Chapter 6
Inferential Statistical Analysis of Data

Contents

H. Guerrero, *Excel Data Analysis*, https://doi.org/10.1007/978-3-030-01279-3_6

6.1 Introduction

We introduced several statistical techniques for the analysis of data in Chap. 3, most of which were descriptive or exploratory. But, we also got our first glimpse of another form of statistical analysis known as *Inferential Statistics*. Inferential statistics is how statisticians use inductive reasoning to move from the specific, the data contained in a sample, to the general, inferring characteristics of the population from which the sample was taken.

Many problems require an understanding of population characteristics; yet, it can be difficult to determine these characteristics because populations can be very large and difficult to access. So rather than throw our hands into the air and proclaim that this is an *impossible* task, we resort to a **sample**: a small slice or view of a population. It is not a perfect solution, but we live in an imperfect world and we must make the best of it. Mathematician and popular writer John Allen Paulos sums it up quite nicely—"Uncertainty is the only certainty there is, and knowing how to live with insecurity is the only security."

So, what sort of imperfection do we face? Sample data can result in measurements that are not representative of the population from which they are taken, so there is always uncertainty as to how well the sample represents the population. We refer to these circumstances as **sampling error**: the difference between the measurement results of a sample and the true measurement values of a population. Fortunately, through carefully designed sampling methods and the subsequent application of statistical techniques, statisticians *can* infer population characteristics from results found in a sample. If performed correctly, the sampling design will provide a measure of reliability about the population inference we will make.

Let us carefully consider why we rely on inferential statistics:

1. The size of a population often makes it impossible to measure characteristics for every member of the population—often there are just too many members of populations. Inferential statistics provides an alternative solution to this problem.
2. Even if it is possible to measure characteristics for the population, the cost can be prohibitive. Accessing measures for every member of a population can be costly. We call this a census.
3. Statisticians have developed techniques that can quantify the uncertainty associated with sample data. Thus, although we know that samples are not perfect, inferential statistics provides a reliability evaluation of how well a sample measure represents a population measure.

This was precisely what we were attempting to do in the survey data on the four webpage designs in Chap. 5; that is, to make population inferences from the webpage preferences found in the sample. In the descriptive analysis we presented a numerical result. With inferential statistics we will make a statistical statement about our confidence that the sample data is representative of the population. For the numerical outcome, we *hoped* that the sample did in fact represent the population, but it was mere hope. With inferential statistics, we will develop techniques that allow us to *quantify* a sample's ability to reflect a population's characteristics, and this will all be done within Excel. We will introduce some often used and important inferential statistics techniques in this chapter.

6.2 Let the Statistical Technique Fit the Data

Consider the type of sample data we have seen thus far in Chaps. 1–5. In just about every case, the data has contained a combination of quantitative and qualitative data elements. For example, the data for teens visiting websites in Chap. 3 provided the number of page—views for each teen, and described the circumstances related to the page-views, either *new* or *old* site. This was our first exposure to sophisticated statistics and to **cause and effect** analysis-one variable causing an effect on another. We can think of these categories, new and old, as experimental **treatments**, and the page-views as a **response variable**. Thus, the treatment is the assumed cause and the effect is the number of views. To determine if the sample means of the two treatments were different or equal, we performed an analysis called a **paired t-Test**. This test permitted us to consider complicated questions.

So, when do we need this *more* sophisticated statistical analysis? Some of the answers to this question can be summarized as follows:

1. When we want to make a precise mathematical statement about the data's capability to infer characteristics of the population.
2. When we want to determine how closely these data fit some assumed model of behavior.
3. When we need a higher level of analysis to further investigate the preliminary findings of descriptive and exploratory analysis.

This chapter will focus on data that has both qualitative and quantitative components, but we will also consider data that is strictly qualitative (categorical), as you will soon see. By no means can we explore the exhaustive set of statistical techniques available for these data types; there are thousands of techniques available and more are being developed as we speak. But, we will introduce some of the most often used tools in statistical analysis. Finally, I repeat that it is important to remember that the type of data we are analyzing will dictate the technique that we can employ. The misapplication of a technique on a set of data is the most common reason for dismissing or ignoring the results of an analysis; the analysis just does not match the data.

6.3 $\chi 2$—Chi-Square Test of Independence for Categorical Data

Let us begin with a powerful analytical tool applied to a frequently occurring type of data—categorical variables. In this analysis, a test is conducted on sample data, and the test attempts to determine if there is an association or relationship between two

Table 6.1 Results of mutual fund sample

Fund types frequency					
Investor risk preference	Bond	Income	Income/Growth	Growth	Totals
Risk-taker	30	9	45	66	150
Conservative	270	51	75	54	450
Totals	300	60	120	120	600

categorical (**nominal**) variables. Ultimately, we would like to know if the result can be extended to the entire population or is due simply to chance. For example, consider the relationship between two variables: (1) an investor's self-perceived behavior toward investing, and (2) the selection of mutual funds made by the investor. This test is known as the **Chi-square**, or Chi-squared, **test of independence**. As the name implies, the test addresses the question of whether or not the two categorical variables are independent (not related).

Now, let us consider a specific example. A mutual fund investment company samples a total of 600 potential investors who have indicated their intention to invest in mutual funds. The investors have been asked to classify themselves as either *risk-taking* or *conservative* investors. Then, they are asked to identify a single type of fund they would like to purchase. Four fund types are specified for possible purchase and only one can be selected—*bond*, *income*, *growth*, and *income and growth*. The results of the sample are shown in Table 6.1. This table structure is known as a **contingency table**, and this particular contingency table happens to have 2 rows and 4 columns—what is known as a 2 by 4 contingency table. Contingency tables show the frequency of occurrence of the row and column categories. For example, 30 (first row/first column) of the 150 (*Totals* row for risk-takers) investors in the sample that identified themselves as risk-takers said they would invest in a bond fund, and 51 (second row/second column) investors considering themselves to be conservative said they would invest in an income fund. These values are **counts** or the frequency of observations associated with a particular cell.

6.3.1 Tests of Hypothesis—Null and Alternative

The mutual fund investment company is interested in determining if there is a relationship in an investor's perception of his own risk and the selection of a fund that the investor actually makes. This information could be very useful for marketing funds to clients and also for counseling clients on risk-tailored investments. To make this determination, we perform an analysis of the data contained in the sample. The analysis is structured as a test of the null hypothesis. There is also an alternative to the null hypothesis called, quite appropriately, the alternative hypothesis. As the name implies, a test of hypothesis, either null or alternative, requires that a hypothesis is posited, and then a test is performed to see if the null hypothesis can be: (1) rejected in favor of the alternative, or (2) not rejected.

In this case, our null hypothesis assumes that self-perceived risk preference is **independent** of a mutual fund selection. That suggests that an investor's self-description as an investor is not related to the mutual funds he or she purchases, or more strongly stated, does not *cause* a purchase of a particular type of mutual fund. If our test suggests otherwise, that is, the test leads us to **reject the null hypothesis**, then we conclude that it is likely to be **dependent** (related).

This discussion may seem tedious, but if you do not have a firm understanding of tests of hypothesis, then the remainder of the chapter will be very difficult, if not impossible, to understand. Before we move on to the calculations necessary for performing the test, the following summarizes the general procedure just discussed:

1. an assumption (*null* hypothesis) that the variables under consideration are independent, or that they are *not* related, is made
2. an alternative assumption (*alternative* hypothesis) relative to the null is made that there *is* dependence between variables
3. the chi-square test is performed on the data contained in a contingency table to test the *null* hypothesis
4. the results, a statistical calculation, is used to attempt to reject the null hypothesis
5. if the null *is* rejected, then this implies that the alternative is accepted; if the null is *not* rejected, then the alternative hypothesis is rejected

The chi-square test is based on a null hypothesis that assumes independence of relationships. If we believe the independence assumption, then the *overall* fraction of investors in a perceived risk category and fund type should be *indicative* of the entire investing population. Thus, an *expected* frequency of investors in each cell can be calculated. We will have more to say about this later in the chapter. The expected frequency, assuming independence, is compared to the actual (observed) and the variation of expected to actual is tested by calculating a statistic, the χ^2 **statistic** (χ is the lower case Greek letter chi). The variation between what is actually observed and what is *expected* is based on the formula that follows. Note that the calculation squares the difference between the observed frequency and the expected frequency, divides by the expected value, and then sums across the two dimensions of the *i* by *j* contingency table:

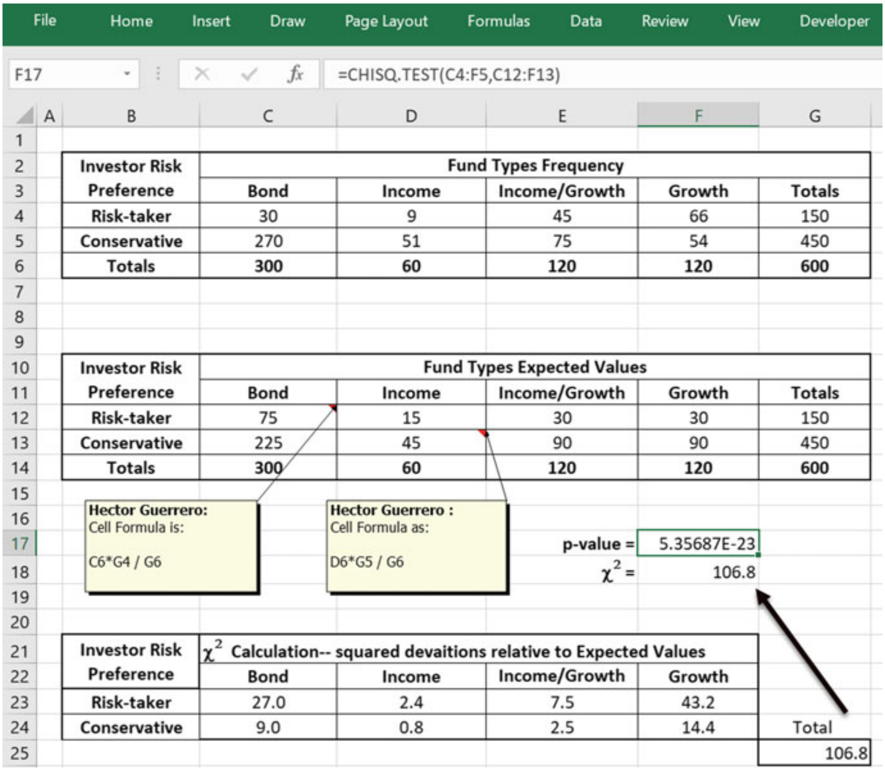
$$\chi^2 = \Sigma_i\ \Sigma j\ \left[\left(obs_{ij} - \exp val_{ij}\right)2/\exp val_{ij}\right]$$

where:

obs_{ij} = frequency or count of observations in the ith row and jth column of the contingency table

$\exp \text{val}_{\text{ij}}$ = expected frequency of observations in the ith row and jth column of the contingency table, when independence of the variables is assumed.[1]

[1] Calculated by multiplying the row total and the column total and dividing by total number of observations—e.g. in Fig. 6.1 expected value for conservative/growth cell is (120 * 450)/600 = 90. Note that 120 is the marginal total Income/Growth and 450 is the marginal total for Conservative.



=CHISQ.TEST(C4:F5,C12:F13)

Investor Risk Preference	Fund Types Frequency				
	Bond	Income	Income/Growth	Growth	Totals
Risk-taker	30	9	45	66	150
Conservative	270	51	75	54	450
Totals	300	60	120	120	600

Investor Risk Preference	Fund Types Expected Values				
	Bond	Income	Income/Growth	Growth	Totals
Risk-taker	75	15	30	30	150
Conservative	225	45	90	90	450
Totals	300	60	120	120	600

Hector Guerrero:
Cell Formula is:

C6*G4 / G6

Hector Guerrero :
Cell Formula as:

D6*G5 / G6

p-value = 5.35687E-23

χ^2 = 106.8

Investor Risk Preference	χ^2 Calculation-- squared devaitions relative to Expected Values				
	Bond	Income	Income/Growth	Growth	
Risk-taker	27.0	2.4	7.5	43.2	
Conservative	9.0	0.8	2.5	14.4	Total
					106.8

Fig. 6.1 Chi-squared calculations via contingency table

Once the χ^2 statistic is calculated, then it can be compared to a benchmark value of χ^2 α that sets a limit, or threshold, for rejecting the null hypothesis. The value of χ^2 α is the limit the χ^2 statistic can achieve before we reject the null hypothesis. These values can be found in most statistics books. To select a particular **χ 2 α**, the **α** (the **level of significance** of the test) must be set by the investigator. It is closely related to the *p-value*—the probability of obtaining a particular statistic value or more extreme by chance, when the null hypothesis is true. Investigators often set α to 0.05; that is, there is a 5% chance of obtaining this χ^2 statistic (or greater) when the null is true. So, our decision-maker only wants a 5% chance of *erroneously* rejecting the null hypothesis. That is relatively conservative, but a more conservative (less chance of erroneously rejecting the null hypothesis) stance would be to set α to 1%, or even less.

Thus, if our χ^2 is greater than or equal to χ^2 α, then we *reject* the null. Alternatively, if the p-value is less than α we *reject* the null. These tests are equivalent. In summary, the rules for rejection are either:

Reject the null hypothesis when $\chi^2 \geq \chi^2$ α
or
Reject the null hypothesis when p-value $\leq$α
(Note that these rules are equivalent)

Figure 6.1 shows a worksheet that performs the test of independence using the chi-square procedure. The figure also shows the typical calculation for contingency table expected values. Of course, in order to perform the analysis, both tables are needed to calculate the χ^2 statistic since both the observed frequency and the expected are used in the calculation. Using the Excel **CHISQ.TEST (actual range, expected range)** cell function permits Excel to calculate the data's χ^2 and then return a p-value (see cell F17 in Fig. 6.1). You can also see from Fig. 6.1 that the *actual range* is C4:F5 and does not include the marginal totals. The *expected range* is C12:F13 and the marginal totals are also omitted. The internally calculated χ^2 value takes into consideration the number of variables for the data, 2 in our case, and the possible levels within each variable—2 for risk preference and 4 for mutual fund types. These variables are derived from the range data information (rows and columns) provided in the *actual* and *expected* tables.

From the spreadsheet analysis in Fig. 6.1 we can see that the calculated χ^2 value in F18 is 106.8 (a relatively large value), and if we assume α to be 0.05, then χ^2 α is approximately 7.82 (from a table in a statistics book). Thus, we can reject the null since $106.8 > 7.82$.[2] Also, the p-value from Fig. 6.1 is extremely small (5.35687E-23)[3] indicating a very small probability of obtaining the χ^2 value of 106.8 when the null hypothesis is true. The p-value returned by the CHISQ.TEST function is shown in cell F17, and it is the only value that is needed to reject, or not reject, the null hypothesis. Note that the cell formula in F18 is the calculation of the χ^2 given in the formula above and is not returned by the CHISQ.TEST function. This result leads us to conclude that the null hypothesis is likely not true, so we reject the notion that the variables are independent. Instead, there appears to be a strong dependence given our test statistic. Earlier, we summarized the general steps in performing a test of hypothesis. Now we describe in detail how to perform the test of hypothesis associated with the χ^2 test. The steps of the process are:

1. Organize the frequency data related to two categorical variables in a contingency table. This shown in Fig. 6.1 in the range B2:G6.
2. From the contingency table values, calculate expected frequencies (see Fig. 6.1 cell comments) under the assumption of independence. The calculation of χ^2 is simple and performed by the *CHISQ.TEST(actual range, expected range)* function. The function returns the p-value of the calculated χ^2. Note that it does not return the χ^2 value, although it does calculate the value for internal use. I have calculated the χ^2 value in cells C23:F24 and their sum in G25 for completeness of calculations, although it is unnecessary to do so.

