# Whatôs Stat (you say?)

**By Dr. Marilyn Simon**

Excerpted from Simon, M. K. (2011). Dissertation and scholarly research: Recipes for success (2011 Ed.). Seattle, WA, Dissertation Success, LLC.
Find this and many other dissertation guides and resources at
www.dissertationrecipes.com

*Statistics is like trying to determine how many different colored m&m's are in a king size bag by looking at only a carefully selected handful.*

The job of a statistician involves: C O A I P ng Data

| | |
|---|---|
| 1/2 cup | C OLLECTING |
| 1/2 cup | O RGANIZING |
| 1 cup | A NALYZING |
| 1-2cups | I NTERPRETING |
| 1 cup | P REDICTING |

Business decision-makers need to systematically test ideas and analyze data that are presented to them. Understanding the basics of statistics, as well as their limitations, are essential skills in this decision- making process.

Statistics can be used to predict, but it is very important to understand that these predictions are not certainties. The fact that conclusions may be incorrect separates statistics from most other branches of mathematics. If a *fair* coin is tossed ten times and ten heads appear, the statistician would incorrectly report that the coin is biased. This conclusion, however, is not certain. It is only a "likely conclusion," reflecting the very low probability of getting ten heads in ten tosses.

After data are collected, they are used to produce various statistical numbers such as means, standard deviations, percentages, etc. These descriptive numbers summarize or describe the important characteristics of a known set of data. In hypothesis testing, descriptive numbers are standardized so that they can be compared to fixed values (found in tables or in computer programs) that indicate how "unusual" it is to obtain the data you collected.

Once data are standardized and significance determined, you may be able to make inferences about an entire population (universe).

Note: This research Cookbook intends to give you a substantial "taste" of statistics so that you will feel comfortable with assessing the value of research you will review.

You might wish to seek further "condiments" to add to the knowledge you will acquire here or consult with a statistician when the information becomes more complex. What you will become, after going through this information, is a *Barefoot Statistician.*

The barefoot statistician is the mathematical equivalent of the barefoot doctor in China. The barefoot doctors were community based health workers, from among the people, educated to promote community health and give health care in cases involving a low level of intervention. They were also educated to recognize when they needed to bring in a more highly trained specialist.

In a similar way the barefoot statistician learns how to work with their local community and workplace. They work with groups to identify issues of common concern, and can help interpret results one the research has been conducted. For example, a group of workers might make the case for the need for a fitness center and then gather data to determine the effectiveness on worker productivity. The barefoot statistician can help design an effective study and once data are collected and analyze, he/she can determine the soundness of the conclusion.

The barefoot statisticians know what statistical tools are available and when and how they could be used. But, they also know enough when to bring in a more competent statistician or other specialist. They have the skills to act at the first level of consultation and understand how statistical hypothesis testing can be used to test a variety of claims. They also can help people more clearly define their own needs and solve their own problems.

Statistical programs or business calculators are now used to perform the tedious computations that often arise during statistical testing. Remember: A business decision-maker is ultimately responsible for the interpreting the study. When presented with a research study, he/she must be able to answer the following questions:

1. Was the research design sound?
2. Why was a certain statistical test chosen?
3. What assumptions were made when a statistical test was selected?
4. How should the results of a test be interpreted?
5. How does this analysis fit in with the intent of the study?
6. Did the conclusion support the hypothesis?

*The Role of Statistics*

Statistics is merely a tool. It is not the be-all and end-all for the researcher. Those who insist that research is not research unless it has a statistical display have a myopic view of the research process. These are often the same folks who are equally adamant that unless research is "experimental research" it is not research.

One cardinal rule applies: The nature of the data, and the type of research questions that are posed, govern the method that is appropriate to interpret the data and the type of research tools required to process those data. A historian seeking answers to problems associated with the assassination of Dr. Martin Luther King, Jr., would be hard put to produce either an experimental design or a statistical display of data, yet the research of the historian can be quite as scholarly and scientifically respectable as that of any quantitative or experimental study.

Statistics many times describes a quasi-world rather than the real world. You might find that the mean grade for a class is 82 but not one student actually received a grade of 82. Consider the person that found out that the average family has 1.75 children and with heartfelt gratitude exclaimed: "Boy, am I grateful that I was the first born!" What is accepted statistically is sometimes meaningless empirically. However statistics is a useful mechanism and *a means of panning precious simplicity from the sea of complexity*. It is a tool that can be applied to practically every discipline! It is a major component in business decision-making.

*Frequently Asked Questions (FAQ's) About Statistics*

1. What is the purpose of statistics?

The purpose of statistics is to collect, organize, and analyze data (from a sample), interpret the results and try to make predictions (about a population). We "do" statistics whenever we COAIP- collect, organize, analyze, interpret, and predict - data. One relies on statistics to determine "how close" to what one anticipated would happen actually did happen.

2. Why and how would one use inferential statistics?

In inferential statistics we compare a numerical result to a number that is reflective of a chance happening, and determine how significant the difference between these two numbers is.

3. Are predictions indisputable in statistics?

Statistics can be used to predict, but these predictions are not certainties. Statistics offers us a "best guess." The fact that conclusions may be incorrect separates statistics from most other branches of mathematics. If a fair coin is tossed ten times and ten heads appear, the statistician would incorrectly report that the coin is biased. This conclusion, however, is not certain. It is only a "likely conclusion," reflecting the very low probability of getting ten heads in ten tosses.

4. What are hypotheses?

Hypotheses are educated guesses (definitive statements) that are derived by logical analysis using induction or deduction from ones knowledge of the problem and from the purpose for conducting a study. They can range from very general statements to highly specific ones. Most research studies focus on the proving or the disproving of hypotheses.

| Broad area | Hypothesis (or hypotheses for the plural) |
|---|---|
| Employee Motivation | The implementation of an attendance bonus is positively related with employee attendance. |
| | There is a positive relationship between providing employeesô specific sales goals and their obtaining these goals. |
| Employee Satisfaction | There is a positive relationship between employee satisfaction and participatory management style. |
| | There is a positive relationship between employee satisfaction and the frequency of communication management delivers to employees. |

| | |
|---|---|
| Marketing | Implementation of the QED teller system has significantly improved customer satisfaction with ABC Bank.<br><br>Customers have a significantly higher preference for the XYZ Bank's location in a grocery store to traditional bank locations. |
| Quality | The customer satisfaction is higher in Division A than in Division B. |

5. What is statistical hypothesis testing?

Statistical Hypothesis Testing, or Tests of Significance, is used to determine if the differences between two or more descriptive statistics (such as a mean, percent, proportion, standard deviation, etc.) are statistically significant or more likely due to chance variations. It is a method of testing claims made about populations by using a sample (subset) from that population.

In hypothesis testing, descriptive numbers are standardized so that they can be compared to fixed values, which are found in tables and in computer programs, which indicate how "unusual" it is to obtain the data collected. A statistical hypothesis to be tested is always written as a null hypothesis (no change). Generally the null hypothesis will contain the symbol ñ=ò to indicate the status quo, or no change. An appropriate test will tell us to either reject the null hypothesis or fail to reject (accept) the null hypothesis.

6. Once I find a test that helps to test my hypothesis is there anything else I need to be concerned about?

Certain conditions are necessary prior to initiating a statistical test. One important condition is the distribution of the data. Once data are standardized and the significance level determined, a statistical test can be performed to analyze the data and possibly make inferences about an entire population (universe).

7. What are "p" values?

A p-value (or probability value) is the probability of getting a value of the sample test statistics that is at least as extreme as the one found from the sample data, assuming the null hypothesis is true. Traditionally, statisticians used "alpha" values that set up a dichotomy: reject/fail to reject conclusion. In contrast, p-values measure how confident we are in rejecting a null hypothesis. If a p-value is less than 0.01 we say this is "highly statistically significant" and there is very strong evidence against the null hypothesis. P-values between 0.01 and 0.05 indicate is that it is statistically significant and adequate evidence against the null hypothesis. For p-values greater than 0.05, there is, generally, insufficient evidence against the null hypothesis.

