Chapter one

Introduction: Statistics as a Research Tool

Initial hurdles

Do Statisticians Have to Be Experts in Mathematics?

Are Computers Making Statisticians Redundant?

Key principles

What is Our Aim in Choosing a Statistic?

What Basic Principles Apply to Different Types of Statistics?

What Are the Different Uses of Statistics in Research?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_1,

THE PURPOSE OF STATISTICAL ANALYSIS is to clarify and not confuse. It is a tool for answering questions. It allows us to take large bodies of information and summarize them with a few simple statements. It lets us come to solid conclusions even when the realities of the research world make it difficult to isolate the problems we seek to study. Without statistics, conducting research about crime and justice would be virtually impossible. Yet, there is perhaps no other subject in their university studies that criminal justice students find so difficult to approach.

A good part of the difficulty lies in the links students make between statistics and math. A course in statistics is often thought to mean long hours spent solving equations. In developing your understanding of statistics in criminal justice research, you will come to better understand the formulas that underlie statistical methods, but the focus will be on concepts and not on computations. There is just no way to develop a good understanding of statistics without doing some work by hand. But in the age of computers, the main purpose of doing computations is to gain a deeper understanding of how statistics work.

Researchers no longer spend long hours calculating statistics. In the 1950s, social scientists would work for months developing results that can now be generated on a computer in a few minutes. Today, you do not need to be a whiz kid in math to carry out a complex statistical analysis. Such analyses can be done with user-friendly computer programs. Why then do you need a course in statistics? Why not just leave it to the computer to provide answers? Why do you still need to learn the basics?

The computer is a powerful tool and has made statistics accessible to a much larger group of criminal justice students and researchers. However, the best researchers still spend many hours on statistical analysis. Now that the computer has freed us from long and tedious calculations, what is left is the most challenging and important part of statistical analysis: identifying the statistical tools that will best serve researchers in interpreting their research for others.

The goal of this text is to provide you with the basic skills you will need to choose statistics for research and interpret them. It is meant for students of criminology and criminal justice. As in other fields, there are specific techniques that are commonly used and specific approaches that have been developed over time by researchers who specialize in this area of study. These are the focus of this text. Not only do we draw our examples from crime and justice issues; we also pay particular attention to the choices that criminal justice researchers make when approaching statistical problems.

Before we begin our study of statistics in criminal justice, it is useful to state some basic principles that underlie the approach taken in this text. They revolve around four basic questions. First, what should our purpose be in choosing a statistic? Second, why do we use statistics to answer research questions? Third, what basic principles apply across very different types of statistics? And finally, what are the different uses for statistics in research?

The Purpose of Statistics Is to Clarify

It sometimes seems as if researchers use statistics as a kind of secret language. In this sense, statistics provide a way for the initiated to share ideas and concepts without including the rest of us. Of course, it is necessary to use a common language to report research results. This is one reason why it is important for you to take a course in statistics. But the reason we use statistics is to make research results easier—not more difficult—to understand. For example, if you wanted to provide a description of three offenders you had studied, you would not need to search for statistics to summarize your results. The simplest way to describe your sample would be just to tell us about your subjects. You could describe each offender and his or her criminal history without creating any real confusion. But what if you wanted to report on 20 offenders? It would take quite a long time to tell us about each one in some detail, and it is likely that we would have difficulty remembering who was who. It would be even more difficult to describe 100 offenders. With thousands of offenders, it would be just about impossible to take this approach.

This is one example of how statistics can help to simplify and clarify the research process. Statistics allow you to use a few summary statements to provide a comprehensive portrait of a large group of offenders. For example, instead of providing the name of each offender and telling us how many crimes he or she committed, you could present a single statistic that described the average number of crimes committed by the people you studied. You might say that, on average, the people you studied committed two to three crimes in the last year. Thus, although it might be impossible to describe each person you studied, you could, by using a statistic, give your audience an overall picture. Statistics make it possible to summarize information about a large number of subjects with a few simple statements.

