UNIT-2 STATISTICAL SAMPLING

Population:

A population consists of all items of interest for a particular decision or investigation—for example, all individuals in the United States who do not own cell phones, all subscribers to Netflix, or all stockholders of Google. A company like Netflix keeps extensive records on its customers, making it easy to retrieve data about the entire population of customers. However, it would probably be impossible to identify all individuals who do not own cell phones.

Sample:

A sample is a subset of a population. For example, a list of individuals who rented a comedy from Netflix in the past year would be a sample from the population of all customers. Whether this sample is representative of the population of customers—which depends on how the sample data are intended to be used—may be debatable.

Measures of Location:

Measures of location provide estimates of a single value that in some fashion represents the "centering" of a set of data. The most common is the average. We all use averages routinely in our lives, for example, to measure student accomplishment in college (e.g., grade point average), to measure the performance of sports teams (e.g., batting average), and to measure performance in business (e.g., average delivery time).

Arithmetic Mean:

The average is formally called the arithmetic mean (or simply the mean), which is the sum of the observations divided by the number of observations. Mathematically, the mean of a population is denoted by the Greek letter m, and the mean of a sample is denoted by x. If a population consists of N observations x1, x2, c, xN, the population mean, m, is calculated as

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The mean of a sample of n observations, $x_1, x_2, \ldots, x_n$, denoted by $\bar{x}$, is calculated as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11

Figure 4.1

Portion of *Purchase Orders* Database

Figure 4.2

Excel Calculations of Mean Cost per Order

	A	B
1	Observation	Cost per order
2	x1	$2,700.00
3	x2	$19,250.00
4	x3	$15,937.50
5	x4	$18,150.00
93	x92	$74,375.00
94	x93	$72,250.00
95	x94	$6,562.50
96	Sum of cost/order	$2,471,760.00
97	Number of observations	94
98		
99	Mean cost/order	$26,295.32
100		
101	Excel AVERAGE function	$26,295.32

Median:

The measure of location that specifies the middle value when the data are arranged from least to greatest is the median. Half the data are below the median, and half the data are above it. For an odd number of observations, the median is the middle of the sorted numbers. For an even number of observations, the median is the mean of the two middle numbers. We could use the Sort option in Excel to rank-order the data and then determine the median. The Excel function MEDIAN(data range) could also be used. The median is meaningful for ratio, interval, and ordinal data. As opposed to the mean, the median is not affected by outliers.

Figure 4.3

Excel Calculations for Median Cost per Order

	A	B	C	D
1	Rank	Cost per order		
2	1	$68.75		
3	2	$82.50		
4	3	$375.00		
5	4	$467.50		
45	44	$14,910.00		
46	45	$14,910.00		
47	46	$15,087.50		
48	47	$15,562.50		$15,562.50
49	48	$15,750.00		$15,750.00
50	49	$15,937.50	Average	$15,656.25
51	50	$16,276.75		
52	51	$16,330.00		

Mode:

A third measure of location is the mode. The mode is the observation that occurs most frequently. The mode is most useful for data sets that contain a relatively small number of unique values. For data sets that have few repeating values, the mode does not provide much practical value. You can easily identify the mode from a frequency distribution by identifying the value having the largest frequency or from a histogram by identifying the highest bar. You may also use the Excel function MODE.SNGL(data range). For frequency distributions and histograms of grouped data, the mode is the group with the greatest frequency.

Some data sets have multiple modes; to identify these, you can use the Excel function MODE.MULT(data range), which returns an array of modal values.

Measures of Dispersion

Dispersion refers to the degree of variation in the data, that is, the numerical spread (or compactness) of the data. Several statistical measures characterize dispersion: the range, variance, and standard deviation.

Range

The **range** is the simplest and is the difference between the maximum value and the minimum value in the data set. Although Excel does not provide a function for the range, it can be computed easily by the formula = MAX(data range) - MIN(data range). Like the midrange, the range is affected by outliers and, thus, is often only used for very small data sets.

Interquartile Range

The difference between the first and third quartiles, Q3 - Q1, is often called the **interquartile range (IQR)**, or the **midspread**. This includes only the middle 50% of the data and, therefore, is not influenced by extreme values. Thus, it is sometimes used as an alternative measure of dispersion.

