is a two-tailed test at the 95% confidence level, the probability under each tail must be $\alpha/2 = 2.5\%$, for a total of 5%. So, referencing the 1-tailed probabilities for Student's *t*-distribution at the back of this book, we find the critical *t*-value (reliability factor) for $\alpha/2 = 0.025$ and df = 29 to be $t_{29,\ 2.5} = 2.045$. Thus, the 95% confidence interval for the population mean is:

$$2\% \pm 2.045\left(\frac{20\%}{\sqrt{30}}\right) = 2\% \pm 2.045(3.6\%) = 2\% \pm 7.4\%$$

Thus, the 95% confidence has a lower limit of –5.4% and an upper limit of +9.4%.

We can interpret this confidence interval by saying that we are 95% confident that the population mean monthly return for McCreary stock is between –5.4% and +9.4%.

Professor's Note: You should practice looking up reliability factors (a.k.a. critical t-values or t-statistics) in a t-table. The first step is always to compute the degrees of freedom, which is n – 1. The second step is to find the appropriate level of alpha or significance. This depends on whether the test you're concerned with is one-tailed (use α) or two-tailed (use $\alpha/2$). In this review, our tests will always be two-tailed because confidence intervals are designed to compute an upper and lower limit. Thus, we will use $\alpha/2$. To look up $t_{29,\ 2.5}$, find the 29 df row and match it with the 0.025 column; t = 2.045 is the result. We'll do more of this in our study of hypothesis testing and regression analysis.

Confidence Interval for a Population Mean When the Population Variance Is Unknown Given a Large Sample From Any Type of Distribution

We now know that the *z*-statistic should be used to construct confidence intervals when the population distribution is normal and the variance is known, and the *t*-statistic should be used when the distribution is normal but the variance is unknown. But what do we do when the *distribution* is *nonnormal*?

As it turns out, the size of the sample influences whether or not we can construct the appropriate confidence interval for the sample mean.

- If the *distribution is nonnormal* but the *population variance is known*, the *z*-statistic can be used as long as the sample size is large ($n \geq 30$). We can do this because the central limit theorem assures us that the distribution of the sample mean is approximately normal when the sample is large.
- If the *distribution is nonnormal* and the *population variance is unknown*, the *t*-statistic can be used as long as the sample size is large ($n \geq 30$). It is also acceptable to use the *z*-statistic, although use of the *t*-statistic is more conservative.

This means that if we are sampling from a nonnormal distribution (which is sometimes the case in finance), *we cannot create a confidence interval if the sample size is less than 30.* So, all else equal, make sure you have a sample of at least 30, and the larger, the better.

Figure 1 summarizes this discussion.

Professor's Note: You should commit the criteria in this table to memory.

Figure 1: Criteria for Selecting the Appropriate Test Statistic

When sampling from a:	*Test Statistic*	
	Small Sample ($n < 30$)	*Large Sample ($n \geq 30$)*
Normal distribution with *known* variance	*z*-statistic	*z*-statistic
Normal distribution with *unknown* variance	*t*-statistic	*t*-statistic*
Nonnormal distribution with *known* variance	not available	*z*-statistic
Nonnormal distribution with *unknown* variance	not available	*t*-statistic*

* The *z*-statistic is theoretically acceptable here, but use of the *t*-statistic is more conservative.

All of the preceding analysis depends on the sample we draw from the population being random. If the sample isn't random, the central limit theorem doesn't apply, our estimates won't have the desirable properties, and we can't form unbiased confidence intervals. Surprisingly, creating a *random sample* is not as easy as one might believe. There are a number of potential mistakes in sampling methods that can bias the results. These biases are particularly problematic in financial research, where available historical data are plentiful, but the creation of new sample data by experimentation is restricted.

LOS 10.k: Discuss the issues regarding selection of the appropriate sample size, data-mining bias, sample selection bias, survivorship bias, look-ahead bias, and time-period bias.

We have seen so far that a larger sample reduces the sampling error and the standard deviation of the sample statistic around its true (population) value. Confidence intervals are narrower (more precise) when samples are larger and the standard errors of the point estimates of population parameters are less.

There are two limitations on this idea of "larger is better" when it comes to selecting an appropriate sample size. One is that larger samples may contain observations from a different population (distribution). If we include observations which come from a different population, that has a different population parameter, we will not necessarily improve the precision of our population parameter estimates. The other consideration is cost. The costs of using a larger sample must be weighed against the value of the increase in precision from the increase in sample size. Both of these factors suggest that the largest possible sample size is not always the most appropriate choice.

Data-Mining Bias, Sample Selection Bias, Survivorship Bias, Look-Ahead Bias, and Time-Period Bias

Data mining occurs when analysts repeatedly use the same database to search for patterns or trading rules until one that "works" is discovered. For example, empirical research has provided evidence that value stocks appear to outperform growth stocks. Some researchers argue that this anomaly is actually the product of data mining. Because the data set of historical stock returns is quite limited, it is difficult to know for sure whether the difference between value and growth stock returns is a true economic phenomenon, or simply a chance pattern that was stumbled upon after repeatedly looking for any identifiable pattern in the data.

Data-mining bias refers to results where the statistical significance of the pattern is overestimated because the results were found through data mining.

When reading research findings that suggest a profitable trading strategy, make sure you heed the following warning signs of data mining:

- Evidence that many different variables were tested, most of which are unreported, until significant ones were found.
- The lack of any economic theory that is consistent with the empirical results.

The best way to avoid data mining is to test a potentially profitable trading rule on a data set different from the one you used to develop the rule (i.e., use out-of-sample data).

Sample selection bias occurs when some data is systematically excluded from the analysis, usually because of the lack of availability. This practice renders the observed sample to be nonrandom, and any conclusions drawn from this sample can't be applied to the population because the observed sample and the portion of the population that was not observed are different.

Survivorship bias is the most common form of sample selection bias. A good example of the existence of survivorship bias in investments is the study of mutual fund performance. Most mutual fund databases, like Morningstar®'s, only include funds currently in existence—the "survivors." They do not include funds that have ceased to exist due to closure or merger.

This would not be a problem if the characteristics of the surviving funds and the missing funds were the same; then the sample of survivor funds would still be a random sample drawn from the population of mutual funds. As one would expect, however, and as evidence has shown, the funds that are dropped from the sample have lower returns relative to the surviving funds. Thus, the surviving sample is biased toward the better funds (i.e., it is not random). The analysis of a mutual fund sample with survivorship bias will yield results that overestimate the average mutual fund return because the database only includes the better-performing funds. The solution to survivorship bias is to use a sample of funds that all started at the same time and not drop funds that have been dropped from the sample.

Look-ahead bias occurs when a study tests a relationship using sample data that was not available on the test date. For example, consider the test of a trading rule that is based on the price-to-book ratio at the end of the fiscal year. Stock prices are available for all companies at the same point in time, while end-of-year book values may not be available until 30 to 60 days after the fiscal year ends. In order to account for this bias, a study that uses price-to-book value ratios to test trading strategies might estimate the book value as reported at fiscal year end and the market value two months later.

Time-period bias can result if the time period over which the data is gathered is either too short or too long. If the time period is too short, research results may reflect phenomena specific to that time period, or perhaps even data mining. If the time period is too long, the fundamental economic relationships that underlie the results may have changed.

For example, research findings may indicate that small stocks outperformed large stocks during the 1980–1985 time period. This may well be the result of time-period bias—in this case, using too short a time period. It's not clear whether this relationship will continue in the future or if it is just an isolated occurrence.

On the other hand, a study that quantifies the relationship between inflation and unemployment (the Phillips Curve) during the period from 1940–2000 will also result in time-period bias—because this period is too long, and it covers a fundamental change in the relationship between inflation and unemployment that occurred in the 1980s. In this case, the data should be divided into two subsamples that span the period before and after the change.

KEY CONCEPTS

1. Simple random sampling is a method of selecting a sample in such a way that each item or person in the population being studied has the same likelihood of being included in the sample.
2. Sampling error is the difference between a sample statistic and its corresponding population parameter (e.g., the sample mean minus the population mean).
3. A sampling distribution is the distribution of all values that a sample statistic can take on when computed from samples of identical size randomly drawn from the same population.
4. Stratified random sampling involves randomly selecting samples proportionally from subgroups that are formed based on one or more distinguishing characteristics.
5. A time-series sample consists of observations taken at specific and equally spaced points in time, while a cross-sectional data sample consists of observations taken at a single point in time.
6. The central limit theorem states that for a population with a mean μ and a finite variance σ^2, the sampling distribution of the sample mean of all possible samples of size n will be approximately normally distributed with a mean equal to μ and a variance equal to σ^2/n.
7. The standard error of the sample mean is the standard deviation of the distribution of the sample means and is calculated as:

   $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where σ, the population standard deviation, is known.

   $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, where s is the sample standard deviation and the population standard deviation is unknown.
8. Point estimates are single value estimates of population parameters, and confidence intervals are ranges of estimated values within which the actual value of the parameter will lie with a given probability.
9. Desirable statistical properties of an estimator include: unbiasedness, efficiency, and consistency.
10. The *t*-distribution is similar, but not identical, to the normal distribution in shape—it is defined by the degrees of freedom, has a lower peak, and has fatter tails.
11. Degrees of freedom for the *t*-distribution is equal to n – 1; Student's *t*-distribution is closer to the normal distribution when df is greater, and confidence intervals are narrower when df is greater.
12. The *t*-distribution is used to construct confidence intervals for the population mean when the population variance is not known. The $(1-\alpha)$ confidence interval for the population mean, μ, is: $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
13. Use the *t*-distribution if:
    Population distribution is normal with an unknown variance (large or small sample).
    Population distribution is nonnormal with unknown variance, but the sample is large (n >30).
14. The standard normal distribution (*z*-distribution) is used to construct confidence intervals for the population mean when the population variance is known. The $(1 - \alpha)$ confidence interval for the population mean, μ, is: $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

15. Use the *z*-distribution if:
    Population distribution is normal with known variance.
    Population distribution is nonnormal and the sample is large ($n \geq 30$).
16. There are a number of potential mistakes in the sampling method that can bias results. These biases include data mining, sample selection bias, look-ahead bias, survivorship bias, and time-period bias.

CONCEPT CHECKERS: SAMPLING AND ESTIMATION

1. Which of the following *most accurately* defines a simple random sample? It is a sample:
   A. that includes every tenth element of an arranged population.
   B. drawn in such a way that each member of the population has some chance of being selected in the sample.
   C. drawn in such a way that each member of the population has an equal chance of being included in the sample.
   D. drawn in such a way that each member of the population has a 1% chance of being included in the sample.

2. Sampling error is defined as:
   A. an error that occurs when a sample of more than 30 elements is drawn.
   B. an error that occurs when a sample of less than 30 elements is drawn.
   C. an error that occurs during collection, recording, and tabulation of data.
   D. the difference between the value of a sample statistic and the value of the corresponding population parameter.

3. The mean age of all CFA candidates is 28 years. The mean age of a random sample of 100 candidates is found to be 26.5 years. The difference, 28 – 26.5 = 1.5, is called the:
   A. random error.
   B. sampling error.
   C. population error.
   D. probability error.

4. If n is large and the population standard deviation is unknown, the standard error of the sampling distribution of the sample mean is *equal* to the:
   A. sample standard deviation divided by the sample size.
   B. population standard deviation multiplied by the sample size.
   C. sample standard deviation divided by the square root of the sample size.
   D. population standard deviation divided by the sample size.

5. The standard error of the sampling distribution of the sample mean for a sample size of n drawn from a population with a mean of μ and a standard deviation of σ is:
   A. sample standard deviation divided by the sample size.
   B. population standard deviation multiplied by the square root of the sample size.
   C. sample standard deviation divided by the square root of the sample size.
   D. population standard deviation divided by the square root of the sample size.

6. To apply the central limit theorem to the sampling distribution of the sample mean, the sample is usually considered to be large if n is *greater* than:
   A. 15.
   B. 20.
   C. 25.
   D. 30.

7. Assume that a population has a mean of 14 with a standard deviation of 2. If a random sample of 49 observations is drawn from this population, the standard error of the sample mean is *closest* to:
   A. 0.04.
   B. 0.29.
   C. 2.00.
   D. 7.00.

8. The population's mean is 30 and the mean of a sample of size 100 is 28.5. The variance of the sample is 25. The standard error of the sample mean is *closest* to:
   A. 0.05.
   B. 0.25.
   C. 0.50.
   D. 2.50.

9. A random sample of 100 computer store customers spent an average of $75 at the store. Assuming the distribution is normal and the population standard deviation is $20, the 95% confidence interval for the population mean is *closest* to:
   A. $69.84 to $80.16.
   B. $71.08 to $78.92.
   C. $73.89 to $80.11.
   D. $74.56 to $79.44.

10. Best Computers, Inc., sells computers and computer parts by mail. A sample of 25 recent orders showed the mean time taken to ship out these orders was 70 hours with a sample standard deviation of 14 hours. Assuming the population is normally distributed, the 99% confidence interval for the population mean is:
   A. 25 ± 6.98 hours.
   B. 70 ± 2.80 hours.
   C. 70 ± 6.98 hours.
   D. 70 ± 7.83 hours.

11. The sampling distribution of a statistic is the probability distribution made up of all possible:
   A. observations from the underlying population.
   B. confidence intervals from sample sizes greater than 30.
   C. sample statistics computed from samples of varying sizes drawn from the same population.
   D. sample statistics computed from samples of the same size drawn from the same population.

12. The sample of debt/equity ratios of 25 publicly traded U.S. banks as of fiscal year-end 2003 is an example of:
   A. a point estimate.
   B. time-series data.
   C. cross-sectional data.
   D. a stratified random sample.

13. Which of the following is NOT a desirable property of an estimate?
   A. Reliability.
   B. Efficiency.
   C. Consistency.
   D. Unbiasedness.

14. If the variance of the sampling distribution of an estimator is smaller than all other unbiased estimators of the parameter of interest, the estimator is:
   A. reliable.
   B. efficient.
   C. unbiased.
   D. consistent.

15. Which of the following is NOT a property of Student's *t*-distribution?
A. It is symmetrical.
B. As the degrees of freedom get larger, the variance approaches zero.
C. It is defined by a single parameter, the degrees of freedom, which is equal to n – 1.
D. It has more probability in the tails and less at the peak than a standard normal distribution.

16. An analyst who uses historical data that was not publicly available at the time period being studied will have a sample with:
A. look-ahead bias.
B. time-period bias.
C. survivorship bias.
D. sample selection bias.

17. The 95% confidence interval of the sample mean of employee age for a major corporation is 19 years to 44 years based on a *z*-statistic. The population of employees is over 5,000 and the sample size of this test is 100. Assuming the population is normally distributed, the standard error of mean employee age is *closest* to:
A. 1.96.
B. 2.58.
C. 6.38.
D. 12.50.

18. Which of the following is *most closely* associated with survivorship bias?
A. Price-to-book studies.
B. Stratified bond sampling studies.
C. Equity-index-linked note studies.
D. Mutual fund performance studies.

19. What is the *most appropriate* test statistic for constructing confidence intervals for the population mean when the population is normally distributed, but the variance is unknown.
A. The *z*-statistic at α with n degrees of freedom.
B. The *z*-statistic with n – 1 degrees of freedom.
C. The *t*-statistic at $\alpha/2$ with n degrees of freedom.
D. The *t*-statistic at $\alpha/2$ with n – 1 degrees of freedom.

20. The *acceptable* test statistic for constructing confidence intervals for the population mean of a nonnormal distribution when the population variance is unknown and the sample size is large (n > 30) is the:
A. *z*-statistic or the *t*-statistic.
B. *z*-statistic at α with n degrees of freedom.
C. *t*-statistic at α with 29 degrees of freedom.
D. *t*-statistic at $\alpha/2$ with n degrees of freedom.

21. Jenny Fox evaluates managers who have a cross-sectional population standard deviation of returns of 8%. If returns are independent across managers, how large of a sample does Fox need so that the standard deviation of sample means is 1.265%?
A. 6.
B. 7.
C. 30.
D. 40.

22. Annual returns on small stocks have a population mean of 12% and a population standard deviation of 20%. If the returns are normally distributed, a 90% confidence interval on mean returns over a 5-year period is:
A. 5.40% to 18.60%.
B. –2.75% to 26.75%.
C. –5.52% to 29.52%.
D. 4.16% to 19.84%.

COMPREHENSIVE PROBLEMS: SAMPLING AND ESTIMATION

1. Using random sampling, a manager wants to construct a portfolio of 50 stocks that will approximate the returns of a broad market index that contains 200 stocks. Explain how he could use simple random sampling and stratified random sampling to select stocks from the index and the possible advantages of stratified random sampling.

2. An analyst has taken a random sample of 50 observations from a population for which she wants to estimate the population mean. She believes this population's distribution is negatively skewed.

   A. Can she use the sample mean to estimate the population mean and construct a confidence interval? Explain.

   B. What are the desirable statistical properties of an estimator?

   C. Which of these properties does the sample mean possess as an estimator of the population mean?

3. A random sample of analyst earnings estimates has a mean of $2.84 and a standard deviation of $0.40. What can we say about the 90% confidence interval for earnings next period if:

   A. the sample size is 20?

   B. the sample size is 40?

   C. What probabilistic statement could we make at the 90% level if the sample size were 15?

   D. What probabilistic statement could we make at the 90% confidence level if the sample size were 60?

ANSWERS – CONCEPT CHECKERS: SAMPLING AND ESTIMATION

1. C In a simple random sample, each element of the population has an equal probability of being selected. Choice D allows for an equal chance, but only if there are 100 elements in the population from which the random sample is drawn.

2. D An example might be the difference between a particular sample mean and the average value of the overall population.