8. What is data mining?

Data mining is an analytic process designed to explore large amounts of data in search for consistent patterns and/or systematic relationships between variables, and then to validate these findings by applying the detected patterns to new subsets of data. There are three basic stages in data mining: exploration, model building or identifying patterns, and validation and verification. If the nature of available data allows, it is typically repeated until a "vigorous" model is identified. However, in business-decision making, options to validate the model are often limited. Thus, the initial results often have the status of general recommendations or guides based on statistical evidence (e.g. soccer moms appear to be more likely to drive a mini-van than an SUV).

9. What are the different levels of measurement?

Data comes in four types and four levels of measurement, which can be remembered by the French word for black:

NOIR - nominal (lowest) ordinal, interval, and ratio highest.

| | |
|---|---|
| Nominal Scale | Measures in terms of name of designations or discrete units or categories. Example: Gender, Color of home, religion, type of business. |
| Ordinal Scale | Measures in terms of such values as more or less, larger or smaller, but without specifying the size of the intervals. Example: Rating scales, ranking scales, Likert-type scales. |
| Interval Scale | Measures in terms of equal intervals or degrees of difference but without a zero point. Ratios do not apply. Example: Temperature, GPA. |
| Ratio Scale | Measures in terms of equal intervals and an absolute zero point of origin. Ratios apply. Example: Height, delay time, weight. |

A general and important guideline is that the statistics based on one level of measurement should not be used for a lower level, but can be used for a higher level. An implication of this guideline is that data obtained from using a Likert-type scale (a scale in which people set their preferences from say 1= totally agree to 10 = totally disagree) should, generally, not be used in parametric tests. The good news is that there is almost always an alternative approach using nonparametric methods.

10. What type of distributions can be found when data are collected?

One of the most important characteristics related to the shape of a distribution is whether the distribution is skewed or symmetrical. Skewness (the degree of asymmetry) is important. A large degree of skewness causes the mean to be less acceptable and useful as the measure of central tendency. To use many parametric statistical tests requires a

normal (symmetrical) distribution of the data. Graphical methods such as histograms are very helpful in identifying skewness in a distribution.

If the mean, median and mode are identical, then the shape of the distribution will be unimodal, symmetric and resemble a normal distribution. A distribution that is skewed to the right and unimodal will have a long right tail, whereas a distribution that is skewed to the left and unimodal will have a long left tail. A unimodal distribution that is skewed has its mean, median, and mode occur at different values. For highly skewed distributions, the median is the preferred measure of central tendency, since a mean can be greatly affected by a few extreme values on one end.

Kurtosis is a parameter that describes whether the particular distribution concentrates its probability in a central peak or in the tails. [How ñpointedò or flatò a distribution looks]. Normal populations lie at 3 on this scale, non-normal populations like on either side of 3.

11. What is the difference between a parametric and a non-parametric test?

Most of the better-known statistical tests use parametric methods. These methods generally require strict restrictions such as:

1. The data should be ratio or interval.
2. The sample data must come from a normally distributed population.

Good things about non-parametric methods:

1. Can be applied to a wider variety of situations and are distribution free.
2. Can be used with nominal and ranked data.
3. Use simpler computations and can be easier to understand.

Not so good things about non-parametric methods:

1. Tend to waste information since most of the information is reduced to qualitative form.
2. Generally less sensitive so stronger evidence is needed to show significance that could mean larger samples are needed.

## How to Exhibit Your Date (a)

Data, which are collected but not organized, are often referred to as "raw" data. It is common to seek a means in which the human mind can easily assimilate and summarize "raw" data. Frequency tables and graphs delectably fulfill this purpose.

FOR YOUR INFORMATION AND EDUCATION:

A frequency table is so named because it lists categories of scores along with their corresponding frequencies. This is an extremely simple and effective means of organizing data for further evaluation. For a large collection of scores, it is best to use a statistical program such as EXCEL, SAS, Statview, SPSS, and GBSTAT, MINITAB etc., where you enter the raw data into your computer and then with the mere press of a button or two, an awesome frequency table is constructed.

If the data you obtained are demographic (about personal characteristics or geographical regions) then it would be beneficial to present the percentages of these characteristics within the sample (e.g. 24% of the subjects studied were Hispanic). If you determine an arithmetic mean in your study, then both the mean and the standard deviation should be presented.

Data are often represented in pictorial form by means of a graph. Some common types of graphs include pie charts; (if you are picturing the relationship of parts to a whole), histograms (if you are displaying the different types of numerical responses with respect to the frequency in which they occur. A histogram is similar to a bar graph which is often used to represent the frequency of nominal data), ogives (if you are displaying cumulative frequencies such as incomes under $10,000), or stem and leaf plots (If you wish that the actual data be preserved and used to form a picture of the distribution).

Note: Statistical computer programs can instantaneously produce frequency tables; compute means, percentages, and standard deviations; and generate suitable graphs reflecting your findings once you have appropriately input raw data. [They never complain about doing any of these things]. By a click of a mouse you can order a statistical test and in under a second know whether to accept or fail to accept your statistical hypothesis!

## Assistant Chefs
## Identifying Your Population and
## Choosing Your Sample

Most chief chefs employ a variety of people to assist them in producing an eloquent banquet. Similarly, most researchers depend upon other people to help them obtain the information that they need to prepare their research report.

If a survey is needed to obtain data for decision-making, one major issue is to get "enough" people whose views count. Usually it is not practical or possible to study the entire universe or population, so you might need to settle for a sample or a subset of the population. In choosing a sample and a method of data collection, you need to determine the answer to the following questions:

1. How quickly are the data needed?
2. What are the resources available?
3. Should probability or non-probability sampling be used?

The three most common methods of sampling are:

1. Simple random sampling.

Simple random sampling; is accomplished when each member of the universe has the exact same chance of being selected. This can be accomplished by first establishing a proper sampling frame or a list of all the members in the universal population being studied. For example, if you plan to sample the pupils at a particular high school, then a collection of the names of all students in the high school would be needed.

The next phase could be to assign a number to all the members of the population. A random number generator either electronic, through dice, or using some set of digits from a telephone directory could be use to generate the sample. Whenever duplicate

numbers come up, they are discarded. Another method of generating a random sample is to put all names in a box, shake the box, and pull out a name. This name should be recorded and then tossed back in the box. Should the same name be selected more than one time before the sample is obtained, it should be disregarded. [This is called "sampling with replacement" and guarantees that the requirement "each member has the same chance of being selected," is met.]

Sometimes it is not possible to gain access to the entire population so that you may need to settle for a smaller sub population. If this is the case, the researcher should try to randomize as much as possible within that sub population.

2. Systematic stratified sampling:

In systematic stratified sampling you draw a sample with a pattern of important characteristics. Members of the population are subdivided into at least two different sub populations or strata e.g. gender. Samples are then drawn from each stratum. If you are surveying a high school whose student population is 40% of Mexican decent, you might wish your sample to reflect that same population by dividing the population into ethnic groups and obtaining a set percentage from each group.

In Cluster sampling - Members of the population are divided into sections (or clusters), randomly select a few of those sections and then choose all the members for the selected sections. For example, in conducting a pre-election poll we could randomly select 30 election precincts and survey all people from those precincts. The results may need to be adjusted to correct for any disproportionate representation of groups. Used extensively by government and private research organizations.

3. Non-probability sampling:

Non-probability sampling is something we all use in every day life. If you want to try out a new brand of crackers, you know that you only need to choose one cracker from one box to decide if you like the cracker because the others are expected to taste pretty much the same. Another common form of non-probability sampling may be carried out when trying to conduct "on the street" interviews. Researchers will often have some bias towards which people they will sample.

In convenience sampling one uses the results that are readily available. Sometimes this is quite good - e.g. a teacher wanting to know about left-handed students needs would not include those who are right-handed. However, such sampling can be seriously biased when the researcher chooses only those they feel comfortable working with.

Sometimes non-probability sampling is done inadvertently. Back in 1933 a telephone poll indicated that Landon would overwhelmingly become our next president. If you have canôt recall president Landon's record, your vexation is justified. What became obvious, after this study was analyzed, was that it failed to take into account that Republicans had most of the phones in 1933 and that Roosevelt's supporters were the majority without telephones.

