*3.4 NOTCHED BOX PLOTS

A pair of box plots for the logarithms of the rainfall data are shown in Figure 3.8. The two box plots depict the data from the control clouds and the seeded clouds, and an important question arises: should the difference in the locations of the distributions — as measured, for example, by the medians — be regarded as meaningful or simply as the result of random fluctuations that occur in data? Put another way, if we were to collect data from another similar set of seeded and control clouds, would we expect the medians to be as different as they were in this experiment?

The box plots as drawn in Figure 3.8 shed no light on this issue, since the answer depends critically on the number of observations that go into making each box, and the number of observations is not depicted. A box plot based on 10 observations would be rather variable from one experiment to another, but one based on 10,000 observations would probably be very stable.

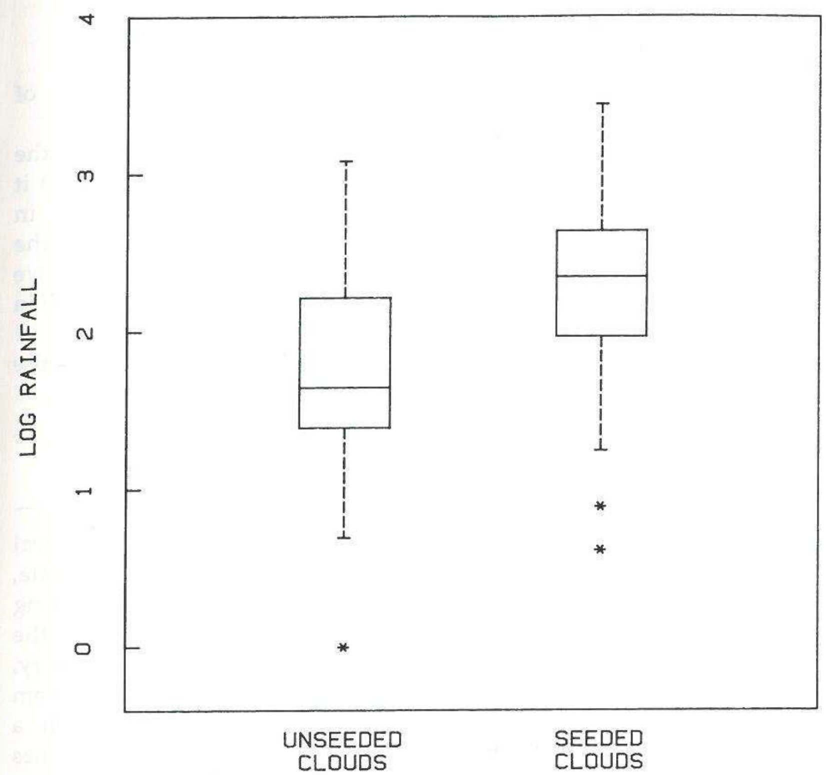

Figure 3.8 Box plots of the logarithms of the rainfall measurements from the unseeded and seeded clouds.

This was not a serious concern in Figure 3.6 because of the smooth time trend in those data. The vertical location of each box plot in Figure 3.6 is reinforced by the location of neighboring box plots and of box plots for the same month in other years; the smooth cyclical trend in that figure is unmistakable.

In applications where comparing locations is important, such as in the cloud-seeding experiment, box plots can be drawn with notches in their sides to help guide our assessment of relative locations. Figure 3.9 shows notched box plots (McGill, Tukey, and Larsen, 1978) for the log rainfall data, with the tops and bottoms of the notches computed from the formula:

$$M \pm 1.57 \times IQR/\sqrt{n}.$$

Here M, IQR and n are the median, interquartile range, and number of observations, respectively, for each subset of the data in turn.

A suitable informal interpretation of the plot with notches is the following: if the notches for two boxes do not overlap, we can regard it as strong evidence that a similar difference in levels would be seen in other sets of data collected under similar circumstances. Since the notches for the two sets of rainfall data do not overlap in Figure 3.9, we have strong evidence that cloud seeding has increased the median rainfall.

In cases where there are three or more box plots, we can use the notches to make a similar judgment about each pair of data sets.

NOTCHES AND HYPOTHESIS TESTING

Although we regard box plots — as well as most of the graphical methods in this book — as informal tools for looking hard at data, readers familiar with classical statistical theory may find it interesting background to know that the formula for notch lengths is based on the formal concept of a hypothesis test. In the language of statistical theory, if the two data sets are independent and identically distributed random samples from two populations with unknown medians but with a normal distributional shape in the central portion, then the notches provide an approximate 95% test of the null hypothesis that the true medians are equal: if the two notches overlap, then we fail to reject the null hypothesis with (approximate) 95% confidence. Alternatively, the difference between the medians could be described as "statistically significant at the .05 level."

We mention this not to encourage readers to regard notches on box plots as a way to make formal hypothesis tests, but to show that statistical theory can help in the design of informal tools for data analysis. The fact is that notches on box plots are useful guides for comparing median levels even when the requirements for the hypothesis test are not strictly met — which is very frequently the case!