[2] Tables of χ^2 α can be found in most statistics texts. You will also need to calculate *the degrees of freedom* for the data: (number of rows–1) × (number of columns–1). In our example: (2–1) × (4–1) =3.

[3] Recall this is a form of what is known as "scientific notation". E-17 means 10 raised to the −17 power, or the decimal point moved 17 decimal places to the left of the current position for 3.8749. Positive (E + 13 e.g.) powers of 10 moves the decimal to the right (13 decimal places).

3. By considering an explicit level of α, the decision to reject the null can be made on the basis of determining if $\chi^2 \geq \chi^2$ α. Alternatively, α can be compared to the calculated p-value: p-value ≤α. Both rules are interchangeable and equivalent. It is often the case that an α of 0.05 is used by investigators.

6.4 z-Test and t-Test of Categorical and Interval Data

Now, let us consider a situation that is similar in many respects to the analysis just performed, but it is different in one important way. In the χ^2 test the subjects in our sample were associated with two variables, both of which were categorical. The cells provided a count, or frequency, of the observations that were classified in each cell. Now, we will turn our attention to sample data that contains categorical *and* interval or ratio data. Additionally, the categorical variable is dichotomous, and thereby can take on only two levels. The categorical variable will be referred to as the experimental *treatment*, and the interval data as the *response* variable. In the following section, we consider an example problem related to the training of human resources that considers experimental treatments and response variables.

6.5 An Example

A large firm with 12,000 call center employees in two locations is experiencing explosive growth. One call center is in South Carolina (SC) and the other is in Texas (TX). The firm has done its own *standard*, internal training of employees for 10 years. The CEO is concerned that the quality of call center service is beginning to deteriorate at an alarming rate. They are receiving many more complaints from customers, and when the CEO disguised herself as a customer requesting call center information, she was appalled at the lack of courtesy and the variation of responses to a relatively simple set of questions. She finds this to be totally unacceptable and has begun to consider possible solutions. One of the solutions being considered is a training program to be administered by an outside organization with experience in the development and delivery of call center training. The hope is to create a systematic and predictable customer service response.

A meeting of high level managers is held to discuss the options, and some skepticism is expressed about training programs in general: many ask the question—Is there really any value in these outside programs? Yet, in spite of the skepticism, managers agree that something has to be done about the deteriorating quality of customer service. The CEO contacts a nationally recognized training firm, EB Associates. EB has considerable experience and understands the concerns of management. The CEO expresses her concern and doubts about training. She is not sure that training can be effective, especially for the type of unskilled workers they hire. EB listens carefully and has heard these concerns before. EB proposes a test to determine if the *special* training methods they provide can be of value for the call center workers. After careful discussion with the CEO, EB makes the following suggestion for testing the effectiveness of *special* (EB) versus *standard* (internal) training:

1. A test will be prepared and administered to all the customer service representatives working in the call centers, 4000 in SC and 8000 TX. The test is designed to assess the *current* competency of the customer service representatives. From this overall data, specific groups will be identified and a sample of 36 observations (test scores) for each group will be a taken. This will provide a baseline call center personnel score, *standard* training.
2. Each customer service representative will receive a score from 0 to 100.
3. A *special* training course by EB will be offered to a selected group of customer service representatives in South Carolina: 36 incarcerated women. The competency test will be *re-administered* to this group after training to detect changes in scores, if any.
4. Analysis of the difference in performance between representatives specially trained and those standard trained will be used to consider the application of the training to all employees. If the special training indicates significantly better performance on the exam, then EB will receive a large contract to administer training for all employees.

As mentioned above, the 36 customer service representatives selected to receive special training are a group of women that are incarcerated in a low security prison facility in the state of South Carolina. The CEO has signed an agreement with the state of South Carolina to provide the SC women with an opportunity to work as customer service representatives and gain skills before being released to the general population. In turn, the firm receives significant tax benefits from South Carolina. Because of the relative ease with which these women can be trained, they are chosen for the special training. They are, after all, a captive audience. There is a similar group of customer service representatives that also are incarcerated woman. They are in a similar low security Texas prison, but these women are not chosen for the special training.

The results of the tests for employees are shown in Table 6.2. Note that the data included in each of five columns is a sample of personnel scores of similar sizes (36): (1) non-prisoners in TX, (2) women prisoners in TX, (3) non-prisoners in SC, (4) women prisoners in SC before special training, and (5) women prisoners in SC after special training. All the columns of data, except the last, are scores for customer service representatives that have only had the internal standard training. The last column is the re-administered test scores of the SC prisoners that received special training from EB. Additionally, the last two columns are the same individual subjects, matched as before and after special training, respectively. The sample sizes for the samples need not be the same, but it does simplify the analysis calculations. Also, there are important advantages to samples greater than approximately 30 observations that we will discuss later.

Table 6.2 Special training and no training scores

Observation	36 Non-prisoner scores TX	36 Women prisoners TX	36 Non-prisoner scores SC	36 Women SC (before special training)∗	36 Women SC (with special training)∗
1	81	93	89	83	85
2	67	68	58	75	76
3	79	72	65	84	87
4	83	84	67	90	92
5	64	77	92	66	67
6	68	85	80	68	71
7	64	63	73	72	73
8	90	87	80	96	98
9	80	91	79	84	85
10	85	71	85	91	94
11	69	101	73	75	77
12	61	82	57	62	64
13	86	93	81	89	90
14	81	81	83	86	89
15	70	76	67	72	73
16	79	90	78	82	84
17	73	78	74	78	80
18	81	73	76	84	85
19	68	81	68	73	76
20	87	77	82	89	91
21	70	80	71	77	79
22	61	62	61	64	65
23	78	85	83	85	87
24	76	84	78	80	81
25	80	83	76	82	84
26	70	77	75	76	79
27	87	83	88	90	93
28	72	87	71	74	75
29	71	76	69	71	74
30	80	68	77	80	83
31	82	90	86	88	89
32	72	93	73	76	78
33	68	75	69	70	72
34	90	73	90	91	93
35	72	84	76	78	81
36	60	70	63	66	68
Averages=	75.14	80.36	75.36	79.08	81.06
Variance=	72.12	80.47	78.47	75.11	77.31
Total TX	74.29		Total TX		
Av = (8000 obs.)			VAR=	71.21	

(continued)

Table 6.2 (continued)

Observation	36 Non-prisoner scores TX	36 Women prisoners TX	36 Non-prisoner scores SC	36 Women SC (before special training)∗	36 Women SC (with special training)∗
Total SC Av=	75.72		Total SC		
(4000 obs.)			VAR=	77.32	
Total Av=	74.77		TX&SC		
(12,000 obs.)			VAR=	73.17	

∗Same 36 SC women prisoners that received training

Every customer service representative at the firm was tested at least once, and the SC women prisoners were tested twice. Excel can easily store these sample data and provide access to specific data elements using the filtering and sorting capabilities we learned in Chap. 5. The data collected by EB provides us with an opportunity to thoroughly analyze the effectiveness of the special training.

So, what are the questions of interest and how will we use inferential statistics to answer them? Recall that EB administered special training to 36 women prisoners in SC. We also have a standard trained non-prisoner group from SC. EB's first question might be—Is there any difference between the *average* score of a randomly selected SC non-prisoner sample with no special training and the SC prisoner's *average* score after special training? Note that our focus is on the aggregate statistic of *average* scores for the groups. Additionally, EB's question involves SC data exclusively. This is done to not confound results, should there be a difference between the competency of customer service representatives in TX and SC. We will study the issue of the possible difference between Texas and SC scores later in our analysis.

EB must plan a study of this type very carefully to achieve the analytical goals she has in mind. It will not be easy to return to these customer representatives and re-administer the competency exams.

6.5.1 z-Test: 2 Sample Means

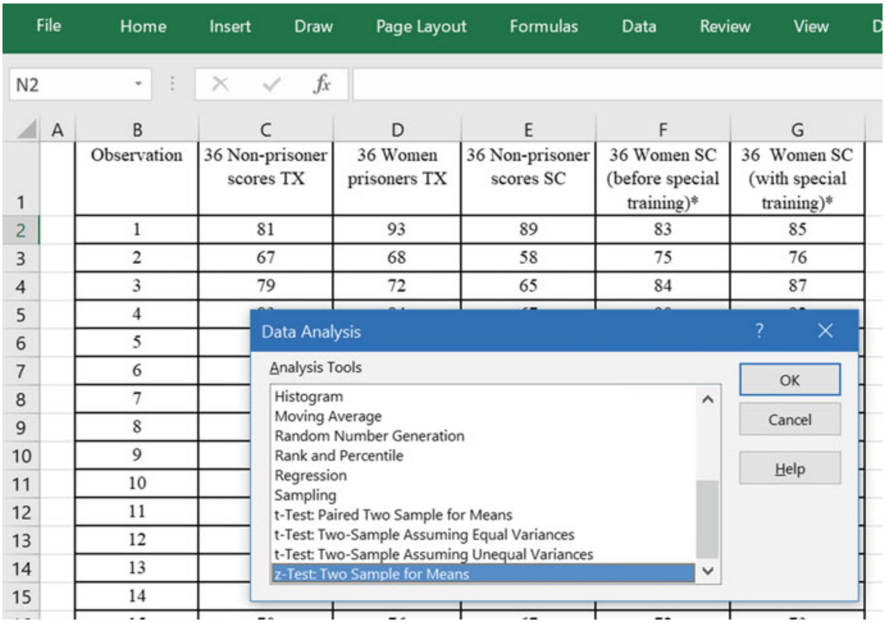
To answer the question of whether or not there is a difference between the average scores of SC non-prisoners *without* special training and prisoners *with* special training, we use the **z-Test: Two Sample for Means** option found in Excel's *Data Analysis* tool. This analysis tests the null hypothesis that there is *no* difference between the two sample means and is generally reserved for samples of 30 observations or more. Pause for a moment to consider this statement. We are focusing on the question of whether two means from sample data are different; different in statistics suggests that the samples come from different underlying populations distributions with different means. For our problem, the question is whether the SC non-prisoner group and the SC prisoner group with special training have different population means for their sample scores. Of course, the process of calculating sample means will very likely lead to different values. If the means are relatively close to one another, then we will conclude that they came from the same population; if the means are relatively different, we are likely to conclude that they are from different populations. Once calculated, the sample means will be examined and a probability estimate will be made as to how likely it is that the two sample means came from the same population. But, the question of importance in these tests of hypothesis is related to the populations—are the averages of the population of SC non-prisoners and of the population of SC prisoners with special training the same, or are they different?

If we reject the null hypothesis that there is no difference in the average scores, then we are deciding in favor of the training indeed leading to a difference in scores. As before, the decision will be made on the basis of a statistic that is calculated from the sample data, in this case the **z-Statistic**, which is then compared to a critical value. The critical value incorporates the decision maker's willingness to commit an error by possibly rejecting a true null hypothesis. Alternatively, we can use the p-value of the test and compare it to the level of significance which we have adopted—as before, frequently assumed to be 0.05. The steps in this procedure are quite similar to the ones we performed in the chi-square analysis, with the exception of the statistic that is calculated, z rather than chi-square.

6.5.2 Is There a Difference in Scores for SC Non-prisoners and EB Trained SC Prisoners?

The procedure for the analysis is shown in Figs. 6.2 and 6.3. Figure 6.2 shows the *Data Analysis* dialogue box in the Analysis group of the Data ribbon used to select the z-Test. We begin data entry for the z-Test in Fig. 6.3 by identifying the range inputs, including labels, for the two samples: 36 SC non-prisoner standard trained scores (E1:E37) and 36 SC prisoners that receive special training (G1:G37). Next, the dialog box requires a hypothesized mean difference. Since we are assuming there is *no* difference in the null hypothesis, the input value is 0. This is usually the case, but you are permitted to designate other differences if you are hypothesizing a specific difference in the sample means. For example, consider the situation in which management is willing to purchase the training, but only if it results in some minimum increase in scores. The desired difference in scores could be tested as the *Hypothesized Mean Difference*.

The variances for the variables can be estimated to be the variances of the samples, if the samples are greater than approximately 30 observations. Recall earlier that I suggested that a sample size of at least 30 was advantageous, *this is why*! We can also use the variance calculated for the entire population at SC (Table 6.2—Total SC VAR =77.32) since it is available, but the difference in the calculated z-statistics is very minor: z-statistic using the sample variance is 2.7375

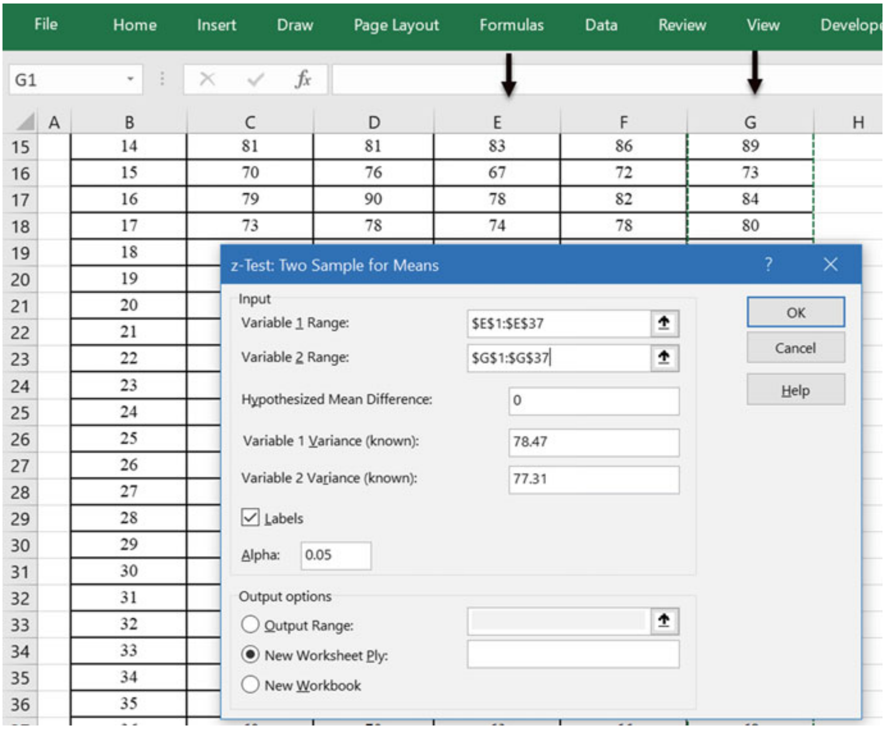


	A	B	C	D	E	F	G
1		Observation	36 Non-prisoner scores TX	36 Women prisoners TX	36 Non-prisoner scores SC	36 Women SC (before special training)*	36 Women SC (with special training)*
2		1	81	93	89	83	85
3		2	67	68	58	75	76
4		3	79	72	65	84	87
5		4					
6		5					
7		6					
8		7					
9		8					
10		9					
11		10					
12		11					
13		12					
14		13					
15		14					

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

Fig. 6.2 Data analysis tool for z-Test



	A	B	C	D	E	F	G	H
15		14	81	81	83	86	89	
16		15	70	76	67	72	73	
17		16	79	90	78	82	84	
18		17	73	78	74	78	80	
19		18						
20		19						
21		20						
22		21						
23		22						
24		23						
25		24						
26		25						
27		26						
28		27						
29		28						
30		29						
31		30						
32		31						
33		32						
34		33						
35		34						
36		35						

Input

Variable 1 Range: $E$1:$E$37

Variable 2 Range: $G$1:$G$37

Hypothesized Mean Difference: 0

Variable 1 Variance (known): 78.47

Variable 2 Variance (known): 77.31

☑ Labels

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Fig. 6.3 Selection of data for z-Test

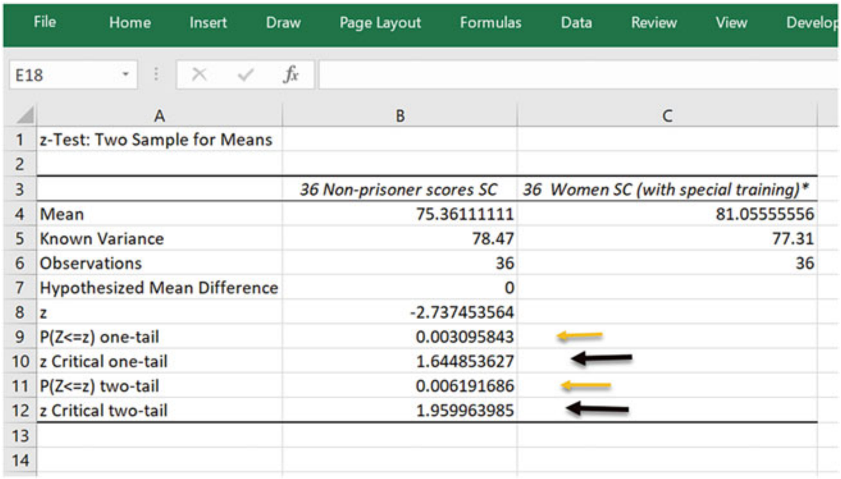
and 2.7475 for the total variance of SC. Next, we choose an α value of 0.05, but you may want to make this smaller if you want to be very cautious about rejecting true null hypotheses. Finally, this test of hypothesis is known as a *two-tail* test since we are not speculating on whether one specific sample mean will be greater than the other mean. We are simply positing a *difference* in the alternative. This is important in the application of a critical z-value for possible rejection of the null hypothesis. In cases where you have prior evidence that one mean is greater than another, then a *one-tail* test is likely appropriate. The critical z-values, **z-Critical one-tail** and **z-Critical two-tail**, and p-values, **P(Z ≤ z) one-tail** and **P(Z ≤ z) two-tail**, are all provided when the analysis is complete. These values represent our test thresholds.