3. B The sampling error is the difference between the population parameter and the sample statistic.

4. C The formula for the standard error when the population standard deviation is unknown is: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

5. D The formula for the standard error when the population standard deviation is known is: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

6. D By definition.

7. B $s_{\bar{X}} = \frac{s}{\sqrt{n}}$. Given $s = 2$, $s_{\bar{X}} = \frac{2}{\sqrt{49}} = \frac{2}{7} = 0.2857$.

8. C $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Given $\sigma^2 = 25$, $s_{\bar{X}} = \frac{5}{\sqrt{100}} = \frac{5}{10} = 0.5$.

9. B Since the population variance is known and $n \geq 30$, the confidence interval is determined as: $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$. $z_{\alpha/2} = z_{0.025} = 1.96$. So, the confidence interval is: $75 \pm 1.96(20/10) = 75 \pm 3.92 = 71.08$ to 78.92.

10. D Since the population variance is unknown and $n < 30$, the confidence interval is determined as: $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$. Look up $t_{\alpha/2}$ and df = n – 1 to get critical t. $t_{0.01/2}$ and df = 24 is 2.797. So, the confidence interval is: $70 \pm 2.797(14/5) = 70 \pm 7.83$.

11. D Suppose you have a population of 10,000 employees. If you take 100 samples of 50 employees each, the distribution of the 100 sample means is the sampling distribution.

12. C Cross-sectional data is a set of data that are all collected as of the same point in time.

13. A Efficiency, consistency, and unbiasedness are all desirable properties of an estimate.

14. B By definition. Efficiency is a desirable property of an estimator.

15. B As the degrees of freedom get larger, the t-distribution approaches the normal distribution. As the degrees of freedom fall, the peak of the t-distribution flattens and its tails get fatter (more probability in the tails—that's why, all else the same, the critical t increases as the *df* decreases).

16. A The primary example of look-ahead bias is using year-end financial information in conjunction with market pricing data to compute ratios like the price/earnings, P/E. The E in the denominator is typically not available for 30-60 days after the end of the period. Hence, data that was available on the test date (P) is mixed with information that was not available (E). That is, the P is "ahead" of the E.

17. C At the 95% level of significance, with sample size n = 100 and mean 31.5 years, the appropriate test statistic is $z_{\alpha/2} = 1.96$. *Note: The mean of 31.5 is calculated as the midpoint of the interval, or (19 + 44) / 2.* Thus, the confidence interval is $31.5 \pm 1.96s_{\bar{X}}$, where $s_{\bar{X}}$ is the standard error of the sample mean. If we take the upper bound, we know that $31.5 + 1.96s_{\bar{X}} = 44$, or $1.96s_{\bar{X}} = 12.5$, or $s_{\bar{X}} = 6.38$ years.

18. D Mutual fund performance studies are most closely associated with survivorship bias because only the better-performing funds remain in the sample over time.

19. D Use the *t*-statistic at $\alpha/2$ and n – 1 degrees of freedom when the population variance is unknown. While the *z*-statistic is acceptable when the sample size is large, sample size is not given here, and the *t*-statistic is always appropriate under these conditions.

20. A When the sample size is large, and the central limit theorem can be relied upon to assure a sampling distribution that is normal, either the *t*-statistic or the *z*-statistic is acceptable for constructing confidence intervals for the population mean. The *t*-statistic, however, will provide a more conservative range (wider) at a given level of significance.

21. D $1.265 = \frac{8}{\sqrt{N}}, N = \left(\frac{8}{1.265}\right)^2 = 40$

22. B With a known population standard deviation of returns and a normally distributed population, we can use the *z*-distribution. The sample mean for a sample of 5 years will have a standard deviation of $\frac{20}{\sqrt{5}} = 8.94\%$. A 90% confidence interval around the mean return of 12% is $12\% \pm 1.65(8.94\%) = -2.75\%$ to 26.75%.

ANSWERS – COMPREHENSIVE PROBLEMS: SAMPLING AND ESTIMATION

1. In simple random sampling, the analyst would select any 50 stocks using a process that gives each stock in the index an equal chance of being chosen.

   Stratified sampling involves dividing a population into subgroups based on key characteristics, selecting random samples from each subgroup in accordance with the proportion of the population contained in each subgroup, and pooling the results. For example, the analyst could divide the index stocks by capitalization and industry to form the subgroups, and then select stocks randomly from each subgroup. In this context, stratified random sampling has the advantage that the sample will have the same proportion of exposure to each industry and firms of, for example, large, small and medium size. If these subgroups successfully capture different risk characteristics, tracking error for the portfolio relative to the index can be reduced.

2. A. She can use the sample mean to estimate the population mean. The central limit theorem states that for a large enough sample size *n* (typically more than 30) from a population with a mean μ and variance σ^2, the probability distribution for the sample mean will be approximately normal with mean μ and variance σ^2/n. The theorem allows us to use the normal distribution to test hypotheses about the population mean, whether the population's distribution is normal or not.

   B. An estimator should be:
   - Unbiased—the expected value of the estimator should be equal to the population parameter.
   - Efficient—the variance of its sampling distribution is smaller than that of all the other unbiased estimators of the parameter.
   - Consistent—the standard error of the estimator should decrease as the sample size increases.

   C. The sample mean has all of these properties.

3. A,B. This is a bit tricky. We have no direct information about the distribution of possible earnings for the next period. We have information about the distribution of *analysts' estimates* of next period earnings. Based on the information given we can make no statement about the 90% confidence interval for earnings next period.

   C. Since we cannot assume that the distribution of analyst estimates is normal, we cannot make any inferences about the mean of the population of analyst estimates with a sample size of only 15.

D. With a sample size of 60 we can make a statement about a confidence interval for the mean of the population of analyst estimates. The t-statistic for a 90% confidence interval with 59 degrees of freedom is approximated by using the value for 60 degrees of freedom, which is 1.671. The confidence interval is $2.84 \pm 1.671\left(\frac{0.40}{\sqrt{60}}\right)$ or \$2.75 to \$2.93. We are 90% confident the true mean of the population of analyst estimates is within this range.

The following is a review of the Quantitative Methods principles designed to address the learning outcome statements set forth by CFA Institute®. This topic is also covered in:

HYPOTHESIS TESTING

Study Session 3

EXAM FOCUS

This review addresses common hypothesis testing procedures. These procedures are used to conduct tests of population means, population variances, differences in means, differences in variances, and mean differences. Specific tests reviewed include the *z*-test, *t*-test, chi-square test, and *F*-test. You should know when and how to apply each of these. A standard hypothesis testing procedure is utilized in this review. Know it! You should be able to perform a hypothesis test on the value of the mean without being given any formulas. Confidence intervals, levels of significance, the power of a test, and types of hypothesis testing errors are also discussed. These are concepts you are likely to see on the exam. Don't worry about memorizing the messy formulas on testing for the equalities and differences in means and variances at the end of this review, but be able to interpret these statistics.

HYPOTHESIS TESTING

Hypothesis testing is the statistical assessment of a statement or idea regarding a population. For instance, a statement could be as follows: "The mean return for the U.S. equity market is greater than zero." Given the relevant returns data, hypothesis testing procedures can be employed to test the validity of this statement at a given significance level.

To illustrate the hypothesis testing concepts and procedures presented in this topic review, an ongoing example pertaining to stock option returns will be used. The background for this example is as follows.

Stock Option Returns: The Common Example

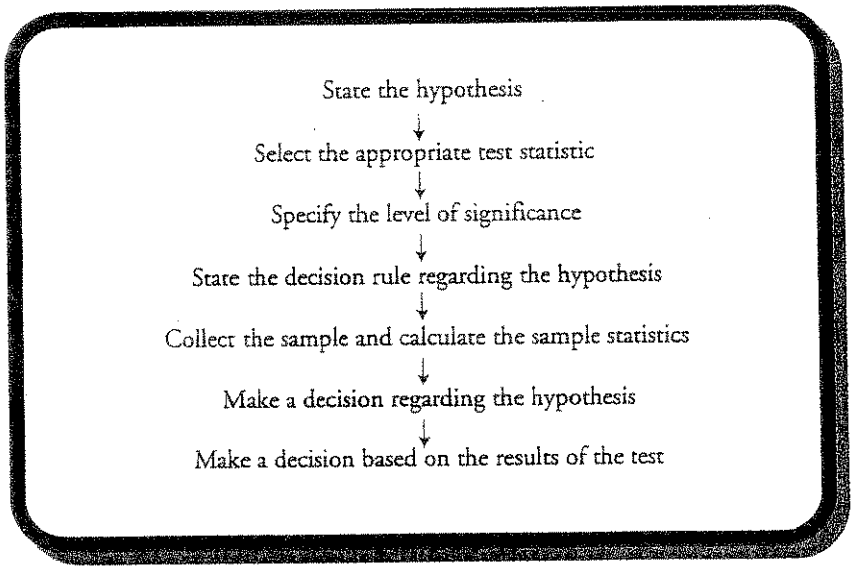
There is an investor who believes that call options should have a mean daily return greater than zero. To empirically assess this belief, she has gathered data on the daily return of a very large portfolio of call options. The mean daily return for the sample portfolio over a period of 250 days is 0.001, or 0.1%, and the sample standard deviation of returns is 0.0025, or 0.25%.

LOS 11.a: Define a hypothesis, describe the steps of hypothesis testing; define and interpret the null hypothesis and alternative hypothesis, discuss the choice of the null and the alternative hypotheses, and distinguish between one-tailed and two-tailed tests of hypotheses.

A **hypothesis** is a statement about the value of a population parameter developed for the purpose of testing a theory or belief. Hypotheses are stated in terms of the population parameter to be tested, like the population mean, μ. For example, a researcher may be interested in the mean daily return on stock options. Hence, the hypothesis may be that the mean daily return on a portfolio of stock options is positive.

Hypothesis testing procedures, based on sample statistics and probability theory, are used to determine whether a hypothesis is a reasonable statement and should not be rejected or if it is an unreasonable statement and should be rejected. The process of hypothesis testing consists of a series of steps shown in Figure 1.

Figure 1: Hypothesis Testing Procedure



* *(Source:* Wayne W. Daniel and James C. Terrell, *Business Statistics, Basic Concepts and Methodology,* Houghton Mifflin, Boston, 1997.)

Professor's Note: You should know this process!

The Null Hypothesis and Alternative Hypothesis

The **null hypothesis**, designated H_0, is the hypothesis that the researcher wants to reject. It is the hypothesis that is actually tested and is the basis for the selection of the test statistics. The null is generally stated as a simple statement about a population parameter. Typical statements of the null hypothesis for the population mean include H_0: $\mu = \mu_0$, H_0: $\mu \leq \mu_0$, and H_0: $\mu \geq \mu_0$, where μ is the population mean and μ_0 is the hypothesized value of the population mean. The null hypothesis always includes the = sign.

The **alternative hypothesis**, designated H_a, is what is concluded if there is sufficient evidence to reject the null hypothesis. It is usually the alternative hypothesis that you are really trying to assess. Why? Since you can never really prove anything with statistics, when the null hypothesis is discredited, the implication is that the alternative hypothesis is valid.

One-Tailed and Two-Tailed Tests of Hypotheses

The alternative hypothesis can be one-sided or two-sided. A one-sided test is referred to as a **one-tailed test**, and a two-sided test is referred to as a **two-tailed test**. Whether the test is one- or two-sided depends on the proposition being tested. If a researcher wants to test whether the return on stock options is greater than zero, a one-tailed test should be used. However, a two-tailed test should be used if the research question is whether the return on options is simply different from zero. Two-sided tests allow for deviation on both sides of the hypothesized value (zero). In practice, most hypothesis tests are constructed as two-tailed tests.

A **two-tailed test** for the population mean may be structured as:

H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$.

Since the alternative hypothesis allows for values above and below the hypothesized parameter, a two-tailed test uses two critical values.

The *general decision rule for a two-tailed test* is:

Reject H_0 if: test statistic > upper critical value or
test statistic < lower critical value

Let's look at the development of the decision rule for a two-tailed test using a *z*-distributed test statistic (a *z*-test) at a 5% level of significance, $\alpha = 0.05$.

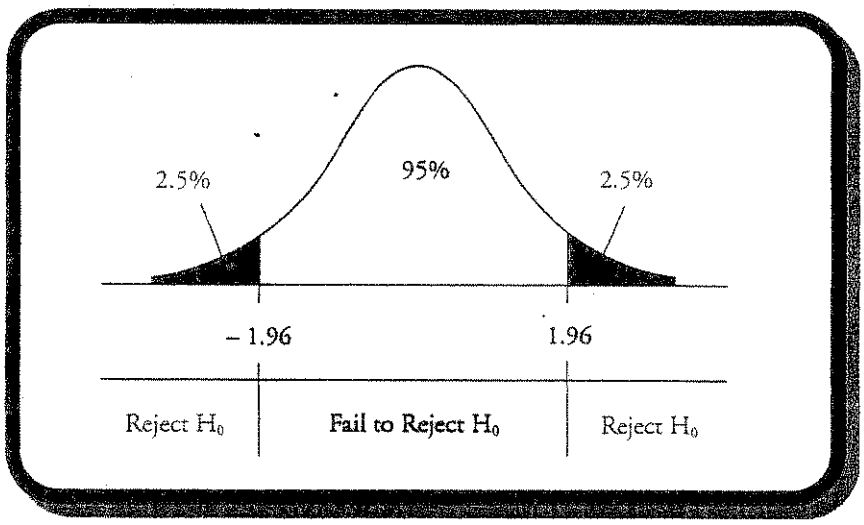
- At $\alpha = 0.05$, the computed test statistic is compared with the critical *z*-values of ±1.96. The values of ±1.96 correspond to $\pm z_{\alpha/2} = \pm z_{0.025}$, which is the range of *z*-values within which 95% of the probability lies. These values are obtained from the cumulative probability table for the standard normal distribution (*z*-table), which is included at the back of this book.
- If the computed test statistic falls outside the range of critical *z*-values (i.e., test statistic > 1.96, or test statistic < –1.96), we reject the null and conclude that the sample statistic is sufficiently different from the hypothesized value.
- If the computed test statistic falls within the range ±1.96, we conclude that the sample statistic is not sufficiently different from the hypothesized value ($\mu = \mu_0$ in this case), and we fail to reject the null hypothesis.

The *decision rule* (rejection rule) *for a two-tailed z-test* at $\alpha = 0.05$ can be stated as:

Reject H_0 if test statistic < –1.96 or if test statistic > 1.96

Figure 2 shows the standard normal distribution for a two-tailed hypothesis test using the *z*-distribution. Notice that the significance level of 0.05 means that there is 0.05 / 2 = 0.025 probability (area) under each tail of the distribution beyond ±1.96.

Figure 2: Two-Tailed Hypothesis Test Using the Standard Normal (*z*) Distribution



Professor's Note: The next two examples are extremely important. Don't move on until you understand them!

Example: Two-tailed test

Referencing our option return data, test the hypothesis that the mean return for options is not zero at the 5% level of significance.

Answer:

Finally, we can perform a hypothesis test for our option return data. Let's start by specifying the null and alternative hypotheses using a two-tailed structure as follows:

H_0: $\mu_0 = 0$ versus H_a: $\mu_0 \neq 0$

At a 5% level of significance, the critical z-values for a two-tailed test are ±1.96, so the decision rule can be stated as:

Reject H_0 if +1.96 < test statistic < −1.96

Our test statistic is $\dfrac{0.001}{\left(\dfrac{0.0025}{\sqrt{250}}\right)} = \dfrac{0.001}{0.000158} = 6.33.$

Since 6.33 > 1.96, we reject the null hypothesis that the mean daily option return is equal to zero. Note that when we reject the null, we conclude that the sample value is significantly different from the hypothesized value. We are saying that the two values are different from one another *after considering the variation in the sample*. That is, the sample mean of 0.001 is statistically different from zero given the sample's standard deviation and size.

For a **one-tailed hypothesis test** of the population mean, the null and alternative hypotheses are either:

Upper tail: H_0: $\mu \leq \mu_0$ versus H_a: $\mu > \mu_0$, or
Lower tail: H_0: $\mu \geq \mu_0$ versus H_a: $\mu < \mu_0$

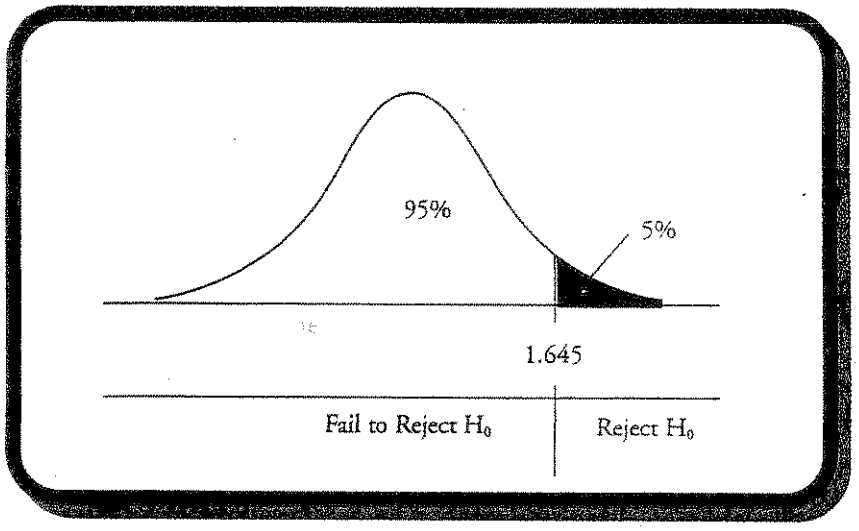
The appropriate set of hypotheses depends on whether we believe the population mean, μ, to be greater than (upper tail) or less than (lower tail) the hypothesized value, μ_0. Using a z-test at the 5% level of significance, the computed test statistic is compared with the critical values of 1.645 for the upper tail tests (i.e., H_a: $\mu > \mu_0$) or −1.645 for lower tail tests (i.e., H_a: $\mu < \mu_0$). These critical values are obtained from a z-table, where $-z_{0.05} = -1.645$ corresponds to a cumulative probability equal to 5%, and the $z_{0.05} = 1.645$ corresponds to a cumulative probability of 95% (1 − 0.05).

