A TOPICAL DICTIONARY OF STATISTICS

Gary L. Tietjen

A TOPICAL DICTIONARY OF STATISTICS

Gary L. Tietjen

CHAPMAN AND HALL
New York London

First published 1986
by Chapman and Hall
29 West 35 Street, New York, N.Y. 10001

Published in Great Britain by
Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data
Tietjen, Gary L.
A topical dictionary of statistics.

Bibliography: p.
Includes index.
1. Mathematical statistics—Terminology. I. Title.
QA276.14.T54 1986 519.5'03'21 86-11716
ISBN 0-412-01201-4

Preface

Statistics is the accepted body of methods for summarizing or describing data and drawing conclusions from the summary measures. Everyone who has data to summarize thus needs some knowledge of statistics. The first step in gaining that knowledge is to master the professional jargon. This dictionary is geared to offer more than the usual string of isolated and independent definitions: it provides also the context, applications, and related terminology.

The intended audience falls into five groups with rather different needs: (1) professional statisticians who need to recall a definition, (2) scientists in disciplines other than statistics who need to know the acceptable methods of summarizing data, (3) students of statistics who need to broaden their knowledge of their subject matter and make constant reference to it, (4) managers who will be reading statistical reports written by their employees, and (5) journalists who need to interpret government or scientific reports and transmit the information to the public.

In every case the word or phrase to be defined should be looked up in the alphabetical index, which will refer the reader to a page in the text. The professional statistician may then find the word and its definition and be finished. Other readers are no doubt looking for *more* information—for background, related words, and an understanding of how this topic fits into the scheme of things. For this purpose the dictionary has been arranged topically rather than alphabetically, and in connected discourse rather than in paragraphs related only by the first few letters of one word. Depending on how much knowledge of the subject is desired, the reader will want to go back a few paragraphs or to the beginning of the chapter and read further. I have even been assured by some professors of statistics that students of statistics will benefit by reading the dictionary from cover to cover. That will give some idea of the "big picture" and help substitute for gaps in educational training.

The first chapter, *Summarizing Data*, has been written particularly for those who have no background in statistics and may be worth their while as an introduction to the book.

In trying to meet the needs of so many, I have tried to stay one level below what a theoretical statistician would like to see. In doing so, I run the risk of falling a little short of absolute rigor, but I have tried not to sacrifice the things that matter in applications. I hope this effort will be worthwhile for the manager and journalist. The question of what to include and what to leave out has arisen many times. There are many specialties in statistics and I have decided to omit terms which are peculiar to a small audience. On the other hand, I have included a number of general but rarely used terms for the sake of completeness, and I can only hope that the reader will bear with me on this matter. There are terms whose use is discouraged and areas where one needs to take great care to avoid misstatements. I have tried to point these out to the reader.

The nonmathematical reader need not be discouraged by the mathematical symbols which are used frequently. They are really a rather simple shorthand. The symbols appearing most often are the following: (1) $\int_a^b f(x)dx$ which is read "the integral of $f(x)$ from a to b" and means the area under the curve $f(x)$ and between the values of $x = a$ and $x = b$ on the x-axis. if the limits are infinite, i.e. if a is $-\infty$ and b is ∞, the area is taken over the entire x-axis (2) the Greek letter Σ means "the sum of" so that $\sum_{i=1}^{N} x_i$ means "the sum of the x_i from $i = 1$ to $i = N$", i.e. $x_1 + x_2 + \cdots + x_N$. For brevity, the subscripts are sometimes omitted when the meaning is clear. Thus the symbol Σx means Σx_i, where i goes from 1 to N unless otherwise indicated. Products are denoted with a Π in place of Σ. The symbol $x!$ means the product of the positive integers from 1 to x, i.e., $5! = 1 \times 2 \times 3 \times 4 \times 5$. The binomial symbol $\begin{bmatrix} N \\ n \end{bmatrix}$ is equal to $N!/(N - n)!n!$ and is the number of ways of selecting a sample of n items from a set of N items (4) the symbol df/dx (read "the derivative of f with respect to x") refers to the slope of the line tangent to the curve $f(x)$ at a point x. (For a given change in x, the slope of a line is the change in y divided by the change in x). If f is a function of more than one variable, $\partial f/\partial x$ is the derivative of f with respect to x while the other variables are held constant (it is read "the partial of f with respect to x").

