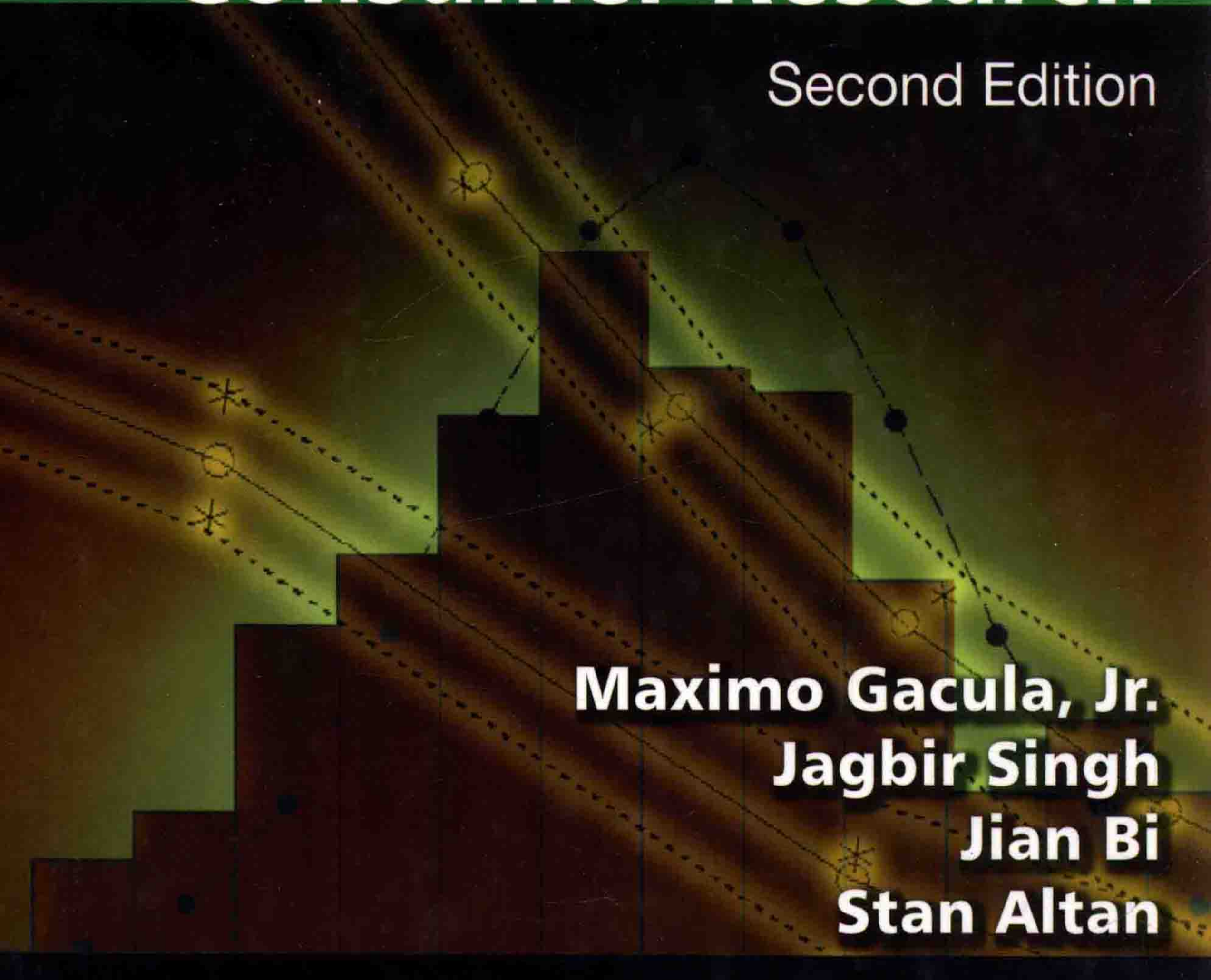
Statistical Methods in

Food and Consumer Research

Second Edition



Maximo Gacula, Jr.
Jagbir Singh
Jian Bi
Stan Altan

Food Science and Technology, International Series

Statistical Methods in Food and Consumer Research
Second Edition

Maximo C. Gacula, Jr.
Gacula Associates Consulting, Scottsdale, Arizona

Jagbir Singh
Department of Statistics, Temple University, Philadelphia, Pennsylvania

Jian Bi
Sensometrics Research and Service, Richmond, Virginia

Stan Altan
J & J Pharmaceutical Research and Development, Raritan, New Jersey

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

Library of Congress Cataloging-in-Publication Data
APPLICATION SUBMITTED

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: 978-0-12-373716-8

For information on all Academic Press publications
visit our Web site at www.elsevierdirect.com

Printed in the United States of America
08 09 10 9 8 7 6 5 4 3 2 1

Statistical Methods in Food and Consumer Research

Second Edition

Food Science and Technology International Series

A complete list of books in this series appears at the end of this volume.

To my wife Mely and daughters Karen, Lisa, Elena.
Maximo C. Gacula, Jr.