Let's use the upper tail test structure where H_0: $\mu \leq \mu_0$ and H_a: $\mu > \mu_0$.

- If the calculated test statistic is greater than 1.645, we conclude that the sample statistic is sufficiently greater than the hypothesized value. In other words, we reject the null hypothesis.
- If the calculated test statistic is less than 1.645, we conclude that the sample statistic is not sufficiently different from the hypothesized value, and we fail to reject the null hypothesis.

Figure 3 shows the standard normal distribution and the rejection region for a one-tailed test (upper tail) at the 5% level of significance.

Figure 3: One-Tailed Hypothesis Test Using the Standard Normal (z) Distribution



Example: One-tailed test

Perform a z-test on our option data to test the proposition that option returns are positive.

Answer:

In this case, we use a one-tailed test with the following structure:

$H_0: \mu \leq 0$ versus $H_a: \mu > 0$

Recalling that $z_{0.05} = 1.645$, the appropriate decision rule for our one-tailed z-test at a significance level of 5% is:

Reject H_0 if test statistic > 1.645

The test statistic is computed the same way regardless of whether we are using a one-tailed or two-tailed test. From the previous example, we know that the test statistic for the option return sample is 6.33. Since 6.33 > 1.645, we reject the null hypothesis and conclude that mean returns are statistically greater than zero at a 5% level of significance.

The Choice of the Null and Alternative Hypotheses

The most common null hypothesis will be an "equal to" hypothesis. Combined with a "not equal to" alternative, this will require a two-tailed test. The alternative is often the hoped-for hypothesis. When the null is that a coefficient is equal to zero, we hope to reject it and show the significance of the relationship.

When the null is less than or equal to, the (mutually exclusive) alternative is framed as greater than, and a one-tail test is appropriate. If we are trying to demonstrate that a return is greater than the risk-free rate, this would be the correct formulation. We will have set up the null and alternative hypothesis so that rejection of the null will lead to acceptance of the alternative, our goal in performing the test.

LOS 11.b: Define and interpret a test statistic, a Type I and a Type II error, and a significance level, and explain how significance levels are used in hypothesis testing.

Hypothesis testing involves two statistics: the test statistic calculated from the sample data and the *critical value* of the test statistic. The value of the computed test statistic relative to the critical value is a key step in assessing the validity of a hypothesis.

A test statistic is calculated by comparing the point estimate of the population parameter with the hypothesized value of the parameter (i.e., the value specified in the null hypothesis). With reference to our option return example, this means we are concerned with the difference between the mean return of the sample (i.e., $\overline{x} = 0.001$) and the hypothesized mean return (i.e., $\mu_0 = 0$). As indicated in the following expression, the test statistic is the difference between the sample statistic and the hypothesized value, scaled by the standard error of the sample statistic.

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the sample statistic}}$$

The standard error of the sample statistic is the adjusted standard deviation of the sample. When the sample statistic is the sample mean, $\overline{x}$, the standard error of the sample statistic for sample size n, is calculated as:

$$\sigma_{\overline{x}} = \sigma / \sqrt{n}$$

when the population standard deviation, σ, *is known*, or

$$s_{\overline{x}} = s / \sqrt{n}$$

when the population standard deviation, σ, *is not known*. In this case, it is estimated using the standard deviation of the sample, s.

Professor's Note: Don't be confused by the notation here. A lot of the literature you will encounter in your studies simply uses the term $\sigma_{\overline{x}}$ for the standard error of the test statistic, regardless of whether the population standard deviation was actually used in its computation.

Example: Test statistic

Compute the test statistic for our option returns example.

Answer:

To compute the test statistic, it is first necessary to calculate the standard error of the sample statistic.

$$s_{\overline{x}} = s / \sqrt{n} = 0.0025 / \sqrt{250} = 0.000158$$

Now, the test statistic can be computed as follows:

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the sample statistic}}$$

$$= \frac{0.001 - 0}{0.000158} = 6.33$$

As you will soon see, a test statistic is a random variable that may follow one of several distributions, depending on the characteristics of the sample and the population. We will look at four distributions for test statistics: the *t*-distribution, the *z*-distribution (standard normal distribution), the chi-square distribution, and the *F*-distribution. The critical value for the appropriate test statistic—the value against which the computed test statistic is compared—is a function of its distribution.

Type I and Type II Errors

Keep in mind that hypothesis testing is used to make inferences about the parameters of a given population on the basis of statistics computed for a sample that is drawn from that population. We must be aware that there is some probability that the sample, in some way, does not represent the population, and any conclusion based on the sample about the population may be made in error.

When drawing inferences from a hypothesis test, there are two types of errors:

- **Type I error**: the rejection of the null hypothesis when it is actually true.
- **Type II error**: the failure to reject the null hypothesis when it is actually false.

The **significance level** is the probability of making a Type I error (rejecting the null when it is true) and is designated by the Greek letter alpha (α). For instance, a significance level of 5% ($\alpha = 0.05$) means that there is a 5% chance of rejecting a true null hypothesis. When conducting hypothesis tests, a significance level must be specified when selecting the critical values to which test statistics are compared.

LOS 11.c: Define and interpret a decision rule and the power of a test, and explain the relation between confidence intervals and hypothesis tests.

The decision for a hypothesis test is to either reject the null hypothesis or fail to reject the null hypothesis. Note that it is statistically incorrect to say "accept" the null hypothesis, it can only be supported or rejected. The **decision rule** for rejecting or failing to reject the null hypothesis is based on the distribution of the test statistic. For example, if the test statistic follows a normal distribution, the decision rule is based on critical values determined from the standard normal distribution (*z*-distribution). Regardless of the appropriate distribution, it must be determined if a one-tailed or two-tailed hypothesis test is appropriate before a decision rule (rejection rule) can be determined.

A decision rule is specific and quantitative. Once we have determined whether a one- or two-tailed test is appropriate, the significance level we require, and the distribution of the test statistic, we can calculate the exact critical value for the test statistic. Then we have a decision rule of the following form: if the test statistic is (greater, less than) the value X, reject the null.

The Power of a Test

While the significance level of a test is the probability of rejecting the null hypothesis when it is true, the **power of a test** is the probability of correctly rejecting the null hypothesis when it is false. The power of a test is actually one minus the probability of making a Type II error, or 1 – P(Type II error). In other words, the probability of rejecting the null when it is false (power of the test) equals one minus the probability of *not* rejecting the null when it is false (Type II error). When more than one test statistic may be used, the power of the test for the competing test statistics may be useful in deciding which test statistic to use. Ordinarily, we wish to use the test statistic that provides the most powerful test among all possible tests.

Figure 4 shows the relationship between the level of significance, the power of a test, and the two types of errors.

Figure 4: Type I and Type II Errors in Hypothesis Testing

	True Condition	
Decision	H_0 is true	H_0 is false
Do not reject H_0	Correct Decision	Incorrect Decision **Type II Error**
Reject H_0	Incorrect Decision **Type I Error** Significance level, α, = P(Type I Error)	Correct Decision Power of the test = 1 – P(Type II Error)

Note that decreasing the probability of making a Type I error (i.e., decreasing the level of significance of the test), makes it more difficult to reject the null when it is true. All else equal, however, the decrease in the chance of making a Type I error comes at the cost of increasing the probability of making a Type II error because the null is rejected less frequently, even when it is actually false. In addition, as the probability of a Type II error increases, the power of the test declines because it is defined as one minus the probability of a Type II error.

The Relation Between Confidence Intervals and Hypothesis Tests

A confidence interval is a range of values within which the researcher believes the true population parameter may lie.

A confidence interval is determined as:

$$\left\{\left[\begin{matrix}\text{sample}\\ \text{statistic}\end{matrix} - \left(\begin{matrix}\text{critical}\\ \text{value}\end{matrix}\right)\left(\begin{matrix}\text{standard}\\ \text{error}\end{matrix}\right)\right] < \begin{matrix}\text{population}\\ \text{parameter}\end{matrix} < \left[\begin{matrix}\text{sample}\\ \text{statistic}\end{matrix} + \left(\begin{matrix}\text{critical}\\ \text{value}\end{matrix}\right)\left(\begin{matrix}\text{standard}\\ \text{error}\end{matrix}\right)\right]\right\}$$

The interpretation of a confidence interval is that for a level of confidence of, say, 95%, there is a 95% probability that the true population parameter is contained in the interval.

From the expression above, we see that a confidence interval and a hypothesis test are linked by the critical value. For example, a 95% confidence interval uses a critical value associated with a given distribution at the 5% level of significance. Similarly, a hypothesis test would compare a test statistic to a critical value at the 5% level of significance. To see this relationship more clearly, the expression for the confidence interval can be manipulated and restated as:

–critical value < test statistic < +critical value

This is the range within which we fail to reject the null for a two-tailed hypothesis test at a given level of significance.

Example: Confidence interval

Using our option example, construct a 95% confidence interval for the population mean. Use a z-distribution. Decide if the hypothesis $\mu = 0$ should be rejected.

Answer:

Given a sample size of 250 with a standard deviation of 0.25%, the standard error can be computed as $s_{\bar{x}} = s/\sqrt{n} = 0.25/\sqrt{250} = 0.0158\%$.

At the 5% level of significance, the critical z-values for the confidence interval are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$. Thus, given a sample mean equal to 0.1%, the 95% confidence interval for the population mean is:

$0.1 - 1.96(0.0158) < \mu < 0.1 + 1.96(0.0158)$, or

$0.069\% < \mu < 0.1310\%$

Since there is a 95% probability that the true mean is within this confidence interval, we can reject the hypothesis $\mu = 0$ because 0 is not within the confidence interval. Alternatively, the z-statistic is $\frac{0.1}{0.0158} = 6.33$, so we reject $\mu = 0$.

The *p-value* is the probability of obtaining a critical value that would lead to a rejection of the null hypothesis, assuming the null hypothesis is true. It is the smallest level of significance for which the null hypothesis can be rejected. For one-tailed tests, the p-value is the probability that lies above the computed test statistic for upper tail tests or below the computed test statistic for lower tail tests. For two-tailed tests, the p-value is the probability that lies above the positive value of the computed test statistic *plus* the probability that lies below the negative value of the computed test statistic.

THE *t*-DISTRIBUTION

The t-distribution is a symmetrical distribution that is centered about zero. The shape of the t-distribution is dependent on the number of degrees of freedom, and degrees of freedom are based on the number of sample observations. The t-distribution is flatter and has thicker tails than the standard normal distribution. As the number of observations increases (i.e., the degrees of freedom increase), the t-distribution becomes more spiked and its tails become thinner. As the number of degrees of freedom increases without bound, the t-distribution converges to the standard normal distribution (z-distribution). The thickness of the tails relative to those of the z-distribution is important in hypothesis testing because thicker tails mean more observations away from the center of the distribution (i.e., more "outliers"). Hence, hypothesis testing using the t-distribution makes it more difficult to reject the null relative to hypothesis testing using the z-distribution.

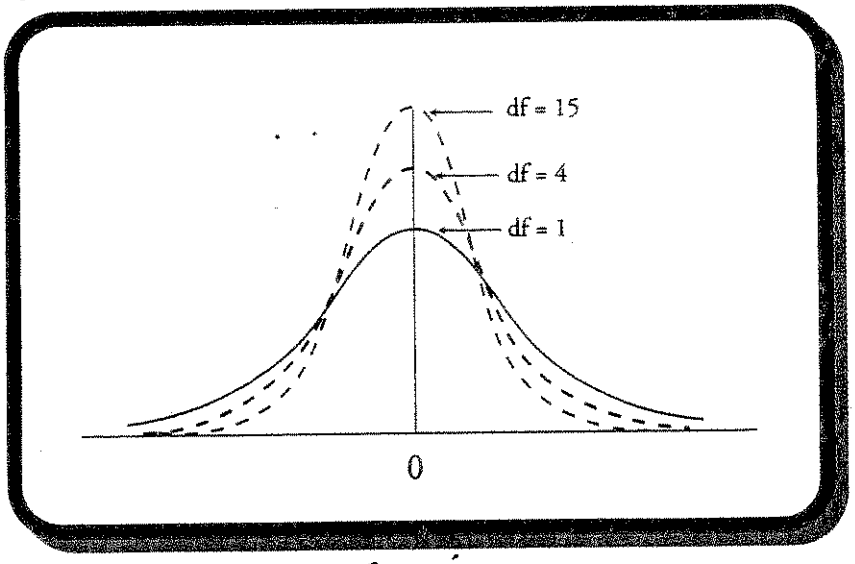
The table in Figure 5 contains one-tailed critical values for the t-distribution at the 0.05 and 0.025 levels of significance with various degrees of freedom (df). Note that, unlike the z-table, the t-values are contained within the table, and the probabilities are located at the column headings. Also note that the level of significance of a t-test corresponds to the *one-tailed probabilities*, p, that head the columns in the t-table.

Figure 6 portrays the different shapes of the *t*-distribution associated with different degrees of freedom and levels of significance. Illustrated in Figure 6 is the tendency for the *t*-distribution to become more peaked and begin to look more and more like the normal distribution as the degrees of freedom increases.

Figure 5: Table of Critical *t*-Values

	One-Tailed Probabilities, p	
df	p = 0.05	p = 0.025
5	2.015	2.571
10	1.812	2.228
15	1.753	2.131
20	1.725	2.086
25	1.708	2.060
30	1.697	2.042
40	1.684	2.021
50	1.676	2.009
60	1.671	2.000
70	1.667	1.994
80	1.664	1.990
90	1.662	1.987
100	1.660	1.984
120	1.658	1.980
∞	1.645	1.960

Figure 6: *t*-Distributions for Different Degrees of Freedom (df)



LOS 11.d: Identify the appropriate test statistic and interpret the results for a hypothesis test concerning 1) the population mean of a normally distributed population with a) known or b) unknown variance; 2) the equality of the population means of two normally distributed populations, based on independent random samples with a) equal or b) unequal assumed variances; and 3) the mean difference of two normally distributed populations (paired comparisons test).

When hypothesis testing, the choice between using a critical value based on the *t*-distribution or the *z*-distribution depends on sample size, the distribution of the population, and whether or not the variance of the population is known.

The *t*-Test

The *t*-test is a widely used hypothesis test that employs a test statistic that is distributed according to a *t*-distribution. Following are the rules for when it is appropriate to use the *t*-test for hypothesis tests of the population mean.

Use the t-test if the population variance is unknown and either of the following conditions exist:

- The sample is large ($n \geq 30$).
- The sample is small (less than 30), but the distribution of the population is normal or approximately normal.

If the sample is small and the distribution is non-normal, we have no reliable statistical test.

The computed value for the test statistic based on the *t*-distribution is referred to as the *t*-statistic. For hypothesis tests of a population mean, a *t*-statistic with $n - 1$ degrees of freedom is computed as:

$$t_{n-1} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where:

$\overline{x}$ = sample mean
μ_0 = hypothesized population mean (i.e., the null)
s = standard deviation of the sample
n = sample size

Professor's Note: This computation is not new. It is the same test statistic computation that we have been performing all along. Note the use of the sample standard deviation, s, in the standard error term in the denominator.

To conduct a *t*-test, the *t*-statistic is compared to a critical *t*-value at the desired level of significance with the appropriate degrees of freedom.

In the real world, the underlying variance of the population is rarely known, so the *t*-test enjoys widespread application.

The *z*-Test

The *z*-test is the appropriate hypothesis test of the population mean when the *population is normally distributed with known variance*. The computed test statistic used with the *z*-test is referred to as the *z*-statistic. The *z*-statistic for a hypothesis test for a population mean is computed as follows:

$$\text{z-statistic} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

where:
$\bar{x}$ = sample mean
μ_0 = hypothesized population mean
σ = standard deviation of the *population*
n = sample size

To test a hypothesis, the *z*-statistic is compared to the critical *z*-value corresponding to the significance of the test. Critical *z*-values for the most common levels of significance are displayed in Figure 7. You should have these memorized by now.

Figure 7: Critical *z*-Values

Level of Significance	*Two-Tailed Test*	*One-Tailed Test*
0.10 = 10%	±1.65	+1.28 or –1.28
0.05 = 5%	±1.96	+1.65 or –1.65
0.01 = 1%	±2.58	+2.33 or –2.33

When the *sample size is large* and the *population variance is unknown*, the *z*-statistic is:

$$\text{z-statistic} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where:
$\bar{x}$ = sample mean
μ_0 = hypothesized population mean
s = standard deviation of the *sample*
n = sample size

Note the use of the sample standard deviation, *s*, versus the population standard deviation, σ. Remember, this is acceptable if the sample size is large, although the *t*-statistic is the more conservative measure when the population variance is unknown.

Example: *z*-test or *t*-test?

Referring to our option return problem once more, explain which test statistic (*z* or *t*) should be used and the difference in the likelihood of rejecting the null with each distribution.

Answer:

The population variance for our sample of returns is unknown. Hence, the *t*-distribution is appropriate. With 250 observations, however, the sample is considered to be large, so the *z*-distribution would also be acceptable. This is a trick question—either distribution, *t* or *z*, is appropriate. With regard to the difference in the likelihood of rejecting the null, since our sample is so large, the critical values for the *t* and *z* are almost identical. Hence, there is almost no difference in the likelihood of rejecting the null.

Example: The *z*-test

When your company's gizmo machine is working properly, the mean length of gizmos is 2.5 inches. However, from time to time the machine gets out of alignment and produces gizmos that are either too long or too short. When this happens, production is stopped and the machine is adjusted. To check the machine, the quality control department takes a gizmo sample each day. Today a random sample of 49 gizmos showed a mean length of 2.49 inches. The population standard deviation is known to be 0.021 inches. Using a 5% significance level, determine if the machine should be shut down and adjusted.

Answer:

Let μ be the mean length of all gizmos made by this machine, and let $\bar{x}$ be the corresponding mean for the sample.

Let's follow the hypothesis testing procedure presented earlier in Figure 1. Again, you should know this process!