The results of our analysis is shown in Table 6.3. Note that a z-statistic of approximately −2.74 has been calculated. We reject the null hypothesis if the test statistic (z) is either:

z ≥ critical two-tail value (1.959963. . ..). . .see cell B12
or
z ≤ − critical two-tail value (−1.959963. . .).

Note that we have two rules for rejection since our test does not assume that one of the sample means is larger or smaller than the other. Alternatively, we can compare the p-value = 0.006191. . . (cell B11) to α = 0.05 and reject if the p-value is ≤α. In this case the critical values (and the p-value) suggest that we *reject* the null hypothesis that the samples means are the same; that is, we have found evidence that the EB training program at SC has indeed had a significant effect on scores for the customer service representatives. EB is elated with this news since it suggests that the training does indeed make a difference, at least at the α = 0.5 level of

Table 6.3 Results of z-test for training of customer service representatives



	A	B	C
1	z-Test: Two Sample for Means		
2			
3		*36 Non-prisoner scores SC*	*36 Women SC (with special training)**
4	Mean	75.36111111	81.05555556
5	Known Variance	78.47	77.31
6	Observations	36	36
7	Hypothesized Mean Difference	0	
8	z	-2.737453564	
9	P(Z<=z) one-tail	0.003095843	
10	z Critical one-tail	1.644853627	
11	P(Z<=z) two-tail	0.006191686	
12	z Critical two-tail	1.959963985	
13			
14			

significance. This last comment is recognition that it is still possible, despite the current results, that our samples have led to the rejection of a true null hypothesis. If greater assurance is required, then run the test with a smaller α, for example 0.01. The results will be the same since a p-value 0.006191794 is less than 0.01. It is not until p-value is 0.001 that we do not reject the null hypothesis in favor of the alternative. This permits only a 1 in 1000 chance of rejecting a true null hypothesis. This is a very conservative test, in that it will only permit a very small type-1 error.

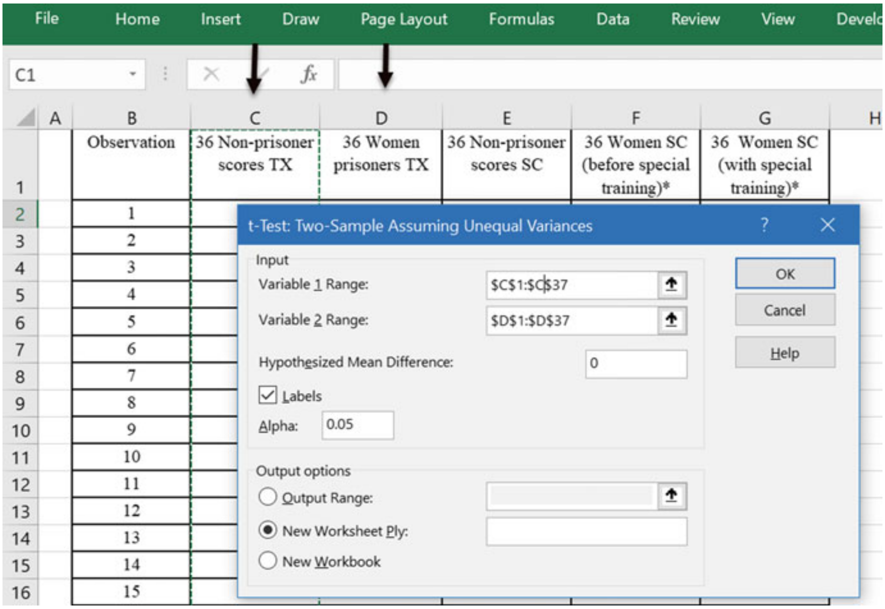
6.5.3 t-Test: Two Samples Unequal Variances

A very similar test, but one that does not explicitly consider the variance of the population to be *known*, is the **t-Test**. It is reserved for small samples, less than 30 observations, although larger samples are permissible. The lack of knowledge of a population variance is a very common situation. Populations are often so large that it is practically impossible to measure the variance or standard deviation of the population, not to mention the possible change in the population's membership. We will see that the calculation of the *t-statistic* is very similar to the calculation of the *z-statistic*.

6.5.4 Do Texas Prisoners Score Higher than Texas Non-prisoners?

Now, let's consider a second, but equally important question that EB will want to answer—Is it possible that women prisoners, ignoring state affiliation, normally score higher than others in the population, and that training is not the only factor in their higher scores? If we ignore the possible differences in state (SC or TX) affiliation of the prisoners for now, we can test this question by performing a test of hypothesis with only the Texas data samples and form a general conclusion. Why might this be an important question? We have already concluded that there is a difference between the mean score of SC prisoners and that of the SC non-prisoners. Before we attribute this difference to the special training provided by EB, let us consider the possibility that the difference may be due to the affiliation with the prison group. One can build an argument that women in prison might be motivated to learn and achieve, especially if they likely to soon be rejoining the general population. As we noted above, we will not deal with state affiliation at this point, although it is possible that one state may have higher scores than another.

To answer this question, we will use the **t-Test: Two Samples Unequal Variances** in the *Data Analysis* tool of Excel. In Fig. 6.4 we see the dialog box associated with the tool. Note that it appears to be quite similar to the z-Test. The difference is that rather than requesting values for known variances the t-Test calculates the sample variances and uses the calculated values in the analysis. The results of the analysis are shown in Table 6.4, and the t-statistic indicates that we



	A	B	C	D	E	F	G	H
1		Observation	36 Non-prisoner scores TX	36 Women prisoners TX	36 Non-prisoner scores SC	36 Women SC (before special training)*	36 Women SC (with special training)*	
2		1						
3		2						
4		3						
5		4						
6		5						
7		6						
8		7						
9		8						
10		9						
11		10						
12		11						
13		12						
14		13						
15		14						
16		15						

t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $C$1:$C$37

Variable 2 Range: $D$1:$D$37

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

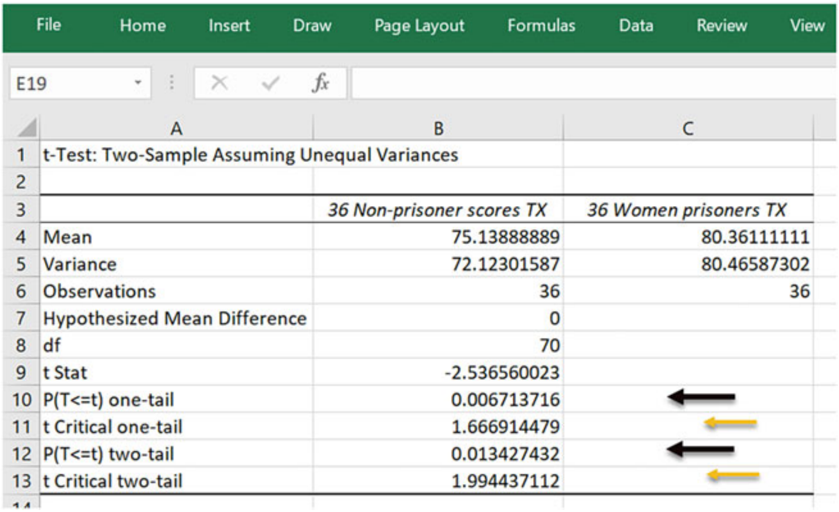
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Fig. 6.4 Data analysis tool for t-Test unequal variances

Table 6.4 Results t-test of prisoner & non-prisoner customer service representatives in TX



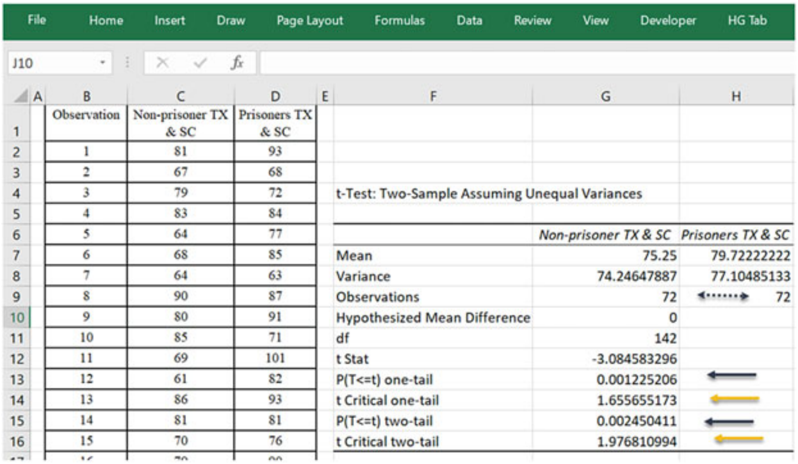
	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*36 Non-prisoner scores TX*	*36 Women prisoners TX*
4	Mean	75.13888889	80.36111111
5	Variance	72.12301587	80.46587302
6	Observations	36	36
7	Hypothesized Mean Difference	0	
8	df	70	
9	t Stat	-2.536560023	
10	P(T<=t) one-tail	0.006713716	
11	t Critical one-tail	1.666914479	
12	P(T<=t) two-tail	0.013427432	
13	t Critical two-tail	1.994437112	

should reject the null hypothesis: means for prisoners and non-prisoners are the same. This is because the t-statistic, −2.53650023 (cell B9), is less than the negative of the critical two-tail t-value, −1.994435479 (negative of cell B13). Additionally, we can see that the p-value for the two-tail test, 0.013427432 (cell B12), is ≤α (0.05). We therefore conclude that alternative hypothesis is likely true—there *is* a difference between the mean scores of the prisoners and non-prisoners. Yet, this could be due to many reasons we have not explored; thus, it might require further investigation.

6.5.5 *Do Prisoners Score Higher Than Non-prisoners Regardless of the State?*

Earlier we suggested that the analysis did not consider state affiliation, but in fact our selection of data has explicitly done so—only Texas data was used. The data is *controlled* for the state affiliation variable; that is, the state variable is *held constant* since all observations are from Texas. What might be a more appropriate analysis if we do not want to hold the state variable constant and thereby make a statement that is not state dependent? The answer is relatively simple: combine the SC and Texas non-prisoner scores in Fig. 6.2 columns C and E (72 observations; 36 + 36) and the SC and Texas Prisoner scores in column D and F (also 72). Note that we use Column F data, rather than G, since we are interested in the standard training only. Now we are ready to perform the analysis on these larger sample data sets, and fortuitously, more data is more reliable. The outcome is now independent of the state affiliation of the observations. In Table 6.5 we see that the results are similar to those in Table 6.4: we reject the null hypothesis in favor of the alternative that there is a difference. A

Table 6.5 Results of t-test scores of prisoner (SC & TX) and non-prisoner (SC & TX)



	A	B	C	D	E	F	G	H
1		Observation	Non-prisoner TX & SC	Prisoners TX & SC				
2		1	81	93				
3		2	67	68				
4		3	79	72		t-Test: Two-Sample Assuming Unequal Variances		
5		4	83	84				
6		5	64	77			*Non-prisoner TX & SC*	*Prisoners TX & SC*
7		6	68	85		Mean	75.25	79.72222222
8		7	64	63		Variance	74.24647887	77.10485133
9		8	90	87		Observations	72	↔ 72
10		9	80	91		Hypothesized Mean Difference	0	
11		10	85	71		df	142	
12		11	69	101		t Stat	-3.084583296	
13		12	61	82		P(T<=t) one-tail	0.001225206	←
14		13	86	93		t Critical one-tail	1.655655173	←
15		14	81	81		P(T<=t) two-tail	0.002450411	←
16		15	70	76		t Critical two-tail	1.976810994	←

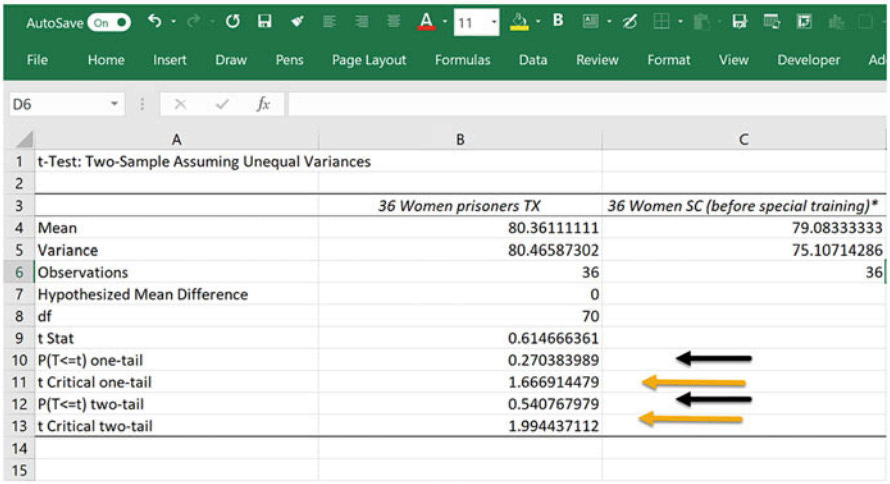
t-statistic of approximately −3.085 (cell G12) and a p-value of 0.0025 (cell G15) is evidence of the need to reject the null hypothesis; −3.085 is less than the critical value −1.977 (cell G16) and 0.0025 is less than α (0.05).

This a broader outcome in that it removes state affiliation, and the increased sample size provides additional assurance that the results are not due to sampling error: the chance of unrepresentative outcomes due to selecting a relatively small random sample. When we discuss confidence intervals later in this chapter we will see the effect of sample size on our confidence in the representative nature of the sample.

