Highway Research Design Guide

WA-RD 154.1

Final Report
April 1988

Washington State Department of Transportation
Planning, Research and Public Transportation Division

in cooperation with the
United States Department of Transportation
Federal Highway Administration

WASHINGTON STATE DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO. WA-RD 154.1	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
4. TITLE AND SUBTITLE HIGHWAY RESEARCH DESIGN GUIDE		5. REPORT DATE April 11, 1988
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Cy Ulberg		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Washington State Transportation Center (TRAC) The Corbet Building, Suite 204 4507 University Way N.E. Seattle, Washington 98105		10. WORK UNIT NO.
		11. CONTRACT OR GRANT NO. WSDOT Y-3399, Task 29
12. SPONSORING AGENCY NAME AND ADDRESS Washington State Department of Transportation Transportation Building, KF-01 Olympia, Washington 98504		13. TYPE OF REPORT AND PERIOD COVERED Final Report
		14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This study was conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration

16. ABSTRACT

This manual is intended for use by state departments of transportation to design and carry out highway research. The examples in the manual cover a wide range of typical research projects, from testing pavement overlays to evaluation of transportation system management experiments.

The manual is divided into two parts. The first deals with issues in research design. Eight potential pitfalls in research are identified and ways to deal with them are discussed. The second section contains a discussion of simple statistical methods likely to be encountered in highway research.

17. KEY WORDS Research design, statistics	18. DISTRIBUTION STATEMENT No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22616

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE

DOT FORM 310-022 (11/86) -1216-

HIGHWAY RESEARCH DESIGN GUIDE

by

Cy Ulberg
Research Engineer
Washington State Transportation Center (TRAC)
The Corbet Building, Suite 204
4507 University Way N.E.
Seattle, Washington 98105

Final Report

Research Project Y-3399
Task 29

Prepared for

Washington State Transportation Commission
Department of Transportation
and in cooperation with
U.S. Department of Transportation
Federal Highway Administration

CONTENTS

FIGURES

TABLES

Highway Research Design Guide

by Cy Ulberg
Research Associate
Washington State Transportation Center

INTRODUCTION

This guide is intended for use by the Washington State Department of Transportation (WSDOT) to assist people who wish to conduct highway research projects. The examples used in the guide are typical of WSDOT research interests. Even though the guide is written specifically to apply to highway issues, much of the information is applicable to other types of research.

The guide is not intended to be a complete description of all possible research designs, nor is it intended to replace a statistics text. However, it covers the most important aspects of designing a research project and some of the basic statistical methods needed to answer most research questions.

Reasons for Research

Research is a structured way of asking questions. The questions can be very important ones. They may involve saving human lives. They may involve the expenditure of large amounts of money. The questions may have profound impacts on how people live and work for decades. Researchers should be concerned that the interpretation of research results leads to the best possible actions.

Research requires an investment. Time and effort must be devoted to the design of the research, the collection of data and the interpretation and communication of the results. Before starting any research project, someone should ask the question, "Is the research worth it?" Perhaps the results are already obvious. It could be that the investment in the research is out of proportion to the money that could be saved or the improvements that should be implemented. Maybe there is not enough time to complete the research before a decision is required. All of these questions should be addressed.

On the other hand, research is usually a sound investment. A relatively small amount of time and money invested in a good research project can save a great deal of money in the future or account for saving numerous lives. A well-designed research project can provide information that is persuasive to decision-makers and the general public when some new initiative is promising but requires some promotion.

Research as Hypothesis Testing

The first question you should ask before designing a research project is why are you doing it in the first place. A good understanding and description of research objectives will go a long way toward specifying how the research will be done. Too often, the research question is not well thought out and leads to results that do not answer any important question at all.

Byem Cheap wanted to find out what kind of pickup truck was the "best" for maintenance crews to use.

He took an inventory of all the pickups that had been retired from the fleet and divided their original cost by the number of years that they lasted. He found that Brand A cost $1,200 per year and Brand B cost $1,500 per year. He recommended that WSDOT buy only Brand A pickups. However, since he hadn't clearly defined what he meant by "best," he paid attention only to the initial investment and failed to notice that Brand A pickups were in the shop for maintenance twice as often as Brand B pickups. If he had added in the maintenance costs and other operating costs, he would have found that Brand B actually cost WSDOT $500 *less* per year to operate. Byem's recommendation cost WSDOT $200,000 per year.

The definition of the research question is usually in the form of a hypothesis. The researcher states a hypothesis that some way of doing things is better than some other way. In forming the hypothesis, care must be taken to clearly define what each "way of doing things" is and what "better" means.

Irwin Stopcrash wanted to determine the best way to place reflecting lane markers in construction areas. One way was to place them on the road surface and the other was to place them on the side of Jersey barriers. In order to define the two "ways of doing things," it was important to specify how far apart the reflectors were placed, how far from the side of the lane they were and how high up on the Jersey barrier they were to be placed. It was important to specify what kind of reflectors would be used and how they would be maintained. These and other specifications were critical in defining exactly what was being compared.

Secondly, Irwin had to define what "better" meant. Was cost an issue? If so, was the time for placement and maintenance of the reflectors taken into account? Were the effects on congestion important? Should weather conditions have been taken into account? Did safety considerations outweigh all others?

One way Irwin could have stated the hypothesis in this case is as follows:

Brand A reflectors placed on the roadway one foot from the barrier spaced at 20 foot intervals lead to fewer accidents under all weather conditions than Brand A reflectors placed 3 feet from the road surface on the side of the barriers spaced at 20 foot intervals.

The form and details of the research hypothesis will determine the structure of the research design, the kinds of data to be collected, the time frame for the research and how the results will be evaluated.

Stating the research in the form of an hypothesis may bother some people, because it can appear to be biased in favor of one alternative or another. You should be aware of at least three issues in dealing with this problem:

1) Typically, research is desired to see if some promising new way of doing things is better than the traditional way. For statistical reasons, stating the hypothesis in the form "the new way is better than the old way" is a conservative approach. It favors keeping the traditional methods. Given the risk and expense of changing methods, this is probably advantageous unless the new way can clearly be proven better than the old way.

2) The hypothesis can be stated in either direction. If there is some strategic reason to do so, the hypothesis can be formulated any way one wishes to.

3) The hypothesis formulated by the researcher does not have to appear in that form in the final evaluation of research results. Stating an explicit hypothesis is simply a means to help structure the research.

Research vs. Data Collection

Research is different from data collection. Data collection is almost always part of research, but data collection by itself does not constitute research. Research, by definition, entails the comparison of alternatives and the testing of hypotheses. Data collection is the gathering of information, which may or may not be used in research.

> Clarence Stopwatch collected data on the numbers of people and the amount of traffic at rest areas on the interstate highways. The data included numbers of vehicle, vehicle occupancies, use of the restrooms, water consumed and length of stay. The data were used to plan for new rest areas and determine the sizes required. However, this data collection effort was not research, since it involved no alternatives and tested no hypotheses.

Even though a data collection effort may not be of immediate use in a research project, it is still important that a data collection effort be designed so that it may eventually be used in research. To the extent possible, the data collection method should be consistent and well-documented.

When a research project uses data collected before the project was designed, the data are referred to as "secondary data." Often, a research project uses only secondary data, and the design must take into account the strengths and weaknesses of the previously collected data.

> Hiram Poolcrash wished to conduct a study of the effect of the HOV lanes on SR 520 on traffic safety. He hypothesized that the introduction of the HOV lanes caused an increase in accidents. Since the study was designed after the introduction of the HOV lanes, Hiram could not specify how the accident data were collected. He had to rely on secondary accident and traffic data. Using existing databases on highway accidents, it was often difficult to determine if an accident was the result of HOV lane operation. Therefore, Hiram had to rely on an analysis of total accident rates in the SR 520 corridor. Since many other factors influence accident rates, he used accident rate analysis in other corridors to control for these other factors. The results were not conclusive.

To the extent possible, people who conduct data collection efforts should take into account potential research uses for the data being collected.

Pitfalls in Research

Numerous factors can lead to misinterpretations of research results. Some are obvious and some are subtle. An objective of good research design is to minimize sources of misinterpretation, or at least to be aware of what the problems might be.

> Alfred Killdriver was interested in the importance of automobile size on the survivability of the occupants of automobiles in accidents. He compared the fatality rates in large, medium and small automobiles that were involved in accidents. To his surprise, he found that occupants of the smallest cars were more likely to survive accidents than those of larger cars. Before announcing this result and advocating that people should drive smaller cares, Alfred decided to examine some additional data. He found that younger people tended to buy and drive smaller cars. Since younger people are more resilient than older people, he speculated that similar injuries are less likely to lead to death. When he re-examined the fatality data and controlled for age of the occupant, he found that survivability was indeed better in larger vehicles.

Errors in interpretation are more likely in research conducted with secondary data. However, even when the data collection is designed for a specific research question, it is quite possible to come up with misleading results.

Statistics Need Not Be Complicated

Most people confronted with a statistics textbook are impressed by the myriad of statistical methods available. Maybe "depressed" is a better word. Even for people who are sophisticated in mathematics and technical analysis, statistics can seem like an unfathomable morass of jargon, assumptions and esoteric trivia. It doesn't have to be that complicated.

The audience for most applied research consists of decision-makers who have considerably less understanding of statistics than the people responsible for designing the research. Most often, research results are boiled down to a simple comparison of two numbers by the time a decision is required. The only use of statistical analysis is to determine whether the difference in those two numbers is "statistically significant." Even if the difference is statistically significant, it may not be significant in other ways.

> Roadway buttons had traditionally been placed using Adhesive A. A new product, Adhesive B, was proposed to replace the original adhesive. Horace Buttonplacer designed a research project to test the hypothesis that the use of A was better than the use of B. Ten thousand buttons were placed using Adhesive A and 10,000 using Adhesive B. After one year, 2,100 of the buttons applied with Adhesive A had been knocked off and 2,000 of the buttons using Adhesive B had been knocked off. This is a statistically significant difference. However, Horace's boss decided that the difference wasn't significant enough to bother with the paper work to change a standing order for Adhesive A. All the statistics in the world wouldn't change Horace's boss' mind.

It is usually advantageous to design research using the simplest statistics possible. Not only will it result in the most simple research, but the results will be understood and trusted by a wider audience than if more "sophisticated" or esoteric statistics are used.

Organization of the Guide

The remainder of this guide is divided into two major sections. One deals with research design and covers the major pitfalls that can lead to misinterpretation of research results along with ways to avoid them. The second section discusses basic statistics. It covers all that will be necessary for 90 percent of the research conducted by WSDOT. For the remaining 10 percent, the section discusses ways to avoid esoteric statistics, or as a last resort, how to find what is needed.

RESEARCH DESIGN

Attention paid to research design will help you avoid some of the pitfalls referred to in the introduction to this guide. There is never a guarantee that research results will be free of misinterpretation, but much of it can be avoided. In this section of the guide, eight factors that can affect the validity of research results are described:

1) outside effects,

2) time effects,

3) testing,

4) instrumentation,

5) central tendency,

6) selection,

7) drop-outs, and

8) sample size.