Statement of hypothesis. For the information provided, the null and alternative hypotheses are appropriately structured as:

H_0: $\mu = 2.5$ (The machine does not need an adjustment.)
H_a: $\mu \neq 2.5$ (The machine needs an adjustment.)

Note that since this is a two-tailed test, H_a allows for values above and below 2.5.

Select the appropriate test statistic. Since the population variance is known and the sample size is > 30, the *z*-statistic is the appropriate test statistic. The *z*-statistic is computed as:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Specify the level of significance. The level of significance is given at 5%, implying that we are willing to accept a 5% probability of rejecting the null when the null is true.

State the decision rule regarding the hypothesis. The ≠ sign in the alternative hypothesis indicates that the test is two-tailed with two rejection regions, one in each tail of the standard normal distribution curve. Because the total area of both rejection regions combined is 0.05 (the significance level), the area of the rejection region in each tail is 0.025. You should know that the critical *z*-values for $\pm z_{0.025}$ are ±1.96. This means that the null hypothesis should not be rejected if the computed *z*-statistic lies between −1.96 and +1.96 and should be rejected if it lies outside of these critical values. The decision rule can be stated as:

Reject H_0 if $-z_{0.025} > z\text{-statistic} > z_{0.025}$, or equivalently,

Reject H_0 if: $-1.96 > z\text{-statistic} > +1.96$

Collect the sample and calculate the test statistic. The value of $\bar{x}$ from the sample is 2.49. Since σ is given as 0.021, we calculate the *z*-statistic using σ as follows:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{2.49 - 2.5}{0.021 / \sqrt{49}} = \frac{-0.01}{0.003} = -3.33$$

Make a decision regarding the hypothesis. The calculated value of the *z*-statistic is = –3.33. Since this value is less than the critical value, $-z_{0.025} = -1.96$, it falls in the rejection region in the left tail of the *z*-distribution. Hence, there is sufficient evidence to reject H_0.

Make a decision based on the results of the test. Based on the sample information and the results of the test, it is concluded that the machine is out of adjustment and should be shut down for repair.

Hypothesis Tests Concerning the Equality of the Population Means of Two Normally Distributed Populations, Based on Independent Random Samples With 1) Equal or 2) Unequal Assumed Variances

Up to this point, we have been concerned with tests of a single population mean. In practice, we frequently want to know if there is a difference between the means of two populations. There are two *t*-tests that are used to test differences between the means of two populations. Application of either of these tests requires that we are reasonably certain that our samples are independent and that they are taken from two normally distributed populations. Both of these *t*-tests are used when the population variance is unknown. In one case, the population variances are assumed to be equal, and the sample observations are pooled. In the other case, however, no assumption is made regarding the equality between the two population variances, and the *t*-test uses an approximated value for the degrees of freedom.

When testing differences between the mean of population 1, μ_1, and mean of population 2, μ_2, we may be interested in knowing if the two means are equal (i.e., $\mu_1 = \mu_2$), if the mean of population 1 is greater than that of population 2 (i.e., $\mu_1 > \mu_2$), or if the mean of population 2 exceeds that of population 1 (i.e., $\mu_2 > \mu_1$). These three sets of hypotheses are structured as:

$H_0: \mu_1 - \mu_2 = 0$ versus $H_a: \mu_1 - \mu_2 \neq 0$ (a two-tail test)
$H_0: \mu_1 - \mu_2 \leq 0$ versus $H_a: \mu_1 - \mu_2 > 0$ (a one-tail test)
$H_0: \mu_1 - \mu_2 \geq 0$ versus $H_a: \mu_1 - \mu_2 < 0$ (a one-tail test)

Note that it is also possible to structure other hypotheses, such as $H_0: \mu_1 - \mu_2 = 50$ versus $H_a: \mu_1 - \mu_2 \neq 50$. Regardless of the specific structure, the hypothesis testing procedure is the same.

A pooled variance is used with the *t*-test for testing differences between the means of normally distributed populations with **unknown variances that are assumed to be equal**. Assuming independent samples, the *t*-statistic in this case is computed as:

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

where:

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

s_1^2 = variance of the first sample

s_2^2 = variance of the second sample

n_1 = number of observations in the first sample

n_2 = number of observations in the second sample

Note: The degrees of freedom, df, is $(n_1 + n_2 - 2)$, and for a test of equality of means, $\mu_1 - \mu_2 = 0$.

When testing the hypothesis of equality, $\mu_1 - \mu_2 = 0$ so that the numerator is just the difference between the sample means, $\bar{x}_1 - \bar{x}_2$. Since we assume that the variances are equal, we just add the variances of the two sample means in order to calculate the standard error in the denominator.

The *t*-test for differences between population means when the populations are normally distributed having **variances that are unknown and assumed to be unequal** uses the sample variances for both populations. Assuming independent samples, the *t*-statistic in this case is computed as follows:

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

where:

$$\text{degrees of freedom} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1} + \frac{\left(s_2^2/n_2\right)^2}{n_2}}$$

and where:

s_1^2 = variance of the first sample

s_2^2 = variance of the second sample

n_1 = number of observations in the first sample

n_2 = number of observations in the second sample

Again, a test of equality of means will have only the difference in sample means in the numerator. However, with no assumption of equal variances, the denominator (standard error) is based on the individual sample variances of the means for each sample. You do not need to memorize these two formulas but should understand the numerator, the fact that these are *t*-statistics, and that the variance of the pooled sample is used when the sample variances are assumed to be equal.

Example: Difference between means – equal variances

Sue Smith is investigating whether the abnormal returns that occur in acquiring firms during merger announcement periods differ for horizontal and vertical mergers. She estimated the abnormal returns for a sample of acquiring firms associated with horizontal mergers and a sample of acquiring firms involved in vertical mergers. Her sample findings are reported in Figure 8.

Figure 8: Abnormal Returns During Merger Announcement Periods

	Abnormal Returns Horizontal Mergers	*Abnormal Returns Vertical Mergers*
Mean	1.0%	2.5%
Standard deviation	1.0%	2.0%
Sample size (*n*)	64	81

Assuming the samples are independent, the population means are normally distributed, and the population variances are equal, determine if there is a statistically significant difference in the announcement period abnormal returns for these two types of mergers.

Answer:

State the hypothesis. Since this is a two-sided test, the structure of the hypotheses takes the following form:

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

where:
μ_1 = the mean of the abnormal returns for the horizontal mergers
μ_2 = the mean of the abnormal returns for the vertical mergers

Select the appropriate test statistic. Since we are assuming equal variances, the test statistic is computed using the following formula:

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

where:

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

Specify the level of significance. We will use the common significance level of 5% ($\alpha = 0.05$). In order to look up the critical *t*-value, we also need the degrees of freedom, which in this case is $n_1 + n_2 - 2$, or df = 64 + 81 – 2 = 143.

State the decision rule regarding the hypothesis. We must identify the critical *t*-value for a 5% level of significance and the *closest* degrees of freedom specified in a *t*-table. As you should verify with the partial *t*-table contained in Figure 9, the closest entry for df = 143 is df = 120. At $\alpha/2 = p = 0.025$ with df = 120, the critical *t*-value = 1.980.

Figure 9: Partial *t*-Table

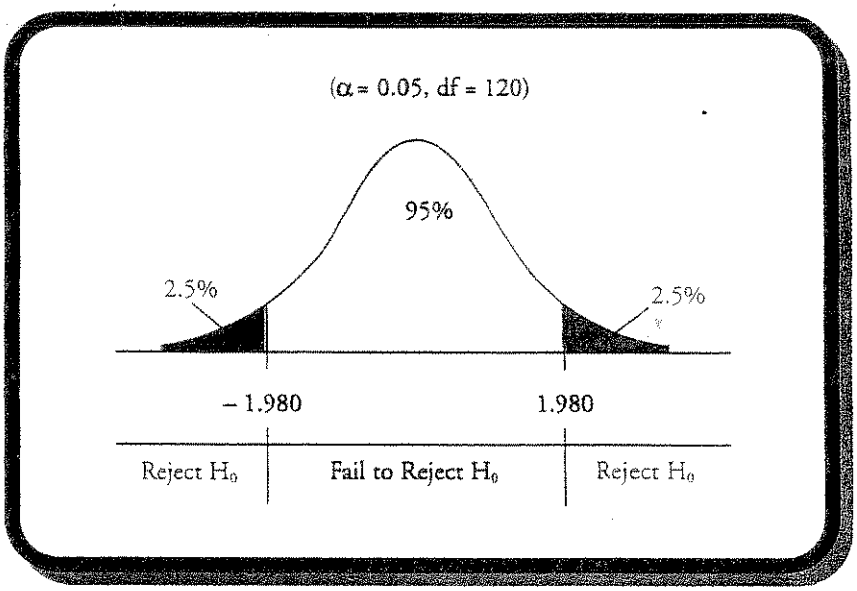
	One-Tailed Probabilities (p)		
df	*p = 0.10*	*p = 0.05*	*p = 0.025*
110	1.289	1.659	1.982
120	1.289	1.658	**1.980**
200	1.286	1.653	1.972

Thus, the decision rule can be stated as:

Reject H_0 if *t*-statistic < –1.980 or *t*-statistic > 1.980

The rejection region for this test is illustrated in Figure 10.

Figure 10: Decision Rule for Two-Tailed *t*-Test



Collect the sample and calculate the sample statistics. Using the information provided, the *t*-statistic can be computed as follows (note that the –0.015 in the numerator equals 0.01 – 0.025, which represents the difference in means) since the hypothesized difference in means $(\mu_1 - \mu_2)$ is zero.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\left(s_p^2/n_1 + s_p^2/n_2\right)^{1/2}} = \frac{-0.015}{0.00274} = -5.474$$

where:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(63)(0.0001) + (80)(0.0004)}{143} = 0.000268$$

Make a decision regarding the hypothesis. Since the calculated test statistic falls to the left of the lowest critical *t*-value, we reject the null hypothesis and conclude that the announcement period abnormal returns are different for horizontal and vertical mergers.

Hypothesis Tests Concerning the Mean Difference of Two Normally Distributed Populations (Paired Comparisons Test)

While the tests considered in the previous section were of the difference between the means of two independent samples, sometimes our samples may be dependent. If the observations in the two samples both depend on some other factor, we can construct a "paired comparisons" test of whether the means of the differences between observations for the two samples are different. Dependence may result from an event that affects both sets of observations for a number of companies or because observations for two firms over time are both influenced by market returns or economic conditions.

For an example of a paired comparisons test, consider a test of whether the returns on two steel firms were equal over a 5-year period. We can't use the difference in means test because we have reason to believe that the samples are not independent. Both will depend to some extent on the returns on the overall market (market risk) and the conditions in the steel industry (industry specific risk). In this case our pairs will be the returns on each firm over the same time periods, so we use the differences in monthly returns for the two companies. The paired

comparisons test is just a test of whether the average difference between monthly returns is significantly different from zero, based on the standard error of the average difference estimated from the sample data.

Remember, the paired comparisons test also requires that the sample data be normally distributed. Although we frequently just want to test the hypothesis that the mean of the differences in the pairs is zero ($\mu_{dz} = 0$), the general form of the test for any hypothesized mean difference, μ_{dz}, is as follows:

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

where:
μ_d = mean of the population of paired differences
μ_{dz} = hypothesized mean of paired differences, which is commonly zero

For one-sided tests, the hypotheses are structured as either:

$H_0: \mu_d \leq \mu_{dz}$ versus $H_a: \mu_d > \mu_{dz}$, or $H_0: \mu_d \geq \mu_{dz}$ versus $H_a: \mu_d < \mu_{dz}$

For the paired comparisons test, the *t*-statistic with n – 1 degrees of freedom is computed as:

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}}$$

where:

$\bar{d}$ = sample mean difference $= \frac{1}{n}\sum_{i=1}^{n} d_i$

d_i = difference between the *i*th pair of observations

$s_{\bar{d}}$ = standard error of the mean difference $= \frac{s_d}{\sqrt{n}}$

s_d = sample standard deviation $= \left(\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1} \right)^{1/2}$

n = the number of paired observations

Example: Paired comparisons test

Joe Andrews is examining changes in estimated betas for the common stock of companies in the telecommunications industry before and after deregulation. Andrews believes that the betas may decline because of deregulation since companies are no longer subject to the uncertainties of rate regulation or that they may increase because there is more uncertainty regarding competition in the industry. The sample information he gathered is reported in Figure 11. Determine whether there is a change in betas.

Figure 11: Beta Differences After Merger Announcement

Mean of differences in betas (before minus after)	0.23
Sample standard deviation of differences	0.14
Sample size	39

Answer:

Once again, we follow our hypothesis testing procedure.

State the hypothesis. There is reason to believe that the mean differences may be positive or negative, so a two-sided alternative hypothesis is in order here. Thus, the hypotheses are structured as:

H_0: $\mu_d = 0$ versus H_a: $\mu_d \neq 0$

Select the appropriate test statistic. As described above, the test statistic for a paired comparisons test is:

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}}$$

Specify the level of significance. Let's use a 5% level of significance.

State the decision rule regarding the hypothesis. There are 39 – 1 = 38 degrees of freedom. Using the *t*-distribution, the two-tailed critical *t*-values for a 5% level of significance with df = 38 is ±2.024. As indicated in the table in Figure 12, the critical *t*-value of 2.024 is located at the intersection of the $p = 0.025$ column and the df = 38 row. The one-tailed probability of 0.025 is used because we need 2.5% in each tail for 5% significance with a two-tailed test.

Figure 12: Partial *t*-Table

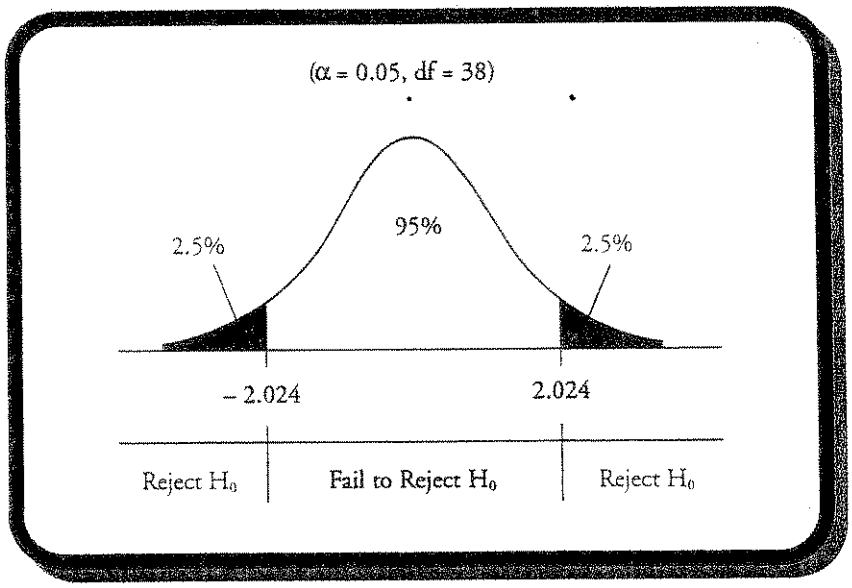
	One-Tailed Probabilities (p)		
df	*p = 0.10*	*P = 0.05*	*p = 0.025*
38	1.304	1.686	**2.024**
39	1.304	1.685	2.023
40	1.303	1.684	2.021

Thus, the decision rule becomes:

Reject H_0 if *t*-statistic < –2.024, or *t*-statistic > 2.024

This decision rule is illustrated in Figure 13.

Figure 13: Decision Rule for a Two-Tailed Paired Comparisons Test



Collect the sample and calculate the sample statistics. Using the sample data provided, the test statistic is computed as follows:

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}} = \frac{0.23}{0.14/\sqrt{39}} = \frac{0.23}{0.022418} = 10.2596$$

Make a decision regarding the hypotheses. The computed test statistic, 10.2596, is greater than the critical *t*-value, 2.024—it falls in the rejection region to the right of 2.024 in Figure 13. Thus we reject the null hypothesis of no difference, concluding that there *is* a statistically significant difference in betas from before to after deregulation.

Make a decision based on the results of the test. We have support for the hypothesis that betas are lower as a result of deregulation, providing support for the proposition that deregulation resulted in decreased risk.

Keep in mind that we have been describing two distinct hypothesis tests: differences between the means of two populations versus the mean of the paired differences from two normal populations. Here are rules for when these tests may be applied:

- The test of the differences in means is used when there are two *independent samples*.
- The test of the mean of the difference is used when the samples are *not independent* but in fact allow *paired comparisons*.

Professor's Note: The LOS here say "Identify the appropriate test statistic and interpret the results..." I can't believe candidates are expected to memorize these formulas (or that you would be a better analyst if you did). The CFA exam is not known for requiring the use of complicated formulas from memory. You should instead focus on the fact that both of these tests involve t-statistics and depend on the degrees of freedom. Also note that when samples are independent you can use the difference in means test and when they are dependent, the statistic is the average difference in (paired) observations divided by the standard error of the average difference.

LOS 11.e: 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.

The *chi-square test* is used for hypothesis tests concerning the variance of a normally distributed population. Letting σ^2 represent the true population variance and σ_0^2 represent the hypothesized variance, the hypotheses for a two-tailed test of a single population variance are structured as:

H_0: $\sigma^2 = \sigma_0^2$ versus H_a: $\sigma^2 \neq \sigma_0^2$

The hypotheses for one-tailed tests are structured as:

H_0: $\sigma^2 \leq \sigma_0^2$ versus H_a: $\sigma^2 > \sigma_0^2$, or
H_0: $\sigma^2 \geq \sigma_0^2$ versus H_a: $\sigma^2 < \sigma_0^2$

Hypothesis testing of the population variance requires the use of a chi-square distributed test statistic, denoted χ^2. The chi-square distribution is asymmetrical and approaches the normal distribution in shape as the degrees of freedom increase.

To illustrate the chi-square distribution, consider a two-tailed test with a 5% level of significance and 30 degrees of freedom. As displayed in Figure 14, the critical chi-square values are 16.791 and 46.979 for the lower and upper bounds, respectively. These values are obtained from a chi-square table, which is used in the same manner as a *t*-table. A portion of a chi-square table is presented in Figure 15.

