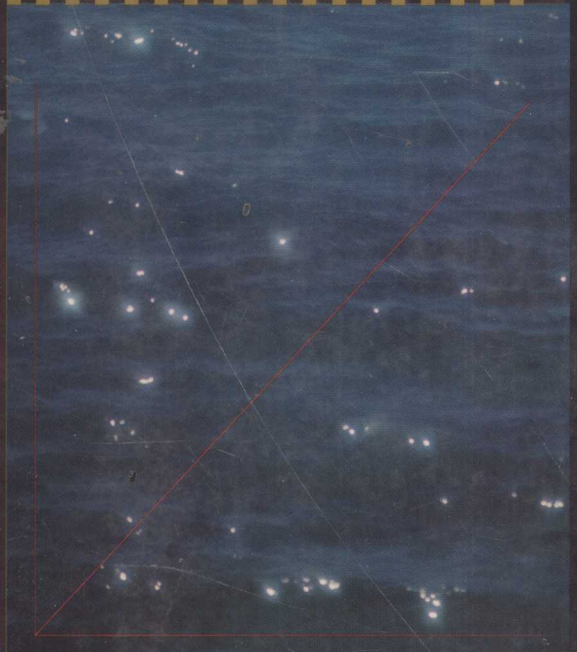
StatConcepts

A VISUAL TOUR OF STATISTICAL IDEAS



H. Joseph Newton

Jane L. Harvill

StataQuest 4 for Windows included

StatConcepts: A Visual Tour of Statistical Ideas

H. JOSEPH NEWTON
Texas A&M University

JANE L. HARVILL
Bowling Green State University

Duxbury Press
An Imprint of Brooks/Cole Publishing Company
I(T)P® An International Thomson Publishing Company

Pacific Grove, CA • Albany, NY • Bonn • Boston • Cincinnati • Detroit • Johannesburg • London
Madrid • Melbourne • Mexico City • New York • Paris • Singapore • Tokyo • Toronto • Washington

Editor: *Curt Hinrichs*
Assistant Editor: *Cynthia Mazow*
Editorial Assistant: *Rita Jaramillo*
Project Editor: *Sandra Craig*
Print Buyer: *Stacey Weinberger*
Permissions Editor: *Peggy Meehan*
Copy Editor: *Thomas L. Briggs*
Cover: *Stuart D. Paterson, Image House Inc.*
Signing Representative: *Ragu Raghavan*
Compositor: *SuperScript*
Printer: *Malloy Lithographing*

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

For more information, contact Duxbury Press at Brooks/Cole Publishing Company, 511 Forest Lodge Road, Pacific Grove, CA 93950, or electronically at http://www.thomson.com/duxbury.html

International Thomson Publishing Europe
Berkshire House 168–173
High Holborn
London, WC1V7AA, England

Thomas Nelson Australia
102 Dodds Street
South Melbourne 3205
Victoria, Australia

Nelson Canada
1120 Birchmount Road
Scarborough, Ontario
Canada M1K 5G4

International Thomson Publishing GmbH
Königswinterer Strasse 418
53227 Bonn, Germany

International Thomson Editores
Campos Eliseos 385, Piso 7
Col. Polanco
11560 México D.F. México

International Thomson Publishing Asia
221 Henderson Road
#05-10 Henderson Building
Singapore 0315

International Thomson Publishing Japan
Hirakawacho Kyowa Building, 3F
2-2-1 Hirakawacho
Chiyoda-hu, Tokyo 102, Japan

International Thomson Publishing Southern Africa
Building 18, Constantia Park
240 Old Pretoria Road
Halfway House, 1685 South Africa

Library of Congress Cataloging-in-Publication Data

Newton, H. Joseph, 1949–
StatConcepts : a visual tour of statistical ideas / H. Joseph Newton, Jane L. Harvill.
p. cm.
Includes index.
ISBN 0-534-26552-9
1. Mathematical statistics—Computer-assisted instruction
I. Harvill, Jane L. II. Title.
QA276.18.N48 1997
519.5'078'553682—dc21 97-1127

To Karah, Tim, Jake, Joseph, Katherine, and Cary

Introduction

Most introductory statistics courses have three parts: (1) *descriptive statistics*, which uses numbers and graphs to summarize the information about a data set, (2) *inferential statistics*, which draws conclusions about numerical characteristics of entire populations of objects from those of samples from the populations, and (3) *statistical concepts*, which are the basic logical and mathematical ideas underpinning descriptive and inferential statistics.

A wide variety of computer programs make it easy for students to accomplish what is required for the first two of these parts, but very little software has been developed for illustrating statistical concepts. That's why we wrote StatConcepts—as a set of "laboratories" for illustrating these concepts.

StatConcepts is actually a collection of programs written in the language of StataQuest, a student version of a program called Stata that is designed to do descriptive and inferential statistics.

StatConcepts is not intended as a text, but as a supplement to introductory statistics texts. Its main focus is on correct intrepretation and understanding of statistical concepts, terminology, and results, not on computation for a given problem. However, StatConcepts does contain some labs that allow students to compute results.

In many ways, the computer is the laboratory for the science of statistics. Most statistical investigations have their roots in a statement along these lines: "If we did this procedure over and over again, then this is what we would see." The only way realistically to do things over and over again is on a computer. In these labs, we have tried to use graphics to show what, in fact, we would see if we did various things over and over again.

We assume that instructors will not incorporate all of the labs in the StatConcepts collection (there are 28 of them!) into their courses, but rather pick and choose those they feel would be most useful in the course (and have time to cover in their already cramped schedule).