I have followed the convention of using capital letters for random variables and the corresponding lower case letters for their realizations (or for nonrandom variables). Greek letters denote parameters of a distribution (for a particular distribution, a parameter is a constant, but it varies from one distribution to another). Parameters are usually unknown constants which have

to be estimated, and their estimates are denoted by the same Greek letter with a carat (^) or a tilde (~) placed above it. A tilde in the middle of the line, however, is read "is distributed as". The symbol $E(X)$ is read "the expected value of X" and is an average of the values taken on by the random variable X. The symbol $\sqrt{x}$ is the positive square root of x and is equivalent to $x^{1/2}$. The use of $\exp(x)$ for e^x is purely for typesetting convenience. The natural logarithm of x (log base e, where e is a constant approximately equal to 2.718) is denoted by $ln(x)$.

This dictionary is by no means a solo production. At one time or another I pressed nearly everyone in the statistics group at Los Alamos into providing some definitions or reading those I had written. Consultants to the group received the same treatment. Dennis Cook generously wrote the major part of the chapter on regression, while George Milliken wrote much of the chapter on Experimental Design. I benefitted especially from a critical review of the first draft by Jay Conover. Ben Duran followed with many valuable suggestions. My good friend S. Juan lent constant support and encouragement. Kay Grady and Corinne Ortiz very competently typed the manuscript and endured revision after revision without complaint. Finally, I am most grateful to my wife who took care of many other duties while I wrote.

Gary Tietjen
Los Alamos
New Mexico
May, 1986

Contents

1

Summarizing Data

This chapter is intended to introduce the basic ideas of statistics to the layman. Statistics is the accepted method of summarizing or describing data and then drawing inferences from the summary measures. Suppose, for example, that a company has made a process change in manufacturing its light bulbs and hopes that the new bulb (Type B) will have a longer lifetime than the old (Type A). Having experimented previously, the company knows that even though the bulbs are treated identically they will vary considerably (60 to 90 hours) in the length of time they will last. That variability, which is a property of almost all manufactured products, is called *inherent variability*. Since we cannot predict how long a bulb will burn, we describe its lifetime as a *random variable*. The company does know that the largest fraction of the bulbs burn about 75 hours, that those with lives of 70 and 80 hours are about equally frequent (but less common than lifetimes of 75 hours), and that those with lives of 65 and 85 hours are even less frequent.

For bulb A we can think of all the past production as a *population* of bulbs with lifetimes that we denote by x. For bulb B the population is mostly conceptual; it consists of bulbs that *will* be produced by the new process. We let y denote the lifetime of a bulb of Type B.

The aim of management in this instance is to evaluate the performance of the Type B bulbs and to compare it with that of Type A. The first thing to do is to picture the situation. It is obviously neither possible nor desirable to test the lifetimes of all the bulbs. Fortunately the company has tested 100 Type A bulbs in the past. The readings range from 60 hours to nearly 100 hours burning time. That portion of the population will ordinarily be called a sample, but the word implies that there are some restrictions in the way the bulbs to be tested are selected. When there are no restrictions, we shall refer to the portion as a *batch*. (In this chapter we are adopting some of the terminology coined recently by John Tukey for the set of techniques that he calls *Exploratory Data Analysis* [*EDA*]. His terminology, while not yet standard, has come into rather wide usage. Exploratory data analysis is a first look—a quick glance—at the data and is usually followed by a *confirmatory data analysis*, using the techniques of classical statistics.)