Note that the chi-square values in the table in Figure 15 correspond to the probabilities in the right tail of the distribution. As such, the 16.791 in Figure 14 is from the column headed *0.975* because 95% + 2.5% of the probability is to the right of it. The 46.979 is from the column headed *0.025* because only 2.5% probability is to the right of it. Similarly, at a 5% level of significance with 10 degrees of freedom, Figure 15 shows that the critical chi-square values for a two-tailed test are 3.247 and 20.483.

Figure 14: Decision Rule for a Two-Tailed Chi-square Test

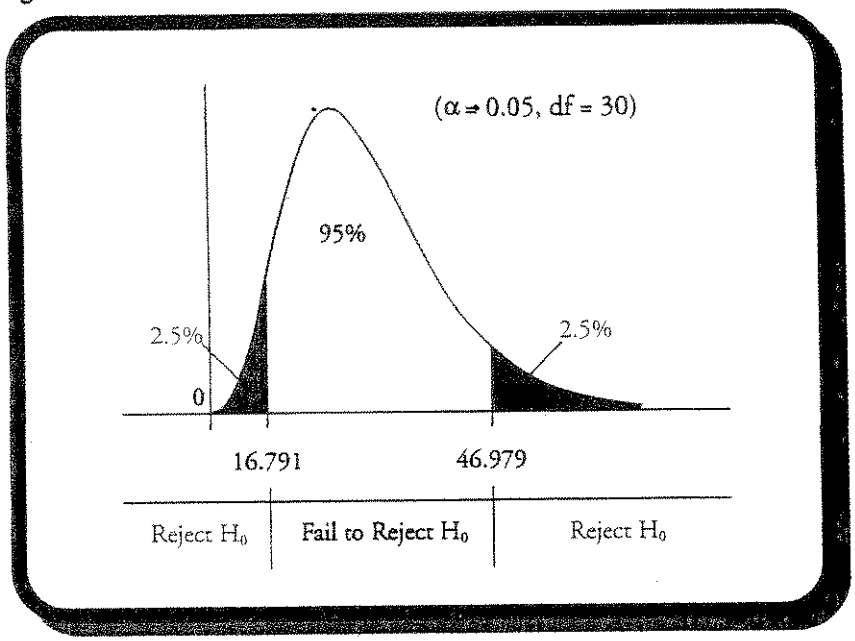


Figure 15: Chi-Square Table

Degrees of Freedom	*Probability in Right Tail*					
	0.975	0.95	0.90	0.1	0.05	0.025
9	2.700	3.325	4.168	14.684	16.919	19.023
10	3.247	3.940	4.865	15.987	18.307	20.483
11	3.816	4.575	5.578	17.275	19.675	21.920
30	**16.791**	**18.493**	20.599	40.256	43.773	**46.979**

The chi-square test statistic, χ^2, with n – 1 degrees of freedom, is computed as:

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

where:
n = sample size
s^2 = sample variance
σ_0^2 = hypothesized value for the population variance.

Similar to other hypothesis tests, the chi-square test compares the test statistic, χ^2_{n-1}, to a critical chi-square value at a given level of significance and n – 1 degrees of freedom. Note that since the chi-square distribution is bounded below by zero, chi-square values cannot be negative.

Example: Chi-square test for a single population variance

Historically, High-Return Equity Fund has advertised that its monthly returns have a standard deviation equal to 4%. This was based on estimates from the 1990-1998 period. High-Return wants to verify whether this claim still adequately describes the standard deviation of the fund's returns. High-Return collected

monthly returns for the 24-month period between 1998 and 2000 and measured a standard deviation of monthly returns of 3.8%. Determine if the more recent standard deviation is different from the advertised standard deviation.

Answer:

State the hypothesis. The null hypothesis is that the standard deviation is equal to 4%, and therefore the variance of monthly returns for the population is $(0.04)^2 = 0.0016$. Since High-Return simply wants to test whether the standard deviation has changed, up or down, a two-sided test should be used. The hypothesis test structure takes the form:

$H_0: \sigma_0^2 = 0.0016$ versus $H_a: \sigma^2 \neq 0.0016$

Select the appropriate test statistic. The appropriate test statistic for tests of variance using the chi-square distribution is computed as follows:

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

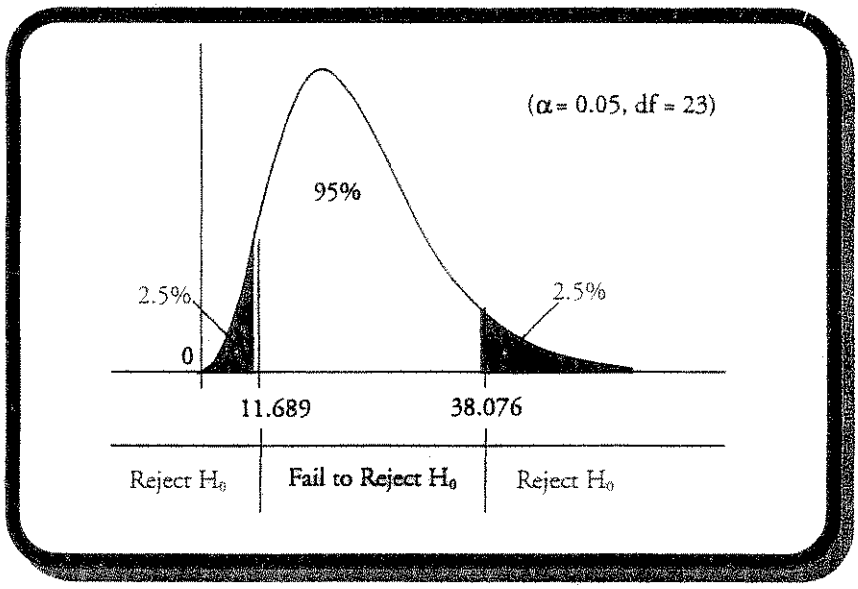
Specify the level of significance. Let's use a 5% level of significance, meaning there will be 2.5% probability in each tail of the chi-square distribution.

State the decision rule regarding the hypothesis. With a 24-month sample, there are 23 degrees of freedom. Using the table of chi-square values at the back of this book, for 23 degrees of freedom and probabilities of 0.975 and 0.025, we find two critical values, 11.689 and 38.076. Thus, the decision rule is:

Reject H_0 if $\chi^2 < 11.689$, or $\chi^2 > 38.076$

This decision rule is illustrated in Figure 16.

Figure 16: Decision Rule for a Two-Tailed Chi-Square Test of a Single Population Variance



Collect the sample and calculate the sample statistics. Using the information provided, the test statistic is computed as:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(23)(0.001444)}{0.0016} = \frac{0.033212}{0.0016} = 20.7575$$

Make a decision regarding the hypothesis. Since the computed test statistic, χ^2, falls between the two critical values, we fail to reject the null hypothesis that the variance is equal to 0.0016.

Make a decision based on the results of the test. It can be concluded that the recently measured standard deviation is close enough to the advertised standard deviation that we cannot say that it is different from 4%, at a 5% level of significance.

Testing the Equality of the Variances of Two Normally Distributed Populations, Based on Two Independent Random Samples

The hypotheses concerned with the equality of the variances of two populations are tested with an *F*-distributed test statistic. Hypothesis testing using a test statistic that follows an *F*-distribution is referred to as the *F*-test. The *F*-test is used under the assumption that the populations from which samples are drawn are normally distributed and that the samples are independent.

If we let σ_1^2 and σ_2^2 represent the variances of normal population 1 and population 2, respectively, the hypotheses for the two-tailed *F*-test of differences in the variances can be structured as:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_a: \sigma_1^2 \neq \sigma_2^2$$

and the one-sided test structures can be specified as:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \text{ versus Ha: } \sigma_1^2 > \sigma_2^2, \text{ or } H_0: \sigma_1^2 \geq \sigma_2^2 \text{ versus Ha: } \sigma_1^2 < \sigma_2^2$$

The test statistic for the *F*-test is the ratio of the sample variances. The *F*-statistic is computed as:

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

where:

s_1^2 = variance of the sample of n_1 observations drawn from population 1

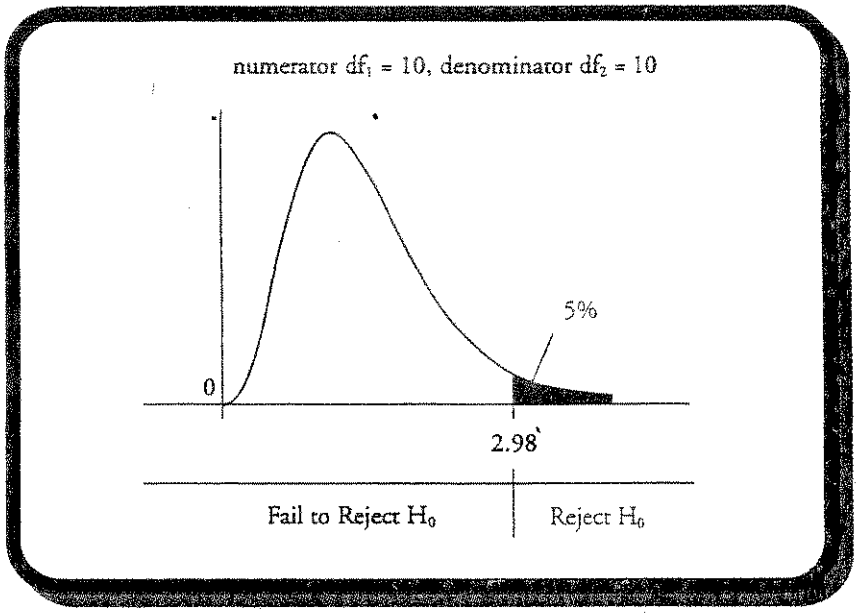
s_2^2 = variance of the sample of n_2 observations drawn from population 2

Note that $n_1 - 1$ and $n_2 - 1$ are the degrees of freedom used when computing s_1^2 and s_2^2, respectively.

Professor's Note: Always put the larger variance in the numerator (s_1^2). Following this convention means we only have to consider the critical value for the right-hand tail.

An *F*-distribution is presented in Figure 17. As indicated, the *F*-distribution is right-skewed and is truncated at zero on the left-hand side. The shape of the *F*-distribution is determined by *two separate degrees of freedom*, the numerator degrees of freedom, df_1, and the denominator degrees of freedom, df_2. Also shown in Figure 17 is that the *rejection region is in the right-side tail* of the distribution. This will always be the case as long as the *F*-statistic is computed with the largest sample variance in the numerator. The labeling of 1 and 2 is arbitrary anyway.

Figure 17: *F*-Distribution



Example: *F*-test for equal variances

Annie Cower is examining the earnings for two different industries. Cower suspects that the earnings of the textile industry are more divergent than those of the paper industry. To confirm this suspicion, Cower has looked at a sample of 31 textile manufacturers and a sample of 41 paper companies. She measured the sample standard deviation of earnings across the textile industry to be $4.30 and that of the paper industry companies to be $3.80. Determine if the earnings of the textile industry have greater standard deviation than those of the paper industry.

Answer:

State the hypothesis. In this example, we are concerned with whether the variance of the earnings of the textile industry is greater (more divergent) than the variance of the earnings of the paper industry. As such, the test hypotheses can be appropriately structured as:

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

where:

σ_1^2 = variance of earnings for the textile industry

σ_2^2 = variance of earnings for the paper industry

Note: $\sigma_1^2 > \sigma_2^2$

Select the appropriate test statistic. For tests of difference between variances, the appropriate test statistic is:

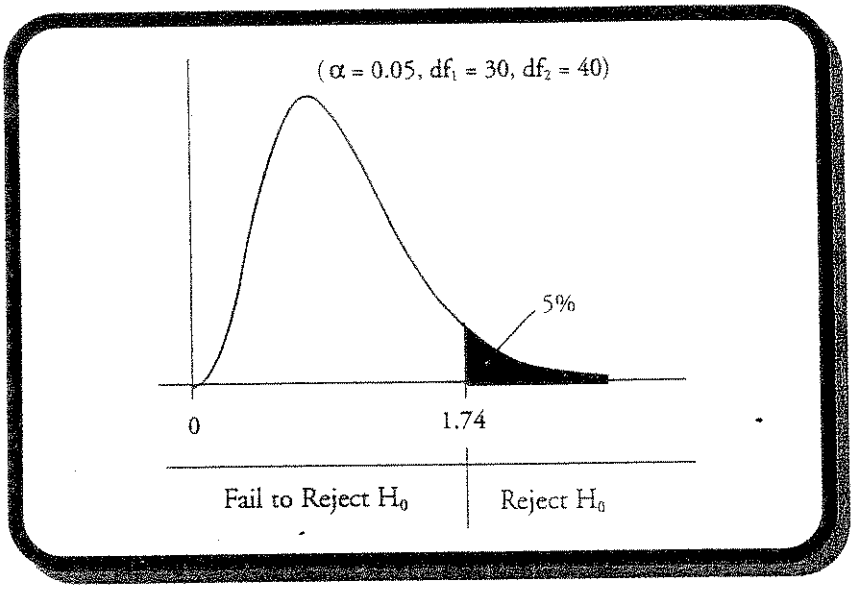
$$F = \frac{s_1^2}{s_2^2}$$

Specify the level of significance. Let's conduct our hypothesis test at the 5% level of significance.

State the decision rule regarding the hypothesis. Using the sample sizes for the two industries, the critical *F*-value for our test is found to be 1.74. This value is obtained from the table of the *F*-distribution at the 5% level of significance with $df_1 = 30$ and $df_2 = 40$. Thus, if the computed *F*-statistic is greater than the critical value of 1.74, the null hypothesis is rejected. The decision rule, illustrated in Figure 18 below, can be stated as:

Reject H_0 if $F > 1.74$

Figure 18: Decision Rule for *F*-Test



Collect the sample and calculate the sample statistics. Using the information provided, the *F*-statistic can be computed as:

$$F = \frac{s_1^2}{s_2^2} = \frac{\$4.30^2}{\$3.80^2} = \frac{\$18.49}{\$14.44} = 1.2805$$

Professor's Note: Remember to square the standard deviations to get the variances.

Make a decision regarding the hypothesis. Since the calculated *F*-statistic of 1.2805 is less than the critical *F*-statistic of 1.74, we fail to reject the null hypothesis.

Make a decision based on the results of the test. Based on the results of the hypothesis test, Cower should conclude that the earnings variances of the industries are not statistically significantly different from one another at a 5% level of significance. More pointedly, the earnings of the textile industry are not more divergent than those of the paper industry.

LOS 11.f: Distinguish between parametric and nonparametric tests and describe the situations in which the use of nonparametric tests may be appropriate.

Parametric tests rely on assumptions regarding the distribution of the population and are specific to population parameters. For example, the *z*-test relies upon a mean and a standard deviation to define the normal distribution. The *z*-test also requires that either the sample is large, relying on the central limit theorem to assure a normal sampling distribution, or that the population is normally distributed.

Nonparametric tests either do not consider a particular population parameter or have few assumptions about the population that is sampled. Nonparametric tests are used when there is concern about quantities other than the parameters of a distribution or when the assumptions of parametric tests can't be supported. They are also used when the data are not suitable for parametric tests (e.g., ranked observations). Nonparametric tests are often used along with parametric tests. In this way, the nonparametric test is a backup in case the assumptions underlying the parametric test do not hold.

KEY CONCEPTS

1. The hypothesis testing process requires a statement of a null and an alternative hypothesis, the selection of the appropriate test statistic, specification of the significance level, a decision rule, the calculation of a sample statistic, a decision regarding the hypotheses based on the test, and a decision based on the test results.
2. The null hypothesis is what the researcher wants to reject. The alternative hypothesis is what the researcher wants to prove, and it is accepted when the null hypothesis is rejected.
3. A two-tailed test results from a two-sided alternative hypothesis (e.g., H_a: $\mu \neq \mu_0$). A one-tailed test results from a one-sided alternative hypothesis (e.g., H_a: $\mu > \mu_0$, or H_a: $\mu < \mu_0$).
4. The decision rule depends on the alternative hypothesis and the distribution of the test statistic.
5. A Type I error is the rejection of the null hypothesis when it is actually true, while a Type II error is the failure to reject the null hypothesis when it is actually false.
6. The significance level can be interpreted as the probability that a test statistic will reject the null hypothesis by chance when it is actually true (i.e., the probability of a Type I error.)
7. The power of a test is the probability of rejecting the null when it is false. The power of a test = 1 – P(Type II error).
8. Hypothesis testing compares a computed test statistic to a critical value at a stated level of significance, which is the decision rule for the test.
9. A hypothesis about a population parameter is rejected when the sample statistic lies outside a confidence interval around the hypothesized value for the chosen level of significance.
10. With unknown population variance, the t-statistic is used for tests of the mean of a normally distributed population: $t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. If the population variance is known, the appropriate test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ for tests of the mean of a population.
11. For two independent samples from two normally distributed populations, the difference in means can be tested with a *t*-statistic. When the two population variances are assumed to be equal, the denominator is based on the variance of the pooled samples, but when sample variances are assumed to be unequal, the denominator is based on a combination of the two samples variances.
12. A paired comparisons test is concerned with the mean of the differences between the paired observations of two dependent, normally distributed samples. A *t*-statistic is used: $t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}}$, where $s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$, and $\bar{d}$ is the average difference of the *n* paired observations.
13. The test of a hypothesis about the population variance for a normally distributed population uses a chi-square test statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$, where *n* is the sample size, s^2 is the sample variance, and σ_0^2 is the hypothesized value for the population variance. Degrees of freedom is n – 1.
14. The test comparing two variances based on independent samples from two normally distributed populations uses an *F*-distributed test statistic: $F = \frac{s_1^2}{s_2^2}$, where s_1^2 is the variance of the first sample and s_2^2 is the (smaller) variance of the second sample.
15. Parametric tests, like the *t*-test, *F*-test, and chi-square tests, make assumptions regarding the distribution of the population from which samples are drawn, while nonparametric tests either do not consider a particular population parameter or have few assumptions about the sampled population.