If research is designed to avoid problems due to each of these factors, the results are not likely to be subject to problems of validity. Sometimes, however, it will not be possible to completely avoid problems due to some of these factors. In those cases, you should at least be aware of the potential problems and interpret results accordingly.

The second part of this section discusses an approach to research design that will minimize problems due to the research pitfalls listed above. An ideal research design is presented. However, in recognition of the fact that the ideal is seldom attainable in applied highway research, some alternative designs are described that can avoid most sources of invalidity.

Sources of Invalidity

This section of the guide should be used as a checklist in the design of highway research. These eight factors cover the vast majority of all problems that may come up in the interpretation of research results.

Outside effects. One of the most common sources of problems in research is when some event or factor that is external to the subject of the research interferes with the results. Especially in applied research, as opposed to laboratory research, this type of **outside effect** is impossible to avoid.

> Jasper Roadfixer wanted to test the efficiency of two different methods of sealing cracks in the pavement. He was careful to define the methods fully and he hypothesized that Method A would be better than Method B. He decided that "better" meant that Method A would last longer than Method B, and, even though it cost more to implement, it would cost less on an annual basis. He selected a certain stretch of I-5 to test the methods on. By using Method A on the northbound roadway and the other on the southbound roadway, he felt that the environments would be the same for each method. Unfortunately, he failed to anticipate a construction project occurring nearby. For several months during the test, trucks full of earth were going northbound in the test section and returning empty southbound. Jasper was unable to tell if the extra damage in the northbound lanes was due to the extra weight of the trucks or to the inferiority of Method A for sealing cracks.

One way to minimize this problem is to try to anticipate every possible outside effect that is likely to occur during the course of a research project. In some cases, this may work. The more control you have over the environment of the research, the more likely you can anticipate and eliminate outside effects. The closer the research resembles laboratory research the easier this is.

However, in real-life applied research, it is virtually impossible to anticipate, much less eliminate, outside effects due to events or factors external to the experiment. The solution to the problem is to design the research so that if such an outside effect occurs, it is likely to occur to an equal extent for all of the alternatives being studied. The research designs described in a later section maximize this likelihood.

Time effects. Another external complication to the interpretation of research results can be the **time effects**. This source of invalidity is similar to outside effects, but not the same. An outside effect is an event or factor external to the subject of the research, while the effects of long-term processes can be intrinsic to the subject of research itself.

> Rheinhold Bridgebuilder was interested in determining if stiffening a bridge by welding on extra supports would reduce pavement cracking and thus save maintenance costs on the bridge. Before adding the supports, he measured the damage to pavement caused by traffic on the bridge and calculated the costs to repair the pavement that year. Three years after the supports were installed, he conducted the same analysis and found that maintenance costs had actually increased, even when inflation was taken into account. What he failed to account for was the long-term increase in traffic going over the bridge. Even though the extra supports were there, the extra traffic caused more damage than the supports were able to prevent.

Dealing with the time effects is similar to dealing with outside effects. One way is to eliminate all the time effects. This will usually be impractical in applied research. However, by making the duration of the research as short as possible, it may be possible to minimize these kinds of effects.

A second way to deal with the time effects is to find some independent way to measure them. In the example, if there had been some known quantitative relationship between traffic volumes and pavement damage on the bridge, it would have been possible to control for the long-term growth in traffic on the bridge. A statistical approach to this type of analysis is discussed later in this guide.

Another way to deal with the time effects is to arrange the research design so that the time effects are the same for all alternatives. Referring again to the example, this would have meant measuring pavement damage on some other bridge that had been experiencing an increase in traffic and did not have extra supports or had some other method of minimizing pavement damage.

Testing. One source of invalidity that is often ignored in the "hard" sciences is the effect of testing. In measuring people's attitudes, motivations or performance, it is easy to understand how the method used to collect data can influence people's behavior. For this reason, one of the requirements of data collection that involves humans is to be as non-obtrusive as possible. The same caution should be exercised in the collection of "hard" data.

> George Weighstation wanted to test the accuracy of a piezo-electric system for measuring the average axle-weight on trucks traveling on I-5 near the Nisqually Delta. He placed the cable two miles preceding a weigh station and recorded the axle-weights of trucks during randomly selected time periods. During those same periods, he recorded the axle-weights of trucks measured at the weigh station. He found that the average axle weight measured by the cable was significantly higher than that measured at the weigh station. he concluded that the cable needed to be recalibrated.
>
> What George failed to take into account was the effect of testing. Truck drivers were unaware that their trucks were being weighed at the cable site. However, since they new that the weigh station was open, some truckers who knew their trucks were overweight bypassed the weigh station or turned off the freeway before

> the weigh station to wait for it to close. Since only heavy trucks bypassed the weigh station, the average weight was biased on the light side. The fact that truckers knew they were being tested influenced their behavior.

Efforts should be made to design data collection methods that cause as little change as possible to whatever is under study. It is usually desirable to choose a more difficult or less accurate method to measure something if the alternative method of measurement will have a strong effect on what is being measured.

If it is impossible to find a method of measurement that won't affect what is being measured, an alternative is to design the research so that the measurement method affects all alternatives equally.

> Otto Roadbuilder was interested in comparing the effects of different kinds of subgrade on the long-term strength of concrete pavement. He had to measure pavement strength over a period of five years. One kind of instrument used to measure the strength of pavement required deforming the pavement. The test method possibly could do minor damage to the pavement and distort the future measurements of pavement strength. However, Otto decided to use the instrument anyway, since it would be used equally on the pavement over each kind of subgrade.

One cautionary note should be made here. Even when the measurement method will be used on all the alternatives equally, there still is a chance that it will have unequal effects on each one. The first priority should be to find a measurement method that will have minimal effects on what is being measured.

Instrumentation. Testing problems refer to the influence of the measurement procedures on whatever is being tested. **Instrumentation** problems refer to problems with the method of measurement itself. If the method employed to collect data varies from location to location or from time to time, the validity of results will surely be questionable.

> Vinton Overweight wanted to determine the effects on overweight violation rates of increasing the enforcement of weight limits on trucks. He used a method to weigh trucks in motion to determine the number of overweight trucks. The method employed a piezo-electric cable imbedded in a groove in the pavement filled in with resin. To his surprise, he found that with strict enforcement, the violation rates actually were increased. What he failed to take into account, however, was that, over time, the surrounding pavement had been worn down, resulting in tires producing more pressure on the cable as time went on. The artificially high weights led him to assume that violation rates had increased.

Often measurement problems can be avoided simply by employing some technique to calibrate instruments. However, some kinds of instrumentation do not lend themselves readily to calibration and other kinds of measurement techniques cannot be calibrated at all. For instance, it is very difficult to calibrate measurement techniques that involve human observers.

Fortunately, there usually are ways to deal with instrumentation problems when measuring tools cannot be satisfactorily calibrated. The solution is in the research design. If the research is designed so that any changes in measurement methods are the same for all alternatives, differences in the alternatives cannot be attributed to differences in measurement methods. If all alternatives being tested are equally likely to experience instrumentation problems, the differences in measurements can be considered random errors and treated statistically like any other random error. The last part of this guide deals with statistical treatment of random error.

Central tendency. Outliers are events or things that are very different from the norm. If a research project involves the observation or measurement of extreme cases, the possibility for error in interpretation is high. For statistical reasons that can best be explained using an illustration, anything that has an extreme measurement one time is more likely to be less extreme the next time than to be even more extreme. This has been called regression to the mean, or **central tendency**.

Ivan Headlight was interested in testing the effect of placing signs telling drivers to turn on their headlights on reducing fatal accidents. He gathered accident statistics for three years and found stretches of interstate where fatal accidents were most prevalent. Signs were placed on some of those stretches. After three years, fatal accidents decreased 50% and Ivan declared the headlight sign program a success.

Ivan failed to notice, however, that the overall rate of fatal accidents had not changed at all and that they had increased in other stretches of road. Even though there is a greater tendency for accidents to occur in some places rather than others, there is a degree of randomness about how frequent they are in any stretch of road for a short amount of time. Ivan's choice of location for the headlight signs was based on places that just happened to have high rates during the three years he looked at. Because of the normal ups and downs of random events, one would expect that the fatal accident rate would be lower the next time the rates were measured there.

This source of invalidity occurs only when the selection of locations or subjects of research is based on extreme measurements. If a whole range of sites are selected, the chances of choosing ones that are high on the scale are balanced out by the chances of choosing ones low on the scale, and the problem will be avoided.

Sometimes, however, there are good reasons to choose extreme sites for study. Ivan's research project, for instance, would probably not be conducted where there never are fatal accidents. However, the researcher must control for the tendency of extremes to move toward the normal. One way to do this is to measure alternative treatments (including no treatment at all) at several extreme sites. A research design that incorporates multiple sites will allow you to measure the natural tendency for extremes to move toward normal measurements and to make valid comparisons.

Selection. Another important source of invalidity in research design occurs in the **selection** of locations or subjects for research. When testing the effectiveness of different ways of doing things, it is critical to know if different results would have occurred in the absence of any difference in the way of doing things.

Ronald Snowplow wanted to find the least expensive way to deal with the problem of snowplows knocking over guideposts on the side of the road. When a wooden post was hit, it almost always had to be replaced. When a new polymer-based type of guidepost was hit, it was more likely to bounce back and continue to be useful. However, since the new type of guidepost was more expensive than the wooden type, Ronald had to demonstrate that its greater survivability was worth the expense.

To test this hypothesis, Ronald placed new guideposts along ten miles of I-90 over Snoqualmie Pass. He alternated types of guideposts every half mile. After a winter's worth of snowplowing and other damage, he went back and found that three times fewer polymer-based guideposts had to be replaced than the wooden ones. Since they were only twice as expensive to install, he concluded that it was cost effective to replace all wooden guideposts with the new ones wherever there would be significant snowplowing.

Ronald went part way in the selection of equivalent sites for testing each kind of guidepost by placing them on alternate half-miles. However, he failed to take into account one of the major factors in guidepost damage, the curves in the road. Snowplows are much more likely to hit guideposts on curves than on the straight-a-ways. By chance, the curves were most prevalent on the every other half mile where the wooden guideposts were placed. The locations were not equivalent and there would have been more guidepost damage in those places, no matter what type of guidepost was placed there.

Research design can help solve the problem of non-equivalence. One way of dealing with non-equivalent selection of research alternative locations or subjects is to assign the alternatives randomly. If there are a large enough number of locations or subjects (which there weren't in Ronald's case), the likelihood of non-equivalence becomes very small.

Another way of dealing with the problem is to measure the factor of interest before assigning different alternatives. If Ronald had measured guidepost loss using all wooden signs, he would have discovered that the two sets of half mile sections were different in the first place. Often, it is not possible to choose research sites or subjects randomly. In these cases, it is imperative that some sort of initial measurement be made in order to understand the underlying differences.

Drop-outs. In many research efforts, interpretation of the findings is hindered by the fact that the subjects of the experiments drop out during the process. This is especially true when the research covers a long time period. It is also true if the research itself causes **drop-outs** to occur. It is obvious how this can occur with human subjects in research. The researcher may simply lose contact with the people or the human subjects may decide that they don't want to continue participating. Even when the subject of research is not human, this source of invalidity may become problematic.