Variance

A more commonly used measure of dispersion is the **variance**, whose computation depends on all the data. The larger the variance, the more the data are spread out from the mean and the more variability one can expect in the observations. The formula used for calculating the variance is different for populations and samples. The formula for the variance of a population is

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Standard Deviation

The **standard deviation** is the square root of the variance. For a population, the standard deviation is computed as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

and for samples, it is

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

Mrs. M. Tulasi
Asst. Professor

The Excel function STDEV.P(data range) calculates the standard deviation for a population 1s2; the function STDEV.S(data range) calculates it for a sample (s).

Standardized Values

A **standardized value**, commonly called a **z-score**, provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement. The z-score for the ith observation in a data set is calculated as follows:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Coefficient of Variation

The **coefficient of variation (CV)** provides a relative measure of the dispersion in data relative to the mean and is defined as

$$\text{CV} = \frac{\text{standard deviation}}{\text{mean}}$$

Sampling Methods

Many types of sampling methods exist. Sampling methods can be subjective or probabilistic. Subjective methods include **judgment sampling**, in which expert judgment is used to select the sample (survey the "best" customers), and **convenience sampling**, in which samples are selected based on the ease with which the data can be collected (survey all customers who happen to visit this month). Probabilistic sampling involves selecting the items in the sample using some random procedure. Probabilistic sampling is necessary to draw valid statistical conclusions.

The most common probabilistic sampling approach is simple random sampling. **Simple random sampling** involves selecting items from a population so that every subset of a given size has an equal chance of being selected. If the population data are stored in a database, simple random samples can generally be easily obtained.

Figure 6.1

Excel *Sampling* Tool Dialog

Systematic (Periodic) Sampling. Systematic, **or periodic**, **sampling** is a sampling plan (one of the options in the Excel Sampling tool) that selects every nth item from the population. For example, to sample 250 names from a list of 400,000, the first name could be selected at random from the first 1,600, and then every 1,600th name could be selected. This approach can be used for telephone sampling when supported by an automatic dialer that is programmed to dial numbers in a systematic manner.

Stratified Sampling. Stratified sampling applies to populations that are divided into natural subsets (called strata) and allocates the appropriate proportion of samples to each stratum. For example, a large city may be divided into political districts called wards. Each ward has a different number of citizens. A stratified sample would choose a sample of individuals in each ward proportionate to its size.

Cluster Sampling. Cluster sampling is based on dividing a population into subgroups (clusters), sampling a set of clusters, and (usually) conducting a complete census within the clusters sampled. For instance, a company might segment its customers into small geographical regions. A cluster sample would consist of a random sample of the geographical regions, and all customers within these regions would be surveyed (which might be easier because regional lists might be easier to produce and mail).

• **Sampling from a Continuous Process.** Selecting a sample from a continuous manufacturing process can be accomplished in two main ways. First, select a time at random; then select the next n items produced after that time. Second, select n times at random; then select the next item produced after each of these times.

Estimating Population Parameters

Sample data provide the basis for many useful analyses to support decision making. **Estimation** involves assessing the value of an unknown population parameter—such as a population mean, population proportion, or population variance—using sample data. **Estimators** are the measures used to estimate population parameters; for example, we use the sample mean x to estimate a population mean m. The sample variance s2 estimates a population variance s2, and the sample proportion p estimates a population proportion p. A **point estimate** is a single number derived from sample data that is used to estimate the value of a population parameter.

Unbiased Estimators

It seems quite intuitive that the sample mean should provide a good point estimate for the population mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

whereas the sample variance is computed by the formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sampling Error:

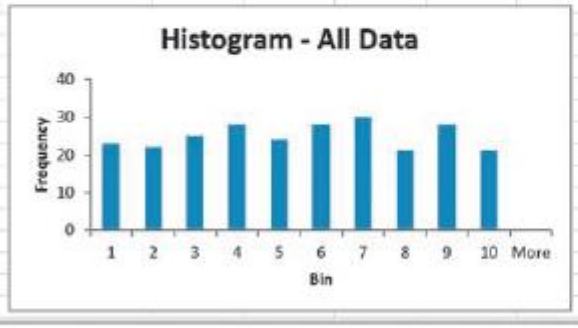
Sampling (statistical) error occurs because samples are only a subset of the total population. Sampling error is inherent in any sampling process, and although it can be minimized, it cannot be totally avoided. Another type of error, called **non sampling error**, occurs when the sample does not represent the target population adequately. This is generally a result of poor sample design, such as using a convenience sample when a simple random sample would have been more appropriate or choosing the wrong population frame. Sampling error depends on the size of the sample relative to the population. Thus, determining the number of samples to take is essentially a statistical issue that is based on the accuracy of the estimates needed to draw a useful conclusion. We discuss this later in this chapter.