Given that statistics should simplify the description of research results, it follows that the researcher should utilize the simplest statistics appropriate for answering the research questions that he or she raises. Nonetheless, it sometimes seems as if researchers go out of their way to identify statistics that few people recognize and even fewer understand. This approach does not help the researcher or his or her audience. There is no benefit in using statistics that are not understood by those interested in your research findings. Using a more complex statistic when a simpler one is appropriate serves no purpose beyond reducing the number of people who will be influenced by your work.

The best presentation of research findings will communicate results in a clear and understandable way. When using complex statistics, the researcher should present them in as straightforward a manner as possible. The mark of good statisticians is not that they can mystify their audiences, but rather that they can communicate even complex results in a way that most people can understand.

Statistics Are Used to Solve Problems

Statistics develop because of a need to deal with a specific type of question or problem. In the example above, you were faced with the dilemma that you could not describe each person in a very large study without creating a good deal of confusion. We suggested that an average might provide a way of using one simple statistic to summarize a characteristic of all the people studied. The average is a statistical solution. It is a tool for solving the problem of how to describe many subjects with a short and simple statement.

As you will see in later chapters, statistics have been developed to deal with many different types of problems that researchers face. Some of these may seem difficult to understand at the outset, and indeed it is natural to be put off by the complexities of some statistics. However, the solutions that statisticians develop are usually based on simple common sense. Contrary to what many people believe, statistics follow a logic that you will find quite easy to follow. Once you learn to trust your common sense, learning statistics will turn out to be surprisingly simple. Indeed, our experience is that students who have good common sense, even if they have very little formal background in this area, tend to become the best criminal justice statisticians. But in order to be able to use common sense, it is important to approach statistics with as little fear as possible. Fear of statistics is a greater barrier to learning than any of the computations or formulas that we will use. It is difficult to learn anything when you approach it with great foreboding. Statistics is a lot easier than you think. The job of this text is to take you step by step through the principles and ideas that underlie basic statistics for criminal justice researchers. At the beginning, we will spend a good deal of time examining the logic behind statistics and illustrating how and why statisticians choose a particular solution to a particular statistical problem. What you must do at the outset is take a deep breath and give statistics a chance. Once you do, you will find that the solutions statisticians use make very good sense.

Basic Principles Apply Across Statistical Techniques

A few basic principles underlie much of the statistical reasoning you will encounter in this text. Stating them at the outset will help you to see how statistical procedures in later chapters are linked one to another. To understand these principles, you do not need to develop any computations or formulas; rather, you need to think generally about what we are trying to achieve when we develop statistics.

The first is simply that *in developing statistics we seek to reduce the level of error as much as possible*. The purpose of research is to provide answers to research questions. In developing these answers, we want to be as accurate as we can. Clearly, we want to make as few mistakes as possible. The best statistic is one that provides the most accurate statement about your study. Accordingly, a major criterion in choosing which statistic to use—or indeed in defining how a statistic is developed—is the amount of error that a statistic incorporates. In statistics, we try to minimize error whenever possible.

Unfortunately, it is virtually impossible to develop any description without some degree of error. This fact is part of everyday reality. For example, we do not expect that our watches will tell perfect time or that our thermostats will be exactly correct. At the same time, we all know that there are better watches and thermostats and that one of the factors that leads us to define them as "better" is that they provide information with less error. Similarly, although we do not expect our stockbroker to be correct all of the time, we are likely to choose the broker who we believe will make the fewest mistakes.

In choosing a statistic, we also use a second principle to which we will return again and again in this text: *Statistics based on more information are generally preferred over those based on less information.* This principle is common to all forms of intelligence gathering and not just those that we use in research. Good decision making is based on information. The more information available to the decision maker, the better he or she can weigh the different options that are presented. The same goes for statistics. A statistic that is based on more information, all else being equal, will be preferred over one that utilizes less information. There are exceptions to this rule, often resulting from the quality or form of the information or data collected. We will discuss these in detail in the text. But as a rule, the best statistic utilizes the maximum amount of information.