CONCEPT CHECKERS: HYPOTHESIS TESTING

1. Which of the following statements about hypothesis testing is TRUE?
   A. A Type II error is rejecting the null when it is actually true.
   B. The significance level equals one minus the probability of a Type I error.
   C. If the alternative hypothesis is H_a: $\mu > \mu_0$, the test is a two-tailed test.
   D. A two-tailed test with a significance level of 5% has *z*-critical values of ±1.96.

2. Which of the following statements is FALSE?
   A. The power of test = 1 – P(Type II error).
   B. A two-tailed test with a significance level of 5% has z-critical values of ±1.96.
   C. If the computed *z*-statistic = –2 and the critical *z*-value = –1.96, the null hypothesis is rejected.
   D. The calculated *z*-statistic for a test of a sample mean when the population variance is known is

$$z = \frac{\overline{X} - \mu}{\frac{\sigma^2}{\sqrt{n}}}.$$

Use the following data to answer Questions 3 through 7.

Austin Roberts wants to determine whether the mean price of houses in the area is greater than $145,000. A random sample of 36 houses in the area has a mean price of $149,750. The population standard deviation is $24,000, and Roberts wants to conduct hypothesis testing at a 1 percent level of significance.

3. The appropriate alternative hypothesis is:
   A. H_a: $\mu <$ $145,000.
   B. H_a: $\mu \pm$ $145,000.
   C. H_a: $\mu \geq$ $145,000.
   D. H_a: $\mu >$ $145,000.

4. The value of the calculated test statistic is *closest* to:
   A. $z = 0.67$.
   B. $z = 1.19$.
   C. $z = 4.00$.
   D. $z = 8.13$.

5. Which of the following *most accurately* describes the appropriate test structure?
   A. *F*-test.
   B. Two-tailed test.
   C. One-tailed test.
   D. Chi-square test.

6. The critical value of the *z*-statistic is:
   A. $z = \pm 1.96$.
   B. $z = +2.33$.
   C. $z = -2.33$.
   D. $z = \pm 2.33$.

7. At a 1% level of significance, Roberts should:
   A. accept the null hypothesis.
   B. reject the null hypothesis.
   C. fail to reject the null hypothesis.
   D. neither reject nor fail to reject the null hypothesis.

Use the following data to answer Questions 8 through 13.

An analyst is conducting a hypothesis test to determine if the mean time spent on investment research is different from 3 hours per day. The test is performed at the 5% level of significance and uses a random sample of 64 portfolio managers, where the mean time spent on research is found to be 2.5 hours. The population standard deviation is 1.5 hours.

8. The appropriate null hypothesis for the described test is:
   A. H_0: $\mu = 3$ hours.
   B. H_0: $\mu \neq 3$ hours.
   C. H_0: $\mu \leq 3$ hours.
   D. H_0: $\mu \geq 3$ hours.

9. This is a:
   A. one-tailed test.
   B. two-tailed test.
   C. chi-square test.
   D. paired comparisons test.

10. The calculated z-statistic is:
   A. –2.13.
   B. –2.67.
   C. +0.33.
   D. +2.67.

11. The critical z-value(s) of the test statistic is (are):
   A. –1.96.
   B. +1.96.
   C. ±1.96.
   D. ±2.58.

12. The 95% confidence interval for the population mean is:
   A. $\{1.00 < \mu < 3.50\}$.
   B. $\{0.54 < \mu < 4.46\}$.
   C. $\{2.13 < \mu < 2.87\}$.
   D. $\{-1.96 < \mu < 1.96\}$.

13. Which of the following decisions is the CORRECT decision for this study?
   A. Reject the null hypothesis.
   B. Fail to reject the null hypothesis.
   C. The sample size is too small, so increase the sample size.
   D. No decision is possible because the sample standard deviation was not given.

14. A study was conducted to determine whether the standard deviation of monthly maintenance costs of a Pepper III aircraft is $300. A sample of 30 Pepper IIIs had a mean monthly maintenance cost of $3,025 and a standard deviation of $325. Using a 5% level of significance, which of the following is the most appropriate conclusion regarding the *difference* between the hypothesized value of the population variance and the sample variance?
A. The difference is not meaningful.
B. The population and sample variances are significantly different.
C. The population and sample variances are not significantly different.
D. There are no tests that may be used to test variance differences in small samples.

Use the following data to answer Questions 15 through 20.

Two samples were drawn from a normally distributed population. For the first sample, the mean was $50 and the standard deviation was $5. For the second sample, the mean was $55 and the standard deviation was $6. The first sample consists of 25 observations and the second sample consists of 36 observations. (Note: In the questions below, the subscripts "1" and "2" indicate the first and second sample, respectively.)

15. Consider the hypotheses structured as H_0: $\mu_1 = \$48$ versus H_a: $\mu_1 \neq \$48$. At a 1% level of significance, the null hypothesis:
A. cannot be rejected.
B. should be rejected.
C. should neither be rejected nor failed to be rejected.
D. cannot be tested using this sample information provided.

16. Using a 5% level of significance and a hypothesis test structure of H_0: $\sigma^2_1 \leq 24$ versus H_a: $\sigma^2_1 > 24$, the null hypothesis:
A. cannot be rejected.
B. should be rejected.
C. should neither be rejected nor failed to be rejected.
D. cannot be tested using this sample information provided.

17. Consider the hypotheses structured as H_0: $\mu_1 \leq \$48$ versus H_a: $\mu_1 > \$48$. At a 5% level of significance, the null hypothesis:
A. cannot be rejected.
B. should be rejected.
C. should neither be rejected nor failed to be rejected.
D. cannot be tested using the sample information provided.

18. Using a 5% level of significance for a test of the null of H_0: $\sigma_1 = \sigma_2$ versus the alternative of H_a: $\sigma_1 \neq \sigma_2$, the null hypothesis:
A. cannot be rejected.
B. should be rejected.
C. should neither be rejected nor failed to be rejected.
D. cannot be tested using the sample information provided.

19. If the significance level of a test is 0.05 and the probability of a Type II error is 0.15, what is the power of the test?
A. 0.015.
B. 0.950.
C. 0.975.
D. 0.850.

20. All of the following are true about the *F*-distribution and chi-square distribution EXCEPT they:
A. are both asymmetrical.
B. are both bound by zero on the left.
C. are both defined by degrees of freedom.
D. both have means that are less than their standard deviations.

21. The appropriate test statistic for a test of the equality of variances for two normally distributed random variables, based on two independent random samples, is the:
A. *t*-test.
B. *F*-test.
C. χ^2 test.
D. *z*-test.

22. The appropriate test statistic for a test that the variance of a normally distributed population is equal to 13, is the:
A. *t*-test.
B. *F*-test.
C. χ^2 test.
D. *z*-test.

23. William Adams wants to test whether the mean monthly returns over the last 5 years are the same for two stocks. If he assumes that the returns distributions are normal and have equal variances, the type of test and test statistic are best described as:
A. paired comparisons test, *t*-statistic.
B. paired comparisons test, *F*-statistic.
C. difference in means test, *t*-statistic.
D. difference in means test, *F*-statistic.

24. Which of the following assumptions is NOT required for the difference in means test based on two samples?
A. The two samples are independent.
B. The two populations are normally distributed.
C. The sample means are approximately normally distributed.
D. The two populations have equal variances.

25. For a hypothesis test with a probability of a Type II error of 60% and a probability of a Type I error of 5%, which of the following statements is TRUE?
A. The power of the test is 40% and there is a 5% probability that the test statistic will exceed the critical value(s).
B. There is a 95% probability that the test statistic will be between the critical values if this is a two-tail test.
C. The power of the test is 55%, and the confidence level is 95%.
D. There is a 5% probability that the null hypothesis will be rejected when actually true, and the probability of rejecting the null when it is false is 40%.

COMPREHENSIVE PROBLEMS: HYPOTHESIS TESTING

1. Ralph Rollins, a researcher, believes that the stocks of firms that have appeared in a certain financial newspaper with a positive headline and story return more on a risk-adjusted basis. He gathers data on the risk-adjusted returns for these stocks over the six months after they appear on the cover, and data on the risk-adjusted returns for an equal-sized sample of firms with characteristics similar to the cover-story firms matched by time period.

   A. State the likely null and alternative hypotheses for a test of his belief.

   B. Is this a one- or two-tailed test?

   C. Describe the steps in testing a hypothesis such as the null you describe.

2. For each of the following hypotheses, describe the appropriate test, identify the appropriate test statistic, and explain under what conditions the null hypothesis should be rejected.

   A. A researcher has returns over 52 weeks for an index of natural gas stocks and for an index of oil stocks and wants to know if the weekly returns are equal. Assume that the returns are approximately normally distributed.

   B. A researcher has two independent samples that are approximately normally distributed. She wishes to test whether the mean values of the two random variables are equal and assumes that the variances of the populations from which the two samples are drawn are not equal.

   C. A researcher has two independent samples that are approximately normally distributed. She wishes to test whether the mean values of the two random variables are equal and assumes that the variances of the populations from which the two samples are drawn are equal. As an additional question here, how should the degrees of freedom be calculated?

   D. A researcher wants to determine whether the variances of two normally distributed random samples are equal. As an additional question here, how should the degrees of freedom be calculated?

   E. A researcher wants to test whether the variance of a normally distributed population is equal to 0.00165. As an additional question here, how should the degrees of freedom be calculated?

ANSWERS – CONCEPT CHECKERS: HYPOTHESIS TESTING

1. D Rejecting the null when it is actually true is a Type I error. A Type II error is failing to reject the null hypothesis when it is false. The significance level equals the probability of a Type I error. If the alternative hypothesis is H_a: $\mu > \mu_0$, then the test is a one-tailed test. A two-tailed test would have an alternative hypothesis of H_a: $\mu \neq \mu_0$.

2. D $z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ (σ^2 is the variance)

3. D H_a: $\mu >$ \$145,000

4. B $z = \dfrac{149{,}750 - 145{,}000}{24{,}000/\sqrt{36}} = 1.1875$

5. C The alternative hypothesis, H_a: $\mu >$ \$145,000, only allows for values greater than the hypothesized value. Thus, this is a one-sided (one-tailed) test.

6. B For a one-tailed z-test at the 1% level of significance, the critical z-value is $z_{0.01} = 2.33$. Since the test is one-tailed on the upper end (i.e., H_a: $\mu > 145{,}000$), we use a positive z-critical value.

7. C The decision rule is to reject H_0 if z-computed > z-critical. Since 1.1875 < 2.33, Roberts will fail to reject the null.

8. A H_0: $\mu = 3$ hours

9. B This is a two-sided (tailed) test. We want to test if the mean "differs from" 3 hours (i.e., H_a: $\mu \neq 3$ hours).

10. B The normally distributed test statistic = $z = \dfrac{(2.5 - 3.0)}{1.5/\sqrt{64}} = -2.67$.

11. C At $\alpha/2 = 0.025$, the critical z-values are: $\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.96$.

12. C The 95% confidence interval is $\{2.5 \pm (1.96)(0.1875)\} = \{2.5 \pm 0.3675\} \rightarrow \{2.1325 < \mu < 2.8675\}$.

13. A Decision rule: reject H_0 if $z_{computed} < -1.96$ or if $z_{computed} > +1.96$. Since $-2.67 < -1.96$, reject H_0.

14. C The wording of the proposition is a little tricky, but the test structure is H_0: $\sigma^2 = 300^2$ versus H_a: $\sigma^2 \neq 300^2$. The appropriate test is a two-tailed chi-square test. The decision rule is to reject H_0 if the test statistic is outside the range defined by the critical chi-square values at $\alpha/2 = 0.025$ with df = 29. The test statistic is $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{(29)(105{,}625)}{90{,}000} = 34.035$. The critical chi-square values are 16.047 on the left and 45.722 on the right. Since the χ^2 falls between these two values, we fail to reject the null hypothesis. This means the population standard deviation is not significantly different than \$300.

15. A A two-tailed t-test is appropriate. The decision rule is to reject H_0 if the t-statistic is outside the range defined by $\pm t$ at $\alpha = 0.01$ with df = 24. The t-statistic = $t_{24} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} = \dfrac{50 - 48}{5/\sqrt{25}} = 2.0$. $\pm t_{24}$ at $\alpha = 0.01 = \pm 2.797$; therefore, H_0 cannot be rejected.

16. A The chi-square test is used to test hypotheses concerning a single population variance. Since this is a one-tailed test, the decision rule is to reject H_0 if $\chi^2 >$ the critical chi-square value at $\alpha = 0.05$ with df = 24.

$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(24)(25)}{24} = 25.0$. The right-tail critical chi-square value is 36.415. Since $\chi^2 = 25 \leq 36.415$, H_0 cannot be rejected.

17. B A one-tailed *t*-test is appropriate. The decision rule is to reject H_0 if the computed *t*-statistic > *t*-critical at $\alpha = 0.05$ with df = 24. The computed value of the *t*-statistic = $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{50-48}{5/\sqrt{25}} = 2.0$, and *t*-critical = $t_{24} = 1.711$. Since $t >$ *t*-critical, H_0 should be rejected.

18. A The *F*-test is appropriate to the equality of population variances. The decision rule is to reject H_0 if the computed test statistic, *F*, exceeds the critical *F*-value at $\alpha/2$. For the information provided, $F = s_1^2/s_2^2 = 36/25 = 1.44$. At a 0.025 level of significance with $d_1 = 35$ and $d_2 = 24$, F-critical = 2.18. Since F < F-critical (1.44 < 2.18), we fail to reject the null hypothesis.

Professor's Note: Many F-tables do not contain numerator df of 35. On the exam, CFA Institute will design problems such that the df are contained directly in the tables that you will be given on the exam. If the tables do not contain the exact df that you need, pick the df that is closest to what you need.

19. D The power of a test is 1 – P(Type II error) = 1 – 0.15 = 0.85.

20. D There is no consistent relationship between the mean and standard deviation of the chi-square distribution or *F*-distribution.

21. B The *F*-test is the appropriate test.

22. C A test of $\sigma^2 = \sigma_0^2$ is a χ^2 test.

23. A Since the observations are likely dependent (both related to market returns), a paired comparisons (mean differences) test is appropriate and is based on a *t*-statistic.

24. D When the variances are assumed to be unequal, we just calculate the denominator (standard error) differently and use both sample variances to calculate the *t*-statistic. The distribution of sample means from a normally distributed population will be at least approximately normal.

25. D A Type I error is rejecting the null hypothesis when it's true. The probability of rejecting a false null is [1 –Prob Type II] = [1 – 0.60] = 40%, which is called the power of the test. A and B are not necessarily true, since the null may be false and the probability of rejection unknown.

ANSWERS – COMPREHENSIVE PROBLEMS: HYPOTHESIS TESTING

1. A. The null hypothesis is typically the one the researcher wants to disprove. In this case, that would be that the mean risk-adjusted return on the cover stocks is less than or equal to the mean risk-adjusted return on the control stocks. The alternative is that the mean risk-adjusted returns on the cover stocks is greater than the mean risk-adjusted return on the control stocks. Rejecting the null will offer statistical support for the proposition the researcher wants to "prove" (the alternative).

   B. This would be a one-tailed test since the alternative is "greater than."

C. The steps in a hypothesis test are:
- State the hypothesis.
- Select the appropriate test statistic.
- Decide on the appropriate level of significance.
- Determine the decision rule.
- Collect the data.
- Calculate the sample statistics.
- Make a decision based on the decision rule for the test.
- Make decisions or inferences based on the results.

2. A. Since these two returns likely exhibit significant correlation and are therefore not independent, a paired comparisons test is appropriate. Differences between the returns on the two indices each week will be used. The standard deviation of the differences is used to construct a t-test of the hypothesis that the mean weekly difference is significantly different from (not equal to) zero. Reject if the t-statistic is greater/less than the positive/negative critical value.

B. This is a test of a difference in means and is a t-test. The test statistic is the difference in means over the square root of the sum of the variances of the two samples divided by their respective sample sizes:

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

Reject if the t-statistic is greater/less than the positive/negative critical value.

C. This is a test of a difference in means and is a t-test. The test statistic is the difference in means over a standard deviation calculated from the pooled variances of the two samples. Reject if the t-statistic is greater/less than the positive/negative critical value. When the variances are assumed to be equal for a difference in means test, we can use the variance of the pooled samples, and the degrees of freedom are simply $n_1 + n_2 - 2$ (total number of observations in both samples minus two).

D. The test statistic is the ratio of the larger sample variance to the smaller sample variance. This statistic follows an F-distribution with n_1-1 and n_2 -1 degrees of freedom. Reject equality if the test statistic exceeds the upper critical value.

E. The test of whether the population variance is equal to a particular value is done with a test statistic with (n-1) times the sample variance in the numerator and the hypothesized variance (0.00165 here) in the denominator.

$$\frac{(n-1)s^2}{s_0^2}$$

The test statistic follows a Chi-square distribution. Reject the null of a population variance equal to 0.00165 if the test statistic is greater than the upper critical value or less than the lower critical value. The degrees of freedom are simply n - 1.

The following is a review of the Quantitative Methods principles designed to address the learning outcome statements set forth by CFA Institute®. This topic is also covered in:

CORRELATION AND REGRESSION

Study Session 3

EXAM FOCUS

Correlation measures the direction and extent of linear association between two variables. Regression is used to summarize the relationship between a dependent variable and one or more independent variables. In addition to calculating and interpreting a correlation coefficient, you should be able to test for its statistical significance with a *t*-test. You will undoubtedly find regression analysis to be more complex than correlation analysis. Focus on the basic interpretation of the regression equation.

