ELEMENTARY BUSINESS STATISTICS

Donald R. Byrkit

ELEMENTARY BUSINESS STATISTICS

Donald R. Byrkit

The University of West Florida

D. VAN NOSTRAND COMPANY
New York Cincinnati Toronto London Melbourne

TO MY WIFE MARNETTE—
whose patience and encouragement
helped make this book possible.

D. Van Nostrand Company Regional Offices:
New York Cincinnati

D. Van Nostrand Company International Offices:
London Toronto Melbourne

Library of Congress Catalog Card Number: 78-62491
ISBN: 0-442-21408-1

Published by D. Van Nostrand Company
135 West 50th Street, New York, N.Y. 10020

10 9 8 7 6 5 4 3 2 1

SYMBOLS USED IN THE TEXT

α	probability of type I error
A	population regression constant
a	sample regression constant
β	probability of type II error
B	population regression coefficient
b	sample regression coefficient
C	cyclical time series component
CV_d	critical value of the difference for Scheffe's *post hoc* test
CV_r	critical value of the difference for the Newman-Keuls *post hoc* test
df	degrees of freedom
E	maximum allowable error with a given probability
E( )	expected value of
E(y|x)	expected value of y for a given value of x in regression
EMV	expected monetary value
EMV*	optimal expected monetary value
EOL	expected opportunity loss
EOL*	minimum expected monetary loss
EVC	expected value under certainty
EVPI	expected value of perfect information
e	error; deviations from the population regression line
F	F-ratio
F_a	value of F for which a is the proportion of the area under the curve to its right
F_{max}	ratio of largest to smallest sample variance
f	frequency
H	Kruskal-Wallis H statistic
H_0	null hypothesis
H_1	alternate hypothesis
I	irregular time series component
μ	population mean; mean of a probability distribution
μ_0	hypothetical population mean
μ_1, μ_2	hypothetical means for two population
μ_d	mean difference, matched populations
μ_s	standard error of standard deviation
μ_x	standard error of the mean
MA	moving average
Md	median
Md_0	hypothetical population median
Mo	mode
MS	mean square
MSA	mean square for treatment A
MSAB	mean square for interaction
MSB	mean square for blocks, or for treatment B
MSD	mean square for deviation from linearity
MSE	mean square for error
MSR	mean square for linear regression
MST	mean square for treatments
MSW	mean square within cells
m	mean of a Poisson distribution
N	population size; total number of observations
n	sample size
ñ	harmonic mean of sample sizes
π	population proportion

π_0	hypothetical population proportion
π_1, π_2	hypothetical proportions for two populations
P( )	probability of
p	sample proportion; probability of success in a binomial experiment
p_1, p_2	sample proportions for two samples
Q	semi-interquartile range
Q_1	first quartile
Q_3	third quartile
q_r	statistic for Newman-Keuls multiple range test
ρ	population coefficient of correlation
ρ^2	population coefficient of determination
R	Spearman correlation coefficient; coefficient of multiple correlation
r	sample correlation coefficient
r^2	sample coefficient of determination
σ	population standard deviation; standard deviation of a probability distribution
σ^2	population variance; variance of a probability distribution
σ_d	standard error of the difference
σ_p	standard error of proportion
σ_s	standard error of standard deviation
$\sigma_{\bar{x}}$	standard error of the mean
s	sample standard deviation
s^2	sample variance
$\hat{s}^2$	pooled variance
s_a	standard error of the regression constant (A)
s_b	standard error of the regression coefficient (B)
s_d	standard error of the difference estimated from sample standard deviations
s_{dp}	standard error of the difference for proportion estimated from sample proportions
s_e	standard error of estimate for sample regression line
$s_{E(y\|x)}$	standard error of the estimated mean
s_f	standard error of forecast
s_p	standard error of proportion estimated from sample proportion
$s_{\bar{x}}$	standard error of the mean estimated from sample standard deviation
T	trend; Wilcoxon T statistic
TSS	total sum of squares
t	Student's t
t_a	value of t for which a is the proportion of the area under the curve to its right
U	Mann-Whitney U statistic
χ^2	Chi-square
χ^2_a	value of Chi-square for which a is the proportion of the area under the curve to its right
x	data point of a random variable
$\bar{x}$	sample mean
$\bar{x}_d$	mean difference of two samples
$\hat{y}$	estimated value of y from sample regression equation
z	standard score
z_a	value of z for which a is the proportion of the area under the normal curve to its right
$\doteq$	is approximately equal to
$<$	is less than
$\leq$	is less than or equal to
$>$	is greater than
$\geq$	is greater than or equal to
$\binom{n}{r}$	combination of n objects taken r at a time
!	factorial
□	decision point
○	probability point