Our third principle relates to a danger that confronts us in using statistics as a tool for describing information. In many studies, there are cases that are very different from all of the others. Indeed, they are so different that they might be termed deviant cases or, as statisticians sometimes call them, "outliers." For example, in a study of criminal careers, there may be one or two offenders who have committed thousands of crimes, whereas the next most active criminal in the sample has committed only a few hundred crimes. Although such cases form a natural part of the research process, they often have very significant implications for your choice of statistics and your presentation of results.

In almost every statistic we will study, outliers present a distinct and troublesome problem. A deviant case can make it look as if your offenders are younger or older than they really are—or less or more criminally active than they really are. Importantly, deviant cases often have the most dramatic effects on more complex statistical procedures. And it is precisely here, where the researcher is often preoccupied with other statistical issues, that deviant cases go unnoticed. But whatever statistic is used, the principle remains the same: *Outliers present a significant problem in choosing and interpreting statistics.*

The final principle is one that is often unstated in statistics, because it is assumed at the outset: *Whatever the method of research, the researcher must strive to systematize the procedures used in data collection and analysis.* As Albert J. Reiss, Jr., a pioneer in criminal justice methodologies, has noted, "systematic" means in part "that observation and recording are done according to explicit procedures which permit replication and that rules are followed which permit the use of scientific inference."[1]

---

[1] A. J. Reiss, Jr., "Systematic Social Observation of Social Phenomenon," in Herbert Costner (ed.), *Sociological Methodology* (San Francisco: Jossey Bass, 1971), pp. 3–33.

While Reiss's comment will become clearer as statistical concepts are defined in coming chapters, his point is simply that you must follow clearly stated procedures and rules in developing and presenting statistical findings. It is important to approach statistics in a systematic way. You cannot be sloppy or haphazard, at least if the statistic is to provide a good answer to the research question you raise. The choice of a statistic should follow a consistent logic from start to finish. You should not jump from statistic to statistic merely because the outcomes are favorable to the thesis you raise. In learning about statistics, it is also important to go step by step—and to be well organized and prepared. You cannot learn statistics by cramming in the last week of classes. The key to learning statistics is to adopt a systematic process and follow it each week.

Statistical procedures are built on all of the research steps that precede them. If these steps are faulty, then the statistics themselves will be faulty. In later chapters, we often talk about this process in terms of the assumptions of the statistics that we use. We assume that all of the rules of good research have been followed up to the point where we decide on a statistic and calculate it. Statistics cannot be disentangled from the larger research process that comes before them. The numbers that we use are only as good as the data collection techniques that we have employed. Very complex statistics cannot hide bad research methods. A systematic approach is crucial not only to the statistical procedures that you will learn about in this text but to the whole research process.

The Uses of Statistics

In the chapters that follow, we will examine three types of statistics or three ways in which statistics are used in criminal justice. The first is called **descriptive statistics,** because it helps in the summary and description of research findings. The second, **inferential or inductive statistics,** allows us to make inferences or statements about large groups from studies of smaller groups, or samples, drawn from them. Finally, we introduce **multivariate statistics** toward the end of the text. Multivariate statistics, as the name implies, allow us to examine a series of variables at one time.

Descriptive Statistics

We are all familiar in some way with descriptive statistics. We use them often in our daily lives, and they appear routinely in newspapers and on television. Indeed, we use them so often that we sometimes don't think of them as statistics at all. During an election year, everyone is concerned about the percentage support that each candidate gains in the primaries. Students at the beginning of the semester want to know what proportion of their grades will be based on weekly exercises. In deciding whether our salaries are fair, we want to know what the average salary is for other people in similar positions. These are all descriptive statistics. They summarize in one simple statement the characteristics of many people. As discussed above in the example concerning criminal histories, descriptive statistics make it possible for us to summarize or describe large amounts of information.

