

NmLcU/IRU New Mexico State Univ



1010513203

Applied Statistical Methods

Irving W. Burr

APPLIED STATISTICAL METHODS

Irving W. Burr

PURDUE UNIVERSITY



ACADEMIC PRESS New York and London

A Subsidiary of Harcourt Brace Jovanovich, Publishers

COPYRIGHT © 1974, BY ACADEMIC PRESS, INC.

ALL RIGHTS RESERVED.

**NO PART OF THIS PUBLICATION MAY BE REPRODUCED OR
TRANSMITTED IN ANY FORM OR BY ANY MEANS, ELECTRONIC
OR MECHANICAL, INCLUDING PHOTOCOPY, RECORDING, OR ANY
INFORMATION STORAGE AND RETRIEVAL SYSTEM, WITHOUT
PERMISSION IN WRITING FROM THE PUBLISHER.**

ACADEMIC PRESS, INC.

111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by

ACADEMIC PRESS, INC. (LONDON) LTD.

24/28 Oval Road, London NW1

Library of Congress Cataloging in Publication Data

Burr, Irving Wingate, Date

Applied statistical methods.

(Operations research and industrial engineering)

Includes bibliographical references.

1. Mathematical statistics. 2. Probabilities.

I. Title

QA276.B863

519.5

73-2077

ISBN 0-12-146150-5

AMS (MOS) 1970 Subject Classifications: 62N99, 62P99, 62N10

PRINTED IN THE UNITED STATES OF AMERICA

Chapter 3

SUMMARIZATION OF DATA BY OBJECTIVE MEASURES

3.1. INTRODUCTION

Although a frequency table and a graph such as a histogram will give one quite a bit of information about the data, most of it is rather subjective. In order to describe the data more objectively and to provide tools for analysis, it is desirable to use what are called "sample statistics." Such statistics are of three kinds: averages, measures of variability, and measures of curve shape. The first two are used with most samples, the last only with large samples.

3.2. SOME AVERAGES

The aim of an average is to give a one-number description of the sample of numbers. If all the numbers were identical, the average would be this number. If the numbers vary, but still show some "central tendency," then it is the object of the average to locate this level. Different kinds of averages have various characteristics: advantages and disadvantages, and each kind has its special use and interpretation. Finally one may say that if the data are so scattered as to exhibit no noticeable central tendency or pattern, then it is doubtful whether any average is worth much.

Certainly the most commonly used and widely understood average is the so-called arithmetic mean. It is simply the sum of the set of numbers,

divided by how many there are. If we call the individual numbers $y_1, y_2, y_3, \dots, y_n$, there being n of them, then their arithmetic mean is defined by

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum y}{n}, \quad (3.1)$$

where the symbol $\sum$ means "sum of." The *sample size* is n . If desirable, we could have shown the limits on the summation as follows:

$$\sum_{i=1}^n y_i = y_1 + y_2 + \dots + y_n. \quad (3.2)$$

The summation symbol is much used in statistics. Three simple laws are as follows (c is a constant)*:

$$\sum_{i=1}^n c = nc \quad \text{or} \quad \sum_{i=a}^b c = (b - a + 1)c \quad (3.3)$$

$$\sum_{i=a}^b cy_i = c \sum_{i=a}^b y_i \quad (3.4)$$

$$\sum_{i=a}^b (y_i \pm x_i) = \left(\sum_{i=a}^b y_i \right) \pm \left(\sum_{i=a}^b x_i \right). \quad (3.5)$$

* The proofs of these laws are almost self-evident, if one merely writes out the expansion as in (3.2) for each $\sum$, and compares.

Some of the desirable properties of the arithmetic mean are as follows:

1. It takes account of the value of each of the numbers and is affected by each.
2. If each y_i is multiplied by the same number k , then the arithmetic mean of the resulting numbers is $k\bar{y}$.
3. If each y_i is changed by the addition of A to it, the arithmetic mean of the resulting numbers is $\bar{y} + A$.
4. If $\bar{y}_1$ is the arithmetic mean of n_1 numbers, and $\bar{y}_2$ of n_2 , then the arithmetic mean of all $n_1 + n_2$ numbers is the "weighted mean"

$$\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}.$$

5. The mean has many desirable properties in the theory and practice of sampling, that is, the relation between sample and population.

As an example consider the following data on amounts of fuel metered by a nozzle in five trials under uniform test condition: y : 96.6, 97.2, 96.4, 97.4, 97.8, $\sum y = 485.4$, $\bar{y} = 97.08$ (all in cubic centimeters).

A second kind of average, the "median," we have already mentioned in Chapter 2. There it was described as the y value for which the cumulative frequency is $n/2$. If n is even and there are no tied values of y in the middle, there will be half of the cases below the median and half above it. Or if n is odd, the median is the midmost number after all the numbers have been arranged in order, that is, in an "array." As many numbers y lie below the median as above it, barring ties in the middle of the array. For the small sample of amounts metered, we first rank in order in an array: 96.4, 96.6, 97.2, 97.4, 97.8 cc, and the median is then the middle number, 97.2 cc. If n is even, then when we rank the numbers into an array, the median is half-way between the *two* middle numbers. Tie scores may make interpretation a little more difficult, but this does not complicate the method of finding the median.

The median has a number of desirable properties. Among those given for the arithmetic mean, it possesses 2 and 3, which should in fact be possessed by nearly any average worth consideration. In regard to property 1, the median is only indirectly affected by the value of each of the n numbers. Thus if the lowest amount metered were 86.4 instead of 96.4, the median would still be 97.2 cc, but the sum would be reduced by 10 cc, and $\bar{y}$ by 2 cc to 95.08 cc. Property 4 definitely does not hold for the median, while regarding 5, the median has less desirable sampling properties than the mean.

A third type of average, the "mode," we have also introduced in Chapter 2. It is the point of greatest concentration of the numbers in the data. It is usually found as the midvalue of the class with greatest frequency. Obviously the mode then depends on whatever classes we have chosen for the tabulation. The mode possesses properties 2 and 3, but not 1 or 4, and it has little use in statistical analysis. Its prime use is in the description of a frequency distribution. For example, a distribution with two isolated high frequencies is described as "bimodal," and we could list two modes.

Since the arithmetic mean is so commonly used, we shall frequently call it simply the "average."

3.3. SOME MEASURES OF VARIABILITY

There are three commonly used measures of the variability in a set of numbers, namely the "range," the "mean deviation," and the "standard deviation." Of these the last is by far the most important.