If a non-probability survey is to be conducted, you must be very careful not to generalize too much from it. It is, however, very useful in the early stages of developing your study in order to get some new ideas and in the development of some interview

questions, in practicing interviews and surveying techniques, or in a pilot study. At times, only a small sample of the population is available to participate in a study.

Non-probability samples are usually easier to obtain but the gains in efficiency are often matched with losses in accuracy and generality.

Sometimes thousands of people are sampled to get the data needed; on other occasions, a sample may be as small as one.

Some factors affecting the size of a sample are:

1. The size of the universe or population being studied,
2. The purpose of the study,
3. The potential application of the result of the study,
4. The type of statistical tests, and
5. The research techniques being used.

By having a relatively large sample you are usually able to see the general, overall pattern but since in many tests the significance of measures is a function of sample size, it is possible to get a statistically significant relation when the strength of the relationship is too small to be used. Under sound statistical practices using simple random samples obtained through probability means, you can often get excellent information from a sample size of 30 or less.

Sometimes a case study of one or two subjects is the most appropriate means of conducting an investigation. This enables you to obtain detailed information about a problem in which there is much interest but little information.

Case studies are usually selected by non-probability sampling according to your judgment about whether the sample is a good representative of the population. For example, most information obtained about the (Idiot) Savant-Syndrome has been obtained through individual case studies of these extraordinary people.

FOR YOUR INFORMATION AND EDUCATION:

Another very important part of sampling is the non-response rate. This sector includes those people who could not be contacted or who refused to answer questions. A general rule is to try to keep the non-response rate under 25%. To keep the non-response rate small, you could ask the assistance of a community leader, and have that person explain the purpose and importance of your study in great detail to the potential respondents.

The size of the survey may be decided with statistical precision. A major concern in choosing a sample size is that it should be large "enough" so that it will be representative of the population from which it comes and from which you wish to make inferences about. It ought to be large enough so that important differences can be found in subgroups such as men and women, democrats and republicans, groups receiving treatment, and control groups, etc.

Two major issues to be considered when using statistical methods to choose sample size are concern with sampling error and confidence levels.

Sampling error: Some small differences will almost always exist among samples and between them and the population from which they are drawn. One approach to measuring sample error is to report the standard error of measurement, which is computed by dividing the population standard deviation (if known) by the square root of

the sample size. Minimizing sampling error helps to maximize the samplesô representativeness.

Example: If the Stanford-Binet IQ test (here standard deviation is 15) is administered to 100 subjects then the standard error of the mean would be: 15/10 or 1.5

Confidence Levels: The researcher needs to decide how "confident" they need to be that the sample is representative of the population. Frequently, the 95% confidence level is chosen. This means that there is a 95% change that the sample and the population will look alike and 5% chance that they will not. The significance depends mostly on the sample size, and how much error can be tolerated. In very large samples, even very small relations between variables will be significant, whereas in very small samples even very large relations cannot be considered reliable (significant).

Suskie (1996) provides a guide to determine how many people you can survey based on sampling error.

| Random Sample Size | Sample Error |
|---|---|
| 196 | 7% |
| 264 | 6% |
| 364 | 5% |
| 1,067 | 4% |
| 2,401 | 2% |
| 9,604 | 1% |

In most studies 5% sampling error is acceptable. Below are the sample sizes you need from a given population to obtain a 5% sampling error.

| Population Size | Sample Size |
|---|---|
| 10,000 | 370 |
| 5,000 | 357 |
| 2,000 | 322 |
| 1,000 | 278 |
| 500 | 217 |
| 250 | 155 |
| 100 | 80 |

- These numbers assume a 100% response rate

Source: Suskie, Linda (1996) Questionnaire Survey Research: What works 2nd edition. Washington, D.C.: Assn for International Research. Assn for Institutional Research; ISBN: 1882393058

STATISTICAL HYPOTHESIS TESTING

1 c Analyzing
1 c Interpreting
1 c Predicting

Featuring:

* (8) Essential Steps in Hypothesis Testing
* How to Choose Desirable Spices (tests)

* Testing Claims About:
means, standard deviations,
proportions, and relations.

In this section we will be exploring Statistical Hypothesis Testing to determine "how close" to what you anticipated would happen actually did happen. Since this contains some technical information it would be wise to read the information slowly and re-read the information a few times. You might also want to refer back to this when analyzing a research report or assisting in the design of a research study.

Hypotheses are educated guesses that are derived by logical analysis using induction or deduction from knowledge of the problem and from the purpose for conducting the study. They can range from very general statements to highly specific ones. Most research studies focus on the proving or the disproving of hypotheses.

After data have been collected and organized in some logical manner, such as a frequency table, and a descriptive statistic (mean, standard deviation, percentage, etc) has been computed, then a statistical test is often utilized to: analyze the data; interpret what this analysis means in terms of the problem; and make predictions based on these interpretations.

Note: When you use statistics you are comparing your numerical results to a number that is reflective of a chance happening and determining how significant the difference between these two numbers is.

It is the intent of this section to familiarize you with the techniques of the statistician and help you determine which statistical tests would work best for a particular study. Remember to keep a positive mental attitude as well as an open and inquisitive mind as you digest the information in this section. When learning technical and analytical techniques it is often necessary to read the material slowly and read it over again several times.

There is a myth that statistical analysis is a difficult process, which requires an advanced degree. This need not be the case. Statistical Hypothesis Testing can be fun and easy.

Although many esoteric tests exists (just as there are many exotic spices in the universe) most researchers use mundane tests (the way most chefs prepare delicious meals with common spices). The mundane "spices" for statistical hypothesis testing are: z-tests, t-tests, chi-square tests, F-tests, and rho-tests.

As you carefully and cheerfully read through this section you will learn which of these "spices" might best compliment a research study? Just as in cooking, sometimes you will find more than one spice that could be apropos and could enhance your meal. In analyzing data you will likely find more than one type of statistical test that would be appropriate for a study, and the choice can be left to the producer of the research.

## TESTING A CLAIM ABOUT A MEAN

The example in this section will be testing a claim about a mean obtained from a sample. If you can answer, ñyesò to one or more of the questions below, an identical statistical test could be used to test a hypothesis presented.

Is the researcher claiming that:

___1. A new product, program, or treatment is better than an existing one?
___2. An existing product, program, or treatment is not what it purports to be?
___3. A group is under (or over) achieving?

Note: The recipe in this section can be used to check *any* statistical hypothesis (not just a statistical hypothesis about a mean), so it would behoove you to read through the following example with eager anticipation and note any similarities between this study and a study or studies you are analyzing.

Example: One of your program managers, Cynthia Rodriguez, has designed a new battery, which she claims is better than the current one you use in your product line. The battery you currently use lasts, on the average, 7.5 months before needing a re-charge. Rodriguez claims that her battery will last significantly longer before a re-charge is necessary.

To test her claim, Rodriguez randomly samples 36 of her new batteries (n= 36), and finds that the mean time between recharging is 7.8 months. However, since the standard deviation of the population of all batteries is 0.76, this could indicate that the sample she selected is just within normal boundaries.

Statistical hypothesis testing will be used to determine if the sample mean score of 7.8 represents a statistically significant increase from the population mean of 7.5, or if the difference is more likely due to chance variation in battery lives.

Before the (8) step statistical test "recipe" is employed, you need to procure preliminary pieces of information - each beginning with the letter s.

(s) What is the substantive hypothesis?

[What does the researcher think will happen?]

*Rodriguez claims that her batteries go longer without needing a charge.*

(s) How large was the **s**ample size?

*Rodriguez sampled 36 batteries (n= 36).*

s) What descriptive **s**tatistic was determined by the sample?

*The mean average of the sample, x-bar, was 7.8.*

s) What is the **s**hape of the distribution.

*Since we are testing a sample mean we do not need to worry about the distribution. However in other tests this will be a concern. Skewness and kurtosis can be determined to make sure the sample is "close enough" to a normal distribution to employ a test you choose.*

Now we are ready to take the information obtained by the four s's and employ an (8) step recipe to create a "delicious" statistical test.

We will determine if Rodriguezôs claim: ñ*The new batteries go longer between recharging"* is statistically significant.