In the chapters that follow, we will be concerned with two types of descriptive statistics: **measures of central tendency** and **measures of dispersion.** Measures of central tendency are measures of typicality. They tell us in one statement what the average case is like. If we could take only one person as the best example for all of the subjects we studied, who would it be? If we could choose only one level of criminal activity to typify the frequency of offending of all subjects, what level would provide the best snapshot? If we wanted to give our audience a general sense of how much, on average, a group of offenders stole in a year, what amount would provide the best portrait? Percentages, proportions, and means are all examples of measures of central tendency that we commonly use. In the coming chapters, you will learn more about these statistics, as well as more complex measures with which you may not be familiar, such as correlation and regression coefficients.

Having a statistic that describes the average case is very helpful in describing research results. However, we might also want to know how typical this average case is of the subjects in our study. The answer to this question is provided by measures of dispersion. They tell us to what extent the other subjects we studied are similar to the case or statistic we have chosen to represent them. Although we don't commonly use measures of dispersion in our daily lives, we do often ask similar questions without the use of such statistics.

For example, in deciding whether our income is fair, we might want to know not only the average income of others in similar positions, but also the range of incomes that such people have. If the range was very small, we would probably decide that the average provides a fairly good portrait of what we should be making. If the range was very large, we might want to investigate more carefully why some people make so much more or less than the average. The range is a measure of dispersion. It tells us about the spread of scores around our statistic. In the chapters that follow, we will look at other measures of dispersion—for example, the standard deviation and variance, which may be less familiar to you. Without these measures, our presentation of research findings would be incomplete. It is not enough simply to describe the typical case; we must also describe to what degree other cases in our study are different from or similar to it.

Inferential Statistics

Inferential statistics allow us to make statements about a population, or the larger group of people we seek to study, on the basis of a sample drawn from that population. Without this very important and powerful tool, it would be very difficult to conduct research in criminal justice. The reason is simple. When we conduct research, we do so to answer questions about populations. But in reality we seldom are able to collect information on the whole population, so we draw a sample from it. Statistical inference makes it possible for us to infer characteristics from that sample to the population.

Why is it that we draw samples if we are really interested in making statements about populations? In good part it is because gaining information on most populations is impractical and/or too expensive. For example, if we seek to examine the attitudes of U.S. citizens toward criminal justice processing, we are interested in how all citizens feel. However, studying all citizens would be a task of gigantic proportion and would cost billions of dollars. Such surveys are done every few years and are called censuses. The last census in the United States took many years to prepare and implement and cost over $5 billion to complete. If every research study of the American population demanded a census, then we would have very few research projects indeed. Even when we are interested in much smaller populations in the criminal justice system, examination of the entire population is often beyond the resources of the criminal justice researcher. For example, to study all U.S. prisoners, we would have to study over 1 million people.

Even if we wanted to look at the 100,000 or so women prisoners, it would likely cost millions of dollars to complete a simple study of their attitudes. This is because the most inexpensive data collection can still cost tens of dollars for each subject studied. When you consider that the National Institute of Justice, the primary funder of criminal justice research in the United States, provides a total of about $100 million a year for all research, it is clear that criminal justice research cannot rely on studies of whole populations.

It is easy to understand, then, why we want to draw a sample or subset of the larger population to study, but it is not obvious why we should believe that what we learn from that sample applies to the population from which it is drawn. How do we know, for example, that the attitudes toward criminal justice expressed by a sample of U.S. citizens are similar to the attitudes of all citizens? The sample is a group of people drawn from the population; it is not the population itself. How much can we rely on such estimates? And to what extent can we trust such statistics? You have probably raised such issues already, in regard to either the surveys that now form so much a part of public life or the studies that you read about in your other college classes. When a news organization conducts a survey of 1,000 people to tell us how all voters will vote in the next election, it is using a sample to make statements about a population.

The criminal justice studies you read about also base their conclusions about populations—whether of offenders, criminal justice agents, crime-prone places, or criminal justice events—on samples. Statistical inference provides a method for deciding to what extent you can have faith in such results. It allows you to decide when the outcome observed in a sample can be generalized to the population from which it was drawn. Statistical inference is a very important part of statistics and one we will spend a good deal of time discussing in this text.