6.5.6 *How Do Scores Differ Among Prisoners of SC and Texas Before Special Training?*

A third and related question of interest is whether the prisoners in SC and TX have mean scores (before training) that are significantly different. To test this question, we can compare the two samples of the prisoners, TX and SC, using the SC prisoners' scores prior to special training. To include EB trained prisoners would be an unfair comparison, given that the special training may influence their scores. Table 6.6 shows the results of the analysis. Again, we perform the t-Test: two-samples unequal variances and get t-statistic of 0.614666361 (cell B9). Given that the two-tail critical value is 1.994437112 (cell B13), the calculated t- statistic is not sufficiently extreme to reject the null hypothesis that there is no difference in mean scores for the prisoners of TX and SC. Additionally, the p-value, 0.540767979, is much larger

Table 6.6 Test of the difference in standard trained TX and SC prisoner scores

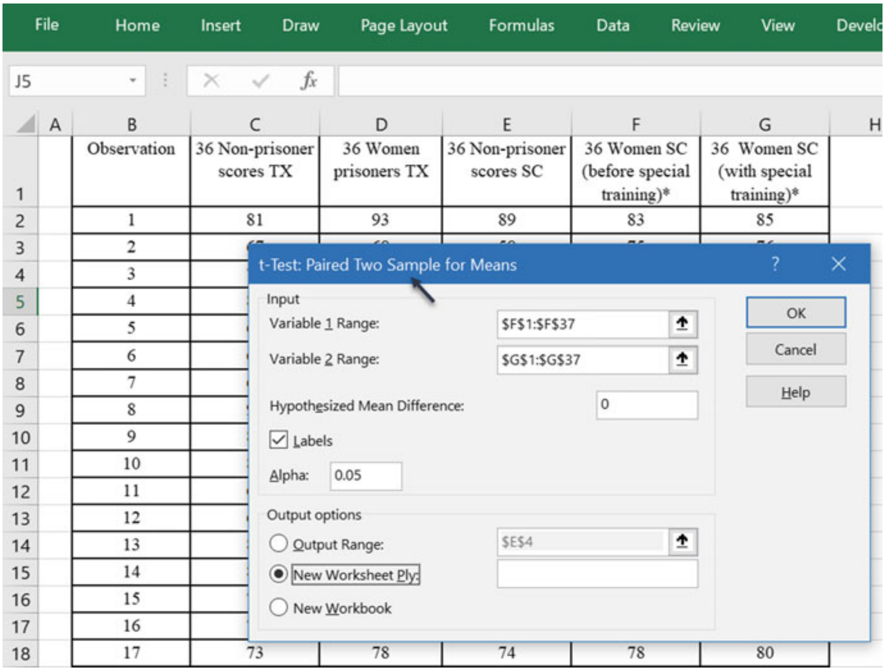


	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*36 Women prisoners TX*	*36 Women SC (before special training)**
4	Mean	80.36111111	79.08333333
5	Variance	80.46587302	75.10714286
6	Observations	36	36
7	Hypothesized Mean Difference	0	
8	df	70	
9	t Stat	0.614666361	
10	P(T<=t) one-tail	0.270383989	
11	t Critical one-tail	1.666914479	
12	P(T<=t) two-tail	0.540767979	
13	t Critical two-tail	1.994437112	
14			
15			

than the α of 0.05. This is not an unexpected outcome given how similar the mean scores, 79.083 and 80.361, were for prisoners in both states.

Finally, we began the example with a question that focused on the viability of special training. Is there a significant difference in scores after special training? The analysis for this question can be done with a specific form of the t-statistic that makes a very important assumption: the samples are **paired** or **matched**. Matched samples simply imply that the sample data is collected from the same 36 observations, in our case the same SC prisoners. This form of sampling *controls* for individual differences in the observations by focusing directly on the special training as a level of treatment. It also can be thought of as a *before-and-after* analysis. For our analysis, there are two levels of training—standard training and special (EB) training. The tool in the Data Analysis menu to perform this type of analysis is **t-Test: Paired Two-Sample for Means**.

Figure 6.5 shows the dialog box for matched samples. The data entry is identical to that of the two-sample assuming unequal variances in Fig. 6.4. Before we perform the analysis, it is worthwhile to consider the outcome. From the data samples collected in Table 6.2, we can see that the average score difference between the two treatments is about 2 points (79.08 before; 81.06 after). More importantly, if you examine the final two data columns in Table 6.2, it is clear that every observation for the prisoners with only standard training is improved when special training is applied. Thus, an informal analysis suggests that scores definitely have improved.



	A	B	C	D	E	F	G	H
1		Observation	36 Non-prisoner scores TX	36 Women prisoners TX	36 Non-prisoner scores SC	36 Women SC (before special training)*	36 Women SC (with special training)*	
2		1	81	93	89	83	85	
3		2						
4		3						
5		4						
6		5						
7		6						
8		7						
9		8						
10		9						
11		10						
12		11						
13		12						
14		13						
15		14						
16		15						
17		16						
18		17	73	78	74	78	80	

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $F$1:$F$37

Variable 2 Range: $G$1:$G$37

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

○ Output Range: $E$4

● New Worksheet Ply:

○ New Workbook

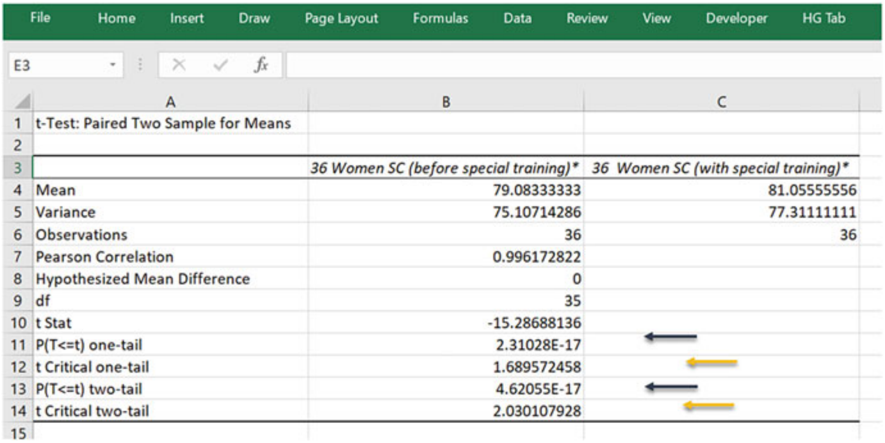
Fig. 6.5 Data analysis tool for paired two-sample means

We would not be as secure in our analysis if we achieved the same sample mean score improvement, but the individual matched scores were not consistently higher. In other words, if we have an improvement in mean scores, but some individual scores improve and some decline, the perception of consistent improvement is less compelling.

6.5.7 *Does the EB Training Program Improve Prisoner Scores?*

Let us now perform the analysis and review the results. Table 6.7 shows the results of the analysis. First, note the Pearson Correlation for the two samples is 99.62% (cell B7). You will note that the Pearson Correlation does not appear in the t-test and z-tests we used before. It is only important in the matched samples t-test. This is a very strong positive correlation in the two data series, verifying the observation that the two scores move together in a very strong fashion; relative to the standard training score, the prisoner scores move in the same direction (positive) after special training. The t-statistic is −15.28688136 (cell B10), which is a very large negative[4] value for the critical two-tale value for rejection of the null hypothesis, 2.030107928 (cell B14). Thus, we reject the null and accept the alternative: training *does* make a difference. The p-value is miniscule, 4.62055E-17 (cell B13), and far smaller than the 0.05 level set for α, which of course similarly suggests rejection of the null. The

Table 6.7 Test of matched samples SC prisoner scores



	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*36 Women SC (before special training)**	*36 Women SC (with special training)**
4	Mean	79.08333333	81.05555556
5	Variance	75.10714286	77.31111111
6	Observations	36	36
7	Pearson Correlation	0.996172822	
8	Hypothesized Mean Difference	0	
9	df	35	
10	t Stat	-15.28688136	
11	P(T<=t) one-tail	2.31028E-17	←
12	t Critical one-tail	1.689572458	←
13	P(T<=t) two-tail	4.62055E-17	←
14	t Critical two-tail	2.030107928	←
15			

[4] −15.28688136 is a negative t-statistic because of the entry order of our data in the Excel dialog box. If we reverse the ranges for variable entry, the result is +15.28688136.

question remains whether an improvement of approximately two points is worth the investment in the training program. This is a cost-benefit tradeoff that must be considered because EB will surely charge for her training services.

6.5.8 What If the Observations Means Are the Same, But We Do Not See Consistent Movement of Scores?

To see how the results will change if consistent improvement in matched pairs does not occur, while maintaining the averages, I will shuffle the data for training scores. In other words, the scores in the *36 Women prisoners SC (trained)* column will remain the same, but they will not be associated with the same values in the *36 Women prisoners SC (before training)* column. Thus, no change will be made in values; only the matched pairs will be changed. Table 6.8 shows the new (shuffled) pairs with the same mean scores as before. Table 6.9 shows the new results. Note that the means remain the same, but the Pearson Correlation value is quite different from before: −0.15617663. This negative value indicates that as one matched pair value increases there is generally a very mild decrease in the other value. Now the newly calculated t-statistic is −0.876116006. Given the critical t-value of 2.030107928, we *cannot* reject the null hypothesis that there is no difference in the means. The results are completely different than before, in spite of similar averages for the matched pairs. Thus, you can see that the consistent movement of matched pairs is extremely important to the analysis.

6.5.9 Summary Comments

In this section, we progressed through a series of hypothesis tests to determine the effectiveness of the EB special training program applied to SC prisoners. As you have seen, the question of the special training's effectiveness is not a simple one to answer. Determining statistically the true effect on the mean score improvement is a complicated task that may require several tests and some personal judgment. We also must make a number of assumptions to perform our tests—do we combine State affiliation (TX and SC), do we include the special training data, etc. It is often the case that observed data can have numerous associated factors. In our example, the observations were identifiable by state (SC or TX), status of freedom (prisoner and non-prisoner), exposure to training (standard or EB special), and finally gender, although it was not fully specified for all observations. It is quite easy to imagine many more factors associated with our sample observations—e.g. age, level of education, etc.

In the next section, we will apply Analysis of Variance (ANOVA) to similar problems. ANOVA will allow us to compare the effects of multiple factors, with each factor containing several levels of treatment on a variable of interest, for

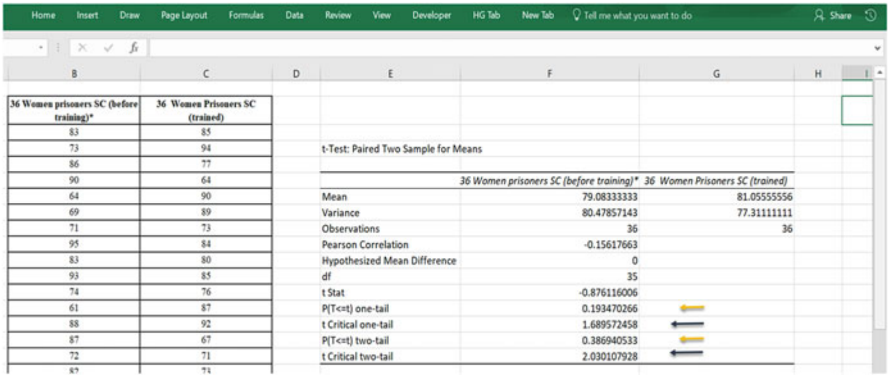
Table 6.8 Scores for matched pairs that have been shuffled

OBS	36 women prisoners SC (before training)[a]	36 women prisoners SC (trained)
1	83	85
2	73	94
3	86	77
4	90	64
5	64	90
6	69	89
7	71	73
8	95	84
9	83	80
10	93	85
11	74	76
12	61	87
13	88	92
14	87	67
15	72	71
16	82	73
17	79	98
18	83	93
19	74	75
20	89	74
21	76	83
22	63	89
23	86	78
24	79	72
25	83	85
26	76	76
27	91	91
28	74	79
29	73	65
30	80	87
31	86	81
32	77	84
33	70	79
34	92	81
35	80	68
36	65	93
Average	79.08	81.06

[a] Same 36 SC women prisoners that received training

example a test score. We will return to our call center example and identify 3 factors with 2 levels of treatment each. If gender could also be identified for each observation, the results would be 4 factors with 2 treatments for each. ANOVA will split our data into components, or groups, which can be associated with the various levels of factors.

Table 6.9 New matched pairs analysis



36 Women prisoners SC (before training)*	36 Women Prisoners SC (trained)
83	85
73	94
86	77
90	64
64	90
69	89
71	73
95	84
83	80
93	85
74	76
61	87
88	92
87	67
72	71
87	73

t-Test: Paired Two Sample for Means

	36 Women prisoners SC (before training)*	36 Women Prisoners SC (trained)
Mean	79.08333333	81.05555556
Variance	80.47857143	77.31111111
Observations	36	36
Pearson Correlation	-0.15617663	
Hypothesized Mean Difference	0	
df	35	
t Stat	-0.876116006	
P(T<=t) one-tail	0.193470266	←
t Critical one-tail	1.689572458	←
P(T<=t) two-tail	0.386940533	←
t Critical two-tail	2.030107928	←

6.6 Confidence Intervals for Sample Statistics

In the prior sections, we applied our analyzes to random samples taken from a population. Our goal was to make a statement about a population from what we discovered in the sample. We will continue with that theme in this section. But rather than focus on inferential statistics, we now turn to an area in statistics called **Estimation**.

This topic is almost always discussed before hypothesis testing, but I present it after with what I believe to be no loss of understanding. You will find that there are numerous similarities in **Confidence Intervals** and hypothesis testing, both in terms of concept and analytical features. This should not be surprising given that they both exist because of sampling. As the name implies, an interval will be created about some value, and a statement of about our confidence that the interval contains something will be made. For example, I attend a very, very large family reunion and my cousin Mario is in attendance. I randomly select 30 of my other relatives to guess Mario's weight—he is a very large man. The average (mean) of the guesses in 205 kilograms (approximately 450 pounds). I then perform an analysis to build a confidence interval. Without getting into the details of my calculations, I tell the attendees that I am 95% confident that Mario weighs between 190 and 220 kilos. This range, 190 to 220 kilos, is my 95% confidence interval, and there is 5% chance that Mario's weight is not in the range. What if you want to be surer about *capturing* Mario's weight in the interval? A 99% confidence interval might be 175 to 235 kilos, and there is a 1% chance that the interval has not captured Mario's true weight. The interval has the same average, but rather than 205 ± 15 kilos, it is now 205 ± 30 kilos. So, to be surer, you must expand the interval about the mean. The details of this example do not exactly fit the notion of a confidence interval, but it is a simple introduction to what's to follow.

6.6.1 What Are the Ingredients of a Confidence Interval?

The area of statistical estimation encompasses interval and point estimates. Both areas will be important to us in estimating the parameters of interest for a population; for example, the population's mean, variance, standard deviation, or proportion. First, we will calculate a point estimate of the population parameter, for example the average computed for cousin Mario's weight. This average is called an unbiased estimator of the populations true average. This translates loosely into the following: if I took many, many samples and calculated their averages, the distribution of these averages would have the same average as the population distribution. Note that we are talking about an average of averages. This is called the sampling distribution. Besides a sampling distribution having an average, it also has a standard deviation, or a measure of the variation of the averages, and it is called the **standard error**.

There is only one term left that is needed to create a confidence interval, the **critical value**. This value will either be a z^*_α or $t^*_{\alpha/2,\, n-1}$ value. The z* will be used if the standard deviation of the population is known; the t* will be used if the standard deviation is unknown. These values should be familiar to you. We used them in the tests of hypothesis we performed earlier, and they were used under the same circumstances about our knowledge of the standard deviation of the population. These values can be found in statistical tables when we provide our level of confidence, 1_α, for both z* and t* and the additional degrees of freedom, n_1, for t*. The value of n is the number of our sample observations, 30 in the case of cousin Mario.

Now, let us refine what is meant by the confidence interval. First, let's understand what it does not mean. It does not mean that a true population parameter has a particular percentage of being contained in the interval derived from the sample. So, for our example of cousin Mario, we cannot say that the interval, 190 to 220 kilos, has a 95% chance of containing Mario's true weight. The interval either *does* or *does not* contain his weight. But, if I repeated the sampling experiment many times, with the procedure described, 95% of the intervals would contain Mario's true weight. The difference in what a confidence interval is and is not may seem subtle, but it is not. The emphasis on what it is, is about the sampling process we use. It is not about the population parameter we are interested in determining. Stated in a slightly different way, the probability statement made (95%) is about the sampling process leading to a confidence interval and not to the population parameter. Maybe the most important way to think about confidence intervals is if the true value of the parameter is *outside* the 95% confidence interval we formed, then a sampling outcome occurred which has a small probability ($\leq$5%) of occurring by mere chance. So, by *inference*, I can feel confident that my population parameter is in the range of confidence interval. This is an indirect, or *back-door*, arrival at confidence about the population parameter.