However, from a practical standpoint, one must also consider the cost of sampling and sometimes make a trade-off between cost and the information that is obtained.

Understanding Sampling Error

Suppose that we estimate the mean of a population using the sample mean. How can we determine how accurate we are? In other words, can we make an informed statement about how far the sample mean might be from the true population mean? We could gain some insight into this question by performing a sampling experiment.

	A	B	C	D	E	F	W	X	Y	Z	AA	AB
1	Observation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 22	Sample 23	Sample 24	Sample 25		
2	1	5.3935	0.8756	9.9338	4.3294	7.1908	8.2244	8.4655	1.9404	9.9133		
3	2	2.8282	2.5047	6.4480	5.9877	7.3946	3.3000	0.3632	9.8871	5.1079		
4	3	5.2715	0.6949	1.5015	8.3935	3.1559	7.1023	9.3628	1.7844	7.3937		
5	4	5.4912	0.7739	8.1466	5.5205	2.4586	8.7262	9.1598	7.5820	1.8513		
6	5	9.3158	4.4591	6.3573	3.8679	1.1493	3.3854	1.2482	1.9391	5.4405		
7	6	7.9745	6.9784	7.9982	2.3157	7.8564	8.9032	3.8716	5.8525	5.4184		
8	7	6.7043	8.4039	5.1088	9.1098	1.1802	2.7732	2.4815	9.0817	4.3889		
9	8	1.3041	2.5678	6.1794	7.8396	6.2709	0.5692	2.5800	1.1911	7.2430		
10	9	0.9870	6.3964	8.2269	9.6112	6.6814	2.8306	4.6004	9.0274	6.1232		
11	10	9.9493	9.3936	4.5015	5.2385	0.6970	3.7074	5.9062	0.6592	7.5021		
12	Sample mean	5.5219	4.3048	6.4400	6.2214	4.4035	4.9522	4.8039	4.8945	6.0380	Average	5.0108
13											Standard Dev.	0.816673