When there are more than two sets of data each pair of notched box plots provides a test of significance for the difference of the medians of the two corresponding distributions. But the notches are not adjusted in any way to take into account that several hypothesis tests are being carried out simultaneously. (If there are p data sets then there are $p(p-1)/2$ pairs of medians being compared by notches.) It should be kept firmly in mind in interpreting the plot that even under the null hypothesis that all of the p medians are equal, the probability that at

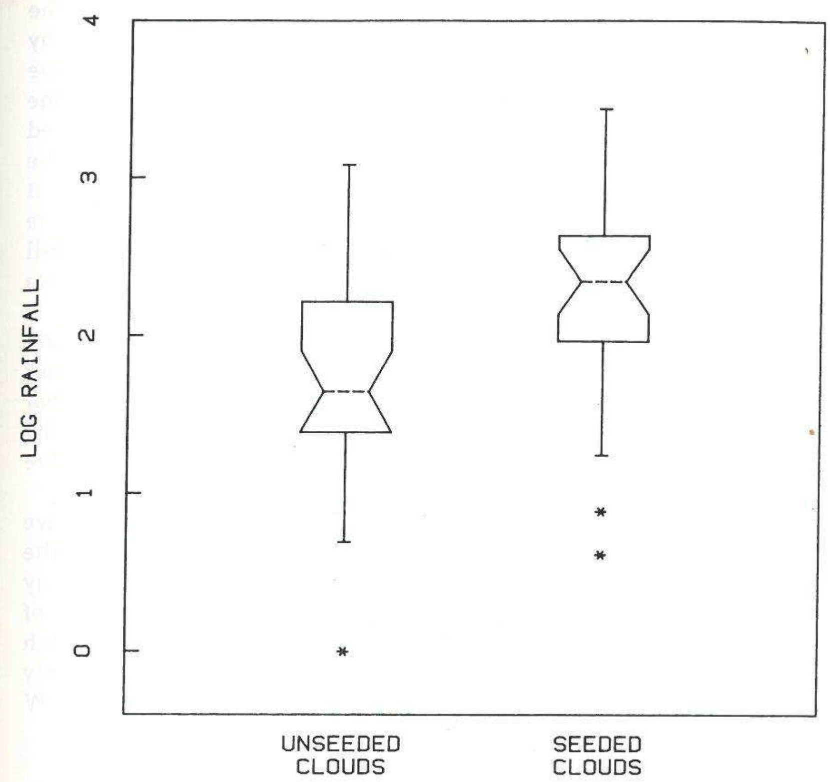

Figure 3.9 Notched box plots of the logarithms of the rainfall measurements from the unseeded and seeded clouds.

least one of the pairs of notches do not overlap will be greater than .05. This is the so-called "multiple comparisons" problem. Technical adjustments are possible, but generally unnecessary, as long as the notched box plots are used in the informal way discussed above.

*3.5 MULTIPLE DENSITY TRACES

If we choose to use density traces to depict data distributions, we can compare two or more data sets by plotting their density traces in the

The Wadsworth Statistics/Probability Series

Series Editors

Peter J. Bickel, University of California
William S. Cleveland, Bell Laboratories
Richard M. Dudley, Massachusetts Institute of Technology

Bickel, P.; Doksum, K.; and Hodges, J.L., Jr.
Festschrift for Erich L. Lehmann.

Chambers, J.; Cleveland, W.; Kleiner, B.; and Tukey, P.
Graphical Methods for Data Analysis.

Graybill, F.
Matrices with Applications in Statistics, Second Edition.

GRAPHICAL METHODS FOR DATA ANALYSIS

John M. Chambers
William S. Cleveland
Beat Kleiner
Paul A. Tukey

Bell Laboratories

WADSWORTH INTERNATIONAL GROUP
Belmont, California

DUXBURY PRESS
Boston

This book is a co-publishing project of Wadsworth International Group and Duxbury Press, divisions of Wadsworth, Inc.

Library of Congress Cataloging in Publication Data

Main entry under title:

Graphical methods for data analysis.

(The Wadsworth statistics/probability series)
Bibliography: p.
Includes index.
1. Statistics—Graphic methods—Congresses.
2. Computer graphics—Congresses. I. Chambers, John M.
II. Series.
QA276.3.G73 1983 001.4'22 83-3660
ISBN 0-534-98052-X

Printed in the United States of America

87 86 85 — 3 4 5 6 7 8 9

ISBN 0-87150-413-8

ISBN 0-534-98052-X INTERNATIONAL

The text for the book was processed at Bell Laboratories on a VAX 11/780 computer running the UNIX* operating system and troff/mm software driving an APS-5 phototypesetter. The font is Palatino. Covers and text printing and binding by The Maple-Vail Book Manufacturing Group.

*UNIX is a trademark of Bell Laboratories, Inc.