If drop-outs occur randomly, the fact that they drop out will not cause problems for the results of the research. However, most often, the characteristics of drop-outs are different from the characteristics of those that remain and may lead to invalid differences in the results.

> Axelrod Reflecto was interested in finding a means to measure the reflectivity of signs that was less expensive and faster than using a reflectometer. He thought that one could train human observers to pick out signs that were below standard in reflectivity and eliminate the necessity for expensive testing. He set up an experiment in which people looked at a large sample of signs that had been tested with a reflectometer. Over a period of five days, the judges looked at signs and were given feedback about their judgments. By the end of five days, judges were very accurate in reproducing the results of the reflectometer, and Axelrod considered his demonstration a success.
>
> Later, however, when people were hired to go out and actually test signs, Axelrod found that their accuracy was not as good as the experiment had shown. What he failed to take into account in his experiment was that some of the experimental judges had dropped out of the testing because they were bored and were not doing well at making the judgment. The small amount they were paid for the experiment was not enough to keep them involved. However, when people were hired at higher wages to do the actual judgments, and they made their judgments out on the highway, they were less likely to get bored and drop out. The people who were not very good at making the judgments of reflectivity continued working in the real case and their ability to detect bad signs showed.

One way to deal with the drop-out factor is to make very strenuous efforts not to allow any drop-outs by investing the time and energy required to find all the cases possible. Often, however, the most strenuous efforts will not yield enough of the drop-outs to avoid validity problems.

Another way to account for drop-outs is to do a special study of the cases that drop out. In the example, for instance, Axelrod could have gone back and searched diligently to find a few of the missing signs and measured their reflectivity to see if it differed from those that remained standing.

The presence of drop-outs is one of the most difficult sources of invalidity to control with appropriate research design. Even if you have information about the drop-outs from some initial measurement, they may be impossible to distinguish from those subjects of research that don't later drop out. By definition, once they have dropped out, they can't be measured again.

Sample size. In general, the more measurements you make, the more sure you can be about research results. However, there are usually fiscal and time constraints on how much measurement is possible. One important issue in designing research is determining how many measurements to make in order to draw appropriate conclusions, yet not spend more money on the research than it is worth.

> Percy Speedbump had noticed that there were several fatal accidents on some long, straight stretches of I-5 in which the drivers had fallen asleep and driven off the side of the road. He thought that placing rumble strips on the side of the roadway might wake up people who were dozing off, yet not interfere with normal driving. He picked a long stretch of road where five such accidents had occurred in the previous three years and had rumble strips placed at one-mile intervals. During the next three years, there was a 40 percent drop in the fatality rate, since only three such accidents occurred. Since the placement of the rumble strips was relatively inexpensive, Percy claimed that the savings of two lives was well worth the cost and effort of the rumble strips.
>
> During the next three years, ten accidents occurred due to drivers falling asleep at the wheel on that same stretch of highway. Percy had read too much into the results of his research. On such a small stretch of road, the number of fatal accidents went up and down considerably due to many factors not influenced by the rumble strips. In order to validly test the ability of rumble strips to save lives, they would have had to be placed on ten times as much highway to have had a sample size large enough to draw the valid conclusion that a 40 percent drop was likely to occur again wherever rumble strips were placed.

The required sample size can be reduced through research design. Stratification, or drawing separate samples from separate categories, can reduce the number of required measurements. This technique is discussed in the section on statistics but should be used with caution.

The simplest solution for avoiding invalidity due to inadequate sample size is to have a large enough sample. Ways to estimate the proper sample size are covered in the section on statistics. The question of the costs and benefits of adequate sample size needs to be addressed during the research design phase, not afterwards when it is too late.

Research Design

Research design consists of several elements. Critical elements of the design include the following:

1) defining the alternatives,

2) timing the application of the alternatives,

3) deciding on measurement tools, and

4) timing the measurements.

Elements 1) and 3) have been discussed in the preceding section under hypothesis testing. Clearly stating a research hypothesis means that the alternatives (or different "ways of doing things") are clearly defined. The choice of appropriate measurement tools will be determined by the definition of what "better" means when the hypothesis is stated as "Method A is a better way of doing things than Method B."

The timing and arrangement of the alternatives and the measurements are crucial to the design of the research. The following discussion concentrates on these elements of research design. In this section, each way of doing something (the alternatives) will be represented with an "X." When more than one way of doing something is being discussed, they will be referred to as "X_1," "X_2," and so on. The current way of doing things, or "base case," will be referred to with a "C," for "current" or "control."

Measurements will be designated with an "O" for "observation." Multiple observations will be referred to as "O_1," "O_2," and so forth. The time frame in the illustrations will be from left to right. Successive lines refer to different research locations or subjects. An "R" in front of a line means that the location or subject has been chosen randomly.

Classic research design. The classic research design employs before and after measurements on all alternatives and on a "control group" in which no change occurs. The locations or subjects for the alternatives and the control group are chosen randomly. Using the representation described above, the classic research design looks like this:

R	O_1	X_1	O_2
R	O_3	X_2	O_4
R	O_5	C	O_6

In this example, two alternatives are tested and compared with a control group.

> Jasper Roadfixer wanted to compared two new ways of sealing cracks in the pavement with the current way of doing it. He was interested in minimizing the moisture in the subgrade. He wanted to employ the classic research design.
>
> First, he chose three hundred separate, fifty-yard-long sections of roadway and measured the moisture content in the subgrade in the middle of each section. He randomly picked three groups of one hundred sections. In one group, the traditional crack sealing method was used. In the other two groups, the two new methods were applied. After three months, he went back and measured the moisture content in the subgrade in the middle of each section. He computed the changes in each section and the average changes for each group. He found that one of the new methods was significantly better than the old method and the other new method and suggested using that method from now on.

In this example the application of the two new methods of crack sealing are X_1 and X_2. The old method is C (the old method could have been to do nothing at all, if that were a reasonable alternative). O_1, O_3 and O_5 represent the measurements taken before the application of the crack sealing and O_2, O_4 and O_6 are the measurements three months later. The differences (O_2 minus O_1, O_4 minus O_3 and O_6 minus O_5) are the main criteria by which to determine the best method. In the example, the biggest difference was O_2 minus O_1.

The classic research design will help avoid most sources of invalidity discussed in the section above. The sample size should be large enough that a predetermined difference among the alternatives can be detected. How to choose that will be discussed in the last major section of this guide. All of the other sources of invalidity are minimized because the application of the alternatives was randomly assigned. If there were any differences due to those sources they could be detected with the before and after measurements at each site.

It is not always possible to use the classic design for research, especially in applied field research. The rest of this section discusses alternatives when elements of the classic design are not possible to achieve.

Lack of random assignment. Often it is not possible to assign different alternatives randomly to different locations. This is especially true when the experimental treatment is costly or has some political factors involved. *It is still important to have a comparison group and to have before and after measurements to minimize problems of interpretation.*

> Gomer Rideshare was interested in seeing if the introduction of HOV lanes was a significant factor in motivating more people to form carpools. There were only two locations in the region where HOV lanes seemed both necessary and possible to build. For reasons outside of Gomer's control, the decision had already been made to build an HOV lane at Site A.
>
> Gomer knew that there were many differences between Site A and Site B, but he wanted to design his research as close to the classic form as possible. He took counts of the number of carpools at both Site A and Site B before the HOV lane was built. He found that the rate of carpooling was about the same at both places. After construction, and some time after people had a chance to change their commuting patterns, he measured the number of carpools at both sites again, and found that the per-

centage of carpools had remained the same at Site B, but had increased at Site A. He concluded that the HOV lanes had increased the motivation for people to form carpools. Since there were only very questionable counter-hypotheses for the increase in carpools at Site A, Gomer was able to convince policy makers and members of the public that HOV lanes were successful in promoting carpooling.

The research that Gomer designed could be diagramed like this:

O_1	X	O_2
O_3	C	O_4

X represents the construction of HOV lanes and C represents the absence of HOV lanes. If the O's are the percentages of carpools observed, O_1 and O_3 are equal in this example and O_2 is greater than O_4.

Even though the initial observations are equal, there is no guarantee that the sites are equivalent. Many aspects of the sites would have to be measured, such as traffic volume, road geometries, numbers of access points and the like, in order to convincingly state that the sites are equivalent, and they probably will not be equal in all respects. However, the more that you know about the sites, the more accurately you can interpret the research results. Even if the initial observations of carpool percentages, O_1 and O_3, are unequal, the values help to interpret differences between the final observations, O_2 and O_4.

Lack of "before" measurements. Sometimes it will be impossible or unwise to collect data before an alternative way of doing something is introduced. Through randomization, you can still design a research project that avoids validity problems.

Cedric Testtube wanted to test whether the irradiation of concrete pavement would improve the long-term ability of the concrete to withstand heavy concussions. Unfortunately, the only way to test improvement in strength of the blocks was to destroy them. He couldn't conduct before and after measurements. He had 100 test blocks and assigned 50 at random to be irradiated and 50 to have nothing done to them except be exposed to the weather for one year. At the end of the year, he tested all the blocks and found that the threshold for crushing the irradiated blocks was significantly higher than the others. Cedric recommended using irradiated concrete for certain critical parts of bridges.

Cedric's research could be diagramed like this:

R	X	O_1
R	C	O_2

where X represents the irradiation of the blocks and O_1 and O_2, the measurement of their strength.

Even though this type of design doesn't involve initial measurements that *guarantee* that the groups are equivalent, random assignment means that, in all likelihood, they are equivalent. The larger the sample sizes involved, the more likely that is to be the case.

There is actually an advantage to this type of design over the classic research design. That is that it avoids testing problems. By not performing the initial measurements, the researcher doesn't risk biasing the results. This is especially true when human subjects are involved.

If testing problems are an important issue in the research, this design can be combined with the classic control group design to actually measure the effects of testing. The design would look like this:

R	O_1	X	O_2
R	O_3	C	O_4
R		X	O_5
R		C	O_6

Comparison of O_2 with O_5 and O_4 with O_6 can determine if the measurement method is influencing the results of the research.

Lack of a "control group." A third deviation from the classic research design is when a comparison, or control, group is not possible to find. If observations are being made at just one location or on just one subject or group of subjects, it is very difficult to determine whether or not there are problems of outside effects, time effects, testing, instrumentation, or drop-outs. If the research looked like this:

$$O_1 \quad X \quad O_2$$

it would be very difficult to determine what caused any differences between O_1 and O_2. There are at least two ways to deal with this situation.

> Hiram Poolcrash wanted to understand the impact of a specific HOV lane on traffic safety. In the year before the HOV lane was introduced, there had been 55 accidents in the corridor they were located. In the year after their introduction, there were 65. Hiram was not convinced that the HOV lanes had caused the increase in accidents. He examined the accidents in the three years before and the three years after the introduction of the HOV lanes and found the numbers to be 35, 45, 55 and 65, 75, 85, respectively. He interpreted the original increase he found as just a part of a long-term pattern (and he was probably right).