Preface

This text is the outgrowth of the author's experience in teaching a beginning course in statistics to students whose primary area of study is business management. Mathematical exposition, therefore, is kept to a minimum and the major emphasis is on reasonable demonstration by example.

The material included in the text consists of all the basic material normally expected in an introductory course and a selection of supplementary material as well. The approach to probability is simplified, consisting solely of the minimum amount necessary to facilitate understanding of the subsequent material in inferential statistics. It has been the author's experience that probability is the primary stumbling block for business students, and most texts present more material than is strictly needed for the work that follows. Since probability is a valuable and interesting topic on its own merits, a more complete exposition has been included in Appendix B.

In addition to basic material, several sections have been designated as optional and may be omitted without disturbing continuity. Technical notes are inserted occasionally to add supplementary material where logical; these may also be omitted. Special features of the text include confidence intervals for differences and for standard deviation, analysis of variance in greater depth than is usual for a first course, an introduction to decision theory, and an early exposure to hypothesis testing.

Chapters 1–3 cover descriptive statistics and probability distributions and are an essential introduction to the remainder of the text. Chapters 4 and 5 cover the binomial, hypergeometric, and Poisson (discrete) distributions and the normal distribution. Hypothesis testing is introduced for the first time in Sections 4.4 and 5.4, but it may be omitted here, if desired, and discussed briefly in connection with Section 7.1, where the topic is discussed in detail. Chapters 6 and 7 cover sampling distributions and some applications to confidence intervals and hypothesis testing for means and proportions. Additional topics in statistical inference include the Chi-Square distribution (Chapter 8) and a detailed study of analysis of variance (Chapter 9), which includes the Hartley test for homogeneity of variance and the Newman-Keuls *post hoc* testing technique as well as the randomized complete-block and the two factor design. Some inference is also included in Chapter 10, which treats linear regression in detail, including analysis of variance, with an introductory look at curvilinear and multiple regression. This leads naturally into time series analysis in Chapter 11, which concludes with an introduction to index numbers. Chapter 12 presents several of the most widely used nonparametric tests, and the text closes with elementary decision theory in Chapter 13. A review of mathematics needed to understand the text is included as Appendix A.

Each chapter section is followed by a set of problems that provide practice in the statistical techniques presented in the section and examples of the application of these techniques to practical business problems. A brief summary, glossary of terms and symbols, and additional problems conclude each chapter. Answers to approximately half the problems are provided. Answers to the remainder of the problems as well as detailed solutions to most of the problems are given in the accompanying instructor's manual.

Widespread use of hand calculators has eliminated the need for concession to human frailty insofar as tedious arithmetic is concerned, and raw data formulas are emphasized. (Coding, no longer important even in time series analysis, is not discussed here.)

