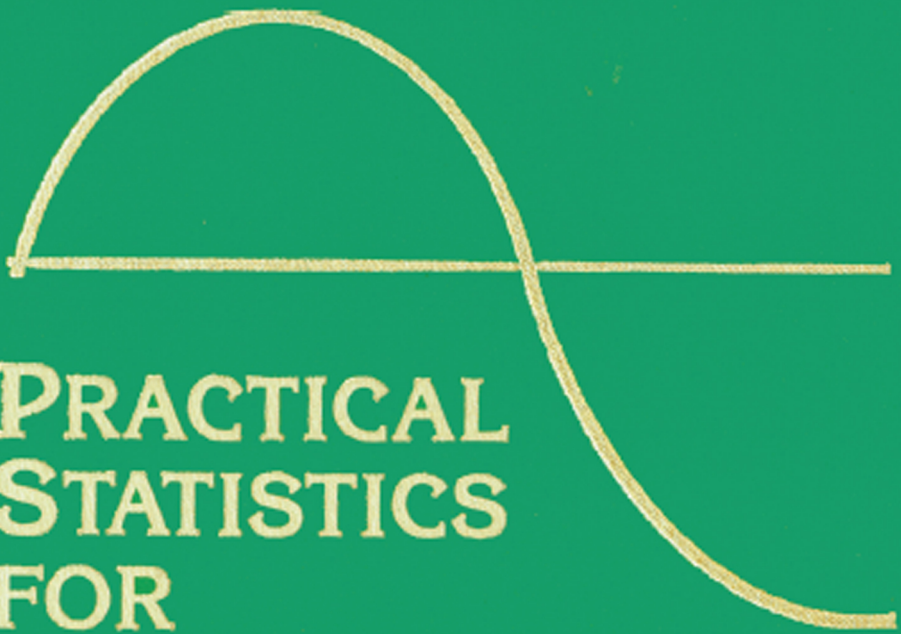
CRC Press
Taylor & Francis Group



PRACTICAL STATISTICS FOR ENGINEERS AND SCIENTISTS

NICHOLAS P. CHEREMISINOFF

Practical Statistics
for Engineers and Scientists

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

PRACTICAL STATISTICS FOR ENGINEERS AND SCIENTISTS

NICHOLAS P. CHEREMISINOFF

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

First published 1987 by Technomic Publishing Company, Inc.

Published 2019 by CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

First issued in paperback 2019

ISBN-13: 978-0-367-45137-0 (pbk)
ISBN-13: 978-0-87762-505-6 (hbk)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Main entry under title:
Practical Statistics for Engineers and Scientists

Bibliography: p. 201
Includes index p. 205

Library of Congress Card No. 86-72352

TABLE OF CONTENTS

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

PREFACE

This monograph is intended as a handy reference of standard statistical analyses and data regression techniques. It is written in a short format to enable the user to obtain a working knowledge of important statistical tests. Each method is fully illustrated with practical problems drawn from engineering and science-oriented applications. The volume does not provide rigorous treatment of statistical theorems and theory, but lists useful references for those readers who desire more in-depth understanding of the mathematical bases. The book also provides direction in constructing regression routines that can be used with worksheet software on personal computers. A software program called DATA-FIT developed for simple linear and non-linear regressions and capable of handling up to three variables is described in this volume. The program is designed to run with LOTUS 1-2-3 spreadsheet on an IBM compatible personal computer. Appendix A provides a full description of the program and can serve as the user's guide. A floppy disc of this user friendly program can be obtained by using the convenient tear-out order form in the back of this volume. In addition, the volume provides a listing of various commercial software programs and the statistical analyses capabilities associated with each.

Science and engineering students should find this to be a good supplemental reference in course work and in graduate studies. Process engineers, chemists and researchers involved in experimental programs and pilot plant operations may find this to be a handy guide in designing experiments and analyzing data.

Special thanks are extended to the staff of Technomic Publishing Co. for the production of this volume, and their continued interest and attention.