Taking into Account Competing Explanations: Multivariate Statistics

Multivariate statistics allow us to solve a different type of problem in research. It is often the case that the issue on which we want to focus is confounded by other factors in our study. Multivariate statistics allow us to isolate one factor while taking into account a host of others. For example, a number of criminal justice studies examine the impact of imprisonment on the future criminal behavior of offenders. In general, they compare offenders who are found guilty in court and sentenced to prison with those who are found guilty but do not receive a prison sanction.

Such studies focus on whether the criminal behavior of prisoners, once they are released into the community, is different from that of nonprisoners. Researchers conducting these studies face a very difficult research problem. Prisoners and nonprisoners are often very different types of people, and some of these differences are likely to affect their criminal behavior in the community. For example, prisoners are more likely than nonprisoners to have been arrested before, since a prior arrest is often an important factor in the judge's decision to incarcerate a convicted offender in the first place. And we know from research about criminal careers that people with a prior history of arrest are much more likely than people without such a history to commit a crime in the future. Accordingly, prisoners are more likely to commit a crime in the future, irrespective of the fact that they have served a prison sentence. This makes it very difficult to assess the impact of imprisonment on future offending. If we discover that prisoners, once released into the community, are more likely than nonprisoners to commit a crime, how can we tell whether this was a result of the experience of imprisonment? It might be due to the simple fact that prisoners are more likely than nonprisoners to commit crimes in the first place. Their more serious arrest histories would predict this result.

The complex task facing the criminal justice researcher is to isolate the specific impact of imprisonment itself from all of the other possible explanations for differences in reoffending between prisoners and nonprisoners. Multivariate analysis provides a statistical solution to this problem. It allows the criminal justice researcher to isolate the impact of one factor—in this case, imprisonment—from those of other factors, such as prior criminal history, that might confound the researcher's conclusions.

Chapter Summary

Statistics seem intimidating because they are associated with complex mathematical formulas and computations. Although some knowledge of math is required, an understanding of the concepts is much more important than an in-depth understanding of the computations. Today's computers, which can perform complex calculations in a matter of seconds or fractions of seconds, have drastically cut the workload of the researcher. They cannot, however, replace the key role a researcher plays in choosing the most appropriate statistical tool for each research problem.

The researcher's aim in using statistics is to communicate findings in a clear and simple form. As a result, the researcher should always choose the simplest statistic appropriate for answering the research question.

Statistics offer commonsense solutions to research problems. The following principles apply to all types of statistics: (1) In developing statistics, we seek to reduce the level of error as much as possible. (2) Statistics based on more information are generally preferred over those based on less information. (3) Outliers present a significant problem in choosing and interpreting statistics. (4) The researcher must strive to systematize the procedures used in data collection and analysis.

There are three principal uses of statistics discussed in this book. In **descriptive statistics,** the researcher summarizes large amounts of information in an efficient manner. Two types of descriptive statistics that go hand in hand are **measures of central tendency,** which describe the characteristics of the average case, and **measures of dispersion,** which tell us just how typical this average case is. We use **inferential statistics** to make statements about a population on the basis of a sample drawn from that population. Finally, in **multivariate statistics,** we isolate the impact of one factor from others that may distort our results.

Key Terms

descriptive statistics A broad area of statistics that is concerned with summarizing large amounts of information in an efficient manner. Descriptive statistics are used to describe or represent in summary form the characteristics of a sample or population.

inferential, or inductive, statistics A broad area of statistics that provides the researcher with tools for making statements about populations on the basis of knowledge about samples. Inferential statistics allow the researcher to make inferences regarding populations from information gained in samples.

measures of central tendency Descriptive statistics that allow us to identify the typical case in a sample or population. Measures of central tendency are measures of typicality.

measures of dispersion Descriptive statistics that tell us how tightly clustered or dispersed the cases in a sample or population are. They answer the question "How typical is the typical case?"

multivariate statistics Statistics that examine the relationships among variables while taking into account the possible influences of other confounding factors. Multivariate statistics allow the researcher to isolate the impact of one variable from others that may distort his or her results.