A time-honored method of graphically portraying the data in the batch of 100 units is to construct a *histogram* of the data. This is done by dividing the interval of possible lifetimes (60 to 100 hours) into k subintervals of equal width called *class intervals*. There is no prescribed way of deciding how many intervals to use, but perhaps 10 will do here: 60–65, 65–70, . . . We then count the number of units (n_1, n_2, . . .) that fall into each interval. We next draw a series of adjacent rectangles, using the class intervals as widths and the frequencies n_1, n_2, . . . , n_k as heights. The histogram shows the *distribution of frequencies* of the lifetimes. The graph can be improved somewhat by using *relative frequencies* (n_1/T, n_2/T, . . . n_k/T, where $T = n_1 + n_2 + \ldots + n_k$) as the heights of the rectangles. Much can be inferred from the histogram. We can decide, for example, what percentage of the bulbs in the sample have lifetimes of less than 70 hours, what percentage burn between 80 and 90 hours, etc. If we took larger and larger samples, we could make the class intervals narrower and narrower until the tops approached a smooth, continuous curve. If such a curve were drawn across the midpoints of the tops of the rectangles, it would represent the *frequency distribution* or *probability density function* (pdf) for the population. We see that the relative frequencies sum to 1; hence it is not hard to believe that the area under the frequency distribution is 1. Further, the area under the curve and within an interval (a,b) of lifetimes is approximately the relative frequency of lifetimes within the interval. The area between a and b is, in fact, the limiting value of the relative frequency and is called the *probability* that the lifetime is between a and b. We thus see that a random variable has a distribution of probabilities associated with it.

The EDA version of a histogram may be quicker to construct and is called a *stem-and-leaf plot*. It will be lying on its side. The stems replace class intervals and in this case would be the first digit of the lifetime. The second

digit constitutes the leaves. By tallying the leaves to the right of their stem as we come to them in the data set, we get the stem and leaf plot of the data (provided that we give the same width to each leaf). If we decide that we do not have enough stems, a period following the stem can represent a stem with leaves 0–4, while an asterisk following the stem accompanies leaves 5–9. Three-digit numbers can be represented by dropping the last digit or by using 2-digit stems.

There is a way of abbreviating a histogram even further. Let us first *rank* the n data points in ascending order so that the smallest point has rank 1, the second-smallest rank 2, and so on to the largest, which has rank n. Now we rank the data in descending order so that the smallest point has rank n and the largest rank 1. The *depth* of a data point is the minimum of the 2 ranks the data point can have. The largest and smallest points, called *extremes*, have depth 1. The middle observation or *median* has depth $(1+n)/2$. When the depth is not an integer, we average the 2 data points with depths on either side of the indicated 1. Thus, if there is an even number of points, the median is the average of the 2 middle points. The *hinges* are halfway between the extremes and the median; they are the points with depth $(1+m)/2$, where m is the integer part of the depth of the median. Similarly the *eighths* are points with depth $(1+h)/2$, where h is the integer part of the depth of the hinges.

A rather neat summary of the histogram is made by plotting the extremes, hinges, and median on a vertical line. A "long, thinnish box" (about ⅛ inch wide) is drawn so that the hinges are at the top and bottom of the box. A horizontal line through the box marks the location of the median. A vertical line connects the extremes with the hinges. That 5-point summary is called a *box-and-whisker plot*. The middle 50 percent of the data lie inside the box; the lower 25 percent of the data are in the lower whisker and the upper 25 percent in the upper whisker. The lower half of the data are below the median and the other half above it.

Another useful plot, very similar to a box-and-whisker plot, is a *schematic plot*. The *H-spread* is the distance between the hinges. A *step* is 1.5 times that distance. An *inner fence* is 1 step beyond the hinges, and an *outer fence* is 2 steps beyond the hinges. The data point closest to the inner fence but inside of it is an *adjacent* point. *Outside* values are those between the inner and outer fences, while those beyond the outer fence are *far out* points. The box for the schematic plot is constructed as before, but the whiskers are dashed and extended only to the adjacent values and end with a short dashed horizontal line. The outside values are labeled separately and the far out values labeled "impressively."