The mathematical formulas that describe a confidence interval are simple:

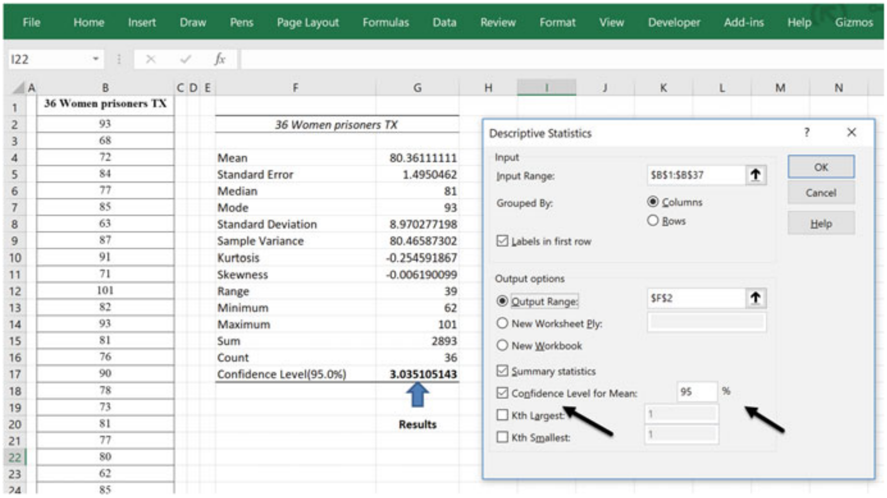
$\bar{x} \pm z^*_\alpha\ (\sigma/\sqrt{n})$ . . ..CI for known variance of the population distribution
$\bar{x} \pm t^*_{\alpha/2,\ n\text{-}1}\ (s/\sqrt{n})$. . ..CI for unknown variance of the population distribution
(we use the sample variance for the standard error)

Where:

- n is the number of observations in the sample
- σ is population standard deviation
- σ/√n is the standard error when population variance is known
- s is sample standard deviation
- s/√n if the standard error when we estimate of the population variance based on the sample variance

6.6.2 A Confidence Interval Example

Let us now consider a problem that uses confidence interval analysis. We will use a portion of the data from our call center problem—the sample of *36 women prisoners in Texas*. Our goal is the answer the following question about the population of all Texas women prisoners working in our call center: Can I construct a range of values that suggests that the true population parameter is inside that range. There are several methods to achieve this in Excel that are relatively simple, when compared to our formulas. The first, which is shown in Fig. 6.6, is to utilize the *Data Analysis* tools in the *Data* ribbon. In the tools we have previously used the *Descriptive Statistics* tool provide a summary of our data. One option we did not exercise what the *Confidence Level for Mean*. Simply check that option box and provide a confidence level (default is 95%), and the value returned can then be added and subtracted from the mean to create the range of the confidence interval:



	A	B	C	D	E	F	G
1		36 Women prisoners TX					
2		93				*36 Women prisoners TX*	
3		68					
4		72				Mean	80.36111111
5		84				Standard Error	1.4950462
6		77				Median	81
7		85				Mode	93
8		63				Standard Deviation	8.970277198
9		87				Sample Variance	80.46587302
10		91				Kurtosis	-0.254591867
11		71				Skewness	-0.006190099
12		101				Range	39
13		82				Minimum	62
14		93				Maximum	101
15		81				Sum	2893
16		76				Count	36
17		90				Confidence Level(95.0%)	**3.035105143**
18		78					
19		73					
20		81					
21		77					
22		80					
23		62					
24		85					

Descriptive Statistics ? ×

Input
Input Range: $B$1:$B$37
Grouped By: ● Columns ○ Rows
☑ Labels in first row

Output options
● Output Range: $F$2
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
☑ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

OK
Cancel
Help

Fig. 6.6 Confidence level calculation in data analysis tools descriptive statistics

Lower Limit.... $80.3611111 - 3.0351051 = 77.326\ldots$
Upper Limit..... $80.3611111 + 3.0351051 = 83.396\ldots$

The assumption that is made regarding the standard error is that the population variance is unknown; therefore, the sample variance is used to calculate the standard error. This is overwhelmingly the case in most problems, so you can generally use this approach to find your confidence interval. So, given our previous convoluted discussion on what the confidence interval means, I feel confident that the population of women prisoners in Texas has an average score between 77.326 and 83.396.

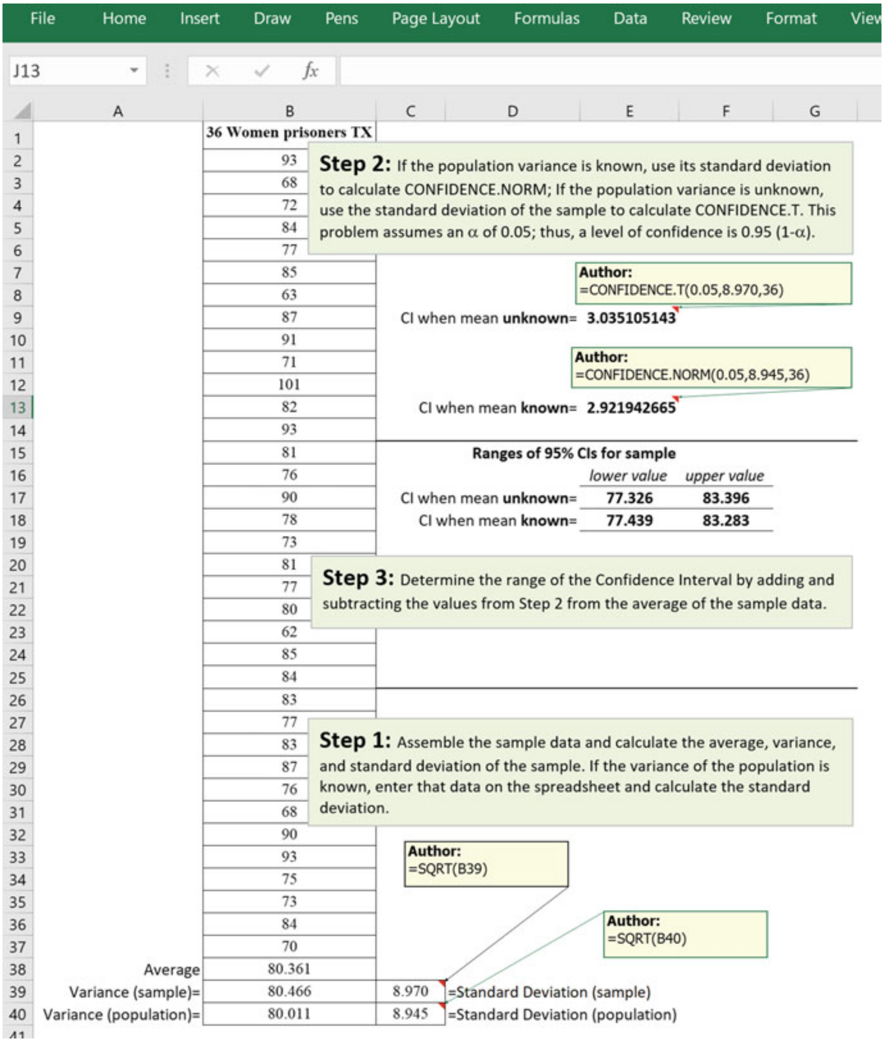
There will often be situations where you don't know the variance of the population, but there may be a situation where you know the variance of the population. In Fig. 6.7 I introduce the Excel functions CONFIDENCE.NORM CONFIDENCE.T that permit you to perform the calculations for both situations. The three steps necessary for the analysis are provided. Cell B40 indicates the variance that is assumed to be known, and in cell C40 we see the standard deviation (the square root of variance). Notice that the range of the confidence interval with unknown variance in cells E17:F17 is exactly what we saw in Fig. 6.6.

After the analysis in Fig. 6.6, the assistant warden of the prison program in charge of the call center makes a bold statement. She says that she is absolutely certain that the average scope for the prisoner population working with the call center is at least 90. Is the statement possibly correct? Will the analysis we performed in Fig. 6.6 provide some insight into the veracity of her statement?

First, we observe that the confidence interval (77.326–83.396) from the sample does not include the Assistant Warden's value ($\geq$90). This suggests that she is not correct in her assertion. Second, the distance of her value from the upper limit of the confidence interval is substantial ($90 - 83.396 = 6.604$); in fact, it is more than twice the 3.0351... value calculated in Fig. 6.6. My response to her would be an emphatic, but polite—"Assistant warden, the probability of you being correct is very, very small." Based on the result that the confidence interval for the sample did not capture her asserted value, I would feel very confident that she is incorrect. The second result is even stronger evidence to the contrary of her assertion.

6.6.3 *Single Sample Hypothesis Tests Are Similar to Confidence Intervals*

There is one important hypothesis test that we have not yet discussed: single sample hypothesis test. We discuss it now because of its similarity to the procedures we used in building confidence intervals. Consider the confidence interval construction where we compared a calculated 95% interval to a value posited by the Assistant Warden. This analysis could have just as easily been performed as a single sample test of hypothesis. We will see that this procedure will require a bit more work that the others we have encountered. This is because the Data Analysis tools do not



	A	B	C	D	E	F
1		36 Women prisoners TX				
2		93				
3		68				
4		72				
5		84				
6		77				
7		85				
8		63				
9		87		CI when mean **unknown**=	**3.035105143**	
10		91				
11		71				
12		101				
13		82		CI when mean **known**=	**2.921942665**	
14		93				
15		81		**Ranges of 95% CIs for sample**		
16		76			*lower value*	*upper value*
17		90		CI when mean **unknown**=	**77.326**	**83.396**
18		78		CI when mean **known**=	**77.439**	**83.283**
19		73				
20		81				
21		77				
22		80				
23		62				
24		85				
25		84				
26		83				
27		77				
28		83				
29		87				
30		76				
31		68				
32		90				
33		93				
34		75				
35		73				
36		84				
37		70				
38	Average	80.361				
39	Variance (sample)=	80.466	8.970	=Standard Deviation (sample)		
40	Variance (population)=	80.011	8.945	=Standard Deviation (population)		

Step 2: If the population variance is known, use its standard deviation to calculate CONFIDENCE.NORM; If the population variance is unknown, use the standard deviation of the sample to calculate CONFIDENCE.T. This problem assumes an α of 0.05; thus, a level of confidence is 0.95 (1-α).

Author:
=CONFIDENCE.T(0.05,8.970,36)

Author:
=CONFIDENCE.NORM(0.05,8.945,36)

Step 3: Determine the range of the Confidence Interval by adding and subtracting the values from Step 2 from the average of the sample data.

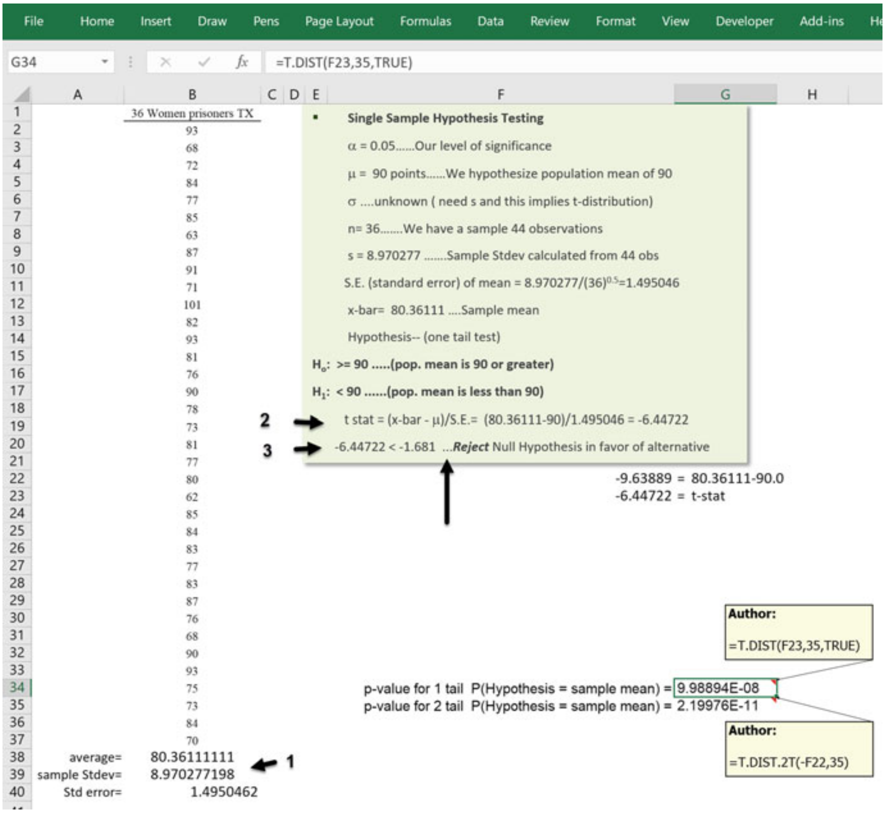
Step 1: Assemble the sample data and calculate the average, variance, and standard deviation of the sample. If the variance of the population is known, enter that data on the spreadsheet and calculate the standard deviation.

Author:
=SQRT(B39)

Author:
=SQRT(B40)

Fig. 6.7 General confidence level calculations using CONFIDENCE.NORM and CONFIDENCE.T

contain a tool for single sample tests, as they did for our two sample tests. So, how do we test the hypothesis that a single sample mean is significantly different from a posited value for the population mean? Just as before, we will need to calculate a t-statistic, and then determine a critical t-value and compare one to the other. Alternatively, if we can find a p-value to compare to our $\boldsymbol{\alpha}$, then we can reject or not-reject the Null hypothesis on the basis of whether or not the p-value is smaller or larger, respectively. These two methods are precisely the same as before in our two sample tests.



=T.DIST(F23,35,TRUE)

	A	B
1		36 Women prisoners TX
2		93
3		68
4		72
5		84
6		77
7		85
8		63
9		87
10		91
11		71
12		101
13		82
14		93
15		81
16		76
17		90
18		78
19		73
20		81
21		77
22		80
23		62
24		85
25		84
26		83
27		77
28		83
29		87
30		76
31		68
32		90
33		93
34		75
35		73
36		84
37		70
38	average=	80.36111111
39	sample Stdev=	8.970277198
40	Std error=	1.4950462

- **Single Sample Hypothesis Testing**

α = 0.05......Our level of significance

μ = 90 points......We hypothesize population mean of 90

σ ....unknown (need s and this implies t-distribution)

n= 36.......We have a sample 44 observations

s = 8.970277 ....Sample Stdev calculated from 44 obs

S.E. (standard error) of mean = 8.970277/(36)$^{0.5}$=1.495046

x-bar= 80.36111 ....Sample mean

Hypothesis-- (one tail test)

H_o: >= 90 .....(pop. mean is 90 or greater)

H_1: < 90 ......(pop. mean is less than 90)

t stat = (x-bar - μ)/S.E.= (80.36111-90)/1.495046 = -6.44722

-6.44722 < -1.681 ...***Reject*** Null Hypothesis in favor of alternative

-9.63889 = 80.36111-90.0
-6.44722 = t-stat

Author:
=T.DIST(F23,35,TRUE)

p-value for 1 tail P(Hypothesis = sample mean) = 9.98894E-08
p-value for 2 tail P(Hypothesis = sample mean) = 2.19976E-11

Author:
=T.DIST.2T(-F22,35)

Fig. 6.8 Example of a single sample hypothesis test

In Fig. 6.8 we provide the data in Assistant Warden problem. Let's begin with the calculation of the t-statistic that is going to be compared to the critical t-value. Although we did not specifically show the calculation for the t-statistics in the two sample tests (they are complicated to calculate, but provided by the *Data Analysis* tool), the calculation in the one sample test is simple:

$$
\begin{aligned}
t-statistic &= (mean\ of\ the\ sample - hypothesized\ population\ mean)/standard\ error \\
&= (80.36111 - 90.0)/1.49505 = -6.44722
\end{aligned}
$$

The specific calculations for our example are shown in a stepwise fashion in the text box of Fig. 6.8. The calculated t-statistic is −6.4472. When compared to the critical t-value of −1.681 the null hypothesis posited by the Assistant Warden is soundly rejected. Alternatively, we can use the T.DIST() function to determine the p-value for our test: 9.98894E-08. This a very small p-value relative to an assumed α of 0.05; thus, it leads to a very strong rejection of the null hypothesis.