NICHOLAS P. CHEREMISINOFF

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

ABOUT THE AUTHOR

Nicholas P. Cheremisinoff heads the Product Development Group of the Elastomers Technology Division of Exxon Chemical Co. in Linden, New Jersey. Among his areas of interest are multiphase flows and rheological and processing behavior of polymeric materials. He is the author and co-author of many engineering textbooks, and is a member of the AIChE, Tau Beta Pi and Sigma Xi. Dr. Cheremisinoff received his B.S., M.S. and Ph.D. degrees in chemical engineering from Clarkson College of Technology.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Notations and Definitions

USE OF SUBSCRIPTS

Subscripts are used to denote the number of an observation in a set. For Example: x_i is the ith observation in a set. $x_{(i)}$ is the ith ordered observation in a set. Let us review Illustration 1.

Illustration 1

Consider the following data set $\{5, 10, 6, 2\}$. The observations are:

$$x_1 = 5, \quad x_2 = 10, \quad x_3 = 6, x_4 = 2$$

$$x_{(1)} = 2, \quad x_{(2)} = 5, \quad x_{(3)} = 6, \quad x_{(4)} = 10$$

Summations are denoted by the symbol of capital Sigma (Σ) and the operation is performed over the limits of the observation.

Illustration 2

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$$

$$\sum_{i=1}^{n} x_1 = x_1 + x_2 + x_3 + \ldots + x_n$$

$$\sum_{i=1}^{n} x_1^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$$

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_1 = c(x_1 + x_2 + \ldots + x_n), \text{ where } c = \text{constant}$$

$$\sum_{i=1}^{n} c = nc$$

$$\sum_{i=1}^{n} (x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{3} \sum_{j=1}^{2} x_{ij} = \sum_{i=1}^{3} (x_{i1} + x_{i2}) = x_{11} + x_{21} + x_{31} + x_{12} + x_{22} + x_{32}$$

DEFINITIONS OF DISTRIBUTIONS

An important step in analyzing the statistical significance of data and in regression is the layout and tabulation of data values. *Frequency distributions* are a useful format for scanning and reviewing the important features of large bodies of data. In preparing frequency distributions one must select data groups. Important rules-of-thumb to apply when choosing data groups are:

- Find the minimum and maximum values in the set.
- Do not use fewer than 6 or more than 15 groups.
- Select unambiguous grouping criteria.
- Group intervals should be of equal length.
- Select a convenient group midpoint.

An example of preparing frequency distributions using the above rules is shown below.

Illustration 3

Bales of rubber are checked in a warehouse for brown spots. These brown spots are considered as contamination and can be the basis for rejection by customers. The number of brown spots per bale is tabulated in Table 1-1 for 100 data sets.

The minimum and maximum values in the set are identified in Table 1-1. We

Table 1-1. Data Tabulation of Brown Spots.

6	20	6	20	26	15	12	19	20	30
3	23	3	(33)	15	0	23	7	19	8
10	19	7	14	20	18	24	8	22	13
20	17	7	15	18	17	25	8	22	15
22	19	8	12	6	15	25	6	17	9
23	24	10	10	7	16	22	23	18	12
27	12	10	13	7	17	19	24	21	11
2	9	15	15	9	17	18	26	26	15
(0)	7	16	17	10	22	18	23	20	18
15	5	15	19	0	21	17	25	18	17

Table 1-2. Tally of Raw Data to Prepare Frequency Distribution.

Contamination Level (No. Brown Spots per Frequency	Group Frequency	Relative
0 - 5	7	0.07
6 - 11	23	0.23
12 - 17	27	0.27
18 - 23	31	0.31
24 - 29	10	0.10
30 - 35	2	0.02
	100	1.00

can now select groups and tally as shown in Table 1-2. The relative frequency column in Table 1-2 is calculated from the following definition.

$$\text{Relative Frequency} = \text{Group Frequency/Total Data} \quad (1)$$

The frequency distribution can be represented graphically as shown in Figure 1-1. Figure 1-la shows a line chart frequency distribution histogram, where the midpoint of each group set is used for the corresponding group frequency. Figure 1-lb shows a bar chart histogram.

On a personal computer using a spreadsheet program such as Lotus 1-2-3, this exercise should only take moments, once the data have been typed in.