Let us return to our example. A batch of bulbs is taken from the production line and tested. Either box-and-whisker plots or schematic plots are constructed so that the plots for the 2 types of bulbs parallel each other. A visual

comparison of the 2 sets of data can now be made. Symmetry of the histogram (with the median near the middle of the box and the whiskers of about equal length) is a good framework for comparing the medians. More important than symmetry is an approximate equality in spread, as shown by box length and whisker length. If either of those conditions is violated seriously, we will want to *re-express* the data (*transform* the data is classical terminology) before comparing the medians. The schematic plot is preferred to the box-and-whisker plot if there are several outside or far out points. If the 2 distributions are moderately symmetrical and close in spread as judged by the eye, the medians will tell about how far apart the average lifetimes will be. Sometimes the *trimeans* (sum of hinges plus twice the median, all divided by 4) are used as an estimate of the "center" of the distribution.

When strong asymmetry/inequality of spread is present, the re-expression is ordinarily done by transforming the data to logs of the data. If that does not work, a square root transformation is made. Negative reciprocals are the third choice. All of those choices preserve the ranks and depths of the data points. Tukey has suggested a *ladder of transformations* . . . X^3, X^2, X, log X, $-1/X$, $-1/X^2$, $-1/X^3$, . . . , where the transformation most likely to help is chosen by plotting the log H-spread (y) against the log of the median (x) for the several populations being compared. If the slope of a line drawn by eye is close to ½, the square root should help. If the slope is somewhat larger than ½, logarithms will be more likely to be useful. In other words, choose the re-expression more apt to result in a horizontal line through the transformed points.

We now give some thought to a confirmatory or classical approach, which might follow the quick EDA look at the data. The EDA approach should have given us a rather good "feel" of the data and any unusual structure that might be present. From that we may have reached some preliminary conclusions. At other times the differences between the 2 bulbs may have been so obvious that no statistics seem to be needed. Regardless, the investigator needs numbers to put in the report. Just how large *are* the differences between medians? How significant are the results? That is an area for classical statistics.

A great many measurements in nature are approximately *normally distributed* (for a good reason to be discussed later). That means that the *probability density function* (the frequency distribution) is symmetrical and bell-shaped. That distribution is so familiar and so ubiquitous that in many cases the statistician just assumes that the measurements are normal (in most cases it does not matter too much if that assumption is slightly off base). In situations where the assumption seems to be badly off, the statistician may test it with a *goodness-of-fit test*. The "center" of the bell, the place where it has a "peak," is called the *mean* and designated by the greek letter μ. The distance from

μ to the point at which the curvature changes from downward to upward is the *standard deviation*, designed by σ. The area under the curve is 1. The area in the interval $\mu \pm \sigma$ is about 68 percent of the total; the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 95 percent of the total, and the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ is well over 99 percent of the total. The curve is completely characterized by a knowledge of μ and σ.

Returning to the bulbs and assuming a normal distribution for each of the 2 populations, we can condense or summarize the data even further. What single number is typical of or characterizes or summarizes the lifetimes in the sample? The *average* lifetime of the bulbs immediately comes to mind, but what shall we average? We think of the entire past production of bulb A as the *population* of interest and let x_i be the lifetime of the i-th bulb ($i = 1, 2, \ldots N$) produced. The *population average* μ_x would then be the sum of the lifetimes divided by the number of bulbs in the sum $\mu_x = \Sigma x_i/N$. The number N, however, is in the millions, and there was no possible way to have gotten the necessary measurements. From the 100 measurements taken on the Type A bulbs, the *sample average* is $\bar{x} = \sum x_i/n$, where $n = 100$. The sample average is an *estimate* of the population average. The formula $\Sigma x_i/n$ is an *estimator* of μ_x. Having obtained $\bar{x}$, we would like to do the same thing for bulb B, which is just getting into production. How large a sample shall we take? With bulb A there was little choice: We took all the information available at the time. It seems intuitive that the larger the sample size the better the estimate (a property that will later be called *consistency*). We thus take the largest sample—say, m bulbs—we can afford; the cost will also involve the time spent in testing. The sample average is $\bar{y} = \sum y_i/m$. We could now compare $\bar{x}$ and $\bar{y}$ to see which is larger, but we do not have, as yet, a good standard with which to judge the difference. If the sample size m were small and if the difference between $\bar{x}$ and $\bar{y}$ were small, a different sample of bulbs might have yielded an average that would have reversed the order of $\bar{x}$ and $\bar{y}$. Whether normality holds or not, a useful measure of the scatter of the data around the sample mean is the *sample variance* $s_y^2 = \sum(y_i - \bar{y})^2/(n-1)$, which estimates or approximates the population variance $\sigma_y^2 = \sum(y_i - \mu_y)^2/N$, the average squared deviation from the mean. We divide the sample variance by $(n-1)$ instead of n because it can be shown that the average value of s_y^2 (with a divisor of n) is $(n-1)\sigma_y^2/N$ σ_y^2. In other words, s_y^2 would be a *biased estimator* of σ_y^2, and the divisor $(n-1)$ is chosen to unbias it.