To my wife, Veena, without whom my professional accomplishments would have been impossible and with whom each milestone is shared and enjoyed.
Jagbir Singh

To my wife Yulin and daughter Cindy.
Jian Bi

To Sodnam, Yelen and Tracy.
Stan Altan

Preface to the Second Edition

Statistical Methods in Food and Consumer Research, published in 1984, is the first book that dealt with the application of statistics to food science and sensory evaluation. Since then, statistical software has become indispensable in the statistical analysis of research data. The use of multivariate analysis in sensory evaluation has emerged in part due to the availability of statistical software that sensory and research analysts can effectively use. The use of nonparametric statistics in sensory evaluation resulted in several scientific research publications which are included in the second edition.

For updating applications of statistics to advancement of sensory science we believe there is a need for a second edition of the book. We are pleased to have Jian Bi with years of experience in applications of statistics, in particular sensometrics, and Dr. Stan Altan with years of expertise in experimental designs as coauthors. Chapters 9–11 of the book were completely re-written and expanded from the basic principles of several statistical procedures to sensory applications. Two chapters were added—Chapter 12, Perceptual Mapping and Descriptive Analysis, and Chapter 13, Sensory Evaluation in Cosmetic Studies. For data analysis we have provided computer programs or codes in S-Plus® and SAS® (Statistical Analysis System) for Chapters 9–13. For Chapter 7 which deals with Response Surface Designs, Design-Expert® is highly recommended.

The revised book can be used as text for a first course in applied statistics for students majoring in Food Science and Technology, and for other students in agricultural and biological sciences. As a first course, Chapters 1–8 will be used. For graduate students majoring in sensory science, a second course would cover Chapters 9–13.

We thank the academic and industrial professionals who found the book useful and encouraged us to update it. We are grateful to Academic Press/Elsevier, Inc. for their interest in publishing the second edition, in particular to Nancy Maragioglio, Carrie Bolger, and Christie Jozwiak for their continuous and readily available support.

Maximo C. Gacula, Jr.
Jagbir Singh
Jian Bi
Stan Altan

Table of Contents

Chapter 1

Introduction

This introductory chapter is written to familiarize the reader with common symbols, notation, concepts, and some probability tools that are used in the subsequent chapters. This chapter, therefore, is a minimal prerequisite for the remainder of this book. The details of the material presented here can be found in many beginning statistical methods textbooks, such as those by Dixon and Massey (1969) and Huntsberger and Billingsley (1981).

1.1 A BRIEF REVIEW OF TOOLS FOR STATISTICAL INFERENCE

The purpose of this section is to review briefly the tools used in *statistical inference*. Statistical inference is the process of making inductive statements based on *random* samples from target populations. The inference process includes the estimation of population parameters and the formulation and testing of hypotheses about the parameters: a *parameter* is a numerical quantity that describes some characteristics of the population. For instance, the mean μ and the variance σ^2 are examples of population parameters. In the hypothesized simple linear regression model, $Y = \alpha + \beta X + \varepsilon$, the intercept α and the slope β are the parameters of the model.

Because random sampling is a prerequisite for valid statistical inference, it is important that the samples be collected following an accepted random sampling procedure. If this is not done, the inferences about the population are not reliable. A precise and reliable inference can be reached only through the use of an appropriate *experimental design* for random sampling and data collection. The general principles of experimental designs are summarized in Section 1.2. The sample size plays a major role in ensuring that the selected random sample will lead to statistical inferences with a specified amount of confidence.

Notation and Symbolism

The science of statistics is concerned with the collection, summarization, and interpretation of experimental data. The data may be summarized in the form of graphs, charts, or tables. Because of the many statistical operations to which the data are subjected, a general system of notation is used to facilitate the operations. A working knowledge of this system is essential in understanding the computational procedures involved in statistical analyses. A brief description of the notation that will be encountered in succeeding chapters is given here.

A commonly encountered symbol is sigma, Σ, used to denote a sum or total. For example, the sum of n numbers $X_1, X_2, \ldots, X_n$ is denoted by

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n. \qquad (1.1\text{-}1)$$

The lower and upper limits on Σ indicate that the X_i is added from $i = 1$ to $i = n$; for simplicity, the limits on Σ may be omitted if no confusion is likely. Similarly, we have the following:

$$\begin{aligned} \sum X_i^2 &= X_1^2 + X_2^2 + \cdots + X_n^2, \\ \sum X_i Y_i &= X_1 Y_1 + X_2 Y_2 + \cdots + X_n Y_n, \\ \sum cX &= cX_1 + cX_2 + \cdots + cX_n = c\sum_{i=1}^{n} X_i, \end{aligned} \qquad (1.1\text{-}2)$$

where c is a constant. Note that $(\Sigma X_i)^2$, which denotes the square of the sum defined by (1.1-1), is not the same as that given by (1.1-2).