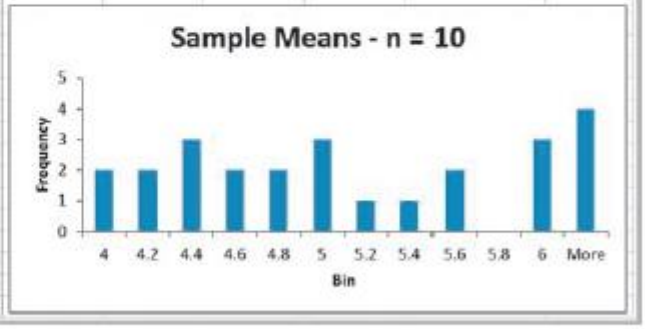


Figure 6.3

Portion of Spreadsheet for *Sampling Experiment*

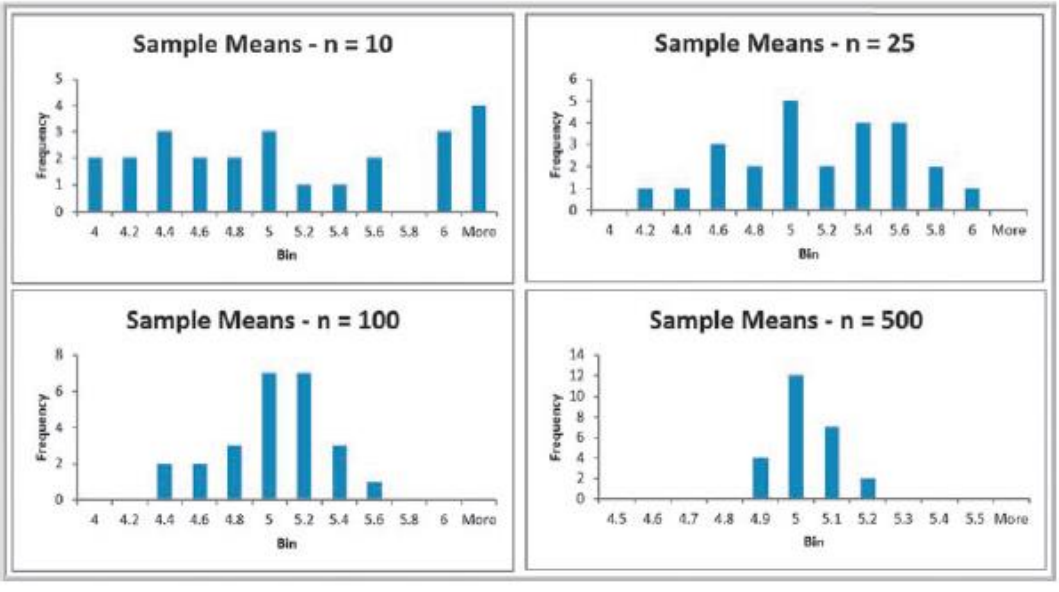
Table 6.1

Results from *Sampling Experiment*

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993

Figure 6.4

Histograms of Sample Means for Increasing Sample Sizes



Sampling Distribution of the Mean

The means of all possible samples of a fixed size n from some population will form a distribution that we call the **sampling distribution of the mean**. The histograms in Figure 6.4 are approximations to the sampling distributions of the mean based on 25 samples. Statisticians have shown two key results about the sampling distribution of the mean. First, the standard deviation of the sampling distribution of the mean, called the **standard error of the mean**, is computed as

$$\text{Standard Error of the Mean} = \sigma/\sqrt{n}$$

where s is the standard deviation of the population from which the individual observations are drawn and n is the sample size.

EXAMPLE 6.5 Computing the Standard Error of the Mean

For our experiment, we know that the variance of the population is 8.33 (because the values were uniformly distributed). Therefore, the standard deviation of the population is $\sigma = 2.89$. We may compute the standard error of the mean for each of the sample sizes in our experiment using formula (6.1). For example, with $n = 10$, we have

$$\text{Standard Error of the Mean} = \sigma/\sqrt{n} = 2.89/\sqrt{10} = 0.914$$

For the remaining data in Table 6.1 we have the following:

Sample Size, *n*	Standard Error of the Mean
10	0.914
25	0.577
100	0.289
500	0.129

Interval Estimate:

An **interval estimate** provides a range for a population characteristic based on a sample. Intervals are quite useful in statistics because they provide more information than a point estimate. Intervals specify a range of plausible values for the characteristic of interest and a way of assessing "how plausible" they

are. In general, a 10011 - a2% **probability interval** is any interval [A, B] such that the probability of falling between A and B is 1 - a. Probability intervals are often centered on the mean or median. For instance, in a normal distribution, the mean plus or minus 1 standard deviation describes an approximate 68% probability interval around the mean. As another example, the 5th and 95th percentiles in a data set constitute a 90% probability interval.

Confidence Interval:

Confidence interval estimates provide a way of assessing the accuracy of a point estimate. A **confidence interval** is a range of values between which the value of the population parameter is believed to be, along with a probability that the interval correctly estimates the true (unknown) population parameter. This probability is called the **level of confidence**, denoted by 1 - a, where a is a number between 0 and 1. The level of confidence is usually expressed as a percent; common values are 90%, 95%, or 99%. (Note that if the level of confidence is 90%, then a = 0.1.) The margin of error depends on the level of confidence and the sample size.

Using confidence intervals for decision making, prediction intervals:

Another type of interval used in estimation is a prediction interval. A **prediction interval** is one that provides a range for predicting the value of a new observation from the same population. This is different from a confidence interval, which provides an interval estimate of a population parameter, such as the mean or proportion. A confidence interval is associated with the sampling distribution of a statistic, but a prediction interval is associated with the distribution of the random variable itself.

When the population standard deviation is unknown, a 10011 - a2% prediction interval for a new observation is

$$\bar{x} \pm t_{\alpha/2,n-1}\left(s\sqrt{1+\frac{1}{n}}\right)$$

Hypothesis testing:

Hypothesis testing involves drawing inferences about two contrasting propositions (each called a **hypothesis**) relating to the value of one or more population parameters, such as the mean, proportion, standard deviation, or variance. One of these propositions (called the **null hypothesis**) describes the existing theory or a belief that is accepted as valid unless strong statistical evidence exists to the contrary. The second proposition (called the **alternative hypothesis**) is the complement of the null hypothesis; it must be true if the null hypothesis is false. The null hypothesis is denoted by H0, and the alternative hypothesis is denoted by H1. Using sample data, we either

1. Reject the null hypothesis and conclude that the sample data provide sufficient statistical evidence to support the alternative hypothesis, or

2. Fail to reject the null hypothesis and conclude that the sample data does not support the alternative hypothesis. If we fail to reject the null hypothesis, then we can only accept as valid the existing theory Or belief, but we can never prove it.

