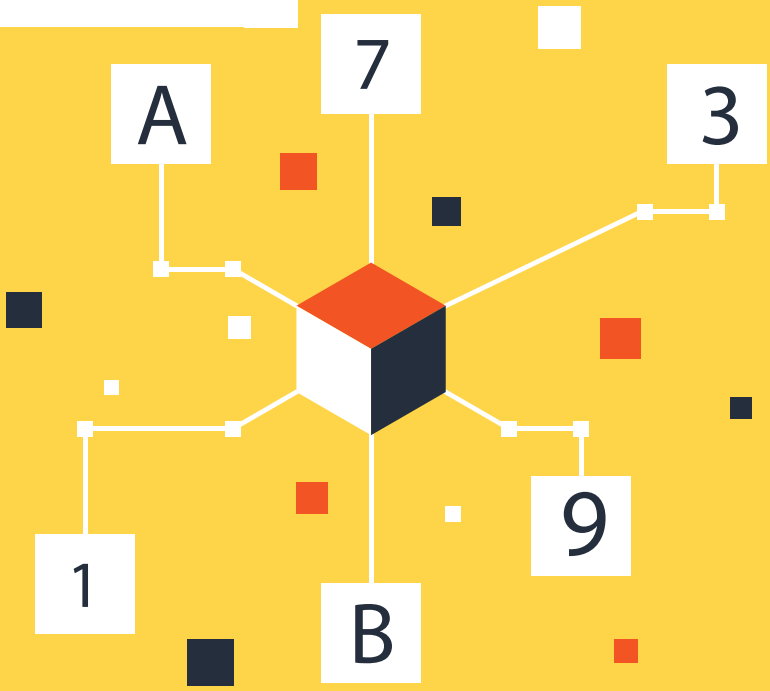
INTRODUCTION TO STATISTICS

COUNTY OF LOS ANGELES



Los Angeles County Human Resources
YOUR CAREER STARTS HERE.

Table of Contents

Introduction

Welcome to learning about basic statistics. The County of Los Angeles Department of Human Resources is pleased to offer this information guide. We hope you find this guide useful in enhancing your understanding of statistics.

Modern organizations rely on massive amounts of data that are used to make important decisions. Collecting data is only the beginning. Organizations must also be able to understand what the data mean in relation to the issue being investigated and how that knowledge can be used to make decisions with a reasonable amount of confidence. Statistical methods are critical to organizations that want to derive meaning from their data and obtain support for important decisions.

Purpose

This guide has been developed to introduce you to the fundamentals of statistics. The primary emphasis of this guide is on *descriptive statistics*, with a brief discussion of *inferential statistics*. The goals of this guide are to present you with **general information** related to this subject and to spark interest in further exploration of the concepts presented here.

Disclaimer

This guide has been developed to serve as an introduction to basic statistics. Because statistics are methods for organizing and analyzing data, readers should first review the County's guide titled *Introduction to Data Collection*. Although this introduction to statistics presents useful and practical information from this subject area, there is **no guarantee** that someone who reads this guide will be able to perform better on the job or on a County examination. By merely using this guide, you consent to understanding and agreeing with this disclaimer.

Note: Information in this guide is current as of April 2007 and will be updated periodically.

Who should use this guide?

This guide is written for those new to statistics or those in need of a refresher. This guide is intended for anyone who collects (or will collect) data and wants to acquire some of the tools necessary to interpret the data and make informed decisions. An understanding of basic mathematical principles (i.e., addition, subtraction, multiplication, division, fractions, square roots, percentages, and exponents) is necessary to fully understand this guide. Readers who are not confident in their understanding of these concepts may want to do a quick review prior to using this guide.

Terminology

Statistics is a field of study that uses specific terminology that may be unfamiliar to some readers. A glossary of terms is provided in the appendix to assist readers with understanding unfamiliar terms (see 25-26). From this point on, words that are italicized are listed in the glossary for easy reference. For the sake of clarity, each term will be italicized the first time it appears and in standard print from there on. Every effort was made to avoid the use of technical terminology whenever possible.

What are statistics?

Statistics are more than just sets of numbers collected from various sources. Statistics is a label that represents a variety of techniques that can be used to help organize data in ways that make them easier to understand. In another guide in this series titled *Introduction to Data Collection*, data were defined as any piece(s) of information, such as numbers, text, graphics (e.g., charts) and verbal exchanges (e.g., interviews). Data are usually collected because they represent something you are interested in studying or knowing more about. You may collect data from hundreds or even thousands of sources. Simply looking at a table full of numbers may not tell you much about what you are trying to learn from your data, and this is where statistics can help.

Statistics provide a means to organize your data and draw conclusions with greater confidence. Statistics help you understand characteristics of your data, such as how spread out they are and where they tend to cluster, and provide a means for understanding possible relationships between sets of data. Hence, statistics can be either *descriptive* or *inferential*.

- **Descriptive statistics** are used to describe the characteristics of your data set.
- **Inferential statistics** provide information to support inferences (predictions) about what you can expect in similar circumstances based on a subset of all possible values (i.e., a data sample).

By now, you may be asking yourself how knowledge of statistics will help you to succeed in the workplace. Statistics help you to evaluate information at a more in-depth level, which can lead to more complete and accurate conclusions or recommendations. The results of statistical analysis can be used to:

- Evaluate a new or existing workplace initiative or program.
- Bolster your recommendation for a course of action.
- Add credibility to presentations you make.
- Enhance your understanding of the information used by managers to make decisions.

Getting Started

What is a variable?

A *variable* is anything that can take on different values (e.g., size, color, quantity, etc.). The data you collect represent things you are interested in learning about. For example, you may want to know how satisfied County residents are with the services they receive, which activities park visitors prefer most, or what types of books library visitors prefer to read. In order for a question to be worthy of investigation, there must be *variability* (i.e., different values) in the data you collect. Without variability you will not be able to identify the differences that can lead to a more complete understanding of the topic under study.

To answer questions about the relationships between different variables, you can apply statistical methods. However, your ability to use different statistical methods depends on how you measure the variables you are interested in learning about. Next, we will examine measurement and how it relates to the types of data you may collect.

Scales of Measurement

Scales of Measurement refers to ways in which variables are defined. "Measurement" is often thought of as measuring something's length (e.g., a ruler) and/or quantity (e.g., a measuring cup). In statistics, measurement is used more broadly to relate to the process of assigning symbols, such as numbers or category labels, to represent quantities (*scaling*) or categories (*classification*) of the important variables under study.

By assigning symbols to your data, you may analyze the data statistically. The statistics available for your use depend on the scale of measurement you use when collecting your data. The three scales of measurement you are most likely to encounter in the workplace are: (a) *nominal*, (b) *ordinal*, and (c) *interval*. The three scales differ in the amount of variability they allow. Nominal measurement scales allow the least variability, followed by ordinal scales. Of the three measurement scales, interval scales are at the highest level of measurement because they allow for the greatest amount of variability. When examining your data, variability is good because it makes it easier to identify important trends in your set of data (discussed later in the section on descriptive statistics), and allows you to make more focused comparisons between multiple sets of data (discussed later in the section on inferential statistics). The three types of measurement scales are discussed in more detail below.