1. Identify the **Claim** to be tested and express it in symbolic form. The claim is about the population, which is reflected by the Greek letter $\mu$.

$$\mu > 7.5$$

That is, Rodriguez claims that her batteries last longer between re-charging than the ones currently used.

2. Express in symbolic form the **Alternative** statement that would be true if the original claim is false. All cases must be covered.

$$\mu \leq 7.5$$

3. Identify the **Null** and alternative hypothesis.

Note: The null hypothesis should be the one that contains no change (an equal sign).

$H_o$: $\mu \leq 7.5$ (Null hypothesis)

$H_1$: $\mu > 7.5$ (Alternative hypothesis)

Note: A statistical test is designed to reject or fail to reject (in essence accept) the statistical null hypothesis being examined.

The null hypothesis will either be true or false. The statistical decision process is set up so that there are no "ties." The null hypothesis is either rejected or not rejected (accepted).

4. **Decide** the level of significance, alpha ($\alpha$) based on the seriousness of a "type I error" which is the mistake of rejecting the null hypothesis when it is in fact true. Make $\alpha$ small if the consequences of rejecting a true $\alpha$ are severe. The smaller the $\alpha$ value, the less likely you will be to reject the null hypothesis. Alpha values 0.05 and 0.01 are very common.

$$\alpha = 0.05$$

FOR YOUR INFORMATION AND EDUCATION:

Some researchers do not use alpha values (which are pre-determined at the beginning of a statistical test and indicate acceptable levels of significance). Instead, they prefer p-values; (which indicate actual levels of significance of a claim and leave the conclusion as to whether this is "significant enough" to the reader). It is possible to do both (set $\alpha$ and compute p) and then compare these two values when reporting the findings.

Before a test of hypotheses is employed we can be certain that only 4 possible things can happen. These are summarized in the table below

| | | Claim is tested | |
|---|---|---|---|
| | | $H_O$ | $H_1$ |
| Decision | $H_O$ | **Correct Acceptance** | Type II Error $\beta$ |
| | $H_1$ | Type I Error $\alpha$ | **Correct Rejection** |

Note that there are two kinds of errors represented in the table. Many statistics textbooks present a point of view that is common in business decision-making: $\alpha$, the type I error, must be kept at or below .05, and that, if at all possible, $\beta$, the Type II error rate, must be kept low as well. "Statistical Power,ò which is equal to 1 - $\beta$, must be kept correspondingly high. Ideally, power should be at least .90 to detect a reasonable departure from the null hypothesis.

5. **Order** a statistical test which is relevant and appropriate for the study (see Table 1)

Since the claim involves a sample mean and n > 30, table (1) assures us that we can compute a z-value and use a Z-test.

That sounds great, you say, but what does it mean, you sigh? A z-value is a number we compute which can then be graphed as a point on the horizontal scale of the standard normal distribution curve. This point indicates how far from the population mean our sample mean is, and thus enables us to determine how "unusual" our research finding are.

FOR YOUR INFORMATION AND EDUCATION:

The Central Limit Theorem implies that for samples of sizes larger than 30, the sample means can be approximated reasonably well by a normal (z) distribution. The approximation gets better as the sample size, n, becomes larger.

When you compute a z-value you are converting your mean to a mean of 0 and your standard deviation to a standard deviation of 1. This allows you to use the standard normal distribution curve and its' corresponding table to determine the significance of your values regardless of the actual value of your mean or standard deviation.

A standard normal probability distribution is a bell-shaped curve (also called a Gaussian curve in honor of its discoverer, Karl Gauss) where the mean, or middle value is 0, and the standard deviation, the place where the curve starts to bend, is equal to 1 on the right and -1 on the left. The area under every probability distribution curves is equal to 1 or 100%). Since a Gaussian curve is symmetric about the mean, it is important to note that the mean divides this curve into 2 equal areas of 50%. Blood cholesterol levels, heights of adult women, weights of 10 year old boys, diameters of apples, scores on standardized test, etc., are all examples of collections of values whose frequency distribution resemble the Gaussian curve.

If you were to record all the possible outcomes from the toss of 100 different coins, by graphing the number of heads that could occur on the horizontal axis (0,1,2,3...100), and the frequency with which each of the number of heads could occur on the vertical axis, you will produce a graph resembling the normal distribution. (The most "frequent" results would cluster around 50 heads and become less and less frequent as you consider values further and further from 50.]

6. Perform the **Arithmetic** to find the test statistic, the critical value(s), and the critical region. [actually the computer will do this for you☺]

The sample mean of 7.8 is equivalent to a z value of 2.37. This z value is the test statistic, and was computed using the following formula:

$$z = (x - \mu) / \sigma / \sqrt{n}$$

x = sample mean,
μ = population mean,
n = size of sample
$\sigma$ = population standard deviation,

Thus, z = (7.8 - 7.5) / (0.76)/6 = 2.37 (to the nearest hundredth)

Note: z numbers usually vary between -3 and +3. If they are outside of this range the null hypothesis will almost always be rejected (or an error was made).
Note: $\sigma / \sqrt{\eta}$ (sigma divided by the square root of n) is often called the standard error of the mean or the standard deviation of the sample means.
Note: Sometimes you can substitute the actual standard deviation of the sample, s, for the population standard deviation, s, if $\sigma$ is unknown.

The $\alpha$ = 0.05 level employs us to find a z-value that will separate the curve into 2 unequal regions; the smaller one with an area of 0.05 (5%), and the larger one with an area of 100% - 5% = 95% (0.95). (Often referred to as a 95% confidence level).

Z-values indicate the percent of area under the bell-shaped curve from the mean (middle) towards the "right tail" of the curve. Thus, for an alpha of 0.05, we need to determine what z value will cut off an area of 45% (0.4500) from the mean towards the "right tail" (we already know that 50% of the area is on the left side of the mean. We obtain 95% by adding 45% to 50%).

"Hunting" through the vast array of four digit numerals in the table, we find our critical value to be between 1.6 (see z column) +. 04 (1.64) which (reading down the .04 column) determines an area of .4495, and 1.6 +. 05 (1.65), which determines an area of .4505. Thus, if we take the mean average of 1.64 and 1.65, we can blissfully determine the critical value to be 1.645. and the critical region to be all Z values greater than 1.645. This determination requires us to reject the null hypothesis if our test statistics (Z value) is greater than 1.645.

Note: Since there is only one alternative hypothesis, $H_1$: $\mu > 7.5$ we call this a one-tailed (right tailed) test. If our alternative hypothesis was $\mu < 7.5$ we would have a left-tailed test, and a two-tailed test would be used since there would be two alternatives: $\mu < 7.5$ or $\mu > 7.5$.

Note: The p value is .0089. We arrived at this value quite easily. Recall, in step 6 we computed the A relation between variables score to be 2.37. If you go down the left side of the table and find a z value of 2.3, then go across to the .07 column, you find, the number .4911. This indicates an area of 49.11% from the mean z value of 0 to the z value of 2.37. Thus, 50%+49.11% or 99.11% of the area of the curve is to the left of 2.37 and the small tail to the right of 2.37 has an area of 100%- 99.11% = 0.89% or 0.0089 which is our p-value.

7. **Look** to reject or fail to reject (accept) the null hypothesis. Reject null hypothesis if the test statistic is in the critical region. Fail to reject the null hypothesis if the test statistic is not in the critical region.

*Ms. Rodriquez's z-value is in the critical region since 2.37 > 1.645. Thus we will reject the null hypothesis.*

8. Restate the previous decision in simple **Lay** non-technical terms.

We have reason to believe that the new batteries reduce the time before re-charging.

FOR YOUR INFORMATION AND EDUCATION:

If your sample size is less than 30 and the population standard deviation is unknown, but there is a normal distribution in the population, then you can compute a "t" statistic, where:

$$t = (x - \mu) / s / \sqrt{n}$$

x is the sample mean, μ is the population mean under contention, s is the standard deviation of the sample, and n is the size of the sample.

Notice that the z and t statistics are computed the same way, the only difference is in their corresponding values of significance when you (or your computer) checks these values on the graph or table.

**Summary**

In this example we tested a claim about a numerical value mean test using ratio data. The sample data came from a population, which was known to have a normal distribution. We were thus able to use parametric methods in our hypothesis testing.