Hiram's research design is called a time series and can be represented like this:

$$O_1 \quad O_2 \quad O_3 \quad X \quad O_4 \quad O_5 \quad O_6$$

By taking *several* measurements before and after the introduction of the alternative being researched, some of the possible sources of invalidity can be minimized. If a difference in observations O_3 and O_4 is not part of a pattern that can be explained by all the other measurements, then X probably caused the difference, rather than something else. On the other hand, if the difference is part of a larger pattern, then X probably didn't cause it (such as in Hiram's case).

There is another way of dealing with a lack of a "control group" under certain circumstances. If the alternatives being tested in the research have only temporary effects, then repeated applications of the alternatives and repeated measures may be appropriate.

> Oscar Stoplight wanted to test the effects of two ways of coordinating signals on traffic flow at a certain difficult intersection. There was no other intersection in the jurisdiction that was anything like this intersection, so all the testing had to be done there. Furthermore, traffic patterns varied so much from day to day and from month to month, Oscar was not sure that he could distinguish the effects of different signal timing methods from different patterns in traffic.
>
> He decided to alter the signal timing every week for six months and measure traffic flow during that time. Even though traffic varied drastically over that six month period, each type of signal timing was applied over the whole range of traffic patterns. He was able to show that one of the signal timing methods improved flow by 20 percent and that the improvement was statistically significant.

Oscar employed a research design that looked like this:

$$X_1 O_1 \quad X_2 O_2 \quad X_1 O_3 \quad X_2 O_4 \ldots\ldots$$

By repeating and alternating the different applications, the different observations of X_1 and X_2 are less likely to be attributed to problems like outside effects, time effects, or instrumentation. In a sense, the research location or subject becomes its own "control group."

Dealing with real life. Applied research very frequently requires compromises from ideal research design. The examples shown above illustrate some specific ways of handling certain exceptions. However, sometimes even the exceptions don't fall into neat packages. You are challenged to find ways to design the research to take the realities into account. The key to doing that is to consider each of the possible sources of invalidity and to "patch up" the research until the problem does not allow a strong hypothesis to compete with the one that you defined in the first place.

STATISTICS

Research usually requires the quantitative comparison of numerical results. The researcher's basic question is whether or not one number is different from another.

> Horace Buttonplacer wanted to know if one method of fastening buttons to the pavement was better than another. He placed one thousand buttons using one method and another thousand using the other method. To measure effectiveness, he returned after three months and counted the number of buttons that had been knocked off. Eighty-five buttons adhered by the first method had been knocked off and 83 buttons using the other method had been knocked off. This didn't seem like much of a difference to Horace.

The likelihood of two measurements being exactly the same is very low. Often, however, results yielding a very small difference means that the numbers might just as well be the same. The important research question is, "When are the differences large enough to be meaningful?" What if Horace had found 150 buttons missing with the first method and only 83 missing with the second? Is that difference large enough to be meaningful? Statistics is a branch of mathematics that allows researchers to make this kind of judgment.

What can be called a "meaningful" difference has different interpretations. One interpretation concerns whether the difference can be replicated. If Horace were to conduct the same test again, is the same result likely to occur? The other interpretation involves the question of whether the difference is large enough to recommend that an action be taken. Even if Horace could reproduce the same results over and over, the improvement that could be gained by switching to the second method might not be important when policy-related considerations were taken into account. Even if the second method produced better adherence, the difficulties of switching to that method might not be worth the effort.

For the most part, the second interpretation of "meaningfulness" is outside the scope of this research guide. However, the first interpretation is the subject matter of statistics. In statistics, this type of "meaningfulness" of a difference is referred to as the **significance** of the difference. Statistical significance is often described by a probability (p). For example, one often sees a notation for significance such as $p<.05$. This notation means that the probability that the difference occurred by chance is 5 percent. In other words, it means that there is a 95 percent chance that a difference in the same direction would occur if the same research were repeated.

Statistical methods can be used to determine the probability of a result recurring. For the vast majority of research likely to be carried out by WSDOT, only a small number of statistical tools will be required. Usually, a research project can be designed so that it requires only the more simple methods of analysis. However, you should be aware of a few precautions when you use only the simplest statistical methods. The objective of this part of the guide is to give you a good basis for understanding the simple statistics and the precautions necessary to use them. In addition, for more complex situations where more sophisticated statistical analyses are required, the guide will suggest what type of statistics might be employed.

The Normal Distribution

One of the basic concepts necessary for understanding the use of simple statistical methods is the

normal distribution. It has been referred to as "the bell-shaped curve," "the standard curve," "the Guassian distribution, "the random distribution" and other names. The curve shown in Figure 1 describes the normal distribution and should be familiar to most people.

The shape of the curve can be defined in mathematical terms. The equation defining the curve was not determined from experiments but was derived mathematically. The actual mathematics and derivation of the formula are unimportant here. It *is* important to understand that the normal distribution is only an approximation to reality, but it is a very good one.

The normal curve shows the probability that observations of random events will fall in a certain range. Figure 2 shows 30 observations of traffic volumes during the peak hour at a hypothetical location. The distribution of observations is shown with a diagram using increments of 100 vehicles per lane-hour. This kind of diagram is called a **histogram**. The continuous line is the normal curve that approximates this distribution. The higher the curve, the more likely the observation will be in that range. In the example, it is very unlikely that less than 1400 or more than 2100 vehicles per hour will be in a lane during the peak hour.

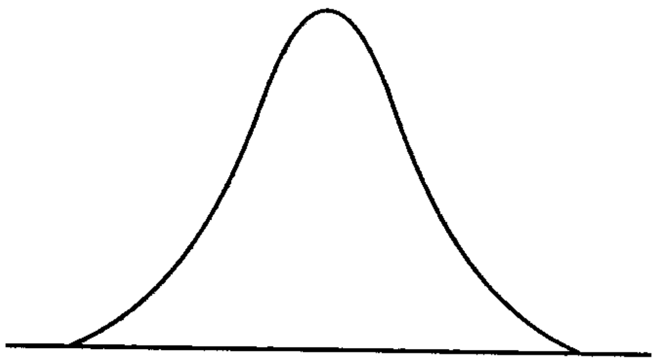


Figure 1. The Normal Distribution

__Precautions.__ For reasons that are beyond the scope of this guide, most events occur with distributions that are very close to the normal curve. Very few phenomena can be said to truly occur at random. Some of the variability can usually be explained. For instance, in the example of distribution of volumes of freeway traffic, we know that the day of the week determines the volume of traffic to some extent. We also know that seasons, weather, pavement condi-

Observations:

1470	1886	1836	1687	1418	1782
1822	1740	1784	1613	1604	1873
1713	1555	1716	1965	1934	1605
1952	1629	1899	1803	1990	2002
1827	1760	1914	1745	1735	1992

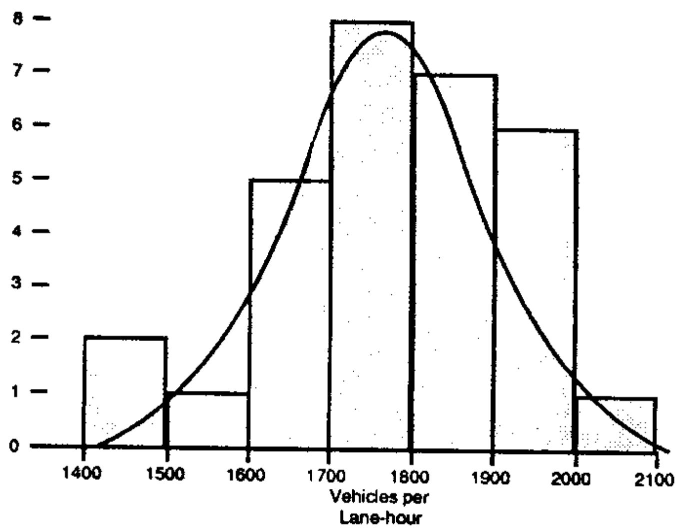


Figure 2. Distribution of Traffic Volumes

tions, and construction projects can all affect traffic volume. The list of known determinants of traffic volume is large. However, even when a number of causes for different traffic volumes are known, when they all act together, the result tends to look random. Even events that are not random tend to occur with a normal distribution.

There are exceptions, however. Some things occur with predictable differences from the normal distribution. For instance, Figure 3 shows the distribution of vehicle lengths on a rural interstate freeway segment. There are many passenger cars and pickup trucks with length under 25 feet and many semi-trucks around 45-50 feet in length, but very few

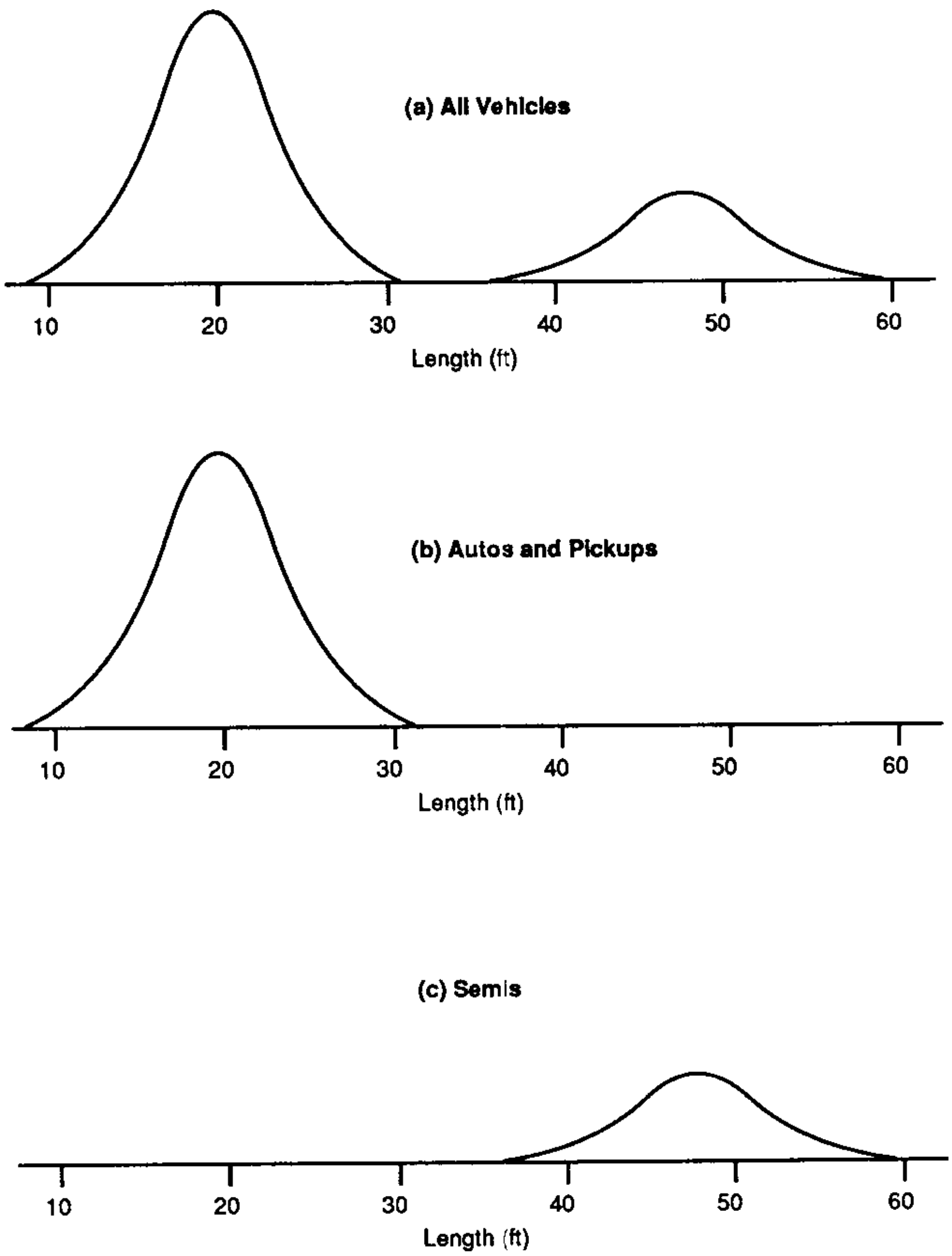


Figure 3. Distribution of Vehicle Lengths

vehicles with lengths in between. This distribution is called **bi-modal**. Statistics assuming a normal distribution may be misleading if applied to this particular kind of measurement. Note, however, that the distributions of passenger car lengths and semi-truck lengths taken individually more closely approximate the normal distribution.