Nominal Measurement Scales

Nominal essentially means naming and it is consistent with the classification process of measurement. Thus, nominal data are used to place data into classes or categories. The categories do not need to be arranged in any specific order. The classification system you adopt depends on what you consider to be the important characteristics of things you are classifying. For example, you might be interested in classifying job applicants by their highest level of formal education as follows:

- Applicants who have less than a high school diploma or G.E.D.
- Applicants who have a high school diploma or G.E.D.
- Applicants who have some college without having earned a degree.
- Applicants who have earned an associate's degree.
- Applicants who have at least a bachelor's degree.

Another characteristic of nominal data is that they are *discrete.* That is, people/objects can be placed in only one of the available categories.

> **Example:** It would be relatively easy to classify anyone as belonging to one, and only one, of the available education level categories we just listed. Notice that the terms "less than" and "at least" include all individuals who are at either end of the education classification range. That is, the first category includes all individuals who have not earned a high school diploma or G.E.D. and the last category includes everyone who has earned a bachelor's degree or higher.

If you find that people/objects may be placed into more than one category, you should revisit your classification system and make clearer distinctions between the classes. Additionally, if some of your people/objects cannot be placed into any of the available categories, you will need to extend your classification system so that it includes all of the people/objects you are interested in classifying.

With nominal measurement scales, the data you will actually collect and analyze will be in the form of *frequencies.* Frequencies are the number of people/objects assigned to each nominal category (i.e., how frequently a certain characteristic is observed in a set of data).

> **Example:** Let us assume you reviewed 250 job applications and sorted them according to the education level of each applicant. Your sorting procedure resulted in the following:

Table 1
Applicant education level.

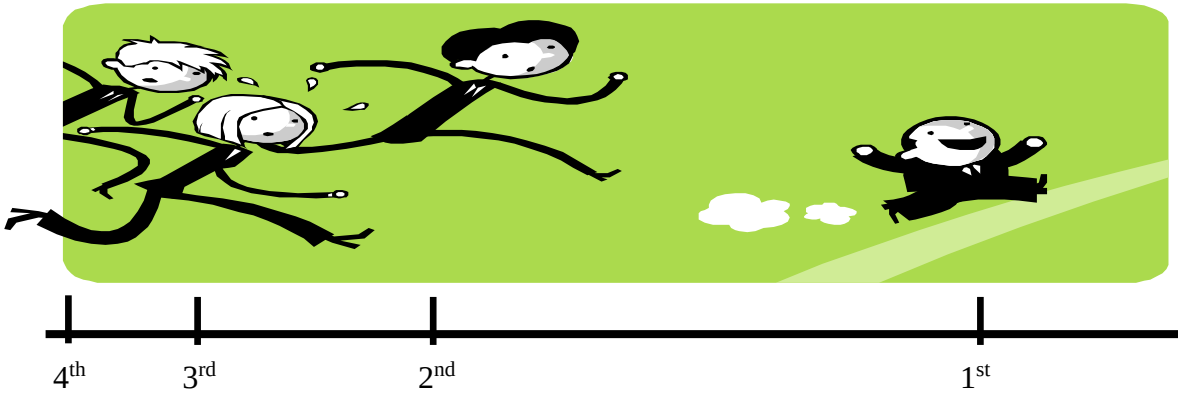
< high school/G.E.D.	high school/G.E.D.	some college	associate's	≥ bachelor's
20	120	80	30	0

> Below each category label in Table 1 is the *frequency* of applicants who were classified as belonging to that category. Hence, 20 applicants have less than a high school diploma, 120 have high school or G.E.D., 80 have some college, 30 have an associate's degree, and none have a bachelor's or higher degree. Notice that if we change the order of the categories we would still get the same frequency for each category. Thus, it is advantageous to order your categories from low to high to assist in the categorizing process, but the category order will make no difference with respect to the meaning of the data.

Ordinal Measurement Scales

Ordinal measurement scales order your data set from the lowest value to the highest value (or from highest to lowest). In other words, the data can be *ranked* according to their standing on some characteristic of interest. Ranking can take on different forms, such as place in a foot race (e.g., first, second, third, etc.), job classification (e.g., Clerk I, Clerk II, Clerk III, etc.), and Sheriff's Deputy rank (e.g., sergeant, lieutenant, captain, etc.). Data collected using an ordinal scale of measurement are also *discrete* because a person/object must be first, second, third, etc. A person/object cannot be both first and third on an ordinal scale. The following example helps to illustrate this point.

> **Example:** A runner in a race may come in first, second, third, or fourth place.



The actual spacing between the values of ordinal data is not specified. That is to say, the scale does not tell you if the number of seconds between first and second place is the same as the number of seconds between second and third place. This scale only tells you the order.

Interval Measurement Scales

Interval scales have equal spacing between values. *Spacing* refers to the distance between points on a scale, which represents the amount of the characteristic being measured. Hence, the spacing between the values 1 and 2 represents the same degree of change in the amount of the characteristic being measured as the spacing between the values 2 and 3.

> **Example:** If you used a survey to collect data from a sample of 200 County employees on their opinions about a new policy, you would most likely be utilizing an interval scale. Suppose the sample of County employees in your survey provided ratings on a 7-point scale, ranging from 1 = *I really don't like the new policy* to 7 = *I really like the new policy.* Suppose your data collection procedure resulted in Table 2.
>
> Table 2
> *County employee opinions about new policy.*
>
1	**2**	**3**	**4**	**5**	**6**	**7**
8	12	20	35	47	50	28
>
> From Table 2, you can see that 8 survey participants provided a rating of 1, 12 provided a rating of 2, and so on. Your question has to do with how much County employees like the new policy. You are not classifying the responses into discrete categories, such as *like* and *dislike.* Furthermore, you are not ranking how much they like the policy in relation to other policies. From the information displayed in Table 2, you can see that each survey participant had seven response choices, which allows for more variability in the responses across all survey participants. It is also possible to compute the *mean* (average) rating across all survey participants. Additionally, you can calculate the degree to which each employee's rating differs from the mean rating of all survey participants. These concepts will be discussed later in the section on descriptive statistics.

The above example illustrates some important characteristics of interval scales. First, they are used to measure the amount of a specific characteristic (i.e., satisfaction, ability, agreement, etc.). Additionally, the data collected using an interval scale can be manipulated in various ways that distinguish interval scales from lower level measurement scales (i.e., nominal and ordinal). Specifically, you can calculate the mean for a set of data measured on an interval scale, calculate statistics that indicate the amount of variability of the individual data points around the mean, and perform a variety of other mathematical and statistical operations. Data measured on an interval scale are also *continuous* because the mean for a particular set of data can assume any number of potential values, based on the unique responses of the participants sampled.

In order to further enhance your understanding of the differences between the three types of measurement scales, consider the folowing examples:

Nominal example: Suppose your department has several branch offices. If each office provides the same set of services, you could ask if people who visit one office utilize more of a particular service than individuals who visit other offices. The services used would be the categories (i.e., classes), and the number of individuals who utilize the various services at each office would be the frequencies. To answer this sort of question statistically, the *chi-square test* (which will be described later in this guide) can be used to determine whether the number of individuals using the services provided in each office is consistent with what you would expect if there was no difference between the offices.