In general, when you test claims with interval or ratio parameters (such as mean, standard deviation, or proportions), and some fairly strict requirements (such as the sample data came from a normally distributed population) are met, you should be able to use parametric methods.

If the necessary requirements for parametric methods are not met, do not despair. It is very likely that there are alternative techniques, named appropriately, non-parametric methods, which could (and should) be used instead.

Since they do not require normally distributed populations, non-parametric tests are often called distribution-free tests. Some advantages to using non-parametric methods include:

1. They can often be applied to nominal data that lack exact numerical values.
2. They usually involve computations that are simpler than the corresponding parametric methods.
3. They tend to be easier to understand.

Unfortunately, there are the following disadvantages:

1. Non-parametric methods tend to "waste" information, since exact numerical data are often reduced to a qualitative form. Some treat all values as either positive (+) or negative (-). [A dieter using this test would count a loss of 60 lbs as the same as a loss of 2 lbs!].
2. They are generally less sensitive than the corresponding parametric methods. This means that you need stronger evidence before rejecting the null hypothesis, or larger sample sizes.

The non-parametric counterpart of both the t and z tests are: the Sign Test

*TABLE 1
Recommended methods of cooking HYPOTHESIS TESTING

| MY CLAIM IS ABOUT A: | CLAIM | ASSUMPTION | PARAMETRIC TEST-/STATISTIC | NONPARAMETRIC TEST/STATISTIC |
|---|---|---|---|---|
| mean IQ than average. | Class A as a higher | $n \geq 30$ or s.d known | Z | Sign Test |
| | n< 30, s.d. unknown | T | | Wilcox/Mann WhitneyU (U) |
| proportion | 75% of voters prefer candidate | np> 5 and nq> 5 | Z | |
| standard deviation | This instrument has fewer errors than others | normal population | x2 | .Kruskal-Wallis H (H) |
| two means | EZ diet is more effective than DF diet? | dependent | T | Sign Test |
| | | independent | T or Z | U |
| | The percent of cures is the same for those using drug A as Drug B. (A low t or U value would indicate that the proportions are similar. ) | | | |
| two standard deviations | The ages of group A are more homogeneous that the ages of group B. | | F | (H) |
| two proportions | There are more Democrats in Chicago than LA. | | Z | Sign Test |
| relationship between 2 variables | Smoking related to cancer. (If r is close to 0, then no relation) | | .Pearson r | Spearman r |
| Are two variables dependent? | | | F | H |
| How close do expected values agree with observed (aka Goodness of fit) | k variables | | $x^2$ df = k-1 | |
| .ANOVA | comparing 3 or more means | (compute: variances between sample means/total variances ) | F | Kruskal-Wallis (H) |
| | df: num:=( k - 1) den = k(n-1);K= no. of groups n= amt in each group | | | |

*Assumptions for ANOVA: normal distribution, equal variances from each sample. However, George E.P. Box demonstrated that as long as the sample sizes are equal (or near equal) the variances can be up to nine times as large and the results from ANOVA will continue to be essentially reliable. However, If the data don't fit these basic assumptions we can always use the non-Parametric version (Kruskal-Wallis).If a significant F ratio is found, another test can be employed to determine where the significance lies. One of these is Tukey's HSD (honestly significant difference).*

**Contingency table** (two ï way ANOVA): A table of observed frequencies where the rows correspond to one variable and the columns another. It is used to see if 2 variables are dependent but can not be used to determine what the relationship is between the two variables.

| | STUDY OF 1000 DEATHS OF MALES | | |
|---|---|---|---|
| | Cancer | Heart Disease | Other |
| Smoking | 135 | 310 | 205 |
| Non-Smoking | 55 | 155 | 140 |

Note: The values in the table are observed values. These would be compared to ñexpectedò values. A $x^2$ statistics would be computed. A large $x^2$ value indicates there is a relation between variables

We are now going to examine some felicitous applications for other statistical tests. You might wish to scan the list and see if you can identify similarities between the examples given and any of the hypotheses that you are planning to test.

TESTING CLAIMS ABOUT 2 MEANS

In this section we will discuss a claim made about two means (that the mean of one group is less than, greater than, or equal to the mean of another group). The researcher will first need to determine if the groups are dependent, i.e. the values in one sample are related to the values in another sample. This includes: before and after tests; tests involving spouses, relationships between an employer and an employee; or if the groups are independent, i.e. the values in one sample are not related to the values in another sample. This includes comparing an experimental group to a control group, or samples from two different populations like the eating habits of people in Michigan vs. Hawaii.

If the researcher can answer yes to one of the questions below, then the identical statistical test described in this section can be employed. Is the researcher claiming that:

___1. One product, program or treatment is better than another?

___2. One group is better (or worse ) than another? (with respect to some variable).

__ 3. An experimental program was effective?

Many real and practical situations involve testing hypotheses made about two population means. For example, a manufacturer may want to compare output on two different machines to see if they obtain the same result. A nutritionist might wish to compare the weight loss that results from patients on two different diet plans to determine which one is more effective. A psychologist may want to test for a difference in mean reaction times between men and women to determine if women respond quicker than men in an emergency situation.

If the two samples (groups) are dependent -- the values in one sample are related to the values in the other in some way, a t- statistic is computed and a simple paired t-test; may be used to test your claim. Computing the differences between the related means and then obtaining the mean of all these differences obtain this t-statistic.

If the two samples are independent i.e. the values in one sample are not related to the values in the other, and the size of each group, n, is 30 or more, or the standard deviations of the population, then a simple z statistics may be computed and a paired z-test could be ordered. In this case, the differences in the population means are computed and subtracted from the differences in the sample means. The result is divided by the square root of the sum of each variance divided by the respective sample size.

BUT if the two samples are: independent, the sample size (n) is less than 30 for each group, and the population standard deviation is not known then whom do you call? The F (team) test. The F test is used first to see if the standard deviations are equal. (A relatively small f value indicates that the standard deviations are the same). If the f-value is "relatively" small, then the researcher, or much more likely a computer, would need to perform a t-test to test the claim. This involves a very hackneyed computation. However, if the f-test was to yield a "relatively" large F value, this would lead to a more benign t-test .

Once the mean and standard deviation is computed for each sample, it is customary to identify the group with the larger standard deviation as group 1 and the other sample as group 2.

The non-parametric counterpart of the paired Z or T-test is the Wilcoxan signed-ranks test, if samples are dependent and the Wilcoxon Rank-Sum Test, if samples are independent

TESTING CLAIMS ABOUT 3 OR MORE MEANS

If the researcher can answer yes to one of the questions below, then the identical statistical test described in this section can be employed. Is there a claim that:

___1. There is a difference between three or more products, programs, or treatments?

___2. There is a different outcome from the same program, product or treatment among three or more different groups?

Claims about 3 or more means, require the creation of an F-statistic and the performing of an F test is on the menu. Here the researcher or most likely the computer, will compare the variances between the samples to the variances within the samples. The nickname for what's happening here is ANOVA; (Analysis of Variance). It is an extension of the t-test, which is used for testing two means. The null hypothesis is that the means are equal.

An F value close to 1 would indicate that there are no significant differences between the sample means. A "relatively large" F value will cause you to reject the null hypothesis and conclude that the means in the samples are not equal. Some important things to know about The F distribution:

1. It is not symmetric, it is skewed to the right.
2. The values of F can be 0 or positive, but not negative.
3. There is a different F distribution for each pair of degrees of freedom for the numerator and denominator.

Note: If you are asking: "Why are we dealing with variances when the claim is about means?" you would be asking a very good question. The answer is that the variance, or the standard deviation squared, is determined by, and dependent on, the mean, so it is actually all in the family!

FOR YOUR INFORMATION AND EDUCATION:

The method of ANOVA owes its beginning to Sir Ronald A. Fisher (1890-1962) and received its early impetus because of its applications to problems in agricultural research. It is such a powerful statistical tool that it has since found applications in just about every branch of scientific research, including economics, psychology, marketing research, and industrial engineering, to mention just a few.

The following example will be testing a claim about three means obtained from a sample Using the ssss candoall model. A report on the findings follows the example.