Another potential problem with assuming the normal distribution is that there is a often a cutoff point at one or the other ends of actual distribution. The normal distribution assumes that the measurement can extend infinitely in either direction. Taking the peak hour vehicle count as an example, we know that the number cannot go below zero. If the speed limit is 55 mph and the average vehicle length is 25 feet, and vehicles are literally bumper-to bumper going at the speed limit, the absolute maximum number of vehicles per lane-hour is 11,616. In this case, the extremes are so far removed from the majority of the observations, that there is little danger in assuming a normal distribution. That is not always the case.

An example illustrating this problem is vehicle occupancy. Figure 4 shows a typical distribution of occupancies. The most frequent number is one, and it is also the minimum possible. This is clearly not shaped like a normal distribution. It also differs from a normal distribution because there are no observations possible between each of the numbers. It is impossible to have 2 1/2 occupants in a vehicle. However, suppose you made several observations of the *average* vehicle occupancy. The results may look like Figure 5. In this case, the distribution is more like a normal one, even though there is a theoretical minimum observation of 1.0. In fact, as will be discussed further in a later section, the *averages* of any kind of distribution are distributed in a way that closely approximates the normal distribution.

Even though most of the statistical methods likely to be used assume normally-distributed observations, that assumption can often be violated without leading to serious problems in the interpretation of results. The researcher should exercise caution, however, and be aware when distributions have peculiar qualities.

Descriptive statistics. A few simple concepts can be used to describe a set of data. They can be used to describe any kind of distribution. One set of concepts is used to describe the *center* of a distribution. Another set describes the *shape* of the distribution.

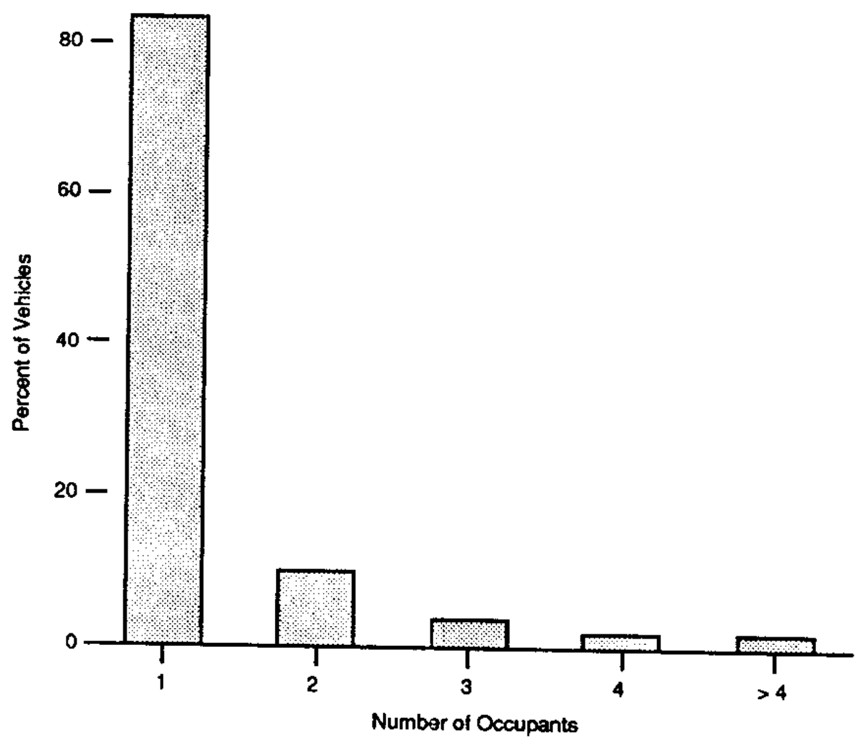


Figure 4. Distribution of Vehicle Occupancies

The most important way to describe the center of a distribution is the **mean**. It is also referred to as the "average." The terms are interchangeable. To compute it, you simply add up all the observations and divide by the number of observations. Another descriptor of the middle of a distribution is the **median**. The median is the point at which half of the observations are below and half above. A third way to describe the middle is the **mode**. This is the most frequent observation.

The most common statistic used to describe the shape of a distribution is the **standard deviation**. The standard deviation can be thought of as the average difference from the mean. It measures how widely the observations are spread. When one subtracts the mean from each of the observations, the result is positive for observations above the mean and negative for those below the mean. If one were to add up these differences, the result would always be zero. Therefore, the average is computed by first squaring all the differences and dividing by the number of observations to get the **average squared difference**. The square root of that result is the standard deviation. The average squared difference is referred to as the **variance**. An alternative way to compute the variance uses the sum of the squares of observations.

Table 1 illustrates how the mean and standard deviation are computed for some observations of traffic volume. Other statistics can be used to describe distributions. They employ the average cubed differences, the average quadrupled differences and so on. However, they are of little practical use in most research.

In computations of the standard deviation, one often sees the sum of squared differences divided by the number of observations *minus one*. For reasons that are unimportant here, that is technically the proper method to compute the standard deviation when one is dealing with a *sample* of all possible observations. However, when the number of observations is large, it makes little practical difference whether you divide by the total number, or one less.

Proportions and probabilities in the normal distribution. Even though the use of the mean and standard deviation is acceptable for any type of distribution, they are especially useful in describing the normal distribution. They can be used to estimate the likelihood that observations will fall into predetermined intervals.

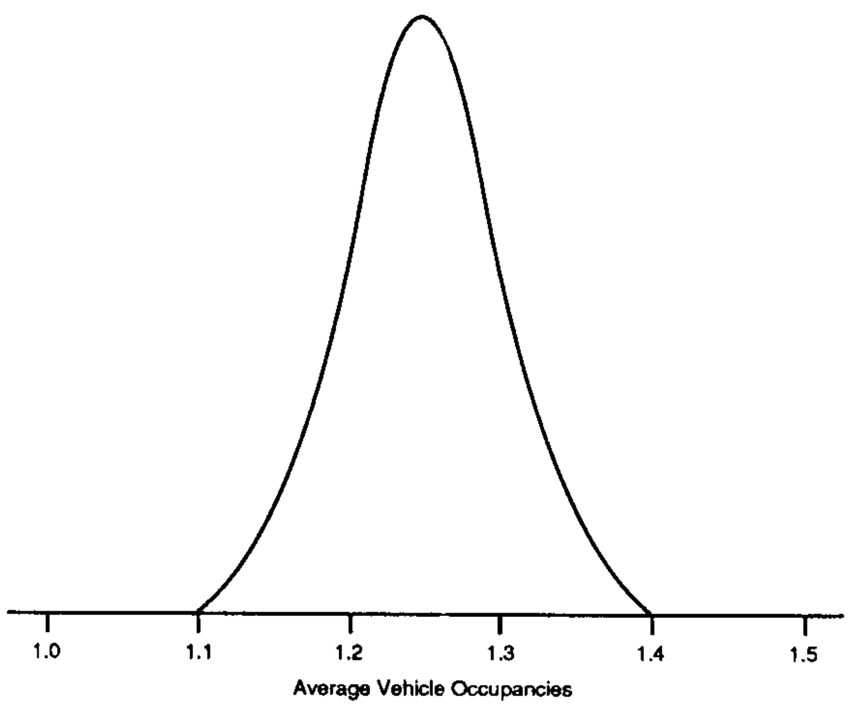


Figure 5. Distribution of Average Vehicle Occupancies

Table 1. Computation of Mean and Standard Deviation

Obs. Number	Traffic Volume	Observation Minus the Mean	$(\text{Observation}-\text{Mean})^2$	$(\text{Observation})^2$
1	1642	-162	26244	2696164
2	1837	33	1089	3374569
3	1942	138	19044	3771364
4	1714	-90	8100	2937796
5	1812	8	64	3283344
6	1546	-258	66564	2390116
7	1972	168	28224	3888784
8	1731	-73	5329	2996361
9	1883	79	6241	3545689
10	1775	-29	841	3150625
11	1664	-140	19600	2768896
12	1942	138	19044	3771364
13	1750	-54	2916	3062500
14	2010	206	42436	4040100
15	1840	36	1296	3385600
sum = Σ	27060	0	247032	49063272

$$\text{mean} = \frac{\Sigma \text{obs}}{N} = \frac{27060}{15} = 1804$$

$$\text{variance} = \frac{\Sigma(\text{obs-mean})^2}{N} = \frac{247032}{15} = 16468.8$$

$$\left(\text{or, } \frac{\Sigma(\text{obs})^2}{N} - (\text{mean})^2 = \frac{49063272}{15} - (1804)^2 = 16468.8\right)$$

$$\text{Standard Deviation} = \sqrt{\text{variance}} = 128.33$$

The area under the normal curve can be used to estimate probabilities. Using the traffic volume data from the example in Table 1, if you were to draw vertical lines around the interval between 1500 and 1700 vehicles per lane-hour (shown in Figure 6), the proportion of the total area covered by the shaded area would be the probability that any given observation will fall between 1500 and 1700 vehicles per lane-hour. Using the mean and standard deviation for this distribution, that probability is .2001. Since the curve is a mathematical function, it is possible to compute the area under any part of the curve using integral calculus. However, that is unnecessary, since tables have been constructed to save that work. A table describing the normal curve is included in the back of the guide.

Table 2 contains a step-by-step description of how the probability of .2001 was arrived at using the example data. The first step is a conversion of the raw numbers (1500 and 1700) to **normal deviates**. This oxymoron (look it up) is another useful statistic. It is also called the "z-score." By subtracting the mean from the observation and dividing by the standard deviation, the result is the number of standard deviations away from the mean that the observation lies. The conversion to the normal deviate allows the use of the normal distribution table, no matter what units are being used and no matter what the mean is. By locating normal deviates on the table, and subtracting, the result is the proportion of the total area under the curve bounded by 1500 and 1700 vehicles per lane-hour. The probability that any observation will fall in that interval is about 20 percent. Figure 7 shows why the subtraction works this way.