Now, what if we are interested in more sophisticated analysis on categorical and interval data that are related? The next technique, analysis of variance (ANOVA), is an extremely powerful tool for a variety of problems and experimental designs.

6.7 ANOVA

In this section, we will use **ANOVA** to find what are known as **main** and **interaction effects** of categorical (nominal) independent variables on an interval, dependent variable. The *main* effect of an independent variable is the *direct* effect it exerts on a dependent variable. The *interaction* effect is a bit more complex. It is the effect that results from the joint interactions of two or more independent variables on a dependent variable. Determining the effects of independent variables on dependent variables is quite similar to the analysis we performed in the sections above. In that analysis, our independent variables were the state (SC or TX), status of freedom (prisoner and non-prisoner), and exposure to training (standard or special). These categorical independent variables are also known as **factors**, and depending on the **level** of the factor, they can affect the scores of the call center employees. Thus, in summary, the levels of the various factors for the call center problem are: (1) *prisoner* and *non-prisoner* status for the freedom factor, (2) *standard* and *special* for the training factor, (3) *SC* and *TX* for state affiliation factor.

Excel permits a number of ANOVA analyses—*single factor*, *two-factor without replication*, and *two-factor with replication*. **Single factor ANOVA** is similar to the t-Tests we previously performed, and it provides an extension of the t-Tests analysis to more that two samples means; thus, the ANOVA tests of hypothesis permit the testing of equality of three or more sample means. It is also found in the *Data Analysis* tool in the Data Ribbon. This reduces the annoyance of constructing many pair-wise t-Tests to fully examine all sample relationships. The two-factor ANOVA, with and without replication, extends ANOVA beyond the capability of t-Tests. Now, let us begin with a very simple example of the use of single factor ANOVA.

6.7.1 ANOVA: Single Factor Example

A shipping firm is interested in the theft and loss of refrigerated shipping containers, commonly called *reefers*, that they experience at three similar sized terminal facilities at three international ports—Port of New York/New Jersey, Port of Amsterdam, and Port of Singapore. Containers, especially refrigerated, are serious investments of capital, not only due to their expense, but also due to the limited production capacity available for their manufacture. The terminals have similar security systems at all three locations, and they were all updated approximately 1 year ago. Therefore, the firm assumes the average number of missing containers at all the terminals should be relatively similar over time. The firm collects data over 23 months at the three locations to determine if the monthly means of lost and stolen reefers at the various sites are significantly different. The data for reefer theft and loss is shown in Table 6.10.

The data in Table 6.10 is in terms of reefers missing per month and represents a total of 23 months of collected data. A casual inspection of the data reveals that the average of missing reefers for Singapore is substantially lower than the averages for Amsterdam and NY/NJ. Also, note that the data includes an additional data element—the security system in place during the month. Security system A was replaced with system B at the end of the first year. In our first analysis of a single factor, we will only consider the Port factor with three levels—NY/NJ, Amsterdam, and Singapore. This factor is the independent variable and the number of missing reefers is the *response*, or dependent variable. It is possible to later consider the security system as an additional factor with two levels, A and B. Here is our first question of interest.

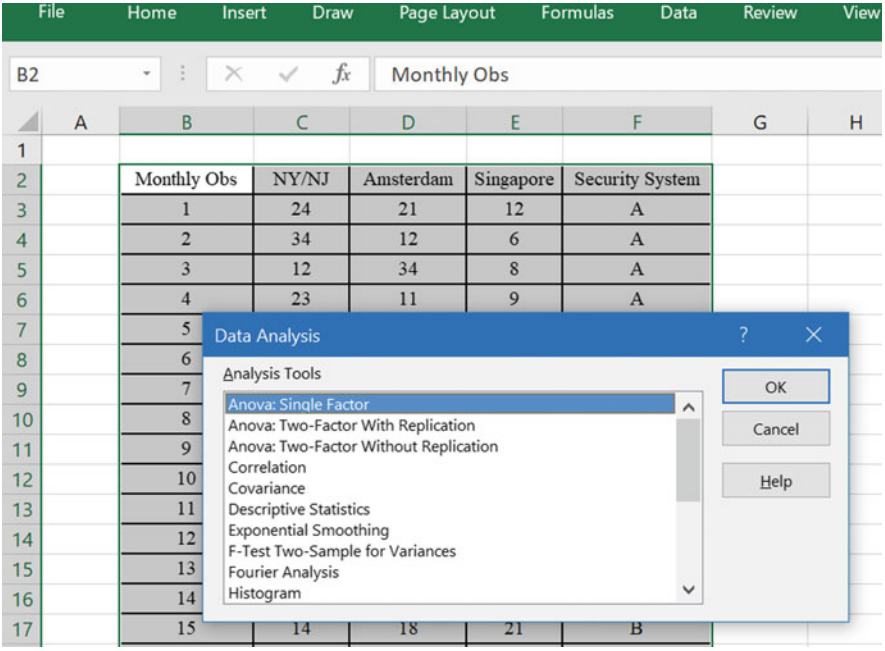
Table 6.10 Reported missing reefers for terminals

Monthly Obs	NY/NJ	Amsterdam	Singapore	Security System
1	24	21	12	A
2	34	12	6	A
3	12	34	8	A
4	23	11	9	A
5	7	18	11	A
6	29	28	3	A
7	18	21	21	A
8	31	25	19	A
9	25	23	6	A
10	23	19	18	A
11	32	40	11	A
12	18	21	4	B
13	27	16	7	B
14	21	17	17	B
15	14	18	21	B
16	6	15	9	B
17	15	7	10	B
18	9	9	3	B
19	12	10	6	B
20	15	19	15	B
21	8	11	9	B
22	12	9	13	B
23	17	13	4	B
Average=	18.78	18.13	10.52	
Stdev=	8.37	8.15	5.66	

6.7.2 *Do the Mean Monthly Losses of Reefers Suggest That the Means Are Different for the Three Ports?*

Now, we consider the application of ANOVA to our problem. In Figs 6.9 and 6.10, we see the dialog box entries that permit us to perform the analysis. As before, we must identify the data range of interest; in this case, the three treatments of the Port factor (C2:E25), including labels. The α selected for comparison to the p-value is 0.05. Also, unlike the t-Test, where we calculate a t-statistic for rejection or non-rejection of the null, in ANOVA we calculate an **F-Statistic** and compare it to a **critical F-value**. Thus, the statistic is different, but the general procedure is similar.

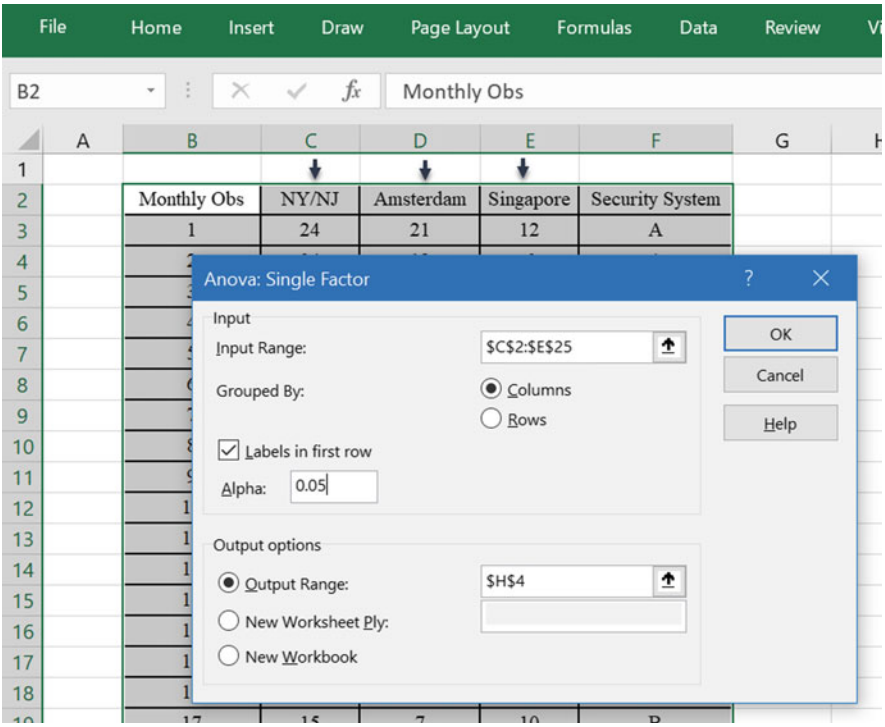
Table 6.11 shows the result of the analysis. Note that the F-statistic, 8.634658 (cell L15) is larger than the critical F-value, 3.135918 (cell N15), so we can reject the null that all means come from the same population of expected reefer losses. Also, if the p-value, 0.000467 (cell M15) is less than our designated α (0.05), which is the case, we reject the null hypothesis. Thus, we have rejected the notion that the average monthly losses at the various ports are similar. At least one of the averages seems to come from a different distribution of monthly losses, and it is not similar to the averages of the other ports. Although the test does not identify the mean that is significantly different, we are certainly capable of noting that it is the mean of



	A	B	C	D	E	F
1						
2		Monthly Obs	NY/NJ	Amsterdam	Singapore	Security System
3		1	24	21	12	A
4		2	34	12	6	A
5		3	12	34	8	A
6		4	23	11	9	A
7		5				
8		6				
9		7				
10		8				
11		9				
12		10				
13		11				
14		12				
15		13				
16		14				
17		15	14	18	21	B

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram

Fig. 6.9 ANOVA: Single factor tool



Monthly Obs	NY/NJ	Amsterdam	Singapore	Security System
1	24	21	12	A

Anova: Single Factor

Input

Input Range: $C$2:$E$25

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

Output options

● Output Range: $H$4

○ New Worksheet Ply:

○ New Workbook

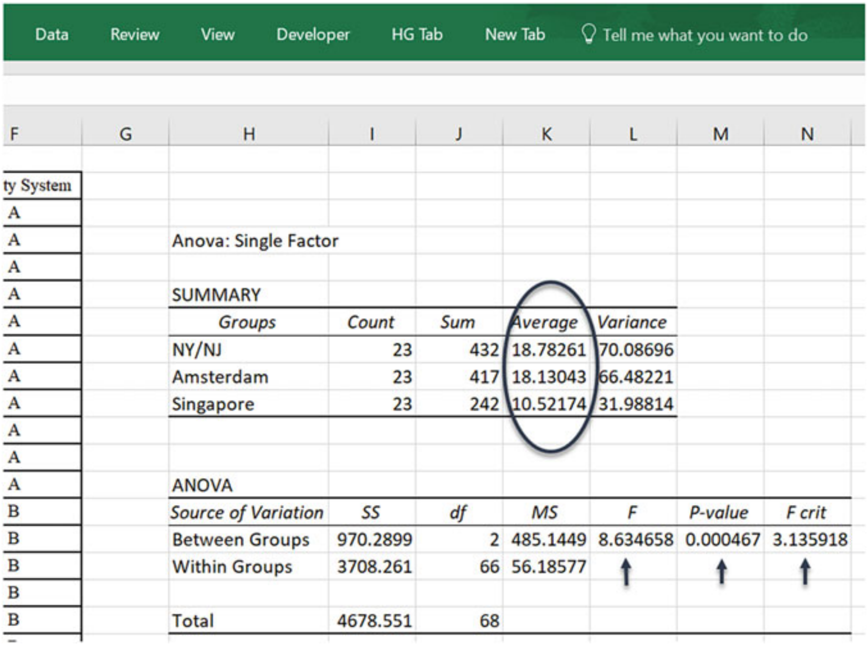
Fig. 6.10 ANOVA: Single factor dialog box

Singapore. We are not surprised to see this result given the much lower average at the Port of Singapore—about 10.5 versus 18.8 and 18.1 for the Port of NY/NJ and Amsterdam, respectively.

6.8 Experimental Design

There are many possible methods by which we conduct a data collection effort. Researchers are interested in carefully controlling and designing experimental studies, not only the analysis of data, but also its collection. The term used for explicitly controlling the collection of observed data is **Experimental Design**. Experimental design permits researchers to refine their understanding of how factors affect the dependent variables in a study. Through the control of *factors* and their levels, the experimenter attempts to eliminate ambiguity and confusion related to the observed outcomes. This is equivalent to eliminating alternative explanations of observed results. Of course, completely eliminating alternative explanations is not possible,

Table 6.11 ANOVA single factor analysis for missing reefers



ty System
A
A
A
A
A
A
A
A
A
A
A
B
B
B
B
B

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
NY/NJ	23	432	18.78261	70.08696
Amsterdam	23	417	18.13043	66.48221
Singapore	23	242	10.52174	31.98814

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	970.2899	2	485.1449	8.634658	0.000467	3.135918
Within Groups	3708.261	66	56.18577			
Total	4678.551	68				

but attempting to control for alternative explanations is the hallmark of a well-conceived study: a *good* experimental design.

There are some studies where we purposefully do not become actively involved in the manipulation of factors. These studies are referred to as **observational studies**. Our refrigerated container example above is best described as an observational study, since we have made no effort to manipulate the study's single factor of concern—Port. These ports simply happen to be where the shipping firm has terminals. If the shipping firm had many terminal locations and it had explicitly selected a limited number ports to study for some underlying reason, then our study would have been best described as an **experiment**. In experiments we have greater ability to influence *factors*. There are many types of experimental designs, some simple and some quite complex. Each design serves a different purpose in permitting the investigator to come to a scientifically focused and justifiable conclusion. We will discuss a small number of designs that are commonly used in analyses. It is impossible to exhaustively cover this topic in a small segment of a single chapter, but there are many good texts available on the subject if you should want to pursue the topic in greater detail.