Example: A study was done to investigate the time in minutes for three police precincts to arrive at the scene of a crime. Sample results from similar types of crimes are:

A: 7 4 4 3
sample size: n = 4, mean. x = 4.5, variance, $s^2 = 3.0$

B: 9 5 7
sample size: n = 3, mean x = 7.0, variance, $s^2 = 4.2$

C: 2 3 5 3 8
sample size: n = 5, mean x = 4.2, variance, $s^2 = 5.7$

At the $\alpha = 0.05$ significance level, test the claim that the precincts have the same mean reaction time to similar crimes.

What are the pre-test s's?

*Assumptions for ANOVA: normal distribution, equal variances from each sample. However, George E.P. Box demonstrated that as long as the sample sizes are equal (or near equal) the variances can be up to nine times as large and the results from ANOVA will continue to be essentially reliable.*

(a) What is the Substantive hypothesis;?
[What does the researcher think will happen?]
The reaction times are similar in the three precincts.

b) How large is the sample size that was studied?
The three groups have sample sizes 4,3, and 5 respectively.

c) What descriptive Statistic was determined by the sample?
The means and variances for each group were determined.

Now we are ready to take the information obtained in (a), (b), and (c) and employ the eight step **CANDOALL** recipe to test this hypothesis.

We will determine if the claim; The reaction times are similar is statistically correct.

1. Identify the **Claim** (C) to be tested and express it in symbolic form.

$$\mu_a = \mu_b = \mu_c$$

That is, there is a claim that the mean reaction time in each precinct is the same

2. Express in symbolic form the **Alternative** (A) statement that would be true if the original claim is false.

$$\mu_a \neq \mu_b \neq \mu_c$$

Remember we must cover all possibilities

3. Identify the **Null (**N) and alternative hypothesis.

Note: The null hypothesis should be the one that contains no change (an equal sign).

$H_o$: $\mu_a = \mu_b = \mu_c$ (Null hypothesis)
$H_1$: $\mu_a \neq \mu_b \neq \mu_c$ (Alternative hypothesis)

Remember: A statistical test is designed to reject or fail to reject (accept) the statistical null hypothesis being examined.

4. **Decide** (D) on the level of significance, alpha ($\alpha$), based on the seriousness of a type I error

Note: This is the mistake of rejecting the null hypothesis when it is in fact true. Make alpha small if the consequences of rejecting a true alpha are severe. The smaller the alpha value, the less likely you will be to reject the null hypothesis. Alpha values 0.05 and 0.01 are very common.

$\alpha = 0.05$

5. **Order** (O) a statistical test and sampling distribution that is relevant to the study.* (see Table 1)

Since the claim involves data from three groups and we wish to test the hypothesis that the differences among the sample means is due to chance, we can use the ANOVA test.

Note: The following assumptions apply when using the ANOVA: The population has a normal distribution, the populations have the same variance (or standard deviation ï or similar sample sizes); the samples are random and independent of each other.

6. Perform the **Arithmetic** (A) and determine: the test statistic, the critical value and the critical region.

Note: It would be best for this to be performed on a computer.

To perform an ANOVA test we need to compute:
The number of samples, k,

k =3

The mean of all the times, X

x = 5.0

The variance between the samples - this is found by subtracting the mean (5.0) from the variance of each sample, squaring the differences, then multiplying each by the sample size and finally adding up the results for each sample.

The variance within the samples - this is found by multiplying the variance of each sample by one less than the number in the sample, adding the results, this equals 39.8, and then dividing by the total population minus the number of samples, 9.

The variance within the samples = 4.4222

The test statistic is F = variance between samples
variance within samples

F = 1.8317

The variance between the samples = 8.1

The degrees of freedom in the numerator = k-1 = 3-1 = 2. The degrees of freedom in the denominator = n-k = 12-3 = 9.

Note: Degrees of freedom are the number of values that are free to vary after certain restrictions have been imposed on all values. For example, if 10 scores must total 80, then we can freely assign values to the first 9 scores, but the tenth score would then be determined so that there would be 9 degrees of freedom. In a test using an F statistic we need to find the degrees of freedom in both the numerator and the denominator.

The critical value of F = 4.2565. (this can be found on a table or from a computer program).

7. **Look** (L) to reject or fail to reject the null hypothesis.

Note: This is a right-tailed test since the F statistic yields only positive values.

Since the test statistic of F = 1.8317 does not exceed the critical value of F = 4.2565, we fail to reject the null hypothesis that the means are equal.

Note: The shape of an F distribution is slightly different for each sample size n. The $\alpha$ = 0.05 level employs us to find an F value that will separate the curve into 2 unequal regions; the smaller one with an area of 0.05 (5%), and the larger one with an area of 100%- 5% or 95% (0.95). (Often referred to as a 95% confidence level).

8. In **Lay** terms (L) write what happened.

There is not sufficient sample evidence to warrant rejection of the claim that the means are equal.

Note: In order for statistics to make sense in research it is important to use a rigorously controlled design in which other factors are forced to be constant. The design of the experiment is critically important, and no statistical calisthenics can salvage a poor design.

**Writing About This Study in a Research Paper**

If this study were to be published in a research journal, the following script could be used to summarize the statistical findings. This information usually appears in the data analysis section of a document but could also be properly placed in the section where the conclusion of the study is found, or even in the methodology section of the paper. This information would also be very appropriate to place in the abstract of the study.

A study was conducted to investigate the time in minutes for three police precincts to arrive at the scene of a crime. Sample results from similar types of crimes were found to be:

A: 7 4 4 3
sample size: n = 4, mean, x, = 4.5, variance, s2 = 3.0
B: 9 5 7
sample size: n = 3, mean, x = 7.0, variance, s2 = 4.2
C: 2 3 5 3 8
sample size: n = 5, mean, x = 4.2 variance, s2 = 5.7

At the $\alpha$ =0.05 significance level, the claim that the precincts had the same mean reaction time to similar crimes was tested. The null hypothesis is the claim that the samples come from populations with the same mean:

$H_o$: $\mu_a = \mu_b = \mu_c$ (Null hypothesis)
$H_1$: $\mu_a \neq \mu_b \neq \mu_c$ (Alternative hypothesis)

To determine if there are any statistically significant differences between the means of the three groups, an ANOVA test was performed. The groups were similar in size and the level of measurement was ratio data. An F distribution was employed to compare the two different estimates of the variance common to the different groups (i.e. variation between samples, and variation within the samples). A test statistic of F = 1.8317 was obtained. With 2 degrees of freedom for the numerator and 9 degrees of freedom for the denominator, the critical F value of 4.2565 was determined. Since the test F does not exceed the critical F value, the null hypothesis was not rejected. There is not sufficient sample evidence to reject the claim that the mean values were equal.

The non-parametric counterpart of ANOVA is: the Kruskal-Wallis Test

TESTING A CLAIM ABOUT PROPORTIONS/PERCENTAGES

If the answer to one of the questions below is ñyesò, then the identical statistical test described in this section could be employed. Is the researcher claiming that:

___1. A certain percent or ratio is higher or lower than what is believed?
___2. There is a characteristic of a group that is actually prevalent in a higher or lower percent

Data at the nominal (name only) level of measurement lacks an real numerical significance and is essentially qualitative in nature. One way to make a quantitative analysis when qualitative data are obtained, is to represent that data in the form of a percentage or a ratio. This representation is very useful in a variety of applications, including surveys, polls, and quality control considerations involving the percentage of defective parts.

A Z- test; will work fine here provided that the size of the population is large enough. The condition is that:

$np \geq 5$ and $nq \geq 5$

where, as always, n = sample size, p = population and q = 1- p

Note: The p in the test of proportions different than the "p- valueò we use to determine significance in hypothesis testing. It is important to be aware that in mathematics often times the same symbol can have more than one interpretation. While doing mathematics keep this in mind and remember to learn the meaning of a symbol in its context.