MEASURES OF LOCATION, MEANS AND DATA CODING

The *measures of location* refer to the centrality or norm that characterizes a typical value of a population or sample population of data. The principle

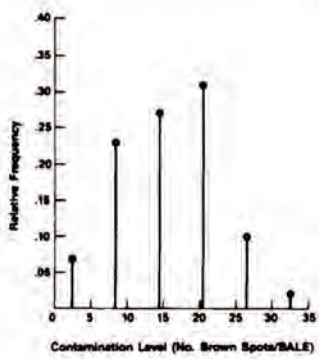


FIGURE 1-la. Line chart histogram.

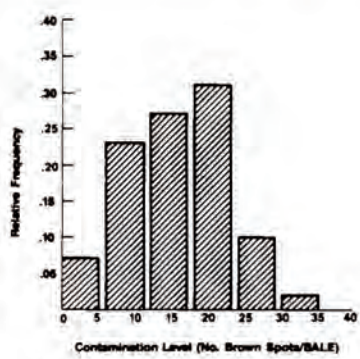


FIGURE 1-lb. Bar chart histogram.

term used to describe this property is the *mean*, of which there are several definitions (average, arithmetic mean, expected value, geometric mean). Symbols used are:

$$\mu = \text{population mean}$$

$$\bar{x} = \text{sample mean}$$

These are the most widely used measure of location and can best be related by analogy to the concept of center of gravity for bodies.

The mean value is computed from a data set $x_1, \ldots x_n$ as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

The population mean, μ, is calculated on a similar basis.

Illustration 4

Calculate the mean value for the following set: {16, 12, 7, 11, 17, 9}:

$$\bar{x} = \frac{16 + 12 + 7 + 11 + 17 + 9}{6} = \frac{72}{6} = 12.0$$

The calculation of the mean of grouped data is more readily performed on the basis of a *weighted average*. Grouping the data will result in some loss of accuracy, compared with averaging ungrouped data.

Illustration 5

Tabulate the k group midpoints, x_i, and group frequencies f_i. Then compute the mean. Use the data set in Illustration 3.

From Table 1-2 we tabulate the following:

x_i	f_i	$f_i x_i$
2.5	7	17.5
8.5	23	195.5
14.5	27	391.5
20.5	31	635.5
26.5	10	265.0
32.5	2	65.0
	100	1570.0

The mean is calculated from:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{1570}{100} = 15.70$$

Note that the mean value of the raw data was *15.47*.

The *median* denotes the 50*th* percentile value (i.e., the middlemost value). The symbol most often used is $\tilde{x}$. It is the second most important measure of location. It does, however, have limitations; it does not use the entire population of data and, hence, is less reliable than the mean as an estimate of μ. A further disadvantage is that it can be a cumbersome calculation for large samples. Consequently, it is most often used for small data samples. Whereas the mean cannot be used for ordered qualitative data, the median can.

The calculation procedure for the median is:

- First, order the data.
- For an odd-numbered sample size n: $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$
- * For an even-numbered sample size n: $\tilde{x} = 1/2\left(x_{\frac{n}{2}} + x_{\frac{n+2}{2}}\right)$

Illustration 6

Using the data set from Illustration 4:

$$x_{(1)} = 7,\quad x_{(2)} = 9,\quad x_{(3)} = 11,\quad x_{(4)} = 12,\quad x_{(5)} = 16,\quad x_{(6)} = 17$$

(a) For $n = 6$, $n/2 = 3$ and $(n + 2)/2 = 4$,

Therefore,

$$\tilde{x} = \frac{1}{2}(x_{(3)} + x_{(4)}) = \frac{1}{2}(11 + 12) = 11.5$$

(b) Assume $n = 5$; then $(n + 1)/2 = 3$

$$\tilde{x} = x_{(3)} = 11$$

The *mode* is a property value that corresponds to the highest frequency in the data. It can be thought of as the most likely value. If the data values are random, the mode may not exist. In addition, it may not be unique;

Table 1-3. Tabulations for Illustration 7.