Now, let us consider in greater detail the use of experimental designs in studies that are *experimental* and not *observational*. As I have stated, it is impossible to consider all the possible designs, but there are three important designs worth considering due to their frequent use. Below I provide a brief description that explains the major features of the three experimental designs:

- **Completely Randomized Design**: This experimental design is structured in a manner such that the treatments that are allocated to the **experimental units** (subjects or observations) are assigned completely at random. For example, consider 20 analysts (our experimental unit) from a population. The analysts will use 4 software products (treatments) for accomplishing a specific technical task. A response measure, the time necessary to complete the task, will be recorded. Each analyst is assigned a unique number from 1 to 20. The 20 numbers are written on 20 identical pieces of paper and placed into a container marked subject. These numbers will be used to allocate analysts to the various software products. Next, a number from 1 to 4, representing the 4 products, is written on 4 pieces of paper and repeated 5 times, resulting in 4 pieces of paper with the number 1, 4 pieces with the number 2, etc. These 20 pieces of paper are placed in a container marked treatment. Finally, we devise a process where we pick a single number out of each container and record the number of the analyst (subject or experimental unit) and the number of the software product (treatment) they will use. Thus, a couplet of an analyst and software treatment is recorded; for example, we might find that analyst 14 and software product 3 form a couplet. After the selection of each couplet, discard the selected pieces of paper (do not return to the containers) and repeat the process until all pieces of paper are discarded. The result is a completely randomized experimental design. The analysts are randomly assigned to a randomly selected software product, thus the description—completely randomized design.
- **Randomized Complete Block Design**: This design is one in which the experimental subjects are grouped (blocked) according to some variable which the experimenter wants to control. The variable could be intelligence, ethnicity, gender, or any other characteristic deemed important. The subjects are put into groups (blocks), with the same number of subjects in a group as the number of treatments. Thus, if there are 4 treatments, then there will be 4 subjects in a block. Next, the constituents of each block are then randomly assigned to different treatment groups, one subject per treatment. For example, consider 20 randomly selected analysts that have a recorded historical average time for completing a software task. We decide to organize the analysts into blocks according to their historical average times. The 4 lowest task averages are selected and placed into a block, the next 4 lowest task averages are selected to form the next block, and the process continues until 5 blocks are formed. Four pieces of paper with a unique number (1, 2, 3, or 4) written on them are placed in a container. Each member of a block randomly selects a single number from the container and discards the number. This number represents the treatment (software product) that the analyst will receive. Note that the procedure accounts for the possible individual differences in analyst capability through the blocking of average times; thus, we are controlling for individual differences in capability. As an extreme case, a block can be comprised of a single analyst. In this case, the analysts will have all four treatments (software products) administered in randomly selected order. The

random application of the treatments helps eliminate the possible interference (learning, fatigue, loss of interest, etc.) of a fixed order of application. Note this randomized block experiment with a single subject in a block (20 blocks) leads to 80 data points (20 blocks × 4 products), while the first block experiment (5 blocks) leads to 20 data points (5 blocks × 4 products).

- **Factorial Design**: A factorial design is one where we consider more than one factor in the experiment. For example, suppose we are interested in assessing the capability of our customer service representatives by considering both training (standard and special) and their freedom status (prisoners or non-prisoners) for SC. Factorial designs will allow us to perform this analysis with two or more factors, simultaneously. Consider the customer representative training problem. It has 2 treatments in each of 2 factors, resulting in a total of 4 unique treatment combinations, sometimes referred to as a cell: prisoner/special training, prisoner/standard training, non-prisoner/special training, and non-prisoner/standard training. To conduct this experimental design, we randomly select an equal number of prisoners and non-prisoners and subject equal numbers to special training and standard training. So, if we randomly choose 12 prisoners and 12 non-prisoners from SC (a total of 24 subjects), we then allocate equal numbers of prisoners and non-prisoners to the 4 treatment combinations—6 observations in each treatment. This type of design results in **replications** for each cell, 6 to be exact. Replication is an important factor for testing the adequacy of models to explain behavior. It permits testing for *lack-of-fit*. Although it is an important topic in statistical analysis, it is beyond the scope of this introductory material.

There are many, many types of experimental designs that are used to study specific experimental effects. We have covered only a small number, but these are some of the most important and commonly used designs. The selection of a design will depend on the goals of the study that is being designed. Now for some examples of experimental design we have discussed.

6.8.1 Randomized Complete Block Design Example

Let us perform one of the experiments discussed above in the Randomized Complete Block Design. Our study will collect data in the form of task completion times from 20 randomly selected analysts. The analysts will be assigned to one of five blocks (A–E) by considering their average task performance times in the past 6 months. The consideration (blocking) of their *average task* times for the previous 6 months is accomplished by sorting the analysts on the *6 Month Task Average* key in Table 6.12. Groups of 4 analysts (A–E) will be selected and blocked until the list is exhausted, beginning with the top 4, and so on. Then analysts will be randomly assigned to one of 4 software products, within each block. Finally, a score will be recorded on their task time and the Excel analysis **ANOVA: Two-Factor without Replication** will be performed. This experimental design and results is shown in Table 6.13.

Table 6.12 Data for four software products experiment

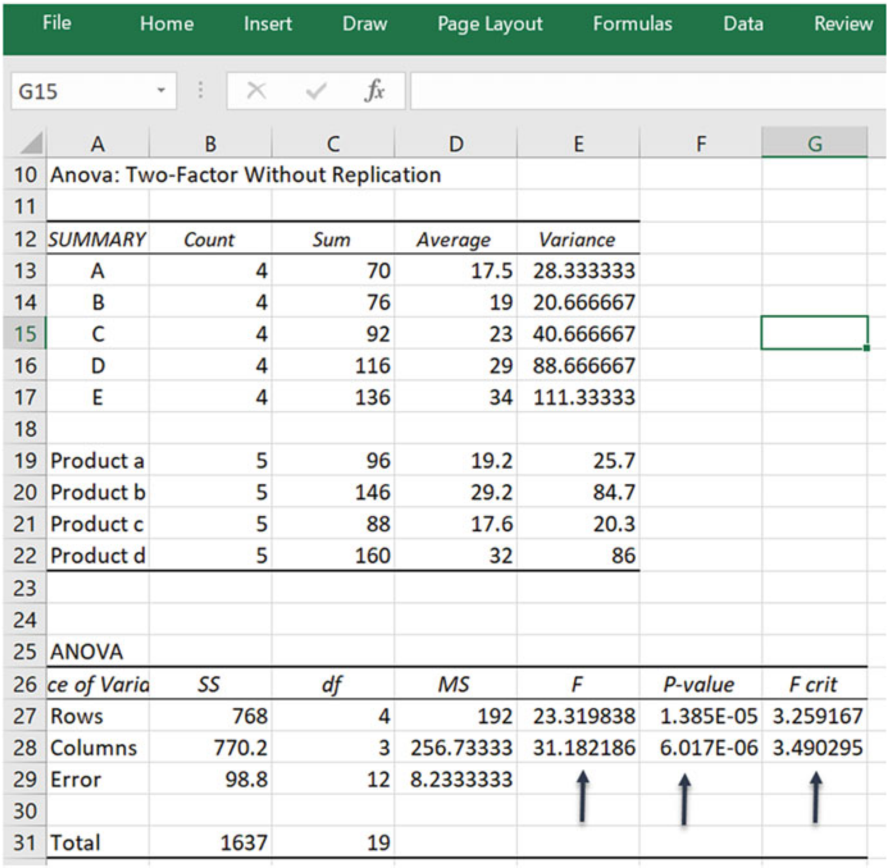
Obs. (Analysts)	6 month task average	Block assignment	Software treatment	Task time
1	12	A	d	23
2	13	A	a	14
3	13	A	c	12
4	13	A	b	21
5	16	B	a	16
6	17	B	d	25
7	17	B	b	20
8	18	B	c	15
9	21	C	c	18
10	22	C	d	29
11	23	C	a	17
12	23	C	b	28
13	28	D	c	19
14	28	D	a	23
15	29	D	b	36
16	31	D	d	38
17	35	E	d	45
18	37	E	b	41
19	39	E	c	24
20	40	E	a	26

Although we are using the Two-Factor procedure, we are interested only in a single factor—the four software product treatments. Our blocking procedure is more an attempt to focus our experiment by eliminating unintended influences (the skill of the analyst prior to the experiment), than it is to explicitly study the effect of more capable analysts on task times. We have is one analyst from each block being counted in the average time for each product; we avoid the possibility that all the analysts evaluating a product could possibly come from single skill block—fastest or slowest. Table 6.12 shows the 20 analysts, their previous 6-month average task scores, the five blocks the analysts are assigned to, the software product they are tested on, and the task time scores they record in the experiment. Figure 6.8 shows the data that Excel will use to perform the ANOVA. Note that *analyst no.* 1 in Block A (see Table 6.12) was randomly assigned *product d*. Cell C8 in Fig. 6.11 represents the score (26) of *analyst no.* 20 *on product a*.

We are now prepared to perform the ANOVA on the data, and we will use the Excel tool *ANOVA: Two-Factor without Replication* to test the null hypothesis that the task completion times for the various software products are no different. Figure 6.12 shows the dialog box to perform the analysis. The *Input Range* is the entire table, including labels, and the level of significance, α, is 0.05. This is the standard format for tables used in this type of analysis.

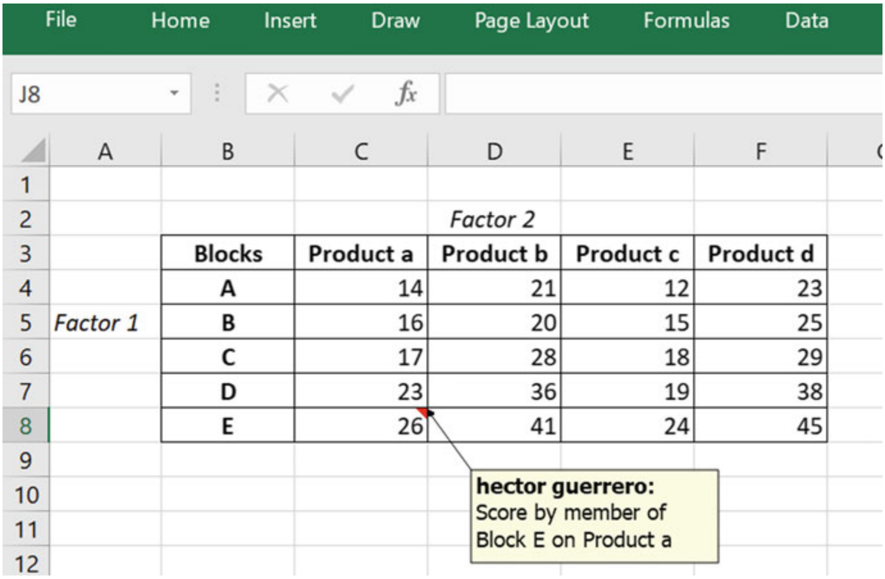
The results of the ANOVA are shown in Table 6.13. The upper section of the output, entitled *SUMMARY*, shows descriptive statistics for the two factors in the

Table 6.13 ANOVA analyst example: Two-factor without replication



	A	B	C	D	E	F	G
10	Anova: Two-Factor Without Replication						
11							
12	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*		
13	A	4	70	17.5	28.333333		
14	B	4	76	19	20.666667		
15	C	4	92	23	40.666667		
16	D	4	116	29	88.666667		
17	E	4	136	34	111.33333		
18							
19	Product a	5	96	19.2	25.7		
20	Product b	5	146	29.2	84.7		
21	Product c	5	88	17.6	20.3		
22	Product d	5	160	32	86		
23							
24							
25	ANOVA						
26	*ce of Varia*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
27	Rows	768	4	192	23.319838	1.385E-05	3.259167
28	Columns	770.2	3	256.73333	31.182186	6.017E-06	3.490295
29	Error	98.8	12	8.2333333			
30							
31	Total	1637	19				

analysis—Groups (A–E) and Products (a–d). Recall that we will be interested only in the single factor, Products, and have used the blocks to mitigate the extraneous effects of skill. The section entitled ANOVA provides the statistics we need to either not-reject or reject the null hypothesis: there is no difference in the task completions times of the four software products. All that is necessary for us to reject the hypothesis is for one of the four software products task completion times to be significantly different from any or all the others. Why do we need ANOVA for this determination? Recall we used the t-Test procedures for comparison of pair-wise differences—two software products with one compared to another. Of course, there are six exhaustive pair-wise comparisons possible in this problem—a/b, a/c, a/d, b/c, b/d, and c/d. Thus, six tests would be necessary to exhaustively cover all possibilities. It is much easier to use ANOVA to accomplish the almost exact analysis as the t-Tests, especially as the number of pairwise comparisons begins to grow large.



Factor 2

	Blocks	Product a	Product b	Product c	Product d
	A	14	21	12	23
Factor 1	B	16	20	15	25
	C	17	28	18	29
	D	23	36	19	38
	E	26	41	24	45

hector guerrero:
Score by member of
Block E on Product a

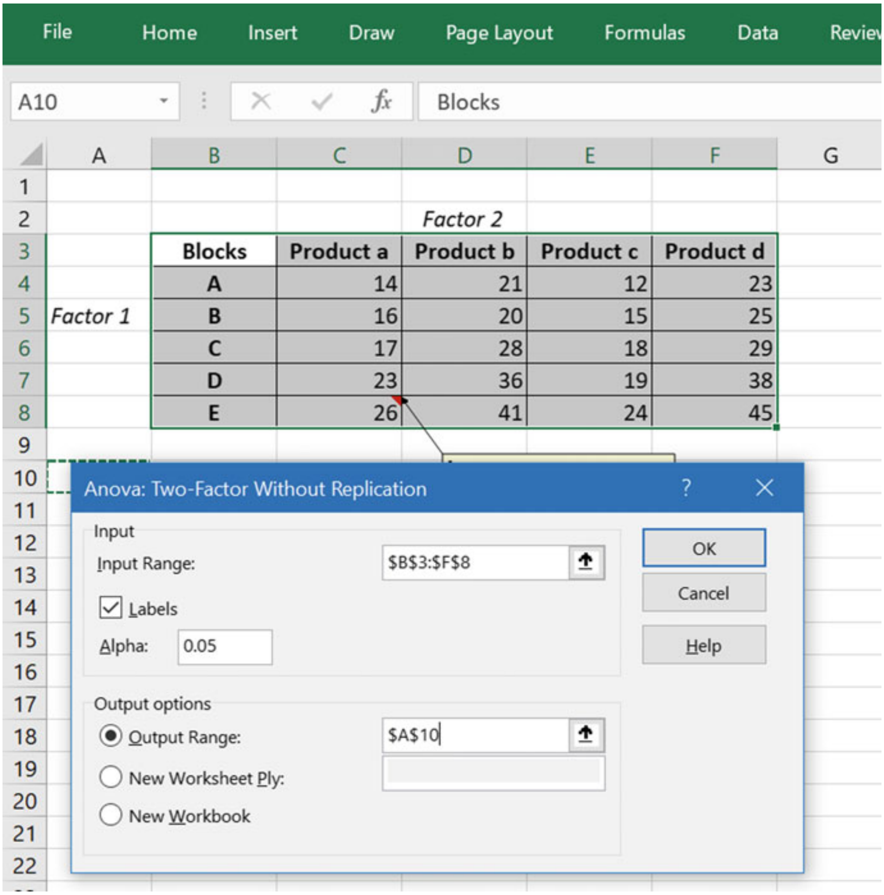
Fig. 6.11 Randomized complete block design analyst example

What is our verdict for the data? Do we reject the null? We are interested in the statistics associated with the sources of variation entitled *columns*. Why? Because in the original data used by Excel, the software product factor was located in the *columns* of the table. Each treatment, Product a–d, contained a column of data for the five block groups that were submitted to the experiment. The average for product a is 19.2 ([14 + 16 + 17 + 23 + 26]/5), and it is much smaller than the average for Product d, 32. Thus, we might expect a rejection of the null.

According to the analysis in Table 6.13, the F-Statistic, 31.182186 (cell E28) is much larger than the F critical, 3.490295 (cell G28). Also, our p-value, 0.00000601728 (cell F28), is much smaller than the assumed α of 0.05. Given the results, we clearly must reject the null hypothesis in favor of the alternative—*at least one* of the mean task completion times is significantly different from the others. If we reexamine the summary statistics in D19:D22 of Table 6.13, we see that at least two of our averages, 29.2 (b) and 32 (d), are much larger than the others, 19.2 (a) and 17.6 (c).