Ordinal example: Suppose your office conducts a survey every two years to determine which benefits employees believe are most important. Employees rank each benefit as being the most important, the second most important, and so on. The responses of all the employees are combined, and the average importance ranking for each of the benefits is determined. What sorts of questions might you be able to answer with this type of data? If you divided the employees that participated in a single survey administration into subgroups (e.g., by gender, marital status, age, etc.), you could ask whether different groups rank the benefits differently in terms of their importance. Additionally, you might want to know if the overall importance employees place on the benefits changes over time. In this case, you would compare the rankings across different two-year administrations to learn whether the rank order changes significantly over the years. Although not presented in this guide, there are statistical methods (usually referred to as *nonparametric* statistics) that can be used to determine whether the relative rank orders differ (statistically) between groups. Interested readers are encouraged to learn more about these statistical methods.

Interval example: Suppose you wanted to know whether job applicants' ratings of how much they like the County employee benefits package is related to whether they are willing to accept a job offer. Your data consist of ratings on a 7-point scale assessing applicants' reactions to the benefits package (i.e., 1 = *Dislike very much* to 7 = *Like very much*) and responses to a 7-point scale that asks to what extent they would be willing to accept a job offer from the County (i.e., 1 = *Definitely would not accept a job offer* to 7 = *Absolutely would accept a job offer*). Both are examples of interval scales of measurement. Each is designed to measure the amount of something (i.e., liking of the benefits package and willingness to accept a job offer). A mean rating across all applicants who responded to each question can be calculated, and the amount of variability in responses around the mean can also be computed. Finally, statistics that determine whether there is a relationship between reactions to the benefit plan and willingness to accept a job offer (which will be described later in this guide) can be computed.

Table 3
Scales of Measurement Summary

Scales of Measurement	Characteristics	Data Collected	Questions
Nominal	Discrete (i.e., can belong to only one category)	Frequencies (i.e., how many individuals belong to each category)	• Are observed differences between categories larger than would be expected by chance?
Ordinal	Discrete (e.g., either first, second, or third, etc.)	Relative rank order (i.e., order of people, data, or things on some characteristic)	• Is the relative rank order the same across groups of people/objects? • Is there a relationship between the rank order on one variable and the rank order on another variable?
Interval	Continuous (i.e., can take on any number of potential values)	Amount of a particular characteristic (e.g., job satisfaction, support for a new strategic plan, support for a change initiative, etc.)	• Do observed differences in the amount of some characteristic indicate that groups of people/objects differ in terms of that characteristic? • Is there a relationship between the amount of one characteristic and the amount of another characteristic? • All questions presented for nominal and ordinal scales.

You are encouraged to collect interval-level data whenever possible because defining intervals allows you to use the most common descriptive and inferential statistics. However, as our examples clearly illustrate, both ordinal and nominal data are capable of yielding information for answering important research questions. In addition to the scale of measurement used to collect your data, another important consideration is how you go about collecting your data.

Sampling

As described in the *Introduction to Data Collection* guide, data may be collected from different sources: (a) document searches, (b) interviews, (c) surveys, and (d) observations. When you want to draw conclusions about a total population of people/objects based on a data sample collected from a subset of the population's members, you also need to think about the technique you use to obtain your data sample. What you are trying to do with statistics is make statements about characteristics of the total population from your data sample. One step necessary to support such statements is obtaining a *random sample* of data from the population of interest. If you use random sampling, you will be able to make accurate predictions about the characteristics of the total population. This is important because you are almost never interested in knowing only about your sample. Let us examine how you might go about obtaining a random sample from each of the four data sources listed above:

1. Document searches might involve reviewing records on file that have accumulated over a period of time. In order to obtain a random sample of such information, you might randomly select months from within a range of years, days within a range of months, and so on. The sampling method will be determined by the period of time you are interested in (i.e., when the documents were produced), the amount of data that can be collected using a particular sampling plan (i.e., how many documents will be available to sample from), and other possible constraints.

2. Interviews may also be conducted at random. You could generate a list of names of all County residents and randomly select a sample of residents to contact for an interview.

3. The same method employed to obtain a random sample of interview participants could also be used to select names of residents to participate in a survey.

4. A random sample of observations could be made to determine whether people who visit County recreational facilities prefer basketball, volleyball, soccer, or baseball. You could select random parks and days of the year to actually go out and count the number of games of basketball, volleyball, soccer, or baseball going on during a specified time period.

The important point with sampling is to select randomly from within the full range of potential values of the variables you are studying.

> **Example:** If you were to go to a park to observe the number of visitors playing sports on only one day of the week or during only one month out of the year, you might obtain data that do not truly reflect the total population of residents who visit County parks. You might expect more people to be playing baseball during the regular baseball season (i.e., the summer months). If your observations only take place during baseball season, you might get the false impression that the majority of park visitors prefer baseball. Different sports teams may also prefer to play on different days of the week, so randomly choosing days is also important.

Note: In some situations it may not be possible or practical to obtain a truly random sample from the population you are interested in learning about. However, making every effort to collect a random sample will allow you to make more accurate predictions from your data. There are other sampling techniques available that are beyond the scope of this guide. Interested readers can find material on quasi-experimental methods that describe non-random sampling techniques that have been found useful in applied settings.

Statistics give you information about how your data are distributed (i.e., arranged around some central point) within a particular population. For instance, you would expect some members of a population to be of average height, some to be above average in height, and some to be below average in height. Because you want to make statements about the population from sample data, it is extremely important that the data you collect accurately represent the population of interest (i.e., have about the same central point and distribution of values around the central point). The following section deals with the topic of data distributions and what they can tell you about populations.

Data Distributions

Because you are dealing with variables, the data you collect will not all be identical; they will be spread out across a range of different values. The way data are distributed within a particular population provides information about the characteristics of that population and helps to distinguish it from other populations. The following example helps to illustrate this point.

Example: After implementing a new procedure, you decide to survey a sample of residents to learn about their perceptions of the quality of services they receive. You might use a rating scale similar to the following:

5 = excellent
4 = good
3 = acceptable
2 = needs improvement
1 = poor

You would not expect every person surveyed to provide the exact same rating to each item in the survey. You are looking for trends that indicate that, on average, County residents view service quality to be at a certain level. Although individual data points can be distributed across a range of values, you can expect most of the ratings to cluster around a certain point in the distribution.

Suppose you obtain the data recorded in Table 4. Notice that the far left column contains numbers representing each person who responded to the survey and the ratings for each of five questions are placed in the remaining columns.

Table 4
Responses to five questions about satisfaction with County services.