The text is designed to be used in its entirety in a year-long course of 5–6 semester hours or 8–10 quarter hours with students of average mathematical ability and no preparation beyond a course in algebra. For a one-term course some selection of topics is necessary. Chapters 1–3 are essential, but Sections 4.4, 5.4, and possibly 4.3 could be eliminated from Chapters 4 and 5. After that, personal preference is the guide. Chapter 7 follows from Chapter 6, Section 11.1 needs Section 10.1, and the various parts of Chapter 12 are linked directly to the parametric tests which they replace. A one-term course preparing for additional statistics courses could consist of Chapters 1–8, while a one-term terminal course for business majors might consist of Chapters 1–3, Sections 4.1–4.3, 5.1, 5.2, 6.1–6.3, 7.1–7.3, 8.1, 8.2, 10.1–10.3, Chapter 11, and possibly Chapter 13 or other topics as desired.

Many sources helped in the writing of this book. The author is especially grateful to Dr. Jean Namias, Phil Desper, Jeanne Libby, and to the following reviewers: Dr. Louis F. Bush of San Diego City College, Dr. James R. McGuigan of Wayne State University, Dr. Jerry L. Hintze of the University of Denver, Dr. Anne B. Koehler of Miami University, Dr. J. Elaine Lockley of Mountain View College, Dr. Roy Mazzagatti of Miami-Dade Junior College, Dr. Buddy L. Myers of Kent State University, Dr. Jean Namias of Montclair State College, Dr. Wayne Stevenson of the University of Utah, and Dr. Michael Umble of Baylor University.

I am indebted to the Biometrika Trustees for permission to use Tables 8, 12, 18, and 31 from *Biometrika Tables for Statisticians*, Volume 1, Third Edition, by Pearson and Hartley; to Lederle Laboratories for permission to reprint data from *Some Rapid Approximate Statistical Procedures*, by Wilcoxon and Wilcox; to the authors and publishers of *Probability: A First Course*, Second Edition, 1970, by Mosteller, Rourke and Thomas (Addison-Wesley), for permission to reprint Table IV, Part B; to the authors and publishers of *Statistical Tables*, by R. R. Sokal and F. J. Rohlf (W. H. Freeman and Company), for permission to reprint part of Table O; to the author and publishers of *Introduction to Statistical Inference* by E. S. Keeping (Van Nostrand Reinhold) for permission to reprint Table B.2; and to the author and publishers of *The Analysis of Variance* by Henry Scheffe (John Wiley & Sons, Inc.) for permission to reprint pp. 434–436.

Donald R. Byrkit

PAA
PAN AM
AIRW

Contents

CHAPTER 1

Organization and Presentation of Data

1.1 FREQUENCY DISTRIBUTIONS

Statistics can be described as the science of classifying and organizing data in order to draw inferences. An important aspect of classifying data is the efficient and effective organization and presentation of data. An unorganized mass of figures is more often confusing than clarifying. This chapter is concerned with the methods of deriving meaning from numerical data.

The term **data**, as used in the preceding paragraph, refers to the set of observations, values, elements, or objects under consideration. The complete set of all possible elements is called a **population** while anything less than the complete set is called a **sample**. Each of the elements is called a **data point**, or **piece of data**. The amount of money spent by a customer in a store on a particular day, for example, is a piece of data, while the collection of all expenditures by all customers on that day would comprise a complete set of data. This latter, in turn, could be considered a sample of the population of all expenditures of all customers on all days in that store.