> Abner Autocount got a reading of 2,280 vehicles per lane-hour one day using his automatic equipment. Using the mean and standard deviation from previous data (the example in Table 2), he computed a z-score of 3.71. He looked in the normal distribution table, and saw that the probability of a reading this high or higher is only .00011 (1-.99989), or about one out of ten thousand. Abner decided to check out his equipment, since a reading this high is so unlikely to occur.

Another feature of the normal distribution table that is useful to keep in mind is that the number of observations that fall within certain limits (expressed as standard deviations) will always be the same. For instance, the number of observations that fall within one standard deviation of the mean is 68.3 percent; within two standard deviations, 95.4 percent; and within three standard deviations, 99.7 percent.

In most research, it is not necessary to compute the probability of observations falling into certain

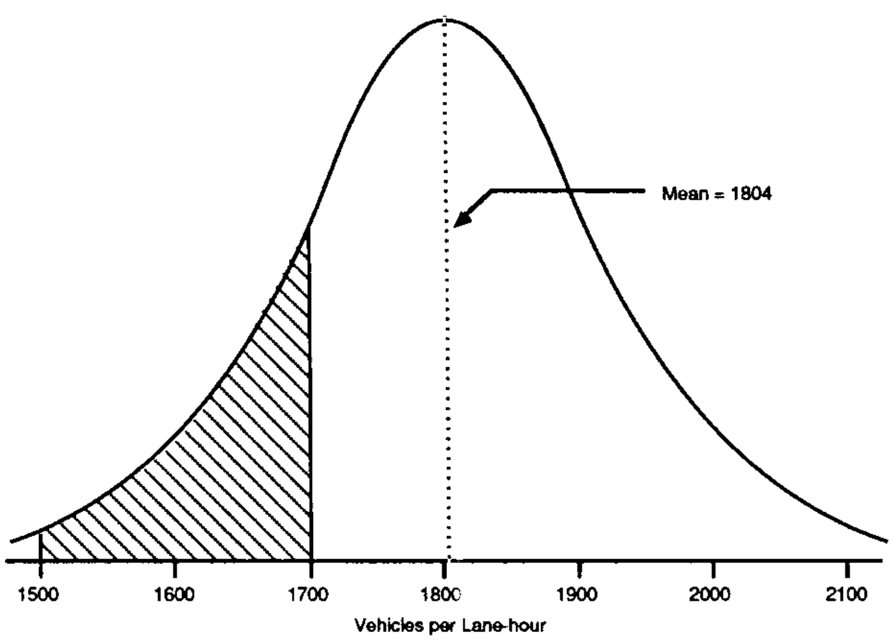


Figure 6. Distribution of Traffic Volume from Table 2

Table 2. Computation of Probability the Traffic Volume Lies Between 1500 and 1700

Traffic Volume	Deviation from Mean	Normal Deviate	Probability from Table
1700	1700 - 1804 = -104	$\frac{-104}{128.33} = -.81$	.2090
1500	1500 - 1804 = -304	$\frac{-304}{128.33} = -2.37$	.0089
		Difference =	.2001

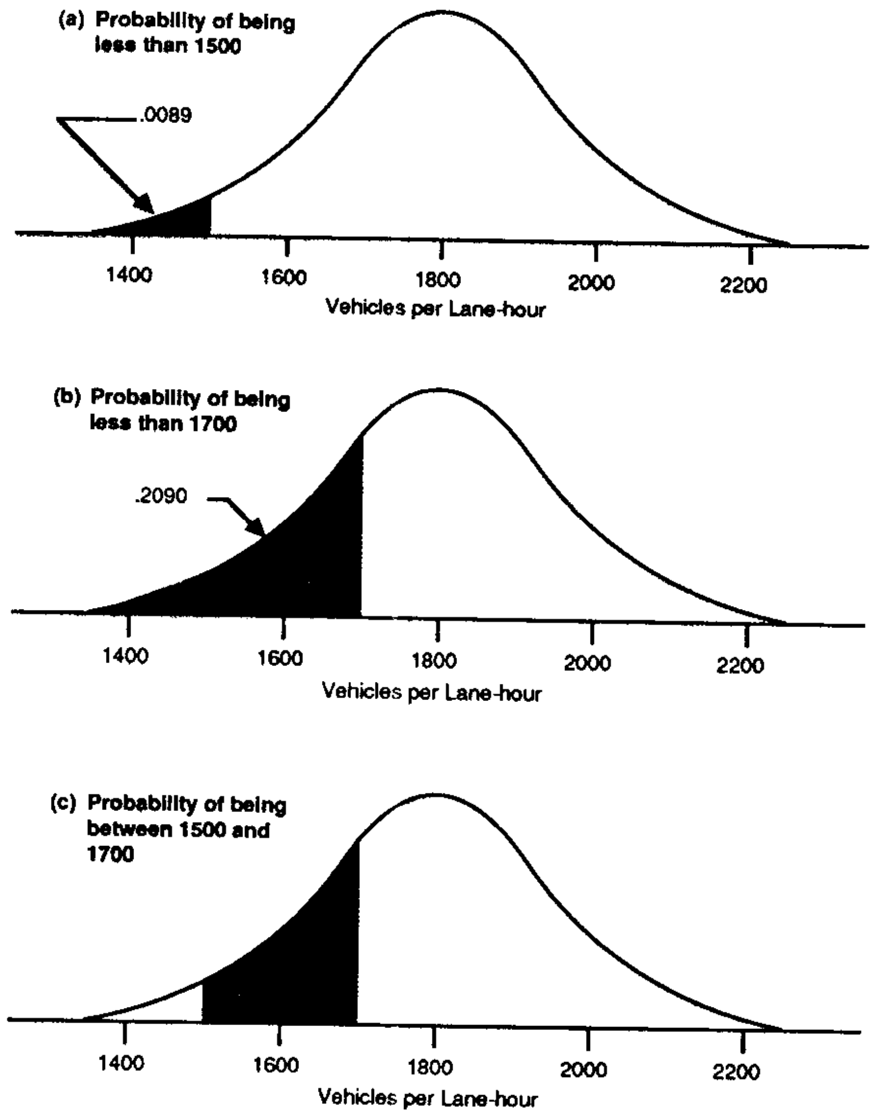


Figure 7. Illustration of the Use of Normal Deviates

intervals, such as in the illustrations in this section. However, it is important to understand the underlying concepts involved, because they have a direct bearing on the determination of the significance of observed differences. Before going on to the next section, make sure that you understand how the area under the normal curve relates to the probability of making observations.

Sampling

It is very rare that we get measurements on all instances of something we are interested in. We may be limited by time or by money. We may be trying to forecast something in the future that is impossible to measure now. "All instances of something we are interested in" is called the **population**. A part of the population is called a **sample**. Some examples of a population are:

1) all the peak hour traffic volumes on a particular freeway segment,

2) all the peak hour traffic volumes on all freeway segments in the county on a particular date,

3) all the peak hour traffic volumes on all freeway segments in the county, and

4) all of the peak hour traffic volumes on all freeway segments in the last five years.

The population can be defined any way you want. However, the definition of the population implies some restrictions. First, the definition determines how the sample can be picked. For instance, in the above examples, definitions (2) and (4) allow the possibility of measuring the whole population. All the others include measurements at future times that could not be made now. Secondly, and most important, you can generalize from a sample only *to the population from which that sample was chosen.*

Selecting the sample. Two basic types of samples are possible. One is called a "convenience" or "judgment" sample. In this type of sample, a part of the population is selected because it is easier to measure, because someone believes that it represents the range contained in the whole population, or because it guarantees that certain examples in the whole population are represented. For example, in choosing a sample of population definition (3) above, one may pick only freeway segments that are close to the office to avoid a lot of travel. Samples chosen on these bases or similar ones do not represent the whole population of interest and are rarely acceptable for research.

The second kind of sample is a random sample. There are many ways to choose a random sample. The primary requirement is that every member of the population be equally likely at the outset to become part of the sample. The simplest way to choose a random sample is to list the entire population and use random numbers to pick items from the list. For example, for population (3), one could list all the freeway segments in the county and, using a random number list, pick the sample.

Sometimes, however, a researcher may wish to make sure that all different types within a population are represented. In this case, it may be appropriate to choose a **stratified random sample**. In this method, the population is sorted into categories (**strata**), and a separate random sample is selected from each category. For instance, the researcher in the example may divide all the freeway segments in the county into two categories: those within city limits and those outside city limits. A separate sample could be drawn randomly from each list.

If one wishes to generalize about the whole population from a stratified random sample, however, it is important to take into account the relative sizes of the strata. In the example, if there are twice as many freeway segments in cities as those outside the cities, the sample from the cities should be twice as large as the sample from outside the cities, or the results for each category should be weighted appropriately.

Interpretation of the sample mean. No matter how a sample is selected, the objective is to draw conclusions about the whole population. The larger the sample is, the more confident one can be that the measurement of a sample represents the whole population. In fact, one can quantify how representative it is by computing **confidence limits**. The confidence limits are the range within which you can be fairly sure that the measurement of the population falls.

Suppose that the observations of traffic volumes in the above example constitutes a random sample. The population is defined as all the traffic volumes on all lanes at a particular freeway segment during the morning peak hour on workdays during the current year. There were 252 workdays and 4 lanes, for a total population of 1008 measurements. We could have obtained all 1008 measurements but chose to use a sample of 15 measurements. Our best estimate of the population mean is the sample mean, or 1804 vehicles per lane-hour. But how good is that estimate?

We can calculate the confidence interval using something called the **standard error**. The standard error is the best estimate of how much the sample means would vary if we chose several samples from the total population. Figure 8 shows the distribution of all 1008 observations of traffic volumes. It also shows the distribution of all samples of 15 measurements at a time that we could choose. You can see that the sample means do not vary as much as the observations themselves. One way to understand this is to think of how the variations in daily observations cancel each other out when means are computed for 15 measurements.

The standard error is computed by dividing the standard deviation of the population by the square root of the sample size. In our example:

$$\sigma_{\bar{x}} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}} = \frac{128.33}{\sqrt{15}} = 33.13$$

This result means that if we collected several different samples (n=15) of traffic volume data, the average sample mean would still be 1804, but 68 percent of the means would be within 33.13 of 1804, or between 1770.87 and 1837.13. Another way of stating this is that we can be 68 percent sure that any sample mean will be within 33.13 of the population mean.