We can now express our uncertainty in the location of the sample mean μ_y by making an *interval estimate* of μ_y, which has a "high probability" of containing μ_y in the following sense: If a large number of such intervals were constructed from different samples, 95 percent of them would contain the

population mean. That interval is called a 95 percent *confidence interval* for μ_y, as opposed to the *point estimate* $\bar{y}$ of μ_y. The interval is $\bar{y} \pm t\, s_y/\sqrt{n}$, where t is a number that depends upon the sample size n and is taken from tabled values of the Student's t distribution, a connection we will explore later.

Finally, we can test whether the population means for the 2 types of bulbs differ significantly, our hypothesis being that $\mu_y = \mu_x$. The sample means clearly differ only if the difference between them is larger than the variability within the measurements that make up the means. The difference $\bar{x} - \bar{y}$ is distributed with mean zero and standard deviation of $(s_y^2/m + s_x^2/n)^{1/2}$ if and only if $\mu_x = \mu_y$. Differences of about 2 of those standard deviations are not so unusual, but differences much larger than that are rare—so rare that we are willing to take a small risk (of the order of 5 percent) of being wrong and declare that μ_x is not equal to μ_y. In other words, we decide that one mean is greater than the other. The actual number of standard deviations by which $\bar{x}$ and $\bar{y}$ can differ without differing significantly is again found in the tables of the Student's t distribution and depends on n,m, and the size of the small risk we take of being wrong. It is one example of *hypothesis-testing*, a very valuable tool.

In this chapter we have touched on random variables and their probability distributions. We have seen how a histogram approximates the probability density function. Those matters are covered in Chapter 2. Lifetimes of electrical components frequently have distributions other than normal. A guide to other distributions and their uses is found in Chapter 3. We have examined estimators and estimates of the mean and variance of the normal distribution, and we have given some thought to the desirable properties of an estimator (consistency and unbiasedness). We have touched on the topic of hypothesis testing. The areas of estimation and hypothesis testing are taken up in detail in Chapter 4.

Suppose now that a special coating of the filament was the design change that resulted in the Type B bulbs. Some thought has been given to whether the life of the bulb might increase directly with the thickness of that coating. Some experiments are carried out using various thicknesses of the coating and testing of the life of each bulb. The data are plotted as lifetime (y) versus thickness (x), and it appears there is a linear relationship. The plot of the data points is called a *scattergram* or *scatterplot*. How to fit a straight line to the data is a problem in estimating the parameters (the slope and intercept) of a straight line. Again, hypothesis testing is used to decide whether the slope is zero (no change in lifetime with thickness) or not. Those matters form the content of Chapter 5 on *Regression*.

It may be that 3 different coatings can be applied, each differing in its composition. In order to test which of the 3 gives the longest average lifetime,

we would design an experiment in which we would measure the lifetime of k bulbs with each coating. The analysis of those data would involve the *Analysis of Variance*, a subject taken up in Chapter 6: The Design of Experiments and Analysis of Variance.

The probability that a bulb will perform its function (burn) for a given length of time under given circumstances is the *reliability* of the bulb. The whole area of estimation and testing of lifetime data is treated in Chapter 7: Reliability and Survival Analysis.

It may have been of interest to estimate the intensity of light from the new bulbs as a function of the age of the bulb. If we had a continuous record of the intensity (or a test every 30 minutes, say) with time, we would have a *time series* and the analysis of the data would come under the chapter on Time Series and its parent: Stochastic Processes.