Column No.	(I) x_i	(II) c	(III) z_i	(IV) f_i	(V) $f_i z_i$
	2.5	*14	*11.5	7	* 80.5
	8.5	*14	* 5.5	23	*126.5
	14.5	*14	0.5	27	13.5
	20.5	*14	6.5	31	201.5
	26.5	*14	12.5	10	125.0
	32.5	*14	18.5	2	37.0
					170.0

that is, two or more values may exist with equal frequency. It can be used for qualitative data having no natural ordering and in samples above the median point.

Manipulating or *coding* of data is a useful technique for simplifying statistical calculations. This is done by subtracting a number from each data value. The number chosen should be rounded off and approximately in the center of the data in order to obtain the coded data. The formula is:

$$z_i = x_i - c \tag{3}$$

Illustration 7

Using the data in Illustration 3, subtract a round number c approximately in the center of the set to obtain the coded data. Using the midpoint values x_i and from the frequency distribution in Table 1-2, we calculate the coded midpoint z_i (refer to columns (I) - (III) in Table 1-3).

Next, we compute the coded mean $\bar{z}$ *using the coded midpoints,* z_i

$$\bar{z} = \frac{\sum_{i=1}^{k} f_i z_i}{\sum_{i=1}^{k} f_i}$$

$f_i z_i$ and frequency in columns (IV) and (V) of Table 1-3. Therefore,

$$\bar{z} = \frac{170}{100} = 1.70$$

We now decode the value to obtain the true mean:

$$\bar{x} = \bar{z} + c$$
$$\bar{x} = 1.7 + 14.0 = 15.7$$

VARIABILITY OF DATA

The spread or dispersion of a data set provides a measure of the variability of the population about the mean. The parameter that describes this property is the *standard deviation* (also called the root mean square deviation). The symbols used are σ for the population standard deviation, and s for the sample standard deviation.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \tag{4}$$

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{n - 1}} \tag{5}$$

The units of σ and s are the same as the units of the data. The quantity $(n - 1)$ is referred to as the *degrees of freedom*. The deviations squared, σ^2 and s^2, are called the *variances* of the population and sample, respectively. An alternate formula for the sample standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n - 1}} \tag{6}$$

Illustration 8

For the data $x_1, \ldots x_n$ in Illustration 4, calculate the standard deviation s. From the data, $\bar{x} = 12.0$. The following values are tabulated:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
16	4	16
12	0	0
7	-5	25
11	-1	1
17	5	25
9	-3	9
		76

By using formula (5), we get:

$$s = \sqrt{\frac{76}{(6-1)}} = 3.90$$

Or, with formula (6):

$$\sum_{i=1}^{n} x_i^2 = 256 + 144 + 49 + 121 + 289 + 81 = 940$$

$$\sum_{i=1}^{n} x_i = 72$$

$$s = \sqrt{\frac{940 - \frac{(72)^2}{6}}{5}} = \sqrt{\frac{940 - 864}{5}} = 3.90$$

The *sample range*, R, represents the difference between the largest and smallest value in the sample. It is most often used as a cross-check of the σ calculation. R does not use all the data and, hence, is not as accurate as s. Usefulness is limited to sample sizes less than 15.

Illustration 9

By using the data from Illustration 4, we obtain:

$$R = x_{(max)} - x_{(min)}$$

$$x_{(max)} = 17, \; x_{(min)} = 7$$

$$R = 17 - 7 = 10$$

The relative variation of data is called the *coefficient of variation*. It is defined as the ratio of the standard deviation to the mean and is expressed as a percentage:

$$r = \frac{\sigma}{\mu} \cdot 100 \qquad (7)$$

Since it is a dimensionless quantity, it can be used to compare variability among different units of scale.