Whereas the population mean is symbolized by μ, the mean of a sample $X_1, X_2, \ldots, X_n$ of n items is denoted by $\overline{X}$, where

$$\overline{X} = \sum X_i / n. \qquad (1.1\text{-}3)$$

Similarly, the population variance is denoted by σ^2, where the sample variance is denoted by S^2, where

$$S^2 = \sum (X_i - \overline{X})^2 / n - 1. \qquad (1.1\text{-}4)$$

Whenever the population parameters are unknown, as is usually the case, they are estimated by appropriate sample quantities called *statistics*. Thus, if population mean μ and variance σ^2 are unknown, they may be estimated by statistics $\overline{X}$ and S^2, respectively. At times an estimate of a parameter, say θ, is denoted by $\widehat{\theta}$. Thus, an estimate of the population μ can be denoted by $\widehat{\mu}$, which may or may not be $\overline{X}$.

Depending on the populations under investigation, appropriate symbols can be made more explicit. For example, if more than one population is under investigation in a study, their means can be denoted by μ_1, μ_2, and so on.

In a multiclassified data structure, the dot notation is used to indicate totals. For example, let X_{ijk} denote the kth observation in the (i, j)th cell. Then $X_{ij\cdot}$ can be used to denote the sum of n_{ij} observations in the (i, j)th cell. That is $X_{ij\cdot} = \sum_{k=1}^{nij} X_{ijk}$. If there is only one observation per cell, that is, $n_{ij} = 1$, then there is no use in retaining the subscript k, and one may express the data in a row and column structure. For example, Table 1.1 consists of r rows and c columns; the observation in the ith row and the jth column is denoted by X_{ij}. As noted earlier, when i or j or both are replaced by a dot $(\cdot)$, the dot denotes the sum over the replaced subscript. Therefore,

$$X_{\cdot\cdot} = \sum_{j=1}^{c} \sum_{i=1}^{r} X_{ij} = G$$

is the grand total, and the grand mean is given by

$$\overline{X} = X_{\cdot\cdot}/rc = \widehat{\mu}.$$

It follows that (see Table 1.1) $X_{i\cdot}$ and $X_{\cdot j}$ are the sums of the ith row and jth column, respectively. Thus, $\overline{X}_{i\cdot} = X_{i\cdot}/c$, and $\overline{X}_{\cdot j} = X_{\cdot j}/r$. This system is easily extended to a more complex data structure such as $X_{ijk\ldots p}$, where the number of subscripts corresponds to the number of classifications. If n_{ijk} denotes the number of observations in the (i, j)th cell, the dot notation representation is similarly used. For

Table 1.1 An $r \times c$ Table with One Observation per Cell

	Columns					
Rows	**1**	**2**	**...**	**c**	**Row sums**	**Row means**
1	X_{11}	X_{12}	...	X_{1j}	$X_{1\cdot}$	$\overline{X}_{1\cdot}$
2	X_{21}	X_{22}	...	X_{2j}	$X_{2\cdot}$	$\overline{X}_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	X_{i1}	X_{i2}	...	X_{ij}	$X_{i\cdot}$	$\overline{X}_{i\cdot}$
Column sums	$X_{\cdot 1}$	$X_{\cdot 2}$	...	$X_{\cdot c}$	$X_{\cdot\cdot}$ (grand total) $= G$	
Column means	$\overline{X}_{\cdot 1}$	$\overline{X}_{\cdot 2}$	...	$\overline{X}_{\cdot c}$		$\overline{X}$ (grand mean)

example, $n_{...}$ denotes the total number of observations in the experiment; $n_{i..}$ denotes the number of observations summed over the j and k classifications.

The Normal Distribution

Researchers design and conduct experiments to observe one or more quantities whose numerical values cannot be predicted with certainty in advance. The quantities of interest may be the shelf life or the preference score of one item, or some sensory characteristics of, say, food products. The shelf life, the preference score, etc., of an item is not known in advance but could be any one of a number of possible values. A quantity that can assume any one numerical value from a range of possible values, depending on the experimental outcomes, is called a *random variable* (r.v.) in statistical terminology. Since the numerical values of an r.v. depend on the experimental outcomes, the values of an r.v. are governed by a chance model or a probability distribution, as it is called in statistical language. In short, associated with each r.v. is a probability distribution, and statisticians use this structure to study various aspects of the quantities (r.v.'s) of interest.

Some probability distributions can be described by close mathematical functions called the *probability density functions* (pdf's). Once the pdf of an r.v. X is given, it can be used to calculate probabilities of events of interest in decision making. In what follows now we introduce one important probability distribution through its pdf and a number of others without writing down their pdf's.

The most important probability distribution in applied statistics is the *normal* or *Gaussian probability distribution*. An r.v. X is said to have a normal probability distribution if its pdf is of the mathematical form

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right], \tag{1.1-5}$$

where $-\infty < X < \infty$, $-\infty < \mu < \infty$, and $\sigma^2 > 0$. The graph of $f(X)$ is sketched in Figure 1.1. The graph is a bell-shaped curve that is symmetric around μ. Hence, the parameter μ is the center of the distribution of the values of X. The shape of the curve depends on σ^2. The curve tends to be flat in shape as the shape parameter σ^2 increases. The normal probability distribution is characterized completely by