Respondent	Q1	Q2	Q3	Q4	Q5
1	3	5	2	5	5
2	1	2	5	3	3
3	4	5	5	3	2
4	3	5	4	4	3
5	3	5	1	3	5
6	2	4	3	3	4
7	3	4	2	1	5
8	5	5	5	5	3
9	3	3	5	3	4
10	4	5	2	2	5
11	2	5	3	2	3
12	5	4	3	1	4
13	3	4	1	3	4
14	2	3	1	4	4
15	4	5	1	3	5

Look at the distribution of ratings for the first question (Q1). As you can see, simply looking at the list of numbers tells you little about how residents in general feel about that particular question. However, when you arrange the data in terms of the frequency with which each value occurs, you can begin to see trends in the data. The following arrangement of numbers from the responses indicated under Q1 in Table 4 is an example of a *frequency distribution* for responses to the first question:

1
2 2 2
3 3 3 3 3 3
4 4 4
5 5

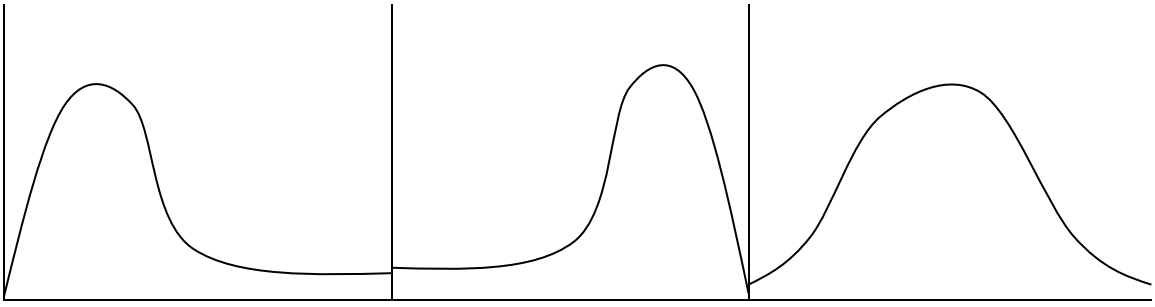
Based on these data, you might conclude that most County residents view the service they receive on the first question as *acceptable*, although scores span the entire range of available options. The distribution in this example appears approximately *normal*, in that most of the ratings tend to cluster around the central value and taper off *symmetrically* in the direction of both ends (i.e., 1 and 5).

Now, let us look at the distribution of responses to the second question (Q2).

2
3 3
4 4 4 4
5 5 5 5 5 5 5

Most of the values for this question tend toward the high end of the scale. Thus, you can easily see that the majority of individuals surveyed view the quality of service they receive on this question as being from *good* to *excellent*. This is called a *skewed* distribution because the majority of ratings tend to cluster near one end of the range of possible ratings. In the present example, you would say that the distribution is *negatively skewed* because the frequency of observations trails off as you approach the low response value end of the scale. If the values tended to cluster at the low response value end of the scale and trailed off toward the high response value end, you would say that the distribution was *positively skewed.* Notice also that, in this example, the range of the data does not include values of 1. This tells you that none of the individuals surveyed viewed the quality of service they received as being *poor*.

Example:



Positively Skewed Negatively Skewed Approximately Normal

If you collect a data sample that represents the population you are interested in learning about, the shape of the distributions of scores provides important information about that popluation. If you have collected your data from a sample that clearly represents the population you are interested in learning about (i.e., a random sample), you can expect the distributions of other samples from the same population to be quite similar.

From your observation of the distribution of scores for the questions (Table 4), you can estimate that, on average, County residents find the quality of service with respect to the first question on the questionnaire to be *acceptable*, but opinions vary from *poor* to *excellent*. On the other hand, the majority of residents viewed the quality of service on the second question to be *excellent* and none felt that the quality of service was *poor*. Knowing where the data tend to cluster and how spread out the data are provides you with important information about the characteristics of a population. This leads us to a discussion of measures of *central tendency* (i.e., where the data tend to cluster) and *dispersion* (i.e., how spread out the data are), which are statistics used to describe populations (i.e., *descriptive statistics).*

Descriptive Statistics

Descriptive statistics describe characteristics of samples. You rarely have every single value in a population available to you. Descriptive statistics obtained from random samples from a population provide the best estimate of the population *parameters* (i.e., true values that describe the characteristics of the population).

Central Tendency

Central tendency refers to the tendency of values in a distribution to cluster around some central value. There are three common measures of central tendency: *mean, median,* and *mode*.

Mean

The *mean* is the arithmetic average, which is calculated by summing the values and dividing by the total number in the sample. In our first example from Table 4 (Q1), the mean would be calculated as:

$$\frac{47}{15} = 3.1$$

Because the sum of the ratings for Q1 is 47 and 15 respondents provided ratings, the mean rating is 3.1 for this sample. If your data are representative of the population (i.e., taken from a random sample), this is the best estimate of the mean rating you would expect from the entire population of County residents.

The mean is important because it estimates the population parameter and sets an anchor around which all the other values in the sample are distributed. Because the mean is only calculated for continuous data sets (i.e., those recorded on an interval scale of measurement), it is most often used with measures of dispersion (discussed in the next section), which tell you whether the data tend to hold tightly to the mean or are spread out fairly widely. The mean, when used with measures of dispersion, helps you identify ratings that are far outside what would be expected if

they were made by members of the same population. If there are too many values that are outside the range you would expect, you might conclude that those people/objects come from a totally different population.

It makes no sense to calculate the mean for any scale of measurement below interval. To illustrate this important point, consider the following example:

> **Example:** If you label cars 1 = red, 2 = blue, 3 = black, 4 = yellow, 5 = green, 6 = brown, it does not make sense to say that the mean car color is 3.5

For data that are measured on lower than an interval scale, it makes more sense to talk about the next two measures of central tendency: *median* and *mode.*

Median

The median is the middle value if you lined up all the ratings in rank order (i.e., from lowest to highest). Hence, your data must be measured on at least an ordinal scale to calculate the median. For the data collected for the second question in our example (Table 4), the median would be calculated as follows:

> 2 3 3 4 4 4 4 **5** 5 5 5 5 5 5 5
>
> To find the central value, calculate the total number of ratings plus one and divide by two. Hence, 15 + 1 = 16 and $\frac{16}{2} = 8$. For our list of data, the median is the 8^{th} rank (see the bolded and underlined 5 above). Since there is an odd number of values in this distribution you would select the single value at the middle of the distribution as the median. Hence, the median in the present case is 5. If you would have been dealing with an even number of values, you would add the two middle values and divide by two to find the median.

The median is a useful measure of central tendency for many of the problems you are likely to face in the workplace. To illustrate this point, examine the distribution of the responses to the second question in Table 4. You probably want most residents to find the service quality to be either good or excellent. Hence, most of the distributions you are likely to be interested in will be skewed distributions. For such skewed distributions, the mean may not be the best estimate of what most residents think because *outliers* (data points far outside the range of typical values) can have a fairly strong impact on the mean, but they have little effect on the median.

In the above example, if you calculated the mean instead of the median you would obtain a value of 4.3. This is because the single 2 rating influences the mean, pulling it away from the center of the distribution. In contrast, the median provides you with a better estimate of what residents think about service quality. The median value indicates that at least half of the residents surveyed find service quality on this item to be excellent, whereas the mean would have only indicated that, on average, service quality is viewed as being good.

Mode

The mode is simply the most frequently occurring value (i.e., the value at the highest point in the distribution). In the sample distribution for the first question (Table 4), the mode is 3 and it is 5 in the distribution for the second question. If two adjacent values have the same frequency, the conventional approach is to take the average of the two values as the mode. If two non-adjacent values have the same frequency, you would say that the distribution is bimodal (i.e., has two modes).

In some situations the mode is all you care about. If you offer users of County services several filing options to obtain a permit (e.g., mail, Internet, email, etc.), you might only be concerned with the one chosen most often. The result could help you decide where to expend more resources to enhance the service option.