Illustration 10

Using the data from Illustration 8, calculate r for $\bar{x} = 12.0$, $s = 3.90$. Hence,

$$r = \frac{s}{\bar{x}} \cdot 100 = \frac{3.90}{12} \times 100 = \underline{3.25\%}$$

TYPES OF DISTRIBUTIONS

Previously we introduced distributions of discrete variates that take on a limited number of distinct values. Now we direct attention to variables that are continuous, such as height, weights, concentrations, or yields. The variable continues without a break from one value to the next with no limit to the number of distinct values. Consider the histogram of contamination levels in Figure 1-1b. Imagine that the size of the sample is increased without limit and the class intervals on the horizontal axis are decreased correspondingly. Figure 1-1b would gradually become a continuous curve. Continuous variables are distributed in a number of ways; we first consider the *normal distribution*.

The normal or Gaussian distribution is a symmetrical, bell-shaped curve, and is entirely determined by its mean μ and standard deviation σ. With a continuous variable x, we may define a function $f(x)$ that states that the height or ordinate of the continuous curve at the abscissa value x. The height $f(x)$ at the value x is:

$$Y = f(x) = [1/\sigma\sqrt{2\pi}]e^{-(x-\mu)^2/2\mu^2} \tag{8}$$

The theoretical Gaussian distribution curve is shown in Figure 1-2. The areas corresponding to the intervals shown are as follows:

Interval	Area (%)	Interval	Area (%)
$\mu \pm \sigma$	68.27	$\mu \pm 0.674\sigma$	50
$\mu \pm 2\sigma$	95.45	$\mu \pm 1.645\sigma$	90
$\mu \pm 3\sigma$	99.73	$\mu \pm 1.960\sigma$	95
$\mu \pm 4\sigma$	99.994	$\mu \pm 2.576\sigma$	99

We can interpret the Gaussian distribution in the following manner. If a very

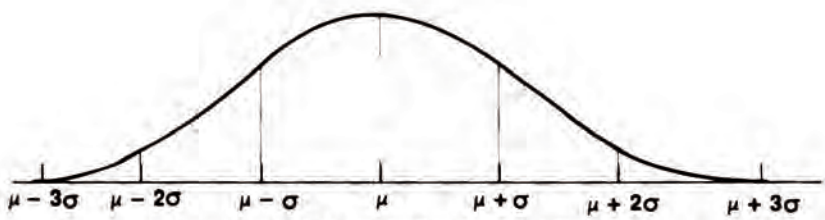


FIGURE 1-2. Gaussian distribution curve.

large sample were drawn from this population, we could expect 95.45% of the values to lie within the limits of $\mu \pm 2\sigma$.

The importance of the Gaussian distribution is as follows:

1 Many distributions of variables of natural phenomenon are approximately normal; examples include bubble sizes in fluidized reactors, natural soil particle sizes, pollutant concentrations in the atmosphere.

2 For measurements whose distributions are not normal, a simple transformation of the scale of measurements may approximate normality. These transformations often occur as a square root, $\sqrt{x}$, and the logarithm, $\ln x$.

3 The normal distribution provides a simple mathematical relation, thus facilitating analytical solutions to many problems. This approach often provides reasonable approximations within engineering standards even when samples are derived from non-normal populations. This approach is often feasible even when handling populations with discrete variables.

4 Even when the distribution in the original population is highly non-normal, the distribution of the sample means $\bar{x}$ tends to approach normal under random sampling, as the sample size increases. This in itself is justification for the importance of the Gaussian distribution.

Since the normal curve depends on the two paramaters μ and σ, there are many different normal curves. Standard tables of this distribution given in statistics textbooks such as Snedecor and Cochran (1980), Walpole and Myers (1972), and Fisher (1950), are for the distribution with $\mu = 0$ and $\sigma = 1$. Therefore, for a measurement x with mean μ and standard deviation σ, in order to use a table of the normal distribution, one must rescale x so that the mean becomes 0 and the standard deviation becomes unity. The rescaled measurement can be calculated from:

$$z' = (x - \mu)/\sigma \tag{9}$$

Variable z' is called the *standard normal variate* (also the standard normal deviate or normal variate in standard measure). Values can be transformed back to the x-scale by:

$$x = \mu + \sigma z \tag{10}$$

A *skewed* distribution is one that is non-symmetrical. Skewness can be measured by the following formula:

$$\Sigma(x - \bar{x})^3/(\Sigma(x - \bar{x})^2)^{3/2} \tag{11}$$