In the production of the new bulbs, it may be desirable or necessary for the manufacturer to assure himself continually that the thickness of the coating is uniform. To do that, he may check the thickness of 3 bulbs from each day's production. The techniques for obtaining such assurance are given in the chapter on *Quality Control*.

If we have 2 characteristics of interest, say lifetime and intensity, then there are 2 random variables to be considered simultaneously. That is a problem in *Multivariate Analysis*, which falls under the chapter of that name.

REFERENCES

For the EDA techniques in this chapter see Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley. The classical techniques will be explained later in detail.

2

Random Variables and Probability Distributions

When the average citizen sees a game of dice, he knows intuitively that the outcome of any 1 roll of the dice is unpredictable—that he is faced with a "chance" or "random" phenomenon. He sees quickly that there are 36 possible outcomes (for each of the 6 sides of die #1 he can get any of the 6 sides of die #2). He nevertheless realizes the possibilities of betting on the outcome when he sees that there is only 1 outcome, (1,1), which gives a "2," while 6 outcomes, (1,6), (2,5), (3,4), (4,3), (5,2), (6,1), give a "7." Thus the "probability" of a 2 is 1/36 and that of a 7 is 6/36. The set of outcomes or "scores" with their associated probabilities constitutes a *probability distribution* and is his best aid to intelligent betting. In that case the dice were treated or "shaken" alike. Individual outcomes differ, but in the "long-run" one can predict how often each outcome will occur. What the layman may not realize is that even in a very precise chemical experiment the outcome is random. The possible outcomes may lie within a narrow range, but when

seemingly identical units are treated as much alike as possible, they still respond differently; there is still some "variability" in the outcome, and the chemist, with the aid of statistics, summarizes the data by telling his readers what they can bet on. We shall now repeat those ideas with more detail.

In an *experiment* the investigator observes the response to a given set of conditions. In some experiments the response is invariably the same, and we say that there is a *deterministic regularity* in the outcome. In other experiments, such as the toss of a die, the outcome is unpredictable, but the experiment has the next best property: The set of outcomes is known, and each outcome occurs with a certain relative frequency. Those *random experiments* (or *random trials* or *random events*), as they are called, are thus said to have a *statistical regularity* in the outcome. The relative frequency with which each outcome occurs approaches a stable limit, called the *probability* of that random event.

The set of all possible outcomes of a random experiment is called the *sample space* or *outcome space S*, and each outcome is a *sample point* ω in that space. An *event* is a subset of the sample space, but there may be some subsets that are not events. An event consisting of a single point is an *elementary event*. An event E is said to *occur* if the outcome ω is in E. In tossing a pair of dice, let $\omega = (2,3)$ be the outcome of a 2 on the first die and a 3 on the second. The point (2,3) is an elementary event. If E is the event that the total shown on the 2 dice equals 5, E occurs if the outcome is (1,4), (4,1), (2,3), or (3,2).

The outcome of a coin-tossing experiment may be "heads" or "tails." In drawing a colored ball from an urn, the outcome may be "blue." In drawing a man from a group of men, the outcome might be Don or Joe. For the sake of a mathematical treatment (rather than a verbal one) we need to assign a real number to every outcome. The number assigned will depend upon our purpose. We might assign the number 1 to "heads" and 0 to "tails," which is useful if we are counting heads. We could assign 1 to "blue" and 0 to any other color; to each man we could assign his height in inches. Given a set A of "objects," a rule that assigns to each object in A 1 (and only 1) member of a set B is called a *function* with *domain A* and *range B*. A *random variable* is a function in which A is the set of outcomes and B consists of real numbers, including $\pm\infty$. In tossing a pair of die, we assign to each outcome the sum of the number of spots on the upward faces. It is important that the function representing the random variable be single-valued and real-valued. If the outcome is a number x in the interval $(-10, 10)$, say, we cannot let the square root of the outcome be the random variable for 2 reasons: (1) If $x = 4$, both $+2$ and -2 are square roots, and there must be only 1 number