Dispersion

Dispersion has to do with how spread out the values in a set of data are. For instance, if everybody gets a score of 90 or better on a math test, the scores are narrowly dispersed. However, if the scores range from 0 to 100 the dispersion of scores is much greater. As is the case with all measures of dispersion, smaller numbers indicate less variability in the data. There are three measures of dispersion commonly used in statistics: (a) *range,* (b) *standard deviation,* and (c) *variance*.

Range

The range is simply the distance from the lowest to the highest value in the distribution. For the data for the first question in Table 4, the range is 5 – 1 = 4. For the second question, the range is 5 – 2 = 3.

The range can provide you with a limited amount of information about how spread out your set of data is. What it tells you depends on the total number of potential values in relation to what is actually recorded.

> **Example:** In our example of service quality ratings, the scale ranges from 1 to 5. This means you will never be able to obtain a value for the range greater than 4. However, if you are counting the number of people who visit the County's website each day for a week, the range of values could be quite large. If 1,000 people visit the County's website on Monday, 2,600 visit on Tuesday, 1,375 visit on Wednesday, 1,183 visit on Thursday, and 3,249 people visit on Friday, the range is 2,249 (i.e., 3,249 – 1,000).

For the service quality survey data, a range of 4 tells you that opinions of service quality are quite varied. If you obtained a range of only 2, you would conclude that most residents agree on the level of quality of the services they receive.

> **Example:** If your goal is to have most County residents view the quality of service they receive as *excellent*, you have two indicators of whether you have reached your goal. One indicator is a median service rating of 5 from a random sample of County residents (i.e., *excellent*). Another important indicator is how dispersed opinions are. If your range is 4 you know that one or more residents are not as satisfied with County services. However, with a median of 5 and a range of 2 you can confidently state that the majority of residents view service as better than *good.*

Standard Deviation

The standard deviation has to do with how the data are distributed around the *mean* of a distribution. Calculating the standard deviation of a distribution involves more math than was needed to find the range. Because you are working with a mean, you are limited to an interval scale of measurement. Let us use the third column in Table 4 to calculate the standard deviation for the third question (Q3) in our opinion survey.

Note: When calculating standard deviation and other more involved statistical calculations, you will most likely rely on Microsoft Excel or a statistical software program. In the spirit of understanding, we will discuss the standard deviation formula and demonstrate the steps involved in its calculation.

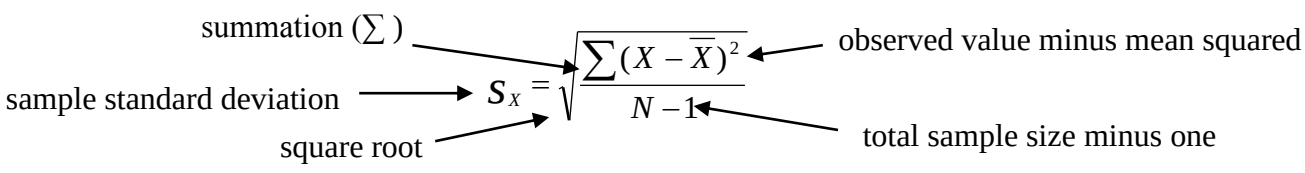
As you look at the data in Table 5, notice that the mean response to this question on your survey is subtracted from each individual's actual response (i.e., deviation from the mean). Then, the deviations are squared to make all the values positive.

Table 5

Deviations, squared deviations, and the sum of the squared deviations for question three.

Respondent	Mean	Observed Value	Deviation	Sq. Deviation
1	2.87	2	-0.87	0.76
2	2.87	5	2.13	4.54
3	2.87	5	2.13	4.54
4	2.87	4	1.13	1.28
5	2.87	1	-1.87	3.50
6	2.87	3	0.13	0.02
7	2.87	2	-0.87	0.76
8	2.87	5	2.13	4.54
9	2.87	5	2.13	4.54
10	2.87	2	-0.87	0.76
11	2.87	3	0.13	0.02
12	2.87	3	0.13	0.02
13	2.87	1	-1.87	3.50
14	2.87	1	-1.87	3.50
15	2.87	1	-1.87	3.50
				35.73

The formula for calculating the sample standard deviation is:



What this formula tells you is that the sample standard deviation is equal to the square root of the sum of the squared deviations divided by the total number of observed values minus one. From Table 5, you can add up the values in the far right column and divide by 14 (15 – 1). The calculation is as follows:

$$S_X = \sqrt{\frac{35.73}{14}} = 1.60$$

The standard deviation is an important concept related to data distributions (see page 9). For example, if your distribution is reasonably normal you can say:

- Approximately 2/3 of all observed values will fall within one standard deviation in either direction from the mean.
- Approximately 95% of all observations will fall within two standard deviations in either direction from the mean.

Because we expect data samples drawn from the same population (based on random sampling) to be similar with respect to their measures of central tendency and dispersion, a sample with a mean that is more than two standard deviations from the mean value obtained from another sample may indicate that the two samples were not drawn from the same population. This is a key concept in the use of *inferential statistics*, which rely on tests of *statistical significance.* These two topics will be explained in more detail in the next section on inferential statistics.

Example: Suppose you collected satisfaction survey data at two different locations within the County. One of the data collection locations (Area 1) is in a shopping mall. The other sample (Area 2) is collected on a corner of a busy intersection where there is considerable foot traffic.

Table 6
County Resident Service Satisfaction Survey Data for Areas 1 and 2.

Sample	Mean Rating	Standard Deviation
Area 1	4.30	0.90
Area 2	2.97	1.35

The two overlapping distributions pictured in Figure 1 help to illustrate the relationship between the two samples. Pay special attention to the locations of the mean on the horizontal line and the amount of overlap between the two distributions.

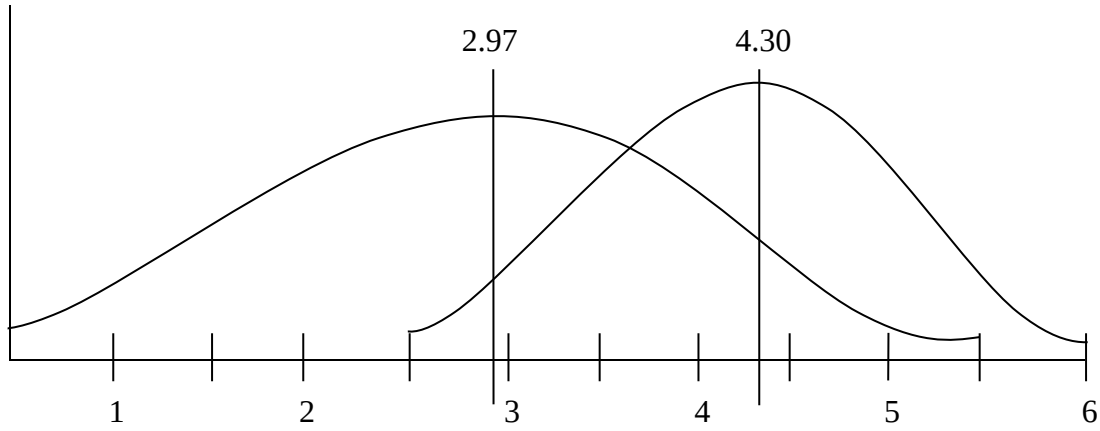


Figure 1
Example of data distributions from two separate samples.

Note: Figure 1 is based on hypothetical data and is for illustrative purposes only. It is not intended to portray any specific set of data with precision.