A **one-sample hypothesis test**

A **one-sample hypothesis test** is one that involves a single population parameter, such as the mean, proportion, standard deviation, and so on. To conduct the test, we use a single sample of data from the population. We may conduct three types of one-sample hypothesis tests:

H0: population parameter Ú constant vs. H1: population parameter 6 constant

H0: population parameter … constant vs. H1: population parameter 7 constant
H0: population parameter = constant vs. H1: population parameter ≠ constant
Notice that one-sample tests always compare a population parameter to some constant. For one-sample tests, the statements of the null hypotheses are expressed as either Ú, …, or =. It is not correct to formulate a null hypothesis using 7, 6, or ≠.
How do we determine the proper form of the null and alternative hypotheses?
Hypothesis testing always assumes that H0 is true and uses sample data to determine whether H1 is more likely to be true. Statistically, we cannot "prove" that H0 is true; we can only fail to reject it. Thus, if we cannot reject the null hypothesis, we have shown only that there is insufficient evidence to conclude that the alternative hypothesis is true. However, rejecting the null hypothesis provides strong evidence (in a statistical sense) that the null hypothesis is not true and that the alternative hypothesis is true.

Understanding Potential Errors in Hypothesis Testing

We already know that sample data can show considerable variation; therefore, conclusions based on sample data may be wrong. Hypothesis testing can result in one of four different outcomes:

1. The null hypothesis is actually true, and the test correctly fails to reject it.

2. The null hypothesis is actually false, and the hypothesis test correctly reaches this conclusion.

3. The null hypothesis is actually true, but the hypothesis test incorrectly rejects it (called **Type I error**).

4. The null hypothesis is actually false, but the hypothesis test incorrectly fails to reject it (called **Type II error**).

The probability of making a Type I error, that is, P(rejecting H0 ·H0 is true), is denoted by a and is called the **level of significance**. This defines the likelihood that you are willing to take in making the incorrect conclusion that the alternative hypothesis is true when, in fact, the null hypothesis is true. The value of a can be controlled by the decision maker and is selected before the test is conducted. Commonly used levels for a are 0.10, 0.05, and 0.01.

The probability of correctly failing to reject the null hypothesis, or P(not rejecting H0 ·H0 is true), is called the **confidence coefficient** and is calculated as 1 - a. For a confidence coefficient of 0.95, we mean that we expect 95 out of 100 samples to support the null hypothesis rather than the alternate hypothesis when H0 is actually true.

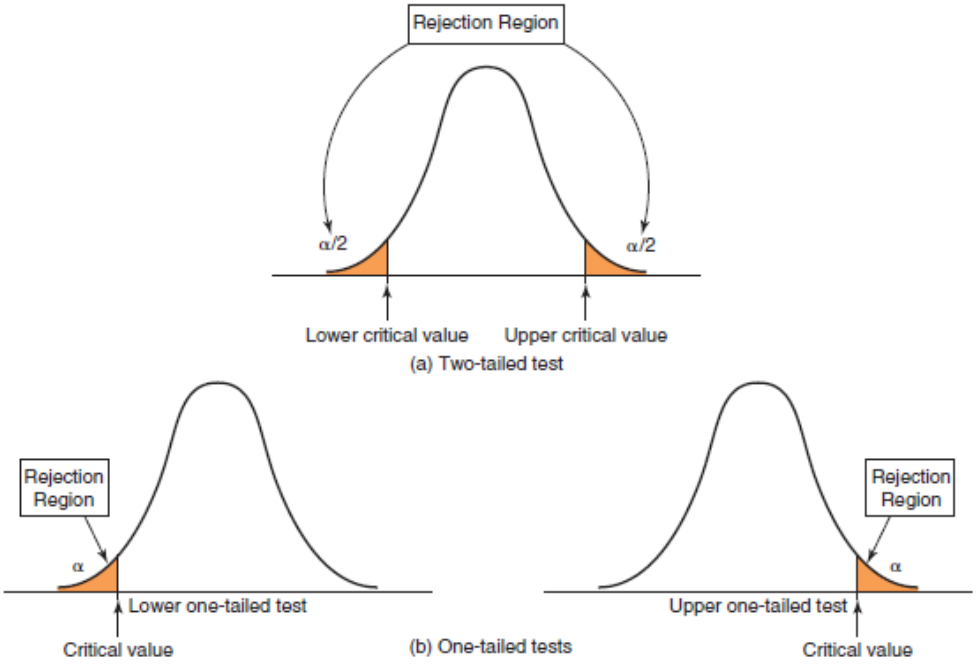
Unfortunately, we cannot control the probability of a Type II error, P(not rejecting H0 ·H0 is false), which is denoted by b. Unlike a, b cannot be specified Selecting the Test Statistic The next step is to collect sample data and use the data to draw a conclusion. The decision to reject or fail to reject a null hypothesis is based on computing a test statistic from the sample data. The test statistic used depends on the type of hypothesis test.