Example: If a manager believes that less than 48% of her employees support the companyôs dress code, the claim can be checked based on the response of a random sample of employees,

If 720 employees were sampled and 54.2% actually favored the dress code, then to check the managerôs claim, the researcher could perform a test of hypothesis to determine if the actual value of 0.542 is significantly different from the value of 0.48. Here, n =720, p = .48, q =0.542. The conditions $np \geq 5$ and $nq \geq 5$ are met since 720(.48) = 345.6 and 720(.542) = 390.24. The z-value would be 3.33. This would lead us to reject the null hypothesis and conclude that this is probably a low estimate.

TESTING CLAIMS ABOUT STANDARD DEVIATIONS AND VARIABILITY

Many real and practical situations demand decisions or inferences about variances and standard deviations. In manufacturing, quality control engineers want to ensure that a product is on the average acceptable but also want to produce items of consistent quality so there are as few defects as possible. Consistency is measured by variances.

FOR YOUR INFORMATION AND EDUCATION:

During World War II, 35,000 American engineers and technicians were taught to use statistics to improve the quality of war material through the efforts of Dr. W. Edwards Deming (born in Sioux City, Iowa, on October 14, 1900). Deming's work was brought to the attention of the Union of Japanese Scientists and Engineers (JUSE). JUSE called upon Deming to help its members increase productivity. Deming convinced the Japanese people that quality drives profits up. The rebirth of Japanese industry and its worldwide success is attributed to the ideas and the teachings of Deming. In gratitude, the late Emperor Hirohito awarded Japanôs Second Order Medal of the Sacred Treasure Deming.

If you can answer yes to one of the questions below, you can use the identical statistical test described in this section. Are you claiming that:

___1. a product, program, or treatment has more or less variability than the standard?

___2. a product, program, or treatment is more or less consistent than the standard?

To test claims involving variability, the researcher usually turns to a Chi-square ( $x^2$) statistic.

FOR YOUR INFORMATION AND EDUCATION:

Both the t and Chi square ( $x^2$) distributions have a slightly different shape depending on n, the number in the sample. For this reason, the researcher needs to determine the "degrees of freedom" to find out what shape curve will be used to obtain the test statistics.

The "degrees of freedom" refer to the number of observations or scores minus the number of parameters that are being studied. (Informally, it is the number of times you can miss a certain targeted number and still have a chance of obtaining that desired outcome). When the researcher uses a sample size of (n) to investigate one parameter, e.g. a mean or standard deviation, the degrees of freedom equals n-1. When investigating a relationship between 2 variables then the degrees of freedom are: ( n-2). The test statistics used in tests of hypothesis involving variances or standard deviations, is chi-square, $X^2$

Example: A supermarket finds that the average check out waiting time for a customer on Saturday mornings is 8 minutes with a standard deviation of 6.2 minutes. One Saturday management experimented with a single queue. They sampled 25 customers and found that the average waiting time remained 8 minutes, but the standard deviation went down to 3.8 minutes.

To test the claim that the single line causes lower variation in waiting time, a computed chi-square value would be: $X^2 = 9.016$ and there would be 24 degrees of freedom since n = 25. The null hypothesis would be that the new line produced a standard deviation of waiting time greater than or equal to 6.2 This would yield a one-tail (left) test. The critical value would be 13.48, and we would reject the null hypothesis if the computed value were less than the critical value. Since $9.016 < 13.48$ we would reject the null hypothesis and conclude that this method seems to lower the variation in waiting time.

TESTING A CLAIM ABOUT THE RELATION BETWEEN TWO VARIABLES (CORRELATION AND REGRESSION ANALYSIS)

Many real and practical situations demand decisions or inferences about how data from a certain variable can be used to determine the value of some other related variable.

For example, a Florida study of the number of powerboat registrations and the number of accidental manatee deaths confirmed that there was a significant positive correlation. As a result, Florida created coastal sanctuaries where powerboats are prohibited so that manatees could thrive.

A study in Sweden found that there was a higher incidence of leukemia among children who lived within 300 meters of a high-tension power line during a 25-year period. This led Sweden's government to consider regulations that would reduce housing in close proximity to high-tension power lines.

If you can answer yes to the questions below, the researcher can use the identical statistical test described in this section. Is the researcher claiming that:

___1. There is a relationship or correlation between two factors, two events or two characteristics **and**

__ 2. The data are at least of the interval measure.

In regression and correlation analysis the data are:

1. Record the information in table form.
2. A scatter diagram is usually created to see any "obvious" relationship or trends.
3. The correlation coefficient r (Rho) aka the Pearson Correlation Coefficient factor, to obtain objective analysis that will uncover the magnitude and significance of the relationship between the variables.
4. A test is performed to determine if r is statistically significant.
5. If r is statistically significant then regression analysis can be used to determine the relationship between the variables.

Example: Suppose randomly selected students are given a standard IQ test and then tested for their levels of math anxiety using a MARS test:

1. Record information in table form:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| IQ(I) | 103 | 113 | 119 | 107 | 78 | 153 | 111 | 128 | 135 | 86 |
| MARS(M) | 75 | 68 | 65 | 70 | 86 | 21 | 85 | 45 | 24 | 73 |

The researchersô hypothesis is that students with higher IQ's have lower levels of math anxiety. [Note: The independent variable (x) is IQ scores, which are being used to predict the dependent variable (y) is MARS scores].

$H_o$: r = 0 (there is no relationship)

$H_1$: r - 0 (there is a relationship)

Note: These will usually be the hypotheses in regression analysis.

2. Draw a scatter diagram:

The points in the figure above seem to follow a downward pattern, so we might conclude that there is a relationship between IQ and levels of Math anxiety, but this is somewhat subjective.

3. Compute r

To obtain a more precise and objective analysis we can compute the linear coefficient constant, r. Computing r is a tedious exercise in arithmetic but practically any

statistical computer program or scientific calculator would willingly help you along. In our example the very friendly program, STATVIEW was used to determine our r = -.882

Some of the properties or this number r are:

1. The computed value of r must be between (-1) and (+1). (If it's not then someone or something messed up.)

2. A strong positive correlation would yield an r-value close to (+1), a strong negative linear correlation would be close to (-1).

3. If r is close to 0 we conclude that there is no significant linear correlation between (X) and (Y).

Checking the table, we find that with a sample size of 10, (n = 10) , the value r = -.882 indicating a strong negative correlation between measures of IQ and measures of math anxiety levels. The r-squared number (.779) indicates that a personôs IQ could explain 77.9% of a persons MARS score.

4. If there is a significant relation, then regression analysis is used to determine what that relationship is. If the relation is linear, the equation of the line of best fit can be determined. [For 2 variables, the equation of a line can be expressed as y = mx + b, where m is the slope and b is the y intercept]

Thus, the equation of the line of best fit would be

$$Y = -.914\ I + 165.04$$

The non-parametric counterpart to r is the Spearman's rank correlation coefficient (rs) or r.

**More on Correlational statistics**

Warning: Correlation does not imply CAUSATION!

The Purpose of Correlational Research is to find Co-relationships between two or more variables with the hope of better understanding the conditions and events we encounter and with the hope of making predictions about the future. [From the Annals of Chaos Theory: Predictions are usually very difficult-especially if they are about the Future; Predictions are like diapers, both need to be changed often and for the same reason!].

As was noted previously, the linear correlation coefficient, r, measures the strength of the linear relationship between two paired variables in a sample. If there is a linear correlation, that is if r is "large enough," between two variables, then regression analysis is used to identify the relationship with the hope of predicting one variable from the other.

Note: If there is no significant linear correlation, then a regression equation cannot be used to make predictions.

A regression equation based on old data is not necessarily valid now. The regression equation relating used car prices and ages of cars is no longer usable if it is based on data from the 1960's. Often a scattergram is plotted to get a visual view of the correlation and possible regression equation.

Note: Nonlinear relationships can also be determined, but due to the fact that more complex mathematics are used to describe and interpret data, they are used considerably less often. The following are characteristics of all linear correlational studies:

1. Main research questions are stated as null hypotheses, i.e. "no" relationship exists between the variables being studied.

2. In simple Correlation, there are two measures for each individual in the sample.
3. To apply parametric methods, there must be at least 30 individuals in the study.
4. Can be used to measure "the degree" of relationships, not simply whether a relationship exists.
5. A perfect positive correlation is 1.00; a perfect negative (inverse) is -1.00.
6. A correlation of 0 indicates no linear relationship exists.
7. If when two variables x and y, are correlated so that r = .5, then we say that (0.5) 2 or 0.25 or 25% of their variation is common, or variable x can predict 25% of the variance in y.