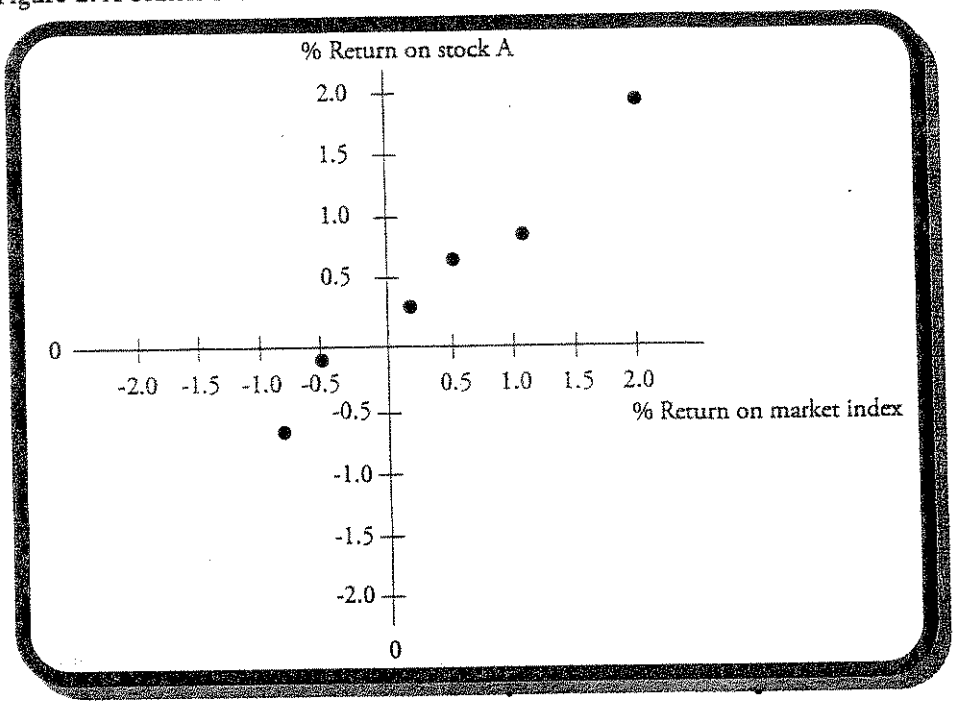
LOS 12.a: Construct and interpret a scatter plot.

A scatter plot is a collection of points on a graph where each point represents the values of two variables (i.e., an X/Y pair). Suppose we wish to graphically represent the data for the returns on Stock A and returns on a market index over the last six months, shown in Figure 1. Figure 2 shows the data graphically with the returns on Stock A shown on the Y-axis and the returns on the market index on the X-axis. Each point of the scatter plot in Figure 2 represents one month of the six in our sample. The rightmost point in the scatter plot is for the month of March, a 2.0% return on the market index and a 1.8% return on Stock A.

Figure 1: Monthly Returns Data

Month	*Return on Stock A*	*Return on Market Index*
Jan	+0.8%	+1.2%
Feb	+0.6%	+0.5%
Mar	+1.8%	+2.0%
Apr	−0.7%	−0.9%
May	+0.3%	+0.2%
June	−0.1%	−0.5%

Figure 2: A Scatter Plot of Returns



LOS 12.b: Calculate and interpret a sample covariance and a sample correlation coefficient.

The **covariance** between two random variables is a statistical measure of the degree to which the two variables move together. The covariance captures the linear relationship between one variable and another. A positive covariance indicates that the variables tend to move together; a negative covariance indicates that the variables tend to move in opposite directions. The **sample covariance** is calculated as:

$$\text{cov}_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

where:

n = sample size

X_i = ith observation on variable X

$\overline{X}$ = mean of the variable X observations

Y_i = ith observation on variable Y

$\overline{Y}$ = mean of the variable Y observations

The actual value of the covariance is not very meaningful because its measurement is extremely sensitive to the scale of the two variables. Also, the covariance may range from negative to positive infinity, and its computation often results in squared units (e.g., percent squared). For these reasons, we calculate the correlation coefficient, which converts the covariance into a measure that is easier to interpret.

Sample Correlation Coefficient

The **correlation coefficient**, r, is a measure of the strength of the linear relationship (correlation) between two variables. The correlation coefficient has no unit of measurement; it is a "pure" measure of the tendency of two variables to move together.

The sample correlation coefficient for two variables, X and Y, is calculated as:

$$r_{XY} = \frac{\text{covariance of X and Y}}{(\text{standard deviation of X})(\text{standard deviation of Y})} = \frac{\text{Cov}(X,Y)}{(\sigma_X)(\sigma_Y)}$$

The correlation coefficient is bounded by positive and negative one (i.e., $-1 \leq r \leq +1$), where a correlation coefficient of +1 indicates that there is a one-for-one movement in the variables. In contrast, if the correlation coefficient is –1, the variables move exactly opposite each other.

The table in Figure 3 provides the data for two variables, X and Y, and shows the calculation of the correlation between X and Y.

Figure 3: Procedure for Computing Correlation

Obs.	X	Y	$X - \bar{X}$	$(X - \bar{X})^2$	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	12	50	–1.50	2.25	8.40	70.56	–12.60
2	13	54	–0.50	0.25	12.40	153.76	–6.20
3	10	48	–3.50	12.25	6.40	40.96	–22.40
4	9	47	–4.50	20.25	5.40	29.16	–24.30
5	20	70	6.50	42.25	28.40	806.56	184.60
6	7	20	–6.50	42.25	–21.60	466.56	140.40
7	4	15	–9.50	90.25	–26.60	707.56	252.70
8	22	40	8.50	72.25	–1.60	2.56	–13.60
9	15	35	1.50	2.25	–6.60	43.56	–9.90
10	23	37	9.50	90.25	–4.60	21.16	–43.70
Sum	135	416	0.00	374.50	0.00	2,342.40	445.00
$\bar{X} = 135 / 10 = 13.5$					$s_X^2 = 374.5 / 9 = 41.611$		
$\bar{Y} = 416 / 10 = 41.6$					$s_Y^2 = 2,342.4 / 9 = 260.267$		

Using the information in Figure 3, the sample correlation coefficient for variables X and Y may be calculated as:

$$r_X = \frac{\frac{445}{9}}{\sqrt{41.611}\sqrt{260.267}} = \frac{49.444}{(6.451)(16.133)} = 0.475$$

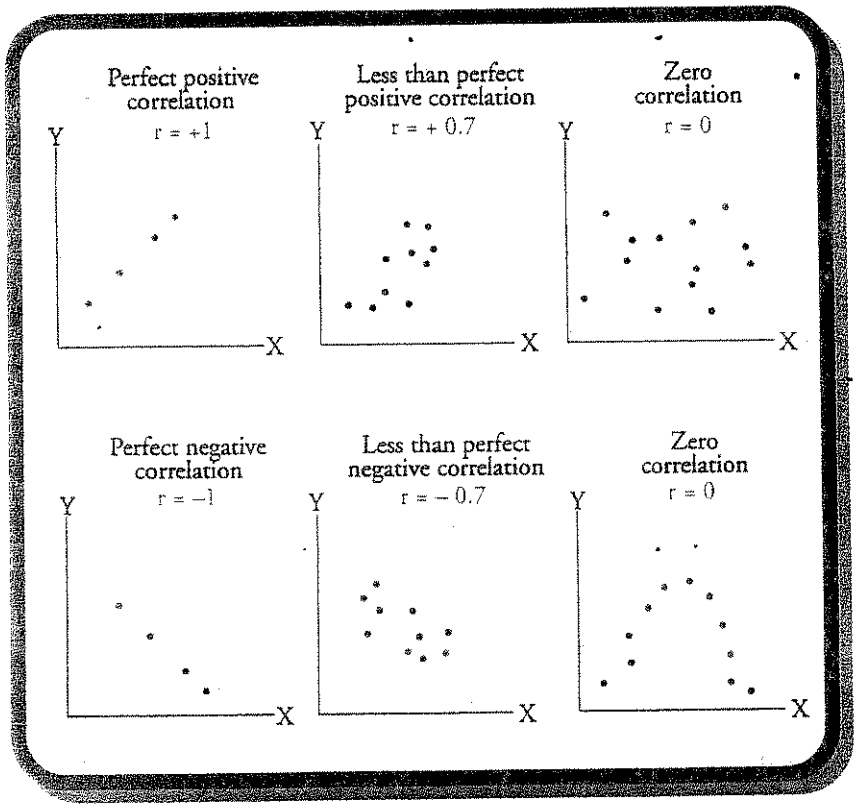
The interpretation of the possible correlation values is summarized in Figure 4.

Figure 4: Interpretation of Correlation Coefficients

Correlation Coefficient (r)	*Interpretation*
$r = +1$	perfect positive correlation
$0 < r < +1$	a positive linear relationship
$r = 0$	no linear relationship
$-1 < r < 0$	a negative linear relationship
$r = -1$	perfect negative correlation

Figure 5 shows several scatter plots for the two random variables X and Y and the corresponding interpretation of correlation. As shown, an upward-sweeping scatter plot indicates a positive correlation between the two variables, while a downward sweeping plot implies a negative correlation. Also illustrated in Figure 5 is that as we move from left to right in the rows of scatter plots, the extent of the linear relationship between the two variables deteriorates, and the correlation gets closer to zero.

Figure 5: Interpretations of Correlation



LOS 12.c: 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.

As indicated earlier, the closer the correlation coefficient is to plus or minus one, the stronger the correlation. With the exception of these extremes (i.e., $r = \pm 1.0$), we cannot really speak of the strength of the relationship indicated by the correlation coefficient without a statistical test of significance. Thus, a hypothesis test is in order.

For our purposes, we want *to test whether the correlation between the population of two variables is equal to zero.* Using the lower case Greek letter rho (ρ) to represent the population parameter, the appropriate null and alternative hypotheses can be structured as a two-tailed test as follows:

$H_0: \rho = 0$ versus $H_a: \rho \neq 0$

Assuming that the two populations are normally distributed, we can use a *t*-test to determine whether the null hypothesis should be rejected. The test statistic is computed using the sample correlation, *r*, with $n - 2$ degrees of freedom (*df*):

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

To make a decision, the calculated *t*-statistic is compared with the critical *t*-value for the appropriate degrees of freedom and level of significance. Bearing in mind that we are conducting a two-tailed test, the decision rule can be stated as:

Reject H_0 if $+t_{critical} < t$, or $t < -t_{critical}$

Example: Test of significance for the correlation coefficient

Using the information from the table in Figure 3, **determine** if the sample correlation is significant at the 5% level of significance.

Answer:

For the sample data in Figure 3, $n = 10$ and $r = 0.475$. Using this information, the test statistic can be computed as:

$$t = \frac{0.475\sqrt{8}}{\sqrt{1-0.475^2}} = \frac{1.3435}{0.88} = 1.5267$$

The two-tailed critical *t*-values at a 5% level of significance with $df = 8$ ($n - 2$) are found in the *t*-table to be ±2.306. (Look in the df = 8 row and match that with the p = 0.05 two-tailed probability column or the p = 0.025 one-tailed probability column.)

Since $-2.306 \leq 1.5267 \leq 2.306$ (i.e., $-t_{critical} \leq t \leq +t_{critical}$), the null cannot be rejected. We conclude that the correlation between variables *X* and *Y* is not significantly different than zero at a 5% significance level.

Example: Test of significance for the correlation coefficient

Suppose the sample correlation between variables *X* and *Y* is 0.2, and the number of sample observations is 32. Using a 5% level of significance, **determine** if this correlation is significantly different from zero.

Answer:

The hypotheses are structured as H_0: $\rho = 0$ versus H_a: $\rho \neq 0$.

The calculated *t*-statistic is $t = \frac{0.2\sqrt{32-2}}{\sqrt{1-0.04}} = \frac{0.2\sqrt{30}}{\sqrt{0.96}} = 1.11803$.

The critical *t*-values at a 5% level of significance with 30 degrees of freedom (32 – 2 = 30) are ±2.042.

Since $-2.042 < 1.11803 < 2.042$ (i.e., $-t_{critical} < t < +t_{critical}$), the null cannot be rejected. We conclude that the correlation between variables *X* and *Y* is not significantly different than zero at a 5% level of significance.

Example: Test of significance for the correlation coefficient

Determine if the correlation of two random variables is significantly different than zero at a 1% level of significance. Assume that a sample correlation has been determined to be 0.80 using a sample with 12 observations.

Answer:

The hypotheses are structured as H_0: $\rho = 0$ versus H_a: $\rho \neq 0$.

The calculated *t*-statistic is $t = \frac{0.80\sqrt{12-2}}{\sqrt{1-0.64}} = \frac{0.80\sqrt{10}}{\sqrt{0.36}} = \frac{2.529822}{0.6} = 4.21637$.

The critical *t*-values at a 1% level of significance with 10 degrees of freedom (12 – 2) are ±3.169.

Since $4.21637 > 3.169$ (i.e., $t > t_{critical}$), the null hypothesis is rejected, and we conclude that there is a significant correlation between the two variables.

LOS 12.d: Differentiate between the dependent and independent variables in a linear regression and explain the assumptions underlying linear regression.

Regression analysis may be used to summarize and explain the nature of the relationship between one variable (a dependent variable) in terms of one or more other variables (independent variables). In this topic review, we will learn how to apply regression techniques to the analysis of the linear relationship between one dependent variable and only one independent variable. This type of application of regression analysis is often referred to as simple linear regression.

The overall purpose of linear regression is to explain the variation in a dependent variable in terms of the variation in the independent variables. Here the term "variation" is interpreted as the degree to which a variable differs from its mean value. Don't confuse variation with variance; they are related but are not the same.

As you progress through the remainder of this review, pay close attention to the following two issues:

- Understanding and interpreting a regression model.
- Using an estimated regression equation to predict a value for a dependent variable.

Professor's Note: Linear Regression is an important topic and is very likely to appear on the exam.

The **dependent variable** is the variable whose variation is explained by the other variable(s). The dependent variable is also referred to as the *explained variable*, the *endogenous variable*, or the *predicted variable*.

The **independent variable** is the variable whose variation is used to explain the variation of the dependent variable. The independent variable is also referred to as the *explanatory variable*, the *exogenous variable*, or the *predicting variable*.

Example: Dependent vs. independent variables

Suppose that you want to predict stock returns with GDP growth. Which variable is the independent variable?

Answer:

Since GDP is going to be used as a predictor of stock returns, stock returns are being *explained* by GDP. Hence, stock returns are the dependent (explained) variable, and GDP is the independent (explanatory) variable.

Assumptions Underlying Linear Regression

Linear regression makes a number of assumptions. Fortunately, the validity of the model is fairly insensitive to minor violations of these assumptions. Most of the major assumptions pertain to the regression model's error term, ε, which is commonly called the residual term, or residual. Memorize the following list.

- A *linear relationship* exists between the dependent and independent variables.
- The *independent variable is uncorrelated with the error term.*
- The expected value of the *error term is zero* ($E(\varepsilon_i) = 0$).
- There is a *constant variance* of the error term (ε_i). In other words, the error terms are homoskedastic. (A violation of this is referred to as *heteroskedasticity*.)
- The *error term is independently distributed*; that is, the error term for one observation is not correlated with that of another observation. (A violation of this is referred to as *autocorrelation*.)
- The *error term is normally distributed.*

LOS 12.e: Interpret a regression coefficient.

The Slope and the Intercept Terms in a Regression

The following linear regression model is used to describe the relationship between two variables, X and Y:

$$Y_i = b_0 + b_1X_i + \varepsilon_i$$

where:

Y_i = ith observation of the dependent variable Y
X_i = ith observation of the independent variable X
b_0 = **intercept with the Y-axis**
b_1 = **slope coefficient**
ε_i = the residual for the ith observation (also referred to as the disturbance term or error term)

The linear regression model says that the value of the dependent variable, Y, is equal to the intercept, b_0, plus the product of the slope coefficient, b_1, and the value of the independent variable, X, plus an error term, ε.

Based on the regression model stated previously, the regression process estimates an equation for a line through a scatter plot of the data that best explains the observed values for *Y* in terms of the observed values for *X*. The linear equation, often called the line of best fit or regression line, takes the following form:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

where:

$\hat{Y}_i$ = the estimated value of Y_i given X_i

$\hat{b}_0$ = the estimated intercept term

$\hat{b}_1$ = the estimated slope coefficient

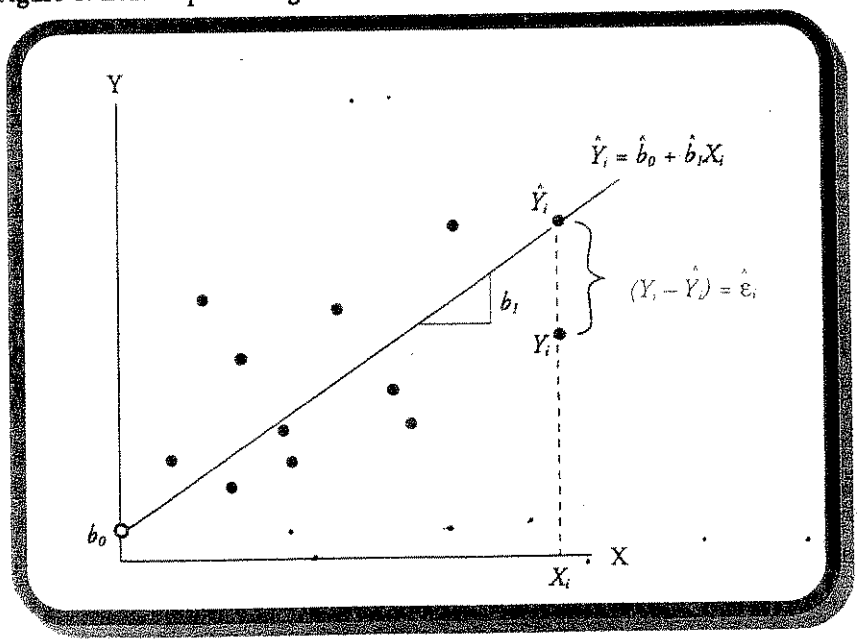
Professor's Note: The hat "^" above a variable or parameter indicates an estimated value.