6.8.2 *Factorial Experimental Design Example*

Now, let us return to our *prisoner/non-prisoner* and *special training/no special training* two factors example. Suppose we collect a *new* set of data for an experimental study—24 observations of equal numbers of prisoners/non-prisoners and not-trained/trained. This implies a selection of two factors of interest: prisoner status



	Factor 2				
	Blocks	Product a	Product b	Product c	Product d
Factor 1	A	14	21	12	23
	B	16	20	15	25
	C	17	28	18	29
	D	23	36	19	38
	E	26	41	24	45

Anova: Two-Factor Without Replication

Input

Input Range: $B$3:$F$8

☑ Labels

Alpha: 0.05

Output options

● Output Range: $A$10

○ New Worksheet Ply:

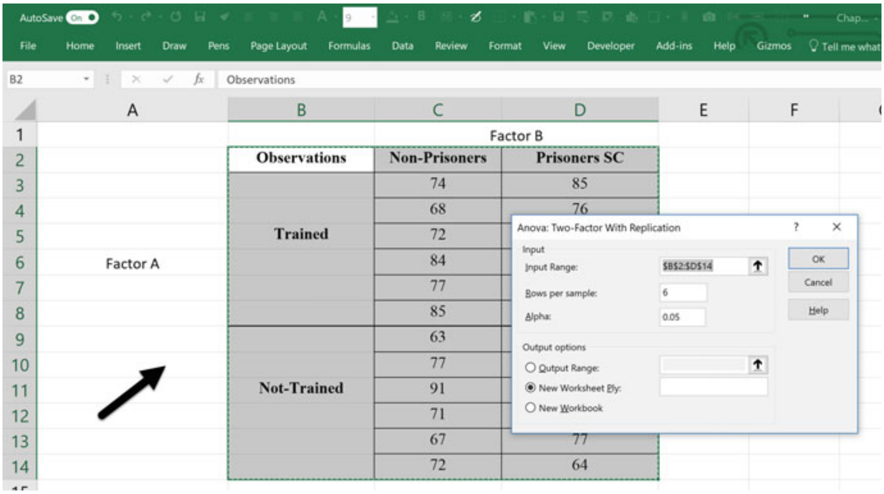
○ New Workbook

Fig. 6.12 Dialog box or ANOVA: Two-factor without replication

and training (see Fig. 6.13). The treatments for prisoner status are *prisoner* and *non-prisoner*, while the treatments for training are *trained* and *not-trained*. The four cells formed by the treatments each contain six replications (unique individual scores) and lead to another type of ANOVA—**ANOVA: Two-Factor with Replication**.

Table 6.14 shows the 24 observations in the two-factor format, and Table 6.15 shows the result of the ANOVA. The last section in Table 6.15, entitled ANOVA, provides the F-Statistics (E40:E42) and p-values (cells F40:F42) to reject the null hypotheses related to the effect of both factors. In general, the null hypotheses states that the various treatments of the factors do not lead to significantly different averages for the scores.

Factor A (Training) and Factor B (Prisoner Status) are represented by the sources of variation entitled *Sample* and *Columns*, respectively. Factor A has an F-Statistic of 1.402199 (cell E40) and a critical value of 4.351244 (cell G40), thus we *cannot*



	A	B	C	D
1			Factor B	
2		Observations	Non-Prisoners	Prisoners SC
3			74	85
4			68	76
5		Trained	72	
6	Factor A		84	
7			77	
8			85	
9			63	
10			77	
11		Not-Trained	91	
12			71	
13			67	77
14			72	64

Anova: Two-Factor With Replication

Input

Input Range: $B$2:$D$14

Rows per sample: 6

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Fig. 6.13 Format for Two-factor with replication analysis

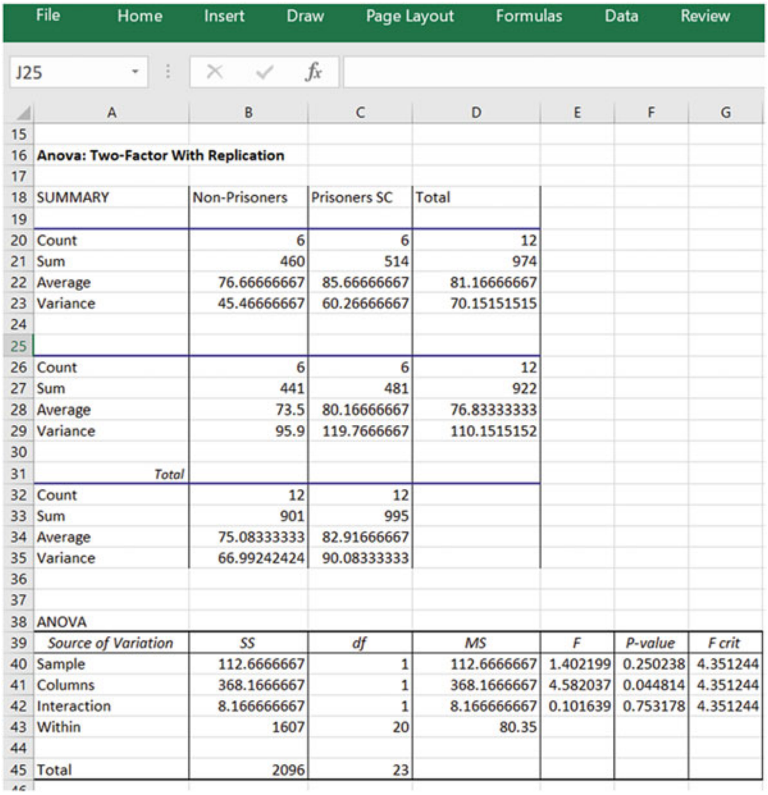
Table 6.14 Training data revisited

		Factor B	
	Observations	Non-prisoners	Prisoners SC
		74	85
		68	76
	Trained	72	87
Factor A	(Special)	84	92
		77	96
		85	78
		63	73
		77	88
	Not-trained	91	85
	(Standard)	71	94
		67	77
		72	64

reject the null. The p-value, 0.250238 (cell F40), is much larger than the assumed α of 0.05.

Factor B has an F-Statistic of 4.582037 (cell E41) that is slightly larger than the critical value of 4.351244 (cell G41). Also, the p-value, 0.044814 (cell F41), is slightly smaller than 0.05. Therefore, for Factor B we can reject the null hypothesis, but not with overwhelming conviction. Although the rule for rejection is quite clear, a result similar to the one we have experienced with Factor B might suggest that further experimentation is in order. Finally, the interaction of the factors does not lead us to reject the null. The F-Statistic is rather small, 0.101639 (cell E42), compared to the critical value, 4.351244 (cell G42).

Table 6.15 ANOVA: Two-factor with replication results



	A	B	C	D	E	F	G
15							
16	**Anova: Two-Factor With Replication**						
17							
18	SUMMARY	Non-Prisoners	Prisoners SC	Total			
19							
20	Count	6	6	12			
21	Sum	460	514	974			
22	Average	76.66666667	85.66666667	81.16666667			
23	Variance	45.46666667	60.26666667	70.15151515			
24							
25							
26	Count	6	6	12			
27	Sum	441	481	922			
28	Average	73.5	80.16666667	76.83333333			
29	Variance	95.9	119.7666667	110.1515152			
30							
31	*Total*						
32	Count	12	12				
33	Sum	901	995				
34	Average	75.08333333	82.91666667				
35	Variance	66.99242424	90.08333333				
36							
37							
38	ANOVA						
39	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
40	Sample	112.6666667	1	112.6666667	1.402199	0.250238	4.351244
41	Columns	368.1666667	1	368.1666667	4.582037	0.044814	4.351244
42	Interaction	8.166666667	1	8.166666667	0.101639	0.753178	4.351244
43	Within	1607	20	80.35			
44							
45	Total	2096	23				

6.9 Summary

The use of inferential statistics is invaluable in analysis and research. Inferential statistics allows us to infer characteristics for a population from the data obtained in a sample. We are often forced to collect sample data because the cost and time required in measuring the characteristics of a population can be prohibitive.

Inferential statistics also provides techniques for quantifying the inherent uncertainty associated with using samples to specify population characteristics. It does not eliminate the uncertainty due to sampling, but it can provide a quantitative measure for the uncertainty we face about our conclusions for the data analysis.

Throughout Chap. 6 we have focused on analyses that involve a variety of data types—categorical, ordinal, interval, and rational. Statistical studies usually involve a rich variety of data types that must be considered simultaneously to answer our questions or to investigate our beliefs. To this end, statisticians have developed a highly structured process of analysis known as tests of hypothesis to formally test the veracity of a researcher's beliefs about behavior. A hypothesis and its alternative are posited then tested by examining data collected in observational or experimental studies. We then construct a test to determine if we can reject the null hypothesis based on the results of the analysis.

Much of this chapter focused on the selection of appropriate analyses to perform the tests of hypothesis. We began with the chi-squared test of independence of variables. This is a relatively simple, but useful, test performed on categorical variables. The z-Test and t-Test expanded our use of data from strictly categorical, to combinations of categorical and interval data types. Depending on our knowledge of the populations we are investigating, we execute the appropriate test of hypothesis, just as we did in the chi-squared. The t-Test was then extended to consider more complex situations through ANOVA. Analysis of variance is a powerful family of techniques for focusing on the effect of independent variables on some response variable. Finally, we discussed how design of experiments helps reduce ambiguity and confusion in ANOVA by focusing our analyses. A thoughtful design of experiments can provide an investigator with the tools for sharply focusing a study, so that the potential of confounding effects can be reduced.

Although application of these statistics appears to be difficult, it is actually very straight forward. Table 6.16 below provides a summary of the various tests presented in this chapter and the rules for rejection of the null hypothesis.

In the next chapter, we will begin our discussion of *Model Building* and *Simulation*—these models represent analogs of realistic situations and problems that we face daily. Our focus will be on *what-if* models. These models will allow us to incorporate the complex uncertainty related to important business factors, events, and outcomes. They will form the basis for rigorous experimentation. Rather than strictly gather empirical data, as we did in this chapter, we will collect data from our models that we can submit to statistical analysis. Yet, the analyses will be similar to the analyses we have performed in this chapter.

Table 6.16 Summary of test statistics used in inferential data analysis

Test statistic	Application	Rule for r*ejecting* null hypothesis
$\chi 2$ – Test of independence	Categorical data	$\chi 2$ (calculated) $\geq \chi 2\ \alpha$ (critical) *or* p-value $\leq \alpha$
z test	Two sample means of categorical and interval data combined	z stat $\geq$ z critical value z stat $\leq -$ z critical value *or* p-value $\leq \alpha$
t test	Two samples of unequal variance; small samples (< 30 observations)	t stat $\geq$ t critical value t stat $\leq -$ t critical value *or* p-value $\leq \alpha$
ANOVA: Single factor	Three or more sample means	F stat $\geq$ F critical value *or* p-value $\leq \alpha$
ANOVA: Two factor without replication	Randomized complete block design	Same as single factor
ANOVA: Two factor with replication	Factorial experimental design	Same as single factor

Key Terms

Sample	Paired or matched
Cause and effect	t-Test: Paired two-sample for means
Response variable	Estimation
Paired t-Test	Confidence intervals
Nominal	Standard error
Chi-square	Critical value
Test of independence	Single sample test of hypothesis
Contingency table	ANOVA
Counts	Main and interaction effects
Test of the null hypothesis	Factors
Alternative hypothesis	Levels
Independent	Single factor ANOVA
Reject the null hypothesis	F-statistic
Dependent	Critical F-value
$\chi 2$ statistic	Experimental design
$\chi 2$ α, α–level of significance	Observational studies
CHITEST(act. range, expect. range)	Experiment
z-Test: Two sample for means	Completely randomized design
z-Statistic	Experimental units
z-Critical one-tail	Randomized complete block design
z-Critical two-tail	Factorial design
$P(Z \leq z)$ one-tail and $P(Z \leq z)$ two tail	Replications
t-Test	ANOVA: Two-factor without
t-Test: Two-samples unequal variances	ANOVA: Two-factor with replication
Variances	

Problems and Exercises

1. Can you ever be totally sure of the *cause and effect* of one variable on another by employing *sampling*?—Y or N
2. Sampling errors can occur naturally, due to the uncertainty inherent in examining less than all constituents of a population—T or F?
3. A sample mean is an estimation of a population mean—T or F?
4. In our webpage example, what represents the treatments, and what represents the response variable?
5. A coffee shop opens in a week and is considering a choice among several brands of coffee, Medalla de Plata and Startles, as their single offering. They hope their choice will promote visits to the shop. What are the treatments and what is the response variable?

6. What does the Chi-square test of independence for categorical data attempt to suggest?
7. What does a contingency table show?
8. Perform a Chi-squared test on the following data. What do you conclude about the null hypothesis?

	Coffee drinks				
Customer type	Coffee	Latte	Cappuccino	Soy-based	Totals
Male	230	50	56	4	
Female	70	90	64	36	
Totals					600

9. What does a particular level of significance, $\alpha = 0.05$, in a test of hypothesis suggest?
10. In a Chi-squared test, if you calculate a p-value that is smaller than your desired α, what is concluded?
11. Describe the basic calculation to determine the expected value for a contingency table cell.
12. Perform tests on Table 6.17 data. What do you conclude about the test of hypothesis?

    (a) z-Test: Two Sample for Means
    (b) t-Test: Two Sample Unequal Variances
    (c) t-Test: Paired Two Sample for Means

13. Perform an ANOVA: Two-Factor Without Replication test of the blocked data in Table 6.18. What is your conclusion about the data?
14. *Advanced Problem*—A company that provides network services to small business has three locations. In the past they have experienced errors in their accounts receivable systems at all locations. They decide to test two systems for detecting accounting errors and make a selection based on the test results. The data in Table 6.19 represents samples of errors (columns 2–4) detected in accounts receivable information at three store locations. Column 5 shows the system used to detect errors. Perform an ANOVA analysis on the results. What is your conclusion about the data?
15. *Advanced Problem*—A transportation and logistics firm, Mar y Tierra (MyT), hires seamen and engineers, international and domestic, to serve on board its container ships. The company has in the past accepted the worker's credentials without an official investigation of veracity. This has led to problems with workers lying about, or exaggerating, their service history, a very important concern for MyT. MyT has decided to hire a consultant to design an experiment to determine the extent of the problem. Some managers at MyT believe that the international workers may be exaggerating their service, since it is not easily verified. A test for first-class engineers is devised and administered to 24 selected workers. Some of the workers are international and some are domestic. Also, some have previous experience with MyT and some do not. The consultant

randomly selects six employees to test in each of the four categories—International/Experience with MyT, International/No Experience with MyT, etc. A Proficiency exam is administered to all the engineers and it is assumed that if there is little difference between the workers scores then their concern is unfounded. If the scores are significantly different (0.05 level), then their concern is well founded. What is your conclusion about the exam data in Table 6.20 and differences among workers?

Table 6.17 Two sample data

Sample 1	Sample 2
83	85
73	94
86	77
90	64
84	90
69	89
71	73
95	84
83	80
93	91
74	76
72	87
88	92
87	67
72	71
82	73
79	98
83	90
74	75
81	74
76	83
63	89
86	78
71	72
83	85
76	76
96	91
77	79
73	65
80	87
86	81
77	84
70	79
92	81
80	68
65	93

Table 6.18 Two factor data

		Factor 2			
	Blocks	W	X	Y	Z
Factor 1	A	14	21	12	23
	B	12	20	15	25
	C	17	18	23	19
	D	23	36	19	38
	E	26	21	24	32

Table 6.19 Three sample data

Obs	Loc. 1	Loc. 2	Loc. 3	Type of system
1	24	21	17	A
2	14	12	6	A
3	12	24	8	A
4	23	11	9	A
5	17	18	11	A
6	29	28	3	A
8	18	21	21	A
9	31	25	19	A
10	25	23	9	A
11	13	19	18	A
12	32	40	11	A
13	18	21	4	B
14	21	16	7	B
15	21	17	17	B
16	14	18	11	B
17	6	15	9	B
18	15	13	10	B
19	9	9	3	B
20	12	10	6	B
21	15	19	15	B
22	12	11	9	B
23	12	9	13	B
24	17	13	9	B

Table 6.20 Multi-factor data

Observations	Foreign	Domestic
	72	82
Previous employment	67	76
With MyT	72	85
	84	92
	77	96
	85	78
	63	73
No previous experience	77	88
With MyT	91	85
	71	94
	67	77
	72	64