Data are of two types, quantitative and qualitative. **Qualitative data**, or **attributes** (sometimes called **categorical data**), result from information which has been sorted into categories. Each piece of data clearly belongs to one classification or category. Automobiles on a parking lot classified by make give one example of attribute data. **Quantitative**, or **variable data**, is data which is a result of counting or measuring. We might count the number of nicks and scratches in the paint of each car, then give the number of cars with 0 scratches, 1 scratch, 2 scratches, and so forth. A car has a whole number of scratches (it cannot have 3.7 scratches, for instance), so there are clear divisions between the values. This type of data is called **discrete**. If we weighed the cars, we could get the weight to the nearest pound, but it would be possible for a car to weigh slightly more or less than we reported. A reported weight of 3,456 pounds, for example, would probably indicate a weight somewhere between 3,455.5 and 3,456.5 pounds. This is an example of a **continuous** variable. Generally, data arising from measurement are continuous, while data arising from counting or arbitrary classification are discrete. An example of the latter is the familiar grading system where the grades are 4.0, 3.5, 3.0, 2.5, 2.0, 1.5, 1.0, and 0 (A, B+, B, C+, C, D+, D, F system). Often there exists an option as to what type of system to use. In the example of the customer in the store, we could classify each customer as to whether he or she bought anything, which would be a qualitative classification; by the number of items bought, which would be variable, but discrete; by the amount of time spent in the store, which would be continuous; or by any of several other methods. A most likely classification would be the amount of money spent. While technically discrete, since the customer does not spend fractional parts of a penny, it is a common practice to consider as "practically" continuous a variable whose unit is quite small in relation to the amounts involved. If we were talking about the different amounts of pennies several children had, the variable would obviously be discrete. In terms of the national debt, the variable could be considered, for all practical purposes, continuous. A matter of judgment is involved.

Suppose you asked someone for data on the height of adult males in a city of some 25,000 population, and the person responded by giving you a list of 8,968 heights. Unless the data were organized in some fashion, this list in its raw form would not be too usable. One of the means employed to organize data is known as the **frequency distribution.** In its simplest form, the frequency distribution consists of listing each possible value the data could have and enumerating the total number, called the **frequency**, for each value. In the height example, if height is measured to the nearest inch, such a frequency distribution might appear as follows:

Height (Inches)	Frequency
83	11
82	44
81	116
80	132
79	157
78	284
77	316
76	388
75	547
74	731
73	783
72	808
71	817
70	931
69	848
68	712
67	604
66	411

Height (Inches)	Frequency
65	206
64	94
63	28
Total	8,968

Such a table tells you at a glance that most of the data points have values from 67 to 75, inclusive, that the number above 78 and below 65 is relatively quite small, and, in short, gives you a very good and accurate profile of this set of data.

It is often useful to determine the **proportion** of cases for each value of the variable. This is also called the **relative frequency**, which is the number of cases (frequency) for a given value divided by the total number of cases (total frequency). Actually, relative frequency can be interpreted as a percent. In the preceding example 931 men were 70 inches tall out of the total of 8,968, so the relative frequency for 70 inches was 931/8,968 or about 0.104, or 10.4%.

It is helpful to have a procedure to follow when constructing a frequency table. For small amounts of data, we can rewrite the data given into ascending (or descending) order. We then have the data placed in order and the construction of the table is relatively simple. Another plan, more useful in cases where a great deal of data is involved, is to find the highest and lowest values, then list these values and all cases between them. In a second column we *tally* the cases by putting a slash (tally) mark for each one as we come to it, usually crossing out the number in the original list, then summarize the results in a frequency table. The relative frequency can be included, if desired. As we shall see in the next chapter, the relative frequency can be highly important if the set of data is a representative sample of some population. The table should be complete, including the title, with enough information to make the table completely self-explanatory if not accompanied with explanatory material.

EXAMPLE 1 As a part of preliminary cost study, the amount of weekly sales at each of a department store chain's 25 outlets was obtained. The data given below represent the average weekly sales per store for the last three months given to the nearest thousand dollars. Construct a table showing frequency and relative frequency for each sales figure.

7	9	8	11	6
13	7	19	9	9
7	13	22	9	12
10	9	13	9	15
7	11	8	9	13

SOLUTION There are just a few pieces of data here, so rewriting these in order would be logical. An example of the tally method might be helpful at this point, however, so it is presented here.

Sales ($000's)	Tally
22	/
21	
20	
19	/
18	
17	
16	
15	/
14	
13	////
12	/
11	//
10	/
9	𝍸 //
8	//
7	𝍸
6	/

Using the tally chart we can construct the following table.