Type of Test	Test Statistic
One-sample test for mean, σ known	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
One-sample test for mean, σ unknown	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Mrs. M. Tulasi
Asst. Professor

Figure 7.2

Illustration of Rejection Regions in Hypothesis Testing



Two-Tailed Test of Hypothesis for the Mean

Basically, all hypothesis tests are similar; you just have to ensure that you select the correct test statistic, critical value, and rejection region, depending on the type of hypothesis. The following example illustrates a two-tailed test of hypothesis for the mean.

EXAMPLE 7.6 Conducting a Two-Tailed Hypothesis Test for the Mean

Figure 7.4 shows a portion of data collected in a survey of 34 respondents by a travel agency (provided in the Excel file *Vacation Survey*). Suppose that the travel agency wanted to target individuals who were approximately 35 years old. Thus, we wish to test whether the average age of respondents is equal to 35. The hypothesis to test is

$$H_0: \text{mean age} = 35$$

$$H_1: \text{mean age} \neq 35$$

The sample mean is computed to be 38.677, and the sample standard deviation is 7.858.

We use the *t*-test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{38.677 - 35}{7.858/\sqrt{34}} = 2.73$$

In this case, the sample mean is 2.73 standard errors above the hypothesized mean of 35. However, because this is a two-tailed test, the rejection region and decision rule are different. For a level of significance α, we reject H_0 if the *t*-test statistic falls either below the negative critical value, $-t_{\alpha/2,n-1}$, or above the positive critical value, $t_{\alpha/2,n-1}$. Using either Table A.2 in Appendix A at the back of this book or the Excel function T.INV.2T(.05,33) to calculate $t_{0.025,33}$, we obtain 2.0345. Thus, the critical values are ±2.0345. Because the *t*-test statistic does *not* fall between these values, we must reject the null hypothesis that the average age is 35 (see Figure 7.5).

p-Values

An alternative approach to comparing a test statistic to a critical value in hypothesis testing is to find the probability of obtaining a test statistic value equal to or more extreme than that obtained from the sample data when the null hypothesis is true.

Two-Sample Hypothesis Tests

Many practical applications of hypothesis testing involve comparing two populations for differences in means, proportions, or other population parameters. Such tests can confirm differences between suppliers, performance at two different factory locations, new and old work methods or reward and recognition programs, and many other situations. Similar to one-sample tests, two-sample hypothesis tests for differences in population parameters have one of the following forms:

1. Lower-tailed test H0: population parameter (1) - population parameter (2) Ú D0 vs. H1: population parameter (1) - population parameter (2) 6 D0. This test seeks evidence that the difference between population parameter (1) and population parameter (2) is less than some value, D0. When D0 = 0, the test simply seeks to conclude whether population parameter (1) is smaller than population parameter (2).

2. Upper-tailed test H0: population parameter (1) - population parameter (2) … D0 vs. H1: population parameter (1) - population parameter (2) 7 D0. This test seeks evidence that the difference between population parameter (1) and population parameter (2) is greater than some value, D0. When D0 = 0, the test simply seeks to conclude whether population parameter (1) is larger than population parameter (2).

3. Two-tailed test H0: population parameter (1) - population parameter (2) = D0 vs. H1: population parameter (1) - population parameter (2) ≠ D0. This test seeks evidence that the difference between the population parameters is equal to D0. When D0 = 0, we are seeking evidence that population parameter (1) differs from parameter (2).