The two distributions pictured in the example show that the two samples differ in terms of both location and shape. The Area 2 sample gave a lower mean rating than the Area 1 sample, and the Area 2 data varied much more than the Area 1 data. Although you would need to conduct a statistical procedure to determine whether the two samples were drawn from two entirely separate populations, simply looking at the distributions gives you some indication that they are different. Statistical tests that indicate whether sets of data are *significantly* related or separate are called *inferential statistics.* These techniques help you to make predictions about the relationships between sets of data within a specified probability (e.g., you could specify that if one distribution overlaps the other by no more than 5 percent, you will conclude that the two distributions probably do not represent the same population).

Variance
Another measure of dispersion is the variance, which is the square of the standard deviation. Although the variance is an important concept, it is seldom used to describe sample data or populations. This is because it is expressed in squared units of the original scale on which the variable was measured, which makes it difficult to interpret. However, it serves as the basis for many analytical techniques in statistics, such as analysis of variance (ANOVA) and multiple regression. These are advanced statistical techniques that are beyond the scope of this guide.

Basic Inferential Statistical Techniques

As a brief introduction, we will discuss two types of inferential statistics: those that indicate whether there is a linear relationship between two variables and those that indicate differences in category membership. We will also look at some ways we can graphically represent such relationships.

Correlation

Correlation coefficients provide you with information about whether two variables are linearly related, which is why they are often referred to as measures of association. By linear, we mean that the pattern of the relationship between the two variables tends to follow a straight line. The illustration in Figure 2 (see page 18) should help you visualize what is meant by a linear relationship between two variables. This is an important concept to grasp because correlation statistics allow you to predict what you can expect to observe in one variable, given information about another variable.

> **Example:** Suppose you are responsible for lifeguard staffing and you are interested in the extent to which the number of beach visitors is affected by the outdoor temperature. You predict that more people will visit County beaches on hot days and less will visit on cold days. Statistics will allow you to test your prediction and make a confident conclusion. If your prediction holds, you will be able to make staffing predictions based, at least partly, on the outdoor temperature forecast for a given day.

The prediction you are making in the above example is that the outdoor temperature and the number of beach visitors are (positively) linearly related. The reason you predict a *positive linear relationship* is because you are saying that as observed values of one variable increase, values of the other variable also increase. If you predicted that as the values of one variable increase the values of another variable decrease, you would be predicting a *negative linear relationship*. Let us now test your prediction using the sample set of data presented in Table 7.

Table 7
Relationship between outdoor temperature and number of beach visitors.

Observation	Outdoor Temperature	Number of Beach Visitors
1	56	20
2	72	33
3	62	27
4	85	48
5	77	40
6	48	10
7	79	45
8	88	55
9	55	17
10	81	45
11	74	36
12	33	5
13	49	15
14	85	51
15	91	60
16	58	27
17	69	28
18	40	11
19	71	40

Let us assume that you randomly selected days during one year to count the number of visitors at County beaches. Additionally, let us say that your temperature readings are in degrees Fahrenheit. One way to easily get an idea about whether there is a linear relationship between your two variables is to represent the values on a graph called a *scatterplot.*

In a scatterplot graph, you place the values of your *predictor* (the variable you believe can be used to predict a change in your variable of interest) along the horizontal (x) axis. In our example, the predictor variable is outdoor temperature. The *criterion* (the variable you are interested in understanding changes in) is scaled along the vertical (y) axis. Hence, the number of beach visitors is recorded on the (y) axis in our example.

The data in Table 7 are plotted on the scatterplot graph that is presented next. Both axes are clearly labeled and scaled in proportion to the original set of data. This type of graph provides a very nice visual representation of your data, and it gives you some important information at a glance.

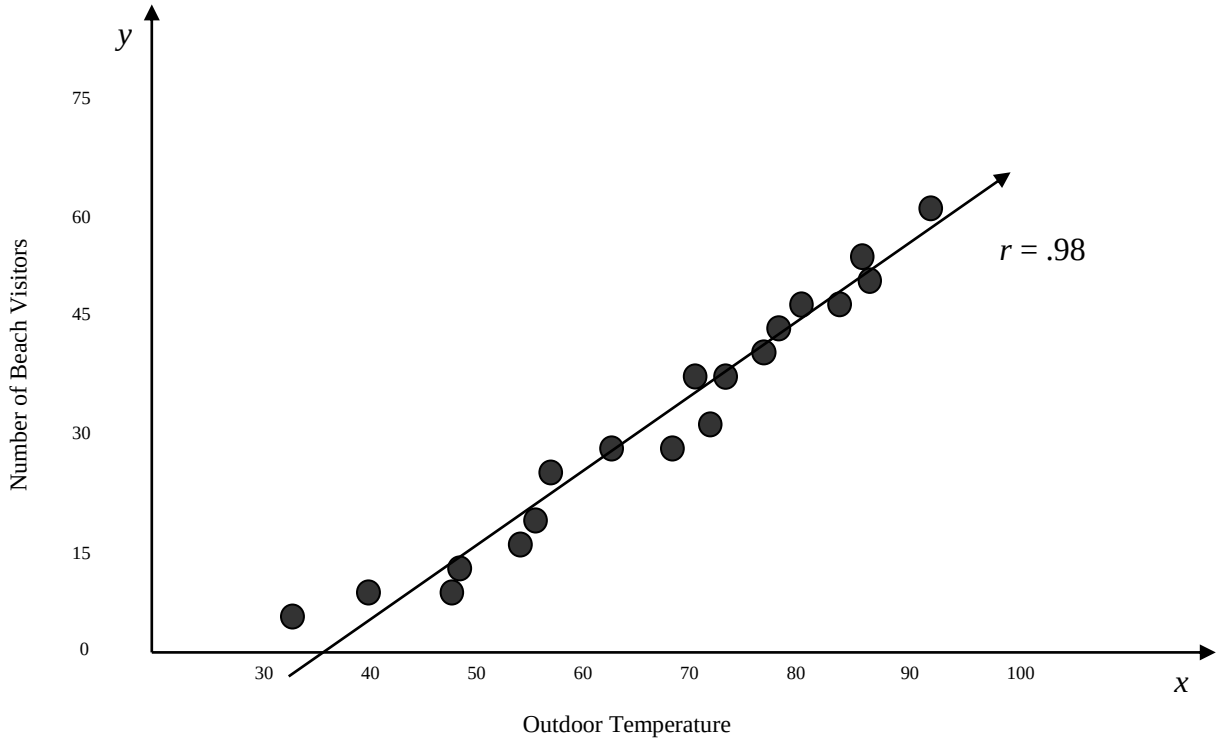


Figure 2
Sample scatterplot diagram.