The regression line is just one of the many possible lines that can be drawn through the scatter plot of X and Y. In fact, the criteria used to estimate this line forms the very essence of linear regression. The regression line is chosen so that the sum of the squared differences (vertical distances) between the Y-values predicted by the regression equation ($\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$) and actual Y-values, Y_i, is minimized. The sum of the squared vertical distances between the estimated and actual Y-values is referred to as the sum of the squared errors, or *residuals*, which is expressed as:

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2 = \sum_{i=1}^{n}\left(\hat{\varepsilon}_i\right)^2 = \text{sum of squared residuals}$$

Thus, the regression line is the line that minimizes the sum of squared residuals. This explains why simple linear regression is frequently referred to as *ordinary least squares* (OLS) regression, and the values estimated by the estimated regression equation, $\hat{Y}_i$, are called least squares estimates. Figure 6 illustrates the concept behind the OLS regression method and shows the estimate of Y, $\hat{Y}_i$, for a specific value of X, X_i.

Figure 6: Least Squares Regression Line



The estimated *slope coefficient* ($\hat{b}_1$) for the regression line describes the change in *Y* for a one-unit change in *X*. It can be positive, negative, or zero, depending on the relationship between the regression variables. The slope term is calculated as:

$$\hat{b}_1 = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

The *intercept* term ($\hat{b}_0$) is the line's intersection with the *Y*-axis at X = 0. It can be positive, negative, or zero.

LOS 12.f: Define, calculate, and interpret the standard error of estimate and the coefficient of determination.

The **standard error of estimate** (SEE) measures the uncertainty about the accuracy of the predicted values of the dependent variable, $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$. In some regressions, the linear relationship between the independent and dependent variables is very strong [e.g., the relationship between 10-year Treasury bond (T-bond) yields and mortgage rates]. In other cases, the relationship is much weaker (e.g., the relationship between stock returns and inflation). SEE will be low (relative to total variability) if the relationship is very strong and high if the relationship is weak.

Formally, SEE is the standard deviation of the differences between the predicted values for the dependent variable (the regression line) and the actual values of the dependent variable. Equivalently, it is the standard deviation of the error terms (residuals) in the regression. As such, SEE is also referred to as the standard error of the residual, or standard error of the regression.

The SEE is easy to calculate. Recall that regression minimizes the sum of the squared vertical distances between the predicted value and actual value for each observation (i.e., prediction errors). Also recall that the sum of the squared prediction errors, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$, is called the sum of squared errors, SSE (not to be confused with SEE). This is the sum of the squared vertical distances between the actual points in the scatter plot and the estimated regression line (i.e., how "spread out" the actual points are around the regression line). Thus, the SEE gives us information about how confident we can be about predictions of Y based on a forecast value of X. If the relationship between the variables in the regression is very close to linear (actual values are close to the line), the prediction errors and the sum of squared residuals will be small. Thus, as shown in the following equations, the standard error of the estimate (SEE) is a function of the sum of squared residuals:

$$\text{SEE} = \sqrt{\frac{\text{sum of squared residuals}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{b}_0 - \hat{b}_i X_i\right)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}{n-2}}$$

where:

$\hat{Y}_i = \hat{b}_0 + \hat{b}_i X_i$ = a point on the regression line corresponding to a value of X_i—it is the expected (predicted) value of *Y*, given the estimated relation between *X* and *Y* and a particular value of *X* (the independent variable)

Similar to the standard deviation for a single variable, SEE measures the degree of variability of the actual *Y*-values relative to the estimated *Y*-values. The SEE gauges the "fit" of the regression line. *The smaller the standard error of the estimate, the better the fit.*

The Coefficient Of Determination

The **coefficient of determination** (R^2) is formally defined as the percentage of the total variation in the dependent variable explained by the independent variable. For example, an R^2 of 0.63 indicates that the variation of the independent variable explains 63% of the variation in the dependent variable.

For simple linear regression (i.e., one independent variable), the coefficient of determination may be computed by simply squaring the correlation coefficient, *r*. In other words, $R^2 = r^2$ for regressions with one independent variable. Unfortunately, this approach is not appropriate when more than one independent variable is used in the regression, as is the case with the multiple regression techniques presented at Level 2.

We now look at a method for measuring the coefficient of determination that may be used for regressions with any number of independent variables. We can calculate R^2 as the proportion of the total variation in the dependent variable (Y) that is explained by the variation in the independent variable (X).

The total variation in the dependent variable (around its mean) can be calculated as:

$$\text{total variation} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

The *total* variation of the dependent variable is the sum of the squared differences between the actual *Y*-values and $\overline{Y}$, the mean of *Y*. SST stands for sum of the squared total variations.

The unexplained variation in the independent variable is:

$$\text{unexplained variation} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The unexplained variation is simply the sum of the squared errors we have been discussing, and we will abbreviate it as SSE (don't confuse this with SEE). It is the sum of the squared vertical distances between the actual *Y*-values, Y_i, and the predicted *Y*-values, $\hat{Y}_i$, on the regression line.

The difference between SST and SSE must be the explained (by the regression) variation, which is:

$$\text{explained variation} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

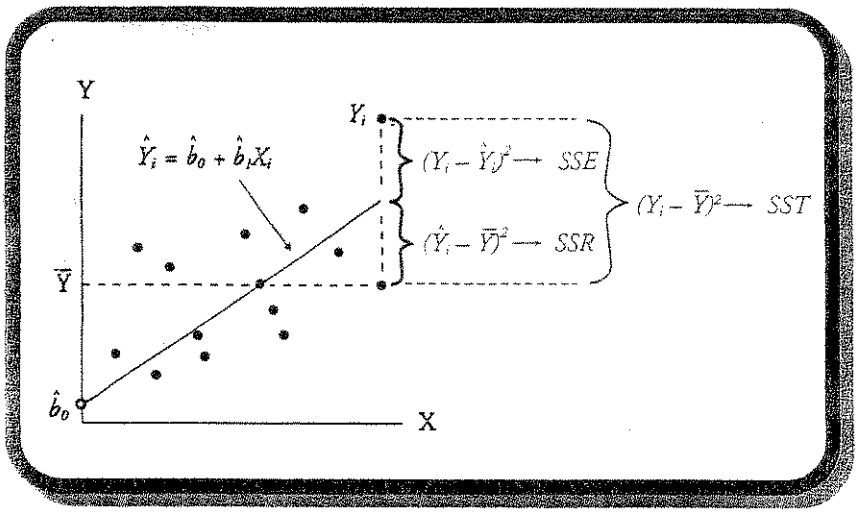
The explained variation is the sum of the squared distances between the predicted *Y*-values and the mean of *Y*. The explained variation is commonly referred to as the sum of the squares regression (SSR).

Thus, total variation = unexplained variation + explained variation, or:

SST = SSE + SSR

Figure 7 illustrates the total variation in the dependent variable.

Figure 7: Components of the Total Variation



The coefficient of determination can be expressed as:

$$R^2 = \frac{\text{total variation} - \text{unexplained variation}}{\text{total variation}} = \frac{\text{explained variation}}{\text{total variation}}$$

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

Example: Coefficient of determination

Given the data in Figure 8, calculate the coefficient of determination.

Figure 8: Components of the Coefficient of Determination R^2

Observation	X_i	Y_i	SST $\left(Y_i-\overline{Y}\right)^2$	$\hat{Y}_i$	$Y_i-\hat{Y}_i$	SSR $\left(\hat{Y}_i-\overline{Y}\right)^2$	SSE $\left(Y_i-\hat{Y}\right)^2$
1	12	50	70.56	39.82	10.18	3.17	103.68
2	13	54	153.76	41.01	12.99	0.35	168.85
3	10	48	40.96	37.44	10.56	17.30	111.49
4	9	47	29.16	36.25	10.75	28.59	115.50
5	20	70	806.56	49.32	20.68	59.65	427.51
6	7	20	466.56	33.88	–13.88	59.65	192.55
7	4	15	707.56	30.31	–15.31	127.43	234.45
8	22	40	2.56	51.70	–11.70	102.01	136.89
9	15	35	43.56	43.38	–8.38	3.18	70.26
10	23	37	21.16	52.89	–15.89	127.43	252.44
Total	135	416	2,342.40	416.00	0.00	528.76	1,813.63

Answer:

$$R^2 = \frac{SSR}{SST} = \frac{528.76}{2,342.40} = 0.2257, \text{or } 22.57\%$$

Alternatively, we can compute R^2 as:

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{1,813.63}{2,342.40} = 0.2257, \text{ or } 22.57\%$$

An R^2 of 0.2257 indicates that the variation of X explains 22.57% of the variation in Y.

KEY CONCEPTS

1. Covariance, COV(X, Y), measures the linear relationship between two random variables and is calculated as:

$$\frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

2. Sample correlation is a measure of the relationship between two variables: $r_{x,y} = \frac{COV(X,Y)}{(\sigma_X)(\sigma_Y)}$, which takes on values from −1.0 to +1.0.
3. A t-test is used to determine if a correlation coefficient, r, is statistically significant:

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Significance is supported if the test statistic is less than $-t_{critical}$ or greater than $t_{critical}$ with $n - 2$ degrees of freedom.
4. Linear regression provides an estimate of the linear relationship between an independent variable (the explanatory variable) and a dependent variable (the predicted variable).
5. The general form of the linear regression model is $Y_i = b_0 + b_1X_i + \varepsilon_i$.
   - Y_i and X_i are the ith observation of the dependent and independent variable, respectively.
   - b_0 = intercept = the value of Y if X is zero.
   - b_1 = slope coefficient = the change in Y for a one-unit change in X.
   - ε_i = residual error for the ith observation.
6. Assumptions made with simple linear regression include:
   - The dependent variable, Y, and independent variable, X, are linearly related.
   - The independent variable is uncorrelated with the error term.
   - The expected value of the error term is zero [$E(\varepsilon_i) = 0$].
   - The variance of the error term is constant for all observations (i.e., $\sigma^2_{\varepsilon 1} = \sigma^2_{\varepsilon 2} = \sigma^2_{\varepsilon 3}$).
   - The error terms are independent.
   - The error term is normally distributed.
7. The regression line, $\hat{Y}_i = \hat{b}_0 + \hat{b}_1X_i$, is the line through the data that minimizes the sum of the squared vertical distances between Y_i and $\hat{Y}_i$ (i.e., minimize $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, or SSE).
8. The standard error of the estimate in a simple linear regression is calculated as:

$$\sqrt{\frac{SSE}{n-2}}$$

9. The coefficient of determination, R^2, is the proportion of the total variation of the dependent variable explained by the regression:

$$R^2 = \frac{SSR}{SST} = \left[1 - \left(\frac{SSE}{SST}\right)\right] = \frac{SST - SSE}{SST}$$

CONCEPT CHECKERS: CORRELATION AND REGRESSION

Use the following data to answer Questions 1 through 3.

An analyst is given the data in Table 1 for the annual sales for Company XYZ, a maker of paper products, and paper product industry sales. The analyst has been asked to develop a model to aid in the prediction of future sales levels for Company XYZ.

Table 1: Regression Output

Parameters	*Coefficient*	*Standard Error of the Coefficient*
Intercept	–94.88	32.97
Slope (industry sales)	0.2796	0.0363

The correlation between company and industry sales is 0.9757. The regression was based on five observations.

1. Which of the following reports the **CORRECT** value and interpretation of the R^2 for this regression? The R^2 is:
   A. 0.048, indicating that the variability of industry sales explains about 4.8% of the variability of company sales.
   B. 0.048, indicating that the variability of company sales explains about 4.8% of the variability of industry sales.
   C. 0.952, indicating that the variability of industry sales explains about 95.2% of the variability of company sales.
   D. 0.952, indicating that the variability of company sales explains about 95.2% of the variability of industry sales.

2. Based on the regression results, XYZ Company's market share of any increase in industry sales is expected to be *closest* to:
   A. 4%.
   B. 28%.
   C. 45%.
   D. 94%.

3. Which of the following **CORRECTLY** represents the *t*-statistic and the statistical significance of the correlation coefficient with 95% confidence?

	t-Statistic	Correlation Coefficient
A.	7.71	not significantly different from zero
B.	7.71	significantly different from zero
C.	60.93	not significantly different from zero
D.	60.93	significantly different from zero

Use the following information to answer Questions 4 and 5.

A study was conducted by the British Department of Transportation to estimate urban travel time between locations in London, England. Data was collected for motorcycles and passenger cars. Simple linear regression was conducted using data sets for both types of vehicles, where Y = urban travel time in minutes and X = distance between locations in kilometers. The following results were obtained:

Regression Results for Travel Times Between Distances in London		
Passenger cars:	$\hat{Y} = 1.85 + 3.86X$	$R^2 = 0.758$
Motorcycles:	$\hat{Y} = 2.50 + 1.93X$	$R^2 = 0.676$

4. Which of the following is the *closest* estimate of the increased travel time for a motorcycle commuter planning to move 8 km further from his workplace in London?
   A. 31 minutes.
   B. 15 minutes.
   C. 5 minutes.
   D. 0.154 hours.

5. Based on the regression results, which model is *more* reliable?
   A. The passenger car model because 3.86 > 1.93.
   B. The motorcycle model because 1.93 < 3.86.
   C. The passenger car model because 0.758 > 0.676.
   D. The motorcycle model because $\sqrt{0.676} < \sqrt{0.758}$.

6. Which of the following is NOT an assumption of simple linear regression analysis?
   A. The residuals are normally distributed.
   B. There is a constant variance of the error term.
   C. The independent variable is uncorrelated with the residuals.
   D. The dependent variable is uncorrelated with the residuals.

COMPREHENSIVE PROBLEMS: CORRELATION AND REGRESSION

1. Use the data in the table below to:
   A. calculate the mean and the sample variance (standard deviation) of X and Y.
   B. calculate the sample covariance and correlation coefficient for X and Y.
   C. calculate the R^2 for a regression of Y on X.
   D. interpret the R^2 calculated in part C in terms of the explanatory power of the regression model.
   E. determine if there is a significant correlation between the independent and dependent variables.

Obs.	X	Y
1	0.4	20
2	0.3	23
3	0.42	18
4	0.34	31
5	0.35	33
6	0.19	19

ANSWERS – CONCEPT CHECKERS: CORRELATION AND REGRESSION

1. C The R^2 is computed as the correlation squared: $(0.9757)^2 = 0.952$.

   The interpretation of this R^2 is that 95.2% of the variation in Company XYZ's sales is explained by the variation in industry sales. Answer D is incorrect because it is the independent variable (industry sales) that explains the variation in the dependent variable (company sales). This interpretation is based on the economic reasoning used in constructing the regression model.

2. B The slope coefficient of 0.2796 indicates that a $1 million increase in industry sales will result in an increase in firm sales of approximately 28% ($279,600) of that amount.

3. B The test of significance for the correlation coefficient is evaluated using the following *t*-statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9757\sqrt{3}}{\sqrt{1-0.952}} = \frac{1.69}{0.219} = 7.72$$

   From the *t*-table, we find that with df = 3 and 95% significance, the two-tailed critical *t* values are ±3.182 (recall that for the *t*-test the degrees of freedom = n – 2). Since the computed *t* is greater than +3.182, the correlation coefficient is significantly different from zero.

4. B The slope coefficient is 1.93, indicating that each additional kilometer increases travel time by 1.93 minutes:

   $1.93 \times 8 = 15.44$

5. C The higher R^2 for the passenger car model indicates that regression results are more reliable. Distance is a better predictor of travel time for cars. Perhaps the aggressiveness of the driver is a bigger factor in travel time for motorcycles than it is for autos.

6. D The model does not assume that the dependent variable is uncorrelated with the residuals.

Comprehensive Problems: Correlation and Regression

1. A. $\frac{\sum X_i}{n} = \frac{2}{6} = 0.33 = \bar{X}$

$$\frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{0.033933}{5} = 0.006787 = \sigma_x^2$$

$$\sigma_x = 0.08238$$

$$\frac{\sum Y_i}{n} = \frac{144}{6} = 24 = \bar{Y}$$

$$\frac{\sum (Y_i - \bar{Y})^2}{n-1} = \frac{208}{5} = 41.6 = \sigma_y^2$$

$$\sigma_y = 6.4498$$

B. $$\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{0.16}{5} = 0.032 = COV_{x,y}$$

$$CORR_{x,y} = \frac{COV_{x,y}}{\sigma_x \sigma_y} = \frac{0.032}{6.4498(0.08238)} = 0.0602$$

C. In a simple regression, R^2 is just $(CORR_{x,y})^2 = 0.0602^2 = 0.003624$.

D. The independent variable only explains approximately 0.36% of the variation in the dependent variable.

E. The *t*-stat for a test of whether the correlation coefficient is equal to zero is:

$$\frac{0.0602\sqrt{6-2}}{\sqrt{1-0.0602^2}} = 0.1206$$

This is not significant given 4 df.

FORMULAS

nominal risk-free rate = real risk-free rate + expected inflation rate

required interest rate on a security = nominal risk-free rate
+ default risk premium
+ liquidity premium
+ maturity risk premium

EAR = (1 + periodic rate)^m − 1

continuous compounding: $e^r - 1 = EAR$

$$PV_{perpetuity} = \frac{PMT}{I/Y}$$

$FV = PV(1 + I/Y)^N$

$$NPV = \sum_{t=0}^{N} \frac{CF_t}{(1+r)^t}$$

general formula for the IRR: $0 = CF_0 + \frac{CF_1}{1+IRR} + \frac{CF_2}{(1+IRR)^2} + \cdots + \frac{CF_N}{(1+IRR)^N}$

bank discount yield: $r_{BD} = \frac{D}{F} \times \frac{360}{t}$

$$HPY = \frac{P_1 - P_0 + D_1}{P_0} = \frac{P_1 + D_1}{P_0} - 1$$

$EAY = (1 + HPY)^{365/t} - 1$

money market yield: $r_{MM} = HPY\left(\frac{360}{t}\right)$

position of the observation at a given percentile, y: $L_y = (n+1)\frac{y}{100}$

population mean: $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$