Type of Test	Excel Procedure
Two-sample test for means, σ^2 known	Excel *z*-test: Two-sample for means
Two-sample test for means, σ^2 unknown, assumed unequal	Excel *t*-test: Two-sample assuming unequal variances
Two-sample test for means, σ^2 unknown, assumed equal	Excel *t*-test: Two-sample assuming equal variances
Paired two-sample test for means	Excel *t*-test: Paired two-sample for means
Two-sample test for equality of variances	Excel *F*-test Two-sample for variances

Selection of the proper test statistic and Excel procedure for a two-sample test for means depends on whether the population standard deviations are known, and if not, whether they are assumed to be equal.

1. Population variance is known. In Excel, choose z-Test: Two-Sample for Means from the Data Analysis menu. This test uses a test statistic that is based on the standard normal distribution.

2. Population variance is unknown and assumed unequal. From the Data Analysis menu, choose t-test: Two-Sample Assuming Unequal Variances. The test statistic for this case has a t-distribution.

3 Population variance unknown but assumed equal. In Excel, choose t-test: Two- Sample Assuming Equal Variances. The test statistic also has a t-distribution, but it is different from the unequal variance case.

Mrs. M. Tulasi
Asst. Professor

Analysis of Variance (ANOVA)

To this point, we have discussed hypothesis tests that compare a population parameter to a constant value or that compare the means of two different populations. Often, we would like to compare the means of several different groups to determine if all are equal or if any are significantly different from the rest.

Assumptions of ANOVA

ANOVA requires assumptions that the m groups or factor levels being studied represent populations whose outcome measures

1. are randomly and independently obtained,

2. are normally distributed, and

3. have equal variances.

Chi-Square Test for Independence

A common problem in business is to determine whether two categorical variables are independent. We introduced the concept of independent events in Chapter 5. In the energy drink survey example (Example 5.9), we used conditional probabilities to determine whether brand preference was independent of gender. However, with sample data, sampling error can make it difficult to properly assess the independence of categorical variables. We would never expect the joint probabilities to be exactly the same as the product of the marginal probabilities because of sampling error even if the two variables are statistically independent. Testing for independence is important in marketing applications. We can test for independence by using a hypothesis test called the chi-square test for independence. The chi-square test for independence tests the following hypotheses:

H0: the two categorical variables are independent

H1: the two categorical variables are dependent

The chi-square test is an example of a nonparametric test; that is, one that does not depend on restrictive statistical assumptions, as ANOVA does. This makes it a widely applicable and popular tool for understanding relationships among categorical data. The first step in the procedure is to compute the expected frequency in each cell of the cross tabulation if the two variables are independent. This is easily done using the following:

$$\text{expected frequency in row } i \text{ and column } j = \frac{(\text{grand total row } i)(\text{grand total column } j)}{\text{total number of observations}}$$

Figure 7.15

Portion of *Energy Drink Survey* and Cross-Tabulation

	A	B	C	D	E	F	G	H	I
1	Energy Drink Survey								
2									
3	Respondent	Gender	Brand Preference						
4	1	Male	Brand 3		Count of Respondent	Column Labels			
5	2	Female	Brand 3		Row Labels	Brand 1	Brand 2	Brand 3	Grand Total
6	3	Male	Brand 3		Female	9	6	22	37
7	4	Male	Brand 1		Male	25	17	21	63
8	5	Male	Brand 1		Grand Total	34	23	43	100
9	6	Female	Brand 2						
10	7	Male	Brand 2						

Figure 7.16

Expected Frequencies for the Chi-Square Test

	E	F	G	H	I	J	K
1	Chi-Square Test						
2							
3	Count of Respondent	Column Labels					
4	Row Labels	Brand 1	Brand 2	Brand 3	Grand Total		
5	Female	9	6	22	37		
6	Male	25	17	21	63		
7	Grand Total	34	23	43	100		
8							
9							
10	Expected Frequency	Brand 1	Brand 2	Brand 3	Grand Total		
11	Female	12.58	8.51	15.91	37		
12	Male	21.42	14.49	27.09	63		
13	Grand Total	34	23	43	100		

Expected frequency of Female and Brand 1 = 37*34/100

Next, we compute a test statistic, called a **chi-square statistic**, which is the sum of the squares of the differences between observed frequency, fo, and expected frequency, fe, divided by the expected frequency in each cell:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

The closer the observed frequencies are to the expected frequencies, the smaller will be the value of the chi-square statistic. The sampling distribution of x2 is a special distribution called the **chi-square** 1x22 **distributions**. The chi-square distribution is characterized by degrees of freedom, similar to the t-distribution.