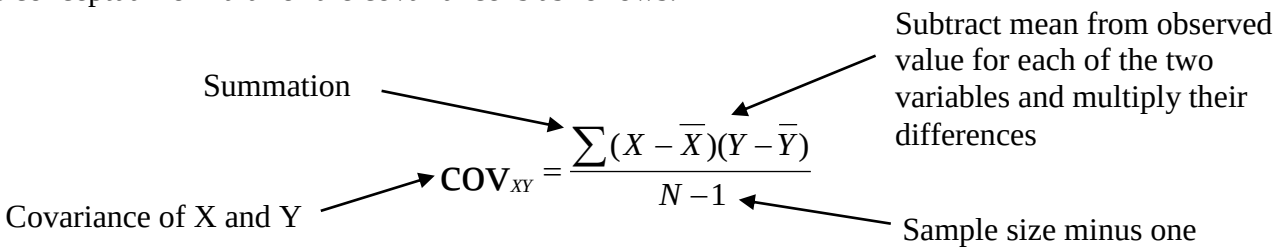
A quick glance at the scatterplot above supports your prediction of a positive linear relationship between increases in the outdoor temperature and the number of beach visitors. The strength of the relationship between the two variables is indicated by how closely the observed values follow a straight line (i.e., cluster tightly along the linear path). Although our example may seem intuitively obvious, some relationships you might be interested in learning about may not be quite as "obvious."

A more accurate way to determine the strength of a linear relationship (assuming your data are on at least an interval scale) is to apply the formula for the *Pearson Product-Moment Correlation Coefficient* (*r*). Values of Pearson's *r* range from -1.0 to 1.0. The values can be either positive or negative, depending on the direction of the relationship. A correlation of 1.0 means there is a perfect positive relationship between your two variables (i.e., there is one unit of measurement increase in one variable for every one unit increase in the other variable). Conversely, a correlation of -1.0 means that as one variable increases by one unit, the other variable decreases by one unit. In our current example, the correlation of .98 in Figure 2 indicates a nearly perfect positive correlation between outdoor temperature and the number of beach visitors.

In order to calculate Pearson's *r,* you need to obtain the three values needed to plug into the formula: (a) the standard deviation of the predictor variable (*Sx*), (b) the standard deviation of the criterion variable (*Sy*), and (c) the *covariance* of the predictor and criterion variables (COV*xy*). We discussed how to calculate the standard deviation for a variable earlier in this guide, but we have not discussed how to calculate the covariance. Calculating the covariance requires a relatively straightforward extension of concepts we have already discussed.

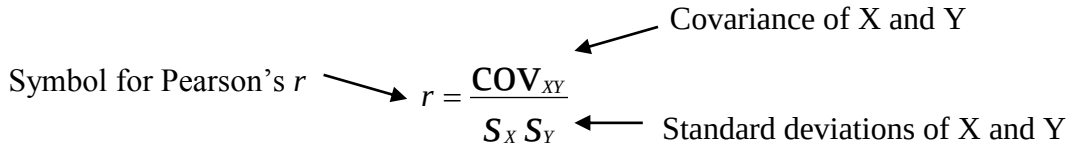
Note: When calculating a correlation coefficient, covariance, and standard deviation you will most likely rely on Microsoft Excel or a statistical software program. In the spirit of understanding, we will discuss the formula for obtaining a correlation coefficient and demonstrate the steps of its calculation.

The conceptual formula for the covariance is as follows:



Notice that this formula is nearly identical to the one used to calculate the standard deviation. You are still calculating deviations of individual observations from the mean of the set of observations. Whereas with the standard deviation you are dealing with how data for a single variable deviate from the mean, with the covariance you are dealing with how data from two variables vary in relation to their respective means and in relation to each other (i.e., co-vary). Also, notice that you do not square the deviations in this formula. Otherwise, this formula is quite similar to the standard deviation formula.

The covariance is used along with the standard deviations of the X and Y variables to produce the formula for Pearson's *r* as follows:



According to this formula, you can find the value of Pearson's *r* by dividing the covariance of *x* and *y* by the product of the standard deviations of *x* and *y*.

Pearson's *r* is used to express the strength and direction of a linear relationship between two continuous variables. Recall that a *continuous* variable can take on any number of different values, whereas discrete variables must take on a specific value (e.g., first, second, or third place in a race). In our beach attendance example, the data are continuous because it could be almost any temperature on a given day and any number of people might decide to go to the beach.

Other measures of association are available for use with data collected using nominal and ordinal measurement scales (e.g, point-biserial correlations and rank-order correlations). Although they will not be discussed in detail in this guide, interested readers are encouraged to learn about these techniques.

In addition to asking whether two sets of data are related (i.e., correlated), you may also want to know if there are differences between sets of data that are not simply due to random variability in the observed values. A statistical procedure that can be used to answer this kind of question with data collected using a nominal scale of measurement is the chi-square statistic.

Chi-Square Statistic

The *chi-square* statistic is very useful for answering questions when your data are purely categorical (i.e., nominal). Recall that categorical data involve frequencies (see page 3). When you use chi-square, you are comparing observed (actual frequency count for a given category) to expected frequencies (what you would expect to count, based on probabilities, if there are no differences).

Note: When calculating the chi-square statistic, you will most likely rely on Microsoft Excel or a statistical software program. In the spirit of understanding, we will discuss the chi-square formula and demonstrate the steps of its calculation.

The basic chi-square formula is as follows:

Chi-square symbol → $\chi^2 = \sum \frac{(O-E)^2}{E}$ ← Observed minus expected frequency squared; ← Expected frequency

Example: Suppose you are interested in whether a job applicant's education level makes a difference in which job search method is used. If there is no difference in job search method by education level, you would expect proportionately equal frequencies for each category. However, if education level affects search method choice, you would expect observed frequencies to depart from the expected frequency. To illustrate, let us construct a chi-square *contingency table.* By "contingency" we mean that the data in each cell of the table are contingent (depend) on the criteria specified by the row and column labels.

Suppose you randomly sampled 100 applicants, collecting information about their education level and whether they used the kiosk at the human resources office or the County Human Resources Department web page to conduct their County job search. Table 8 represents a sample contingency table for these data. Note that the numbers in the very first data cell are contingent on the individual having a high school diploma/G.E.D. and using the kiosk job search method.

Table 8
Education level and job search method.

Job Search Method	High School/ G.E.D.	Some College (less than 26 units)	Bachelor's Degree	Totals
Kiosk	12 (11.88)	17 (21.06)	25 (21.06)	54
Web Page	10 (10.12)	22 (17.94)	14 (17.94)	46
Totals	22	39	39	100

In Table 8, the observed frequencies are at the top of each cell and the expected frequencies are in parentheses. You can calculate the expected frequencies for each cell by multiplying the row and column totals corresponding to the specified cell and dividing by the total number of observations.

For the first cell, representing applicants with a high school/G.E.D. level of education who used the kiosk, the row total is 54 and the column total is 22. You can use the following calculation to find the expected frequency for this cell:

$$\frac{(54)(22)}{100} = \frac{1188}{100} = 11.88$$

Once you have completed the contingency table, you are ready to calculate the chi-square statistic as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(12-11.88)^2}{11.88} + \frac{(17-21.06)^2}{21.06} + \frac{(25-21.06)^2}{21.06} + \frac{(10-10.12)^2}{10.12} + \frac{(22-17.94)^2}{17.94}$$

$$+ \frac{(14-17.94)^2}{17.94} = 3.31$$

Although the calculations can be long and tedious, the math involves only subtraction, squaring numbers, division, and addition. However, you are not through yet. How do you know if the observed chi-square of 3.31 is large enough to indicate that there is a difference between the groups in terms of their education level and chosen job search method? To answer this question, we need to discuss the concept of statistical significance.

