Biostatistical Analysis

JERROLD H. ZAR

Associate Professor
Department of Biological Sciences
Northern Illinois University

PRENTICE-HALL, INC.

Englewood Cliffs, N.J.

Library of Congress Cataloging in Publication Data

ZAR, JERROLD H
Biostatistical analysis.

(Prentice-Hall biological sciences series)
Bibliography: p.
1. Biometry. I. Title. DNLM: 1. Biometry.
2. Statistics. QH 405 Z36b 1974
QH323.5.Z37 574'.01'5195 73-3443
ISBN 0-13-076984-3

10 9 8 7

Printed in the United States of America

PRENTICE-HALL INTERNATIONAL, INC., *London*
PRENTICE-HALL OF AUSTRALIA, PTY. LTD., *Sydney*
PRENTICE-HALL OF CANADA, LTD., *Toronto*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*

4

Measures of Dispersion and Variability

In addition to a measure of central tendency, it is generally desirable to have a *measure of dispersion* of data. A measure of dispersion, or a measure of variability, as it is sometimes called, is an indication of the clustering of measurements around the center of the distribution, or, conversely, an indication of how variable the measurements are. Measures of dispersion of populations are parameters of the population, and the sample measures of dispersion that estimate them are statistics.

4.1 The Range

The difference between the highest and lowest measurements in a group of data is termed the *range*. If sample measurements are arranged in increasing order of magnitude, as if the median were about to be determined, then

$$\text{sample range} = X_n - X_1. \qquad (4.1)$$

Sample 1 in Example 4.1 is a hypothetical set of data in which $X_1 = 1.2$ g and $X_n = 2.4$ g. Thus, the range may be expressed as 1.2 to 2.4 g, or as 2.4 g — 1.2 g = 1.2 g. (We might bear in mind that X_1 is really within the limits of 1.15 to 1.25 g and X_n is really 2.35 to 2.45 g, so that the range of the sample would be expressed by a few authors as 2.45 g — 1.15 g = 1.3 g.) Note that the range has the same units as the individual measurements.

The range is a relatively crude measure of dispersion, inasmuch as it does not take into account any measurements except the highest and the lowest. Furthermore, since it is unlikely that a sample will contain both the highest and lowest values in the population, the sample range usually underestimates the population range; therefore,

Example 4.1 Calculation of measures of dispersion for two hypothetical samples.

Sample 1

X_i (g)	$X_i - \bar{X}$ (g)	$\lvert X_i - \bar{X} \rvert$ (g)	$(X_i - \bar{X})^2$ (g²)
1.2	−0.6	0.6	0.36
1.4	−0.4	0.4	0.16
1.6	−0.2	0.2	0.04
1.8	0.0	0.0	0.00
2.0	0.2	0.2	0.04
2.2	0.4	0.4	0.16
2.4	0.6	0.6	0.36
$\sum X_i = 12.6$ g	$\sum (X_i - \bar{X}) = 0.0$ g	$\sum \lvert X_i - \bar{X} \rvert = 2.4$ g	$\sum (X_i - \bar{X})^2 = 1.12$ g² = "sum of squares"

$$\bar{X} = \frac{12.6 \text{ g}}{7} = 1.8 \text{ g}$$

$$\text{range} = X_7 - X_1 = 2.4 \text{ g} - 1.2 \text{ g} = 1.2 \text{ g}$$

$$\text{mean deviation} = \frac{\sum \lvert X_i - \bar{X} \rvert}{n} = \frac{2.4 \text{ g}}{7} = 0.34 \text{ g}$$

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{1.12 \text{ g}^2}{6} = 0.1867 \text{ g}^2$$

$$s = \sqrt{0.1867 \text{ g}^2} = 0.43 \text{ g}$$

Sample 2

X_i (g)	$X_i - \bar{X}$ (g)	$\lvert X_i - \bar{X} \rvert$ (g)	$(X_i - \bar{X})^2$ (g²)
1.2	−0.6	0.6	0.36
1.6	−0.2	0.2	0.04
1.7	−0.1	0.1	0.01
1.8	0.0	0.0	0.00
1.9	0.1	0.1	0.01
2.0	0.2	0.2	0.04
2.4	0.6	0.6	0.36
$\sum X_i = 12.6$ g	$\sum (X_i - \bar{X}) = 0.0$ g	$\sum \lvert X_i - \bar{X} \rvert = 1.8$ g	$\sum (X_i - \bar{X})^2 = 0.82$ g² = "sum of squares"

$$\bar{X} = \frac{12.6 \text{ g}}{7} = 1.8 \text{ g}$$

$$\text{range} = X_7 - X_1 = 2.4 \text{ g} - 1.2 \text{ g} = 1.2 \text{ g}$$

$$\text{mean deviation} = \frac{\sum \lvert X_i - \bar{X} \rvert}{n} = \frac{1.8 \text{ g}}{7} = 0.26 \text{ g}$$

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{0.82 \text{ g}^2}{6} = 0.1367 \text{ g}^2$$

$$s = \sqrt{0.1367 \text{ g}^2} = 0.37 \text{ g}$$

it is a biased and inefficient estimator. Nonetheless, it is useful in some circumstances to present the sample range as an estimate (although a poor one) of the population range. Taxonomists are frequently concerned, for example, with having an estimate of what the highest and lowest values in a population are expected to be. Whenever the range is specified in reporting data, however, it is usually a good practice to report another measure of dispersion as well. The range is applicable to ordinal, interval, and ratio scale data.

4.2 The Mean Deviation

As is evident from the two samples in Example 4.1, the range conveys no information about how clustered about the middle of the distribution the measurements are. Since the mean is so useful a measure of central tendency, one might express dispersion in terms of deviations from the mean. The sum of all deviations from the mean, i.e., $\sum (X_i - \bar{X})$, will always equal zero, however, so such a summation would be useless as a measure of dispersion (see Example 4.1).

To sum the absolute values of the deviations from the mean results in a quantity that is an expression of dispersion about the mean. Dividing this quantity by n yields a measure known as the *mean deviation*, or *mean absolute deviation* of the sample. In Example 4.1, sample 1 is more variable (or more dispersed, or less concentrated) than sample 2. Although the two samples have the same range, the mean deviation, calculated as

$$\text{sample mean deviation} = \frac{\sum |X_i - \bar{X}|}{n}, \tag{4.2}$$

expresses the differences in dispersion. Mean deviation can also be defined by using the sum of the absolute deviations from the median rather than from the mean.

4.3 The Variance

Another method of eliminating the signs of the deviations from the mean is to square the deviations. The sum of the squares of the deviations from the mean is called the *sum of squares*, abbreviated SS, and is defined as follows:

$$\text{sample SS} = \sum (X_i - \bar{X})^2. \tag{4.3}$$

The mean sum of squares is called the *variance* (or *mean square*, the latter being short for *mean squared deviation*), and for a population is denoted by σ^2 ("sigma squared," using the lowercase Greek letter):

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}. \tag{4.5}$$

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.