Now we have all the elements necessary to interpret sample means. Make sure you understand the following terms and concepts:

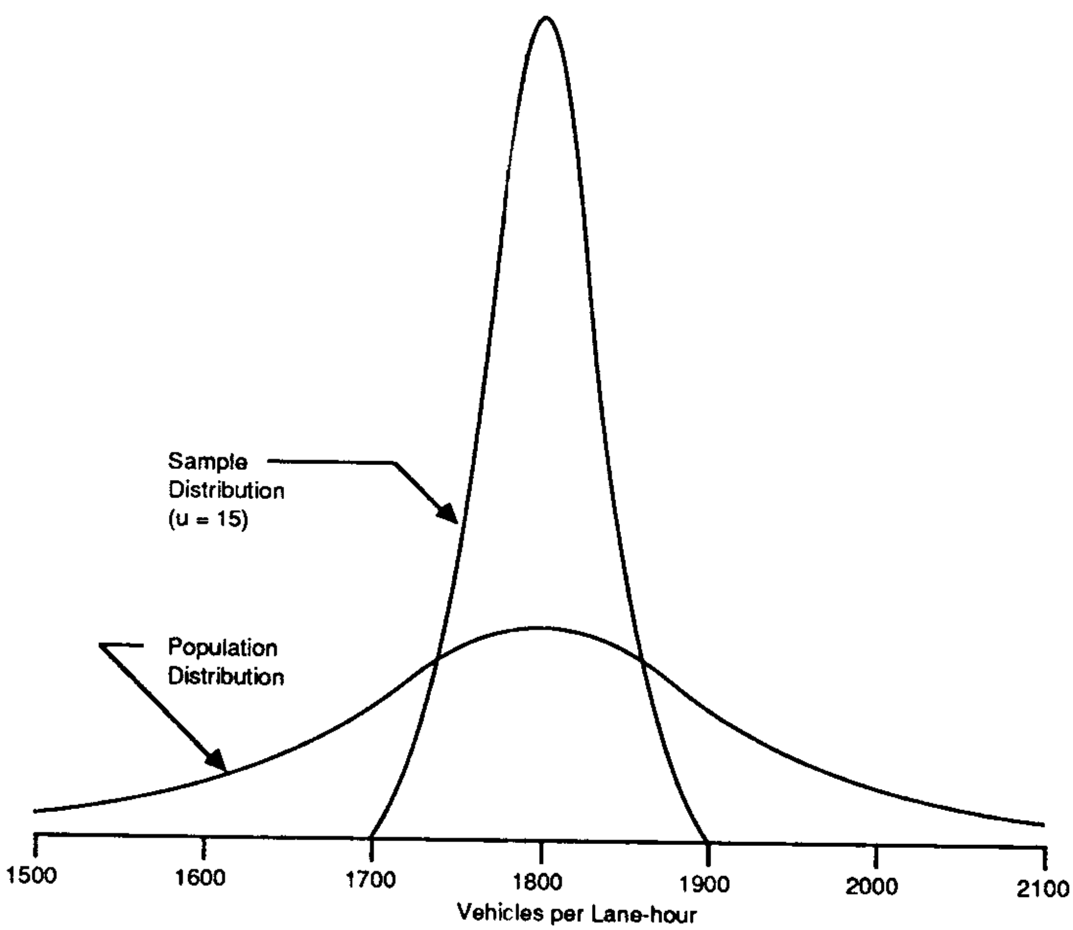


Figure 8. Population and Sample Distribution for Traffic Volumes

- mean,
- standard deviation,
- population,
- sample,
- sample mean and
- standard error.

With a thorough understanding of these concepts, you can interpret most research findings.

Sample proportions. Sometimes the data collected in a research project are not in a form from which a mean can be computed. The results may be proportions or percentages. Much of the preceding discussion of sampling applies to proportions as well as to means of observations. However, some distinctions must be made.

Proportions are used when the researcher is interested in the number of cases that fall into different categories. For instance, vehicle occupancy data may be collected either way. An average vehicle occupancy can be determined by computing the mean number of people in vehicles. On the other hand, the researcher can also summarize results by reporting the proportions or percentages of vehicles with certain numbers of people in them. Table 3 shows results of a vehicle occupancy study reported in both ways. Since the distribution of the observations is not normal, the interpretation of the mean and the standard deviation of the number of people in vehicles can be problematic in some contexts. The *proportion* of SOVs (single occupancy vehicles) does not suffer from this problem.

The proportion of SOVs in the sample is a way of estimating a characteristic of the whole population, just as the average vehicle occupancy is a way to do it. The proportion has a standard error that is interpreted in the same way as the standard error of the mean. It is also easier to compute. The formula for the standard error of a proportion is

Table 3. Vehicle Occupancy Data

	Observations						
		Carpools					
Observation	SOV's	2	3	4	Total	AVO	%SOV's
1	1265	121	32	13	1431	1.157	88.4
2	1296	128	39	13	1476	1.166	87.8
3	1259	145	46	11	1461	1.185	86.2
4	1370	129	37	13	1549	1.156	88.4
5	1362	116	34	12	1524	1.144	89.4
6	1284	114	36	10	1444	1.150	88.9
7	1235	101	43	8	1387	1.152	89.0
8	1240	104	32	5	1381	1.133	89.8
9	1276	150	44	13	1483	1.187	86.0
10	1273	130	32	9	1444	1.153	88.2
11	1236	101	41	11	1389	1.156	89.0
12	1413	134	41	8	1596	1.150	88.5
13	1302	105	40	9	1456	1.146	89.4
14	1297	127	33	9	1466	1.150	88.5
15	1370	136	35	12	1553	1.156	88.2
Average	1299	123	38	10	1469	1.156	884

$$\sigma_p = \sqrt{\frac{p(1-p)}{N}}$$

where **p** is the population proportion and **N** is the total number of observations. In the example shown in Table 3, the standard error of the proportion of SOVs is computed as follows:

$$\sigma_p = \sqrt{\frac{.884(1-.884)}{1469}} = .0084$$

In this computation of the standard error of the proportion, we use the average values. We could have used any of the observations, as well, with similar results.

The confidence interval for a sample proportion is analogous to that for a mean. In this example, we can predict that in 68 percent of all samples of this size collected from this population, the resulting proportion of SOVs will be .88 plus or minus .008. Of the 15 observations in the figure, 10 (or 67 percent) are within the predicted range. If a larger sample were used, the confidence interval would be smaller.

Testing a Difference

This section contains a description of the most commonly used statistics, those that determine whether a difference is statistically significant or not. The first part deals with differences in means and the second with differences in proportions.

Differences in means. One of the most common situations that occurs in research is the determination of whether the difference between two sample means is statistically significant. The researcher may test averages from two different points in time or averages from two different kinds of treatments. The basic question is "Are the averages really different?" The question of whether the means are significantly different can be stated in a different way: "What is the probability that the two samples come from the same population?"

> Abner Autocount was interested in determining if the traffic volumes in the summer are significantly lower than during the winter months. Figure 9 shows distributions of a random sample of traffic volume observations he collected during two different months, August and October. The results looked very different, but he was uncertain that the difference meant anything, since the samples were so small. Strictly speaking, Abner should have collected data randomly from all the winter months and all the summer months if he wished to define populations including whole seasons. Since the samples were chosen randomly from the two months, he could at least draw conclusions about those two months. Generalizing to whole seasons would be a little risky.
>
> There was substantial overlap in the *observations* from each of the months. However, in this case, Abner was interested in the *average* volumes in the two months. To determine if the *averages* could be from the same population, he first had to calculate the standard errors. The standard error for the October sample was 12.3 vehicles per lane-hour. Since Abner read an earlier section of this guide, he knew that there was a 99.7 percent chance that the true population mean for October lay within three standard errors of the sample mean, or somewhere between 1701 and 1775. Using the same kind of calculation, he was 99.7 percent sure that the August mean lay between 1615 and 1695. Since there was no overlap in these two intervals, he knew the likelihood that the true population means for August and October were the same was extremely low.

Usually, the significance of the difference between two means is not so obvious as in Abner's case and the confidence intervals overlap to some extent. Fortunately, there is a simple way to calculate the probability that the two means come from the same population. Think of the difference between the means as one of several possible differences that could be computed by drawing new samples, computing the means and subtracting one from the other. Those differences have a distribution, just like the means themselves. They also have a standard error, just like the means. The best estimate of the population difference is the difference between the two

	Observations	
	Aug.	Oct.
	1684	1684
	1607	1738
	1665	1785
	1680	1677
	1621	1758
	1617	1782
	1748	1703
	1608	1787
	1681	1720
	1673	1747
Mean	1655	1738
Std. Dev.	42	40
Std. Error	13.2	12.3
Conf. Int.	1655±3(13.2)	1738±3(12.3)
(99.7%)	(1615-1695)	(1701-1775)