Bivariate correlation is when there are only 2 variables being investigated. The following definitions help us determine which statistical test can be used to determine correlation and regression.

Continuous scores: Scores can be measured using a rational scale.
Ranked data: Likert Scale, Class Rankings
Dichotomy: subjects classified into two categories- Republican Vs. Democrat
Artificial - pass fail (arbitrary decision); true dichotomy ( male-female).

The Pearson Product Moment Correlation Coefficient (that is a mouthful!) or simply the Pearson r, is the most common measure of the strength of the linear relationship between two variables. It is named for Karl Pearson (1857-1936) who originally developed it. The Spearman Rank Correlation Coefficient, or Spearman r, (which we performed above), used for ranked data or when you have a sample size less than 30 (n<30), is the second most popular measure of the strength of the linear relationship between two variables. To measure of the strength of the linear relationship between test items for reliability purposes, the Cronbach alpha is the most efficient method of measuring the internal consistency. The following is a table to determine what statistical technique is best used with respect to the type of data the researcher collects.

| **Technique** | **Symbol** | **Variable 1** | **Variable 2** | **Remarks** |
|---|---|---|---|---|
| Pearson | r | Continuous | Continuous | Smallest standard of error |
| Spearman Rank | r | Ranks | Ranks | Used also when n< 30 |
| Kendall's tau | t | Ranks | Ranks | Used for n < 10 |
| Biserial Correlation (Cronbach) | a/bis | Artificial Dichotomy | Continuous | Sometimes exceeds 1 often used in item analysis. |
| Widespread biserial correlation | r/wbis | Artificial Dichotomy | Continuous | Looking for extremes on Variable 1 |
| Point- biserial correlation | r/pbis | True dichotomy | Continuous | Yields lower correlation than r/biserial |
| Tetrachoric correlation (example: self-confidence Vs Internal Locus of Control) | r/t | Artificial Dichotomy | Artificial Dichotomy | Used when Var 1 and 2 can be split arbitrarily |
| Phi coefficient | f | True Dichotomy | True Dichotomy | |
| Correlation ratio eta | h | Continuous | Continuous | Nonlinear Relationships |

Multivariate Correlational statistics

If you wish to test a claim that multiple independent variables might be used to make a prediction about a dependent variable, you have several possible tests that can be constructed. Such studies involve Multivariate Correlational statistics

Discriminate Analysis - Used to determine the correlation between two or more predictor variables and a dichotomous criterion variable. The main use of discriminant analysis is to predict group membership (e.g. success/non-success) from a set of predictors. If a set of variables is found which provide satisfactory discrimination, classification equations can be derived, their use checked out through hit/rate tables, and if good, they can be used to classify new subjects who were not in the original analysis. In order to use discriminate analysis certain assumptions (conditions) must be met:

* At least twice the number of subjects as variables in study.
* Each groups has at least n=# of variables.
* Groups have the same variance/covariance structures.
* All variables are normally distributed.

Canonical correlation - Used to predict a combination of several criterion variables from a combination of several predictor variables.

Path Analysis- Used to test theories about hypothesized causal links between variables that are correlated.

Factor Analysis - Used to reduce a large number of variables to a few factors by combining variables that are moderately or highly correlated with one another.

Differential analysis - Used to examine correlation between variables among homogeneous subgroups within a sample, can be used to identify moderator variables that improve a measure's predictive validity.

Multiple Regression - Used to determine the correlation between a criterion variable and a combination of two or more predictor variables. As in any regression method we need the following conditions to be met: We are investigating linear relationships; for each x value, y is a random variable having a normal distribution. All of the y variables have the same variance; for a given value of x, the distribution of y values has a mean that lies on the regression line.

Note: results are not seriously affected if departures from normal distributions and equal variances are not too extreme.

The following example illustrates how a researcher might use different multivariate correlational statistics in a research project.

Example: Suppose a researcher has, among other data, scores on three measures for a group of teachers working overseas:

1. years of experience as a teacher
2. extent of travel while growing up
3. tolerance for ambiguity.

Research Question: Can these measures (or other factors) predict the degree of adaptation to the overseas culture they are working on?

Discriminate Analysis : Hypothesis (1) : The outcome is dichotomous between those who adapted well and those who adapted poorly based on these three measures. Hypothesis (2): Knowing these three factors could be used to predict success.

Multiple Regression: Hypothesis: Some combination of the three predictor measures correlates better with predicting the outcome measure than any one predictor alone.

Canonical Correlation: Hypothesis: several measures of adaptation could be quantified, i.e. adaptation to food, climate, customs, etc. based on these predictors.

Path Analysis - Hypothesis: Childhood travel experience leads to tolerance for ambiguity and desire for travel as an adult, and this makes it more likely that a teacher will score high on these predictors which will lead them to seek an overseas teaching experience and adapt well to the experience.

Factor Analysis - Suppose there are 5 more ( a total of 8) adaptive measures which could be determined. All eight measures can be examined to determine whether they cluster into groups such as: education, experience, personality traits, etc.

ANALYSIS OF COVARIANCE

A "yes" on question one below will lead you to a different type of statistical test involving bivariate data. If the researcher is claiming that:

___1. Two groups being compared come from the same population and contain similar characteristics.

When subjects are divided into 2 groups (perhaps a control group and an experimental group), or if two different treatments on two different groups will be given, then problems in randomization and matching the groups might be a concern.

A statistical process called analysis of covariance (ANCOVA) has been developed to equate the groups on certain relevant variables identified prior to the investigation. Pre-test mean scores are often used as covariates.

The following guidelines should be used in the analysis of covariance:

1. The correlation of the covariate (some variable different than the one you are testing) and the response variable should be statistically and educationally significant.
2. The covariate should have a high reliability.
3. The covariate should be a variable that is measured prior to any portions of the treatments.

4. The conditions of homogeneity of regression should be met. The slopes of the regression lines describing the linear relationships of the criterion variable and the covariate from cell to cell must not be statistically different.

5. All follow-up procedures for significant interaction or "post-hoc comparisons" should be made with the adjusted cell or adjusted marginal means.

ANCOVA is a transformation from raw scores to adjusted scores which take into account the effects of the covariate. ANCOVA allows us to compensate somewhat when groups are selected by non random methods

The non-parametric counterpart of ANCOVA is: the Runs Test.

If your hypothesis is that many variables or factors are contributing to a certain condition, you may wish to use multiple regression analysis. This is similar to linear regression analysis with a significant increase in number crunching. If this is not how you wish to spend several hours of your day, we recommend that you employ a computer to crank out the numerical information necessary to use multiple regression analysis.

## CONTINGENCY TABLES:

Sometimes a researcher is only interested in the following:

___1. Whether or not two variables are dependent on one another, (e.g. are death and smoking dependent variables; are SAT scores and high school grades independent variables?)

To test this type of claim a contingency table could be used, with the null hypothesis being that the variables are independent. Setting up a contingency table is easy; the rows are one variable the columns another. In contingency table analysis (also called two-way ANOVA) the researcher determines how closely the amount in each cell coincides with the expected value of each cell if the two variables were independent.

The following contingency table lists the response to a bill pertaining to gun control.

| | In favor | Opposed |
|---|---|---|
| Northeast | 10 | 30 |
| Southeast | 15 | 25 |
| Northwest | 35 | 10 |
| Southwest | 10 | 25 |

Notice that cell 1 indicates that 10 people in the Northeast were in favor of the bill.

Example: In the previous contingency table, 40 out of 160 (1/4) of those surveyed were from the Northeast. If the two variables were independent, you would expect 1/2 of that amount (20) to be in favor of the amendment since there were only two choices. We would be checking to see if the observed value of 10 was significantly different from the expected value of 20.

To determine how close the expected values are to the actual values, the test statistic chi-square is determined. Small values of chi-square support the claim of independence between the two variables. That is, chi-square will be small when observed and expected frequencies are close. Large values of chi-square would cause the null hypothesis to be rejected and reflect significant differences between observed and expected frequencies.