Average Weekly Sales, Le Chateau Stores[a]

Sales ($000's)	No. of Stores	Relative Frequency
22	1	.04
21	0	0.
20	0	0.
19	1	.04
18	0	0.
17	0	0.
16	0	0.
15	1	.04
14	0	0.
13	4	.16
12	1	.04
11	2	.08
10	1	.04
9	7	.28
8	2	.08
7	4	.16
6	1	.04
Total	25	1.00

[a]Survey covering June-August, 1977

COMMENT When several values of the variable contain no entries or relatively few entries, it is often convenient to combine them into one class. This is often done when these classes fall at one end of the distribution. Here we could write

14–22	3	.12

This does not significantly impair our understanding of the table, but may make it difficult to perform arithmetic on the data. It should never be done, for example, if any further analysis is contemplated. It is useful *only* for presentation, and never for interpretation.

DISCUSSION It generally does not matter whether tables are arranged in ascending or descending order. The guiding principles should be ease and clarity of presentation.

For some purposes a **cumulative frequency table** is useful. Although it does not matter whether the accumulation is done in ascending or descending order, the most frequent uses for such a table are found when the data are arranged in descending order with the cumulative frequency listed. The cumulative frequency is the sum of all frequencies equal to or less than the listed value. The **relative cumulative frequency** is the proportion of all frequencies *equal to or less than* the listed value. It can also be called the **cumulative relative frequency**. A table for the sales data follows.

Sales ($000's)	Number of Stores	Cumulative Frequency	Relative Cumulative Frequency
6	1	1	.04
7	4	5	.20
8	2	7	.28
9	7	14	.56
10	1	15	.60
11	2	17	.68
12	1	18	.72
13	4	22	.88
14–22	3	25	1.00

In general, only those columns which are needed would be listed. In the above table, we would probably have at most three columns, listing the data we wished to present.

A frequency table may also be used to classify attributes. As mentioned before, attributes are a type of data that cannot be measured, but can only be described. For example, a table may classify the automobiles in a parking lot by make.

Make	Number of Cars
Chevrolet	33
Ford	28
Pontiac	11
Volkswagen	9
Plymouth	7
Other makes	4
Total	92

Another example of attributes is the division of data into **strata**, such as low, middle, and high income groups. The difference between the two types of data is that data which is categorical only, without any ordering, is called a **nominal** classification (from the Latin for name), while categorical data in which an ordering is implicit is called an **ordinal** classification (from order). Data which is ranked is also an ordinal classification; for example, the finish of runners in a race is ordinal.

The other two classifications of data are interval and ratio. **Interval** data is data in which the differences between values are meaningful; that is, a difference of ten units means the same thing wherever it occurs. An example is temperature. A rise of twenty degrees means the same whether it is an increase from 80° to 100° or from $-8°$ to $+12°$. **Ratio** data is data which has a natural zero, so that ratios of values are meaningful. Weight and distance are examples of ratio data. A weight of 50 pounds is 5 times as heavy as a weight of 10 pounds and a distance of 65 miles is 13 times as far as a distance of 5 miles. This is not true of all interval data. A temperature of 50° is not 5 times as hot as a temperature of 10°. Such a statement has no meaning whatsoever.

Now suppose that a set of data gives the incomes of families in a city. Such a listing of *all* the values is called a census. It is obvious that a listing of all possible incomes (even to the nearest dollar) would involve a listing of hundreds, even thousands of numbers. One way out of such a difficulty is the method of **grouping data**. Thus we may consider all incomes from, say 4,000 to 7,999 dollars as constituting one **class**.

A few general observations govern the grouping of data into classes. First, we do not want too many or too few classes, since this might result in a distortion of the picture we want to convey. Second, the classes should be of the same width (although this rule may be bent in certain circumstances). Third, the classes should be convenient to handle. To achieve these aims, it has been found that eight to fifteen classes are a reasonable number. To determine class width, the range is divided by the approximate number of classes decided upon, then the number is rounded out to the nearest convenient division. The **range** of a set of data is defined as being the highest value minus the lowest value and is thus the distance between the two extremes. If the data are discrete, however, we add one to this value since both endpoints are included.