When we say an observed value is statistically significant, we mean that it is not likely that you would observe such a value if there were no differences between groups. As we mentioned previously, if the populations are the same you can expect 95% of the values observed to be within two standard deviations of the mean of the distribution. Anything outside this range

would be considered a rare event in probabilistic terms (i.e., less than or equal to a 5% chance of occurrence). Hence, a statistic that has a corresponding probability of less than or equal to 5% is considered "statistically significant."

In the back of most introductory statistics textbooks, you can find tables to help you determine whether the statistic you calculate from your data sample is statistically significant. However, to use the chi-square table, having the value of the test statistic is not enough. You will also need to determine the *degrees of freedom.* The formula for determining the degrees of freedom for the chi-square statistic is as follows:

> $df = (R - 1)(C - 1)$, or the number of rows minus one times the number of columns minus one.

For our example we have two rows and three columns. That means we have $(2 - 1)(3 - 1)$ or $1 \times 2 = 2$ degrees of freedom.

According to the table for a chi-square distribution, for an observed chi-square with two degrees of freedom to be significant it would need to be greater or equal to 5.99. This means that in our study there is not sufficient evidence to conclude that there are differences in job search method used based on an applicant's education level. Therefore, an applicant's education level has no bearing on whether he/she chooses to use the kiosk or the web page to search for a County job.

The chi-square statistic can be used for other purposes not described in this guide. Interested readers are encouraged to learn more about this useful statistic from other available sources. The statistics discussed in this guide are described in greater detail in most published texts on introductory statistics.

Conclusion

Statistics have become a widely used analytical tool to help organizations' derive meaning from large amounts of data to advance decision making. This guide was developed to provide you with a solid introduction to the field of statistics by covering general terminology, scales of measurement, and fundamental statistical calculations used in business environments. Reading this guide is a first step to enhancing your understanding of statistics, and you are encouraged to learn more about this topic through additional readings of relevant sources.

Your Feedback

In order to assist us in enhancing this document, we would greatly appreciate any feedback you would like to provide. Please email any suggestions to testprep@lacdhr.org. In the subject line of your email, please write "Basic Statistics Guide." Thank you in advance for your response.

Bibliography

This guide was developed based on the education and experience of its authors, along with integrating knowledge from the sources listed below. This guide was developed for an applied setting, and we freely share it with all readers who may find its contents of interest.

Books

Howell, D. C. (2002). Statistical methods for psychology (5th Ed.). Pacific Grove, CA: Duxbury.

McGrath, R. E. (1997). Understanding statistics: A research perspective. New York: Addison Wesley Longman, Inc.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory (3rd Ed.). New York: McGraw-Hill.

Links to Internet Sources

http://www.statsoft.com/textbook/stbasic.html

http://www.statsoft.com/textbook.esc.html

For more information about Internet search techniques, go to http://www.searchenginewatch.com

About the Authors

Paul E. Pluta
Human Resources Analyst I – Test Research

Paul is a doctoral candidate who currently possesses an M.A. degree in Industrial/Organizational Psychology. Additionally, he has nearly three years of professional human resources experience in both the private and government sectors, which includes recruitment, selection, classification, training, team building, and performance management. Paul also has over two years experience teaching undergraduate courses in research methods and psychology at a public university in the Southeastern United States. He has presented at regional and international conferences, and has been published in professional journals.

Marc C. Shartzer
Principal Human Resources Analyst – Test Research

Marc possesses an M.S. degree in Industrial/Organizational Psychology and professional certificates in Human Resources Management, Project Management, and Technical Writing. He has over nine years of professional experience in public, private, and consulting organizations. His areas of practice include selection research and test development for entry- through management-level positions, workforce planning and program development, compensation administration, and other human resource activities. He has presented at regional and international conferences and has been published in professional journals.

Additional Contributors

Angela Hunt, B.A.
Human Resources Analyst III – Test Research

Jeffery Kane, Ph.D.
Human Resources Analyst IV – Test Research

Lester Sapitula, M.A.
Human Resources Analyst IV – Test Research

Skye Knighton, B.A.
Administrative Assistant I – Test Research

Sara Lupo, M.S.
Human Resources Analyst I – Test Research

Glossary of Terms

Linguistic Terms

Chi-square test – A statistical test that can be applied to data measured on a nominal scale.

Classification – The arrangement or organization of objects according to a class or category.

Data – Information, especially when organized for analysis or used as the basis for a decision.

Descriptive statistics – Statistics that describe characteristics of a data set (e.g., mean, standard deviation, variance, etc.).

Discrete – Constituting a separate thing; individual, distinct. Consisting of unconnected distinct parts.

Dispersion – The degree of scatter of data, usually about some mean or median value.

Distribution – Set of numbers collected from a well-defined universe of possible measurements arising from a property or relationship under study.

Frequency – The number of times a value occurs within a specified interval.

Frequency Distribution – A set of adjacent intervals into which the range of a statistical distribution is divided, each associated with a frequency indicating the number of measurements in that interval.

Inference – A conclusion based on a premise.

Inferential statistics – Statistics that are used to make predictions, based on sample data, about observations or events outside the range of the sample data set.

Interval – A space between two objects, points, or units.

Mean – The arithmetic average of a series of numbers.

Median – The middle value in a distribution, above and below which lies an equal number of values.

Mode – The value or item occurring most frequently in a series of observations or statistical data.

Nominal – Of, like, pertaining to, or consisting of a name or names.

Nonparametric statistics – Statistical procedures that do not rely on rigid assumptions about population characteristics required to perform other types of statistical tests.

Ordinal – Indicating position in a series or order.

Outliers – Extreme values in a set of data that could be caused by errors, or be representative of some person or object that is not representative of the same population represented by the rest of the data set.

Parameters – Population characteristics (e.g., mean, variance, and standard deviation) that are estimated by statistics calculated from a random sample drawn from the population.

Pearson's Product-Moment Correlation Coefficient – A statistical formula developed by Karl Pearson as an standard index of the extent of linear relationship between two variables.

Random sample – A sample drawn from a population so that each member of the population has an equal chance to be drawn.

Range – A measure of dispersion equal to the difference between the smallest and largest of a set of quantities.

Ranked – Having a particular order or position, relative to other objects in the series, in terms of some value or characteristic.

Scaling – Ordering points at fixed intervals to be used as a reference standard in measurement.

Scatterplot – A diagram that represents individual data points for two or more variables as single points located along two or more coordinate axes.

Standard deviation – The square root of the variance. The square root of the arithmetic average of the squares of the deviations from the mean.

Statistic – An estimation of a population parameter, such as the mean or variance, obtained from a sample.

Statistics – The mathematics of the collection, organization, and interpretation of numerical data; especially the analysis of population characteristics by inference from sampling.

Symmetrically – In exact correspondence of form and constituent configuration on opposite sides of a dividing line.

Variable – A quantity capable of assuming any of a set of values or a symbol representing such a quantity.

Variability – Having the ability to assume any of a set of values.

Variance – The mean of the squares of the variations from the mean of a frequency distribution.

Mathematical Terms

s – Mathematical term that symbolizes the sample standard deviation.

$\sum$ – Summation sign.

N – Mathematical term that symbolizes the total number of objects in a data set.

X and Y – Symbols for variables. X is normally used to symbolize a predictor variable, and Y is most often used to symbolize a criterion variable.

χ^2 – Symbol for the chi-square statistic.