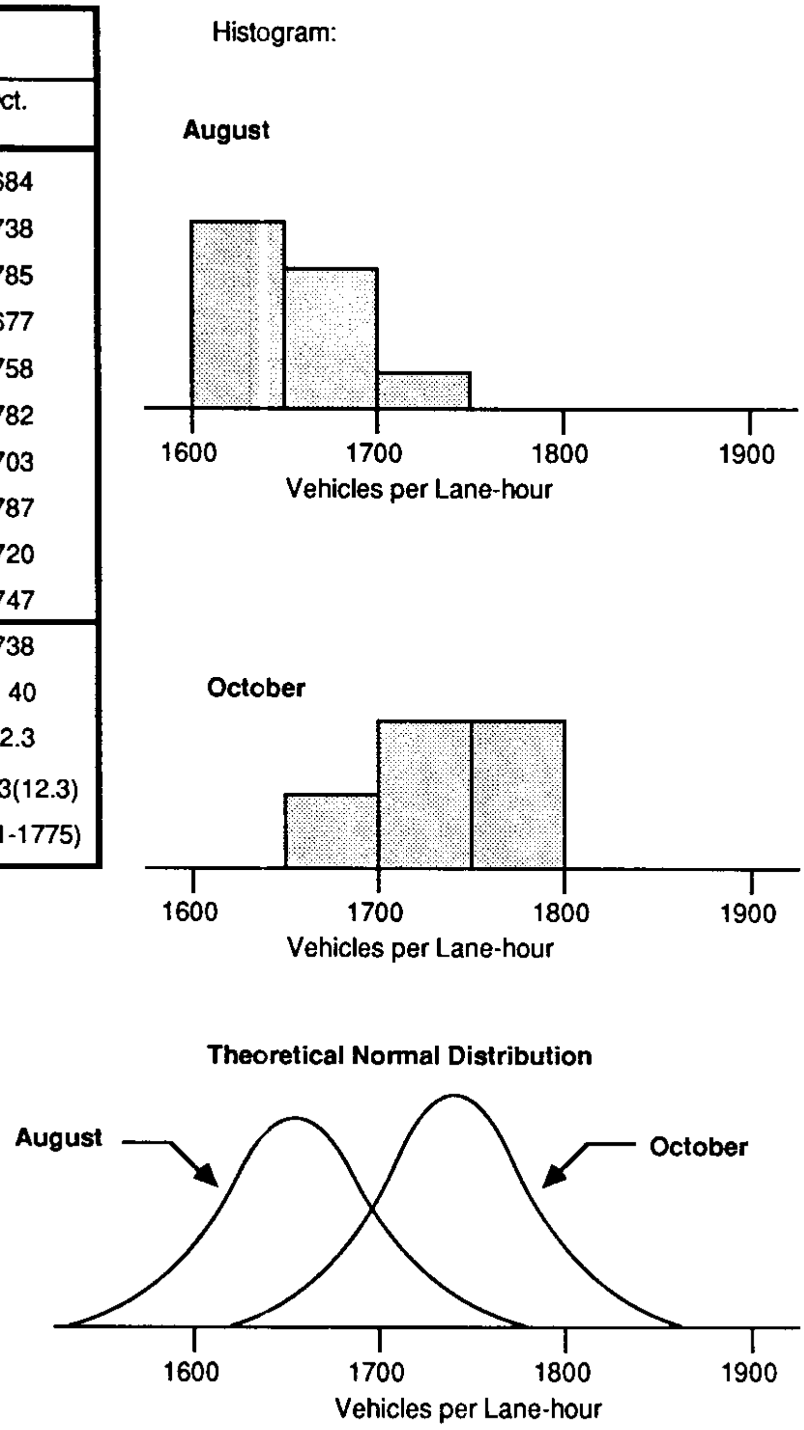


Figure 9. Comparing Traffic Volumes in August and October

sample averages. The standard error of the difference is

$$\sigma_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

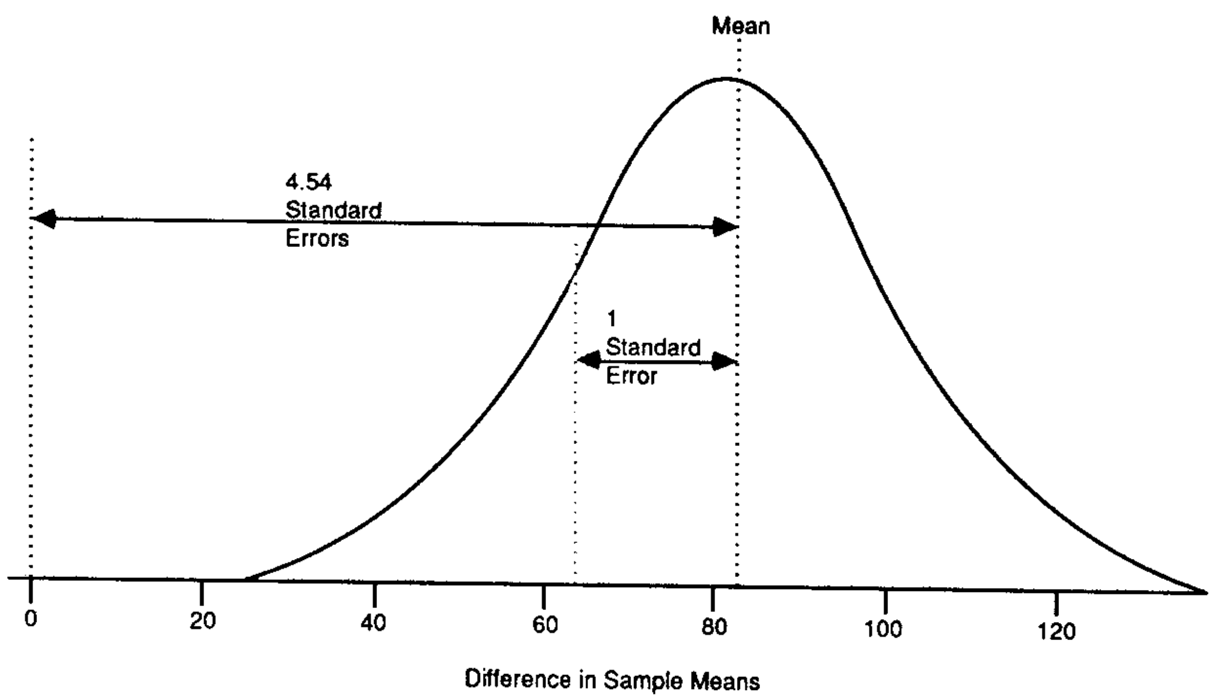
where s_1 and s_2 are the standard deviations of each of the samples and n_1 and n_2 are the numbers of observations in each sample.

In Abner's case, the difference is 83 vehicles per lane-hour and the standard error of the difference is 18.3 vehicles per lane-hour. To test the significance of the observed difference in the two samples, Abner computed the probability of the actual difference being 0 or less using the normal curve. Figure 10 shows what the distribution of differences would be, if Abner had collected all possible samples of 10 for August and October. He could see that 0 lay about 4.5 (83/18.3) standard deviations away from the observed difference. The likelihood of 0 or less being the population difference was extremely small.

Differences of proportions. Differences in proportions are handled analogously to the differences in means. The standard error of the difference in proportions is

$$\sigma_d = \sqrt{p_c(1 - p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where n_1 and n_2 are the numbers of observations in the two samples and p_c is the combined proportion of



$$\text{Standard Error of the Difference} = s_d = \sqrt{\frac{(13.2)^2}{10} + \frac{(12.3)^2}{10}} = 18.34$$

Figure 10. Distribution of the Differences in Sample Means

the two samples, which can be calculated using this formula:

$$P_c = \frac{n_1 p_1 = n_2 p_2}{n_1 + n_2}$$

Abner found that 82 percent (a .82 proportion) of the vehicles in the October survey were SOV's and 78 percent of the vehicles in the August survey were SOVs. He surveyed a total of 5000 vehicles in October and 7000 vehicles in August. He calculated the standard error of the difference in proportions to be

$$P_c = \frac{5000(.82) + 7000(.78)}{5000 + 7000} = .7967$$

$$\sigma_d = \sqrt{.7967(1 - .7967)\left(\frac{1}{7000} + \frac{1}{5000}\right)} = .00745$$

Since the difference in proportions Abner observed in the samples was .04, the chance of the actual population difference being 0 was small. Zero lay (.04 - 0)/.00745, or over 5 standard deviations away from .04. Abner concluded (rightly) that more people tend to drive alone in August than in October.

Testing multiple differences

Often research involves testing more than one difference at a time. Several statistical analysis approaches exist to deal with this situation. Details of these methods are beyond the scope of this guide. However, it is important that someone conducting a research project be aware that, in some cases, more sophisticated tests than have been described here may be warranted. Almost any statistics textbook can provide the details of how to conduct these tests, and there are numerous computer packages available to calculate the statistics.

Analysis of variance. One case is when more than two alternatives are under examination at the same time. For instance, a research project to evaluate three different ways of sealing pavement cracks under two different environmental conditions is such a case. The aim of the research may be two-fold. One aim is to find the best method of sealing cracks under all conditions and another would be to see if the environmental conditions affect which method is the best.

One appropriate approach for this kind of situation is the use of **analysis of variance**. This technique is designed to test multiple alternatives (different crack-sealing methods) at once and to test for interactions among the variables (do environmental conditions affect how well each method works). It is possible to conduct the same analysis by examining the six possible treatments two at a time. However, the more appropriate approach is to use analysis of variance. This technique separates the variability due to different treatments and conditions from that due to random variation.

Regression analysis. A second situation is when the researcher desires to determine how two factors for which there are multiple observations relate to each other. For instance, a researcher may desire to compute a quantitative relationship between accident rates and traffic volumes. If she has measurements of accident rates and traffic volumes for several years and in several locations available, she can develop such a quantitative description of the relationship using **regression analysis**. This type of analysis takes advantage of the existence of several measurements and shows how strongly these two variables are related to each other. It is also possible to determine how multiple factors influence something all at once.

A simpler approach is to test differences in means. The observations of accident levels could be divided into two classes, those with high traffic volumes and those with low traffic volumes. The mean accident levels for each case could be tested using the approach outlined in the previous section on difference testing. In this example, however, the effects of central tendency might be a problem since the two samples would be selected based on extreme values of traffic volumes.

Chi-square test. A third case is testing differences in categorical data. Suppose a researcher wanted to see if the *distribution* of vehicle occupancies in two situations were significantly different. The average vehicle occupancies might be equal, but, in one case, a large number of SOVs might be balanced out by a large number of vans. The differences in proportions of vehicles with each occupancy level could be tested one at a time. However, there are several statistical approaches appropriate for situations like this. This type of analysis is called **non-parametric statistics**. A specific test to use in this example is the **chi-square test**.

Sample Size

The researcher has to choose how many observations to make of each alternative under investigation. As we have seen in all the above examples, the size of the confidence intervals and the degree of certainty with which we can draw conclusions from research results depends on the number of observations that are recorded.

The computation of required sample size can be done simply by working backward from a prediction of the differences one expects and the degree of certainty one requires in the results to the number of observations required. Take, for instance, the equation for the standard error of the mean,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{u}$$

Solving this equation for n, we get

$$n = \frac{\sigma_x^2}{\sigma_{\bar{x}}^{2}}$$

One way to interpret this equation is that if you want your standard error to be a certain level and you estimate that the standard deviation of the observations will be a certain number, you need to have **n** observations. However, the question remains, "How do you determine the standard error and the probable standard deviation of the observations?"

> Abner Autocount wanted to see if the volume of traffic in lane 4 at a particular point on I-5 was significantly different from that in lane 3. He didn't know how often he would have to count traffic to be sure of his results. He decided that if the two samples he collected differed by 30 vehicles per lane-hour during the peak hour, he wanted to be able to say that that difference reflected a *real* difference in traffic volume. Furthermore, he wanted to be 95 percent certain of the finding.
>
> In statistical terms, Abner wanted to be 95 percent sure that a difference of 30 vehicles per lane-hour would indicate a real difference in populations. This means that 30 vehicles per lane-hour would have to have a z-score of 1.96. One standard error would therefore be 30/1.96 or 15.3. Suppose Abner knew from previous research that the standard deviation of the differences in observations would be 100. He computed the required sample size as follows:
>
> $$n = \frac{(100)^2}{(15.3)^2} = 43$$

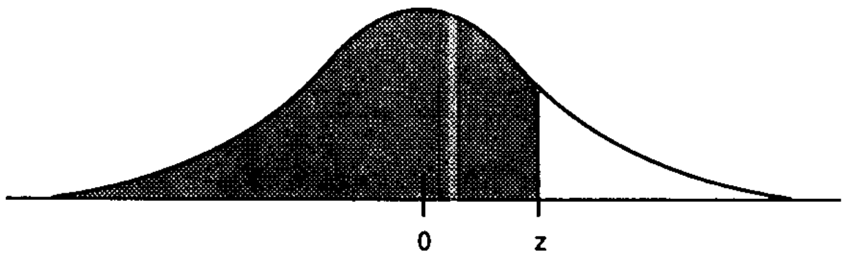
You will not always have previous research from which to estimate the standard deviation of the observations. Another way to estimate the standard deviation is to consider the extreme values you expect to observe and divide by six. For instance, if the highest traffic volume is expected to be 2200 and the lowest, 1600, the estimated standard deviation would be 100. The reason for dividing by six is that plus or minus three standard deviations encompasses 99.7 percent of all observations. The estimate of the standard deviation can be corrected, if necessary, and the number of observations adjusted accordingly, after some of the data have been collected.

In general, then, sample size can be determined if you choose two things,

- the size of difference you want to detect and
- the certainty with which you want to say that the difference is significant.

The computation of the sample size requires knowledge of, or an estimate for, the standard deviation of the observations.

Normal Distribution



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7793	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000