We hope that instructors can show the labs to students using some kind of projection, but each chapter of this book contains a "guided tour" through each lab that a student could read while at a computer. These guided tours cannot totally replace

an instructor, but they can certainly help instructors use the labs as a supplement to their courses.

Although the labs and this book are intended primarily for introductory courses, we have found them very valuable in courses at all levels. We have kept the material as nontechnical as possible, but more advanced students will be able to relate to the graphs and descriptions at a more mathematical level.

Computer Requirements for Using StatConcepts

From a software point of view, StatConcepts is totally self-contained and requires only a computer running Microsoft Windows.

The Structure of the Chapters

Except for the introductory lab and the *Calculating Confidence Intervals*, *Calculating Tests of Hypotheses*, and *Calculating One-way ANOVA* Labs (which are more calculation than concepts oriented), each chapter in this book is structured in the same way:

1 *Introduction:* The first section provides background information needed for the lab or labs in the chapter.

2 *Objectives:* This section briefly summarizes the concepts the lab or labs illustrate.

3 *Description:* This section briefly describes how the lab or labs work and what the items in the dialog box are.

4 *Guided Tour of the Lab:* This section is the heart of each chapter. Although the labs are best used while in front of the computer, we have included enough graphs from the labs to communicate the basic ideas by simply reading the tour. Some of the tours have many stops. Again, we have tried to design them so that readers can visit as many of the stops on a tour as they have the time or interest for.

5 *Summary:* This section summarizes what the guided tour has illustrated.

6 *Lab Exercises:* This section contains a set of exercises that can be used to further illustrate the key ideas.

Acknowledgments

We have received an amazing amount of help from a wide variety of people in creating and finishing this project. First, thanks to the people at Stata Corporation for having written such a nice piece of software, particularly the graphics and dialog boxes that make StatConcepts possible. James Hardin was particularly helpful in answering our many questions.

We also appreciate the comments of the manuscript reviewers: Robert Hale, Pennsylvania State University; Robert Heckard, Pennsylvania State University;

David C. Howell, The University of Vermont; Dennis Jowaisas, Oklahoma City University; Bill Notz, Ohio State University; and Bill Seaver, University of Tennessee, Knoxville.

We are indebted to Stan Loll, Alexander Kugushev, and Curt Hinrichs of Duxbury Press for their encouragement at the beginning of the project and to Sandra Craig, William Baxter, and Thomas Briggs for their yeoman work in the production of the book.

At least 25 different instructors here at Texas A&M have used preliminary versions of these labs in teaching our large elementary statistics courses, and they have given us many wonderful suggestions. We would like to single out Naisyin Wang, Donald Lancon, Andy Liaw, and, in particular, Julie Hagen Carroll for their special help.

Finally, we thank our families, and especially Linda and Marty, for their patience and support.

H. Joseph Newton
Department of Statistics
Texas A&M University
College Station, TX 77843-3143
jnewton@stat.tamu.edu

Jane L. Harvill
Applied Statistics and Operations Research
Bowling Green State University
Bowling Green, OH 43403-0267
jharvil@cba.bgsu.edu

As noted previously, there are 28 labs in all, although there are fewer items on the `Labs` menu because some items have submenus containing more than one lab. There is a chapter in this book for each item on the `Labs` menu, as follows:

1 *Introduction to Concept Labs:* This lab is actually just a greeting and an invitation to look at a help file giving an overview of the entire collection of labs. It also allows users to specify their own random number generator seed.

2 *Random Sampling:* This lab repeatedly shows random sampling without replacement from a population of 100 boxes. It also previews the ideas of sampling distributions and the Central Limit Theorem.

3 *Relative Frequency and Probability:* This lab again illustrates random sampling without replacement using the example of a lottery game. It also illustrates the relative frequency interpretation of probability by repeatedly drawing six winning numbers from the numbers 1–50 and keeping track of the number of draws containing at least two consecutive numbers. Deriving the formula for the probability of this event is beyond the scope of most courses.

4 *How Are Populations Distributed?* This lab shows students that distributions come in all shapes and sizes and in parametric families. The lab graphs densities from 14 different families. It also generates random samples from one member of each family and superimposes the density on the histogram of the sample, thus illustrating variability from one sample to another.

5 *Sampling From 0–1 Populations:* This item actually leads to four different labs:

   (a) *Sampling With and Without Replacement:* The binomial and hypergeometric distributions are illustrated by having the user specify the number of elements in a 0–1 population, the proportion of 1's, and the size of a sample, and then superimposing the probability plot of the number of 1's in the sample under conditions of sampling with and without replacement.

   (b) *The Negative Binomial Distribution:* This lab graphs the negative binomial distribution for user-specified values of the parameters.

   (c) *Poisson Approximation to Binomial:* This lab superimposes the binomial distribution and its Poisson approximation for user-specified values of the

parameters. It makes it easy to see when the Poisson approximation works well and when it doesn't.

(d) *Normal Approximation to Binomial:* This lab superimposes the binomial distribution and its normal approximation for user-specified values of the parameters. It makes it easy to see when the normal approximation works well and when it doesn't.

6 *Bivariate Descriptive Statistics:* This item leads to three different labs:

(a) *Scatterplots I:* This lab shows scatterplots of random samples from a bivariate normal population for 20 different values of the correlation coefficient ranging from -0.9 to 0.9.

(b) *Scatterplots II:* This lab allows the user to generate scatterplots for any sample size and any population correlation coefficient.

(c) *Least Squares:* This lab allows the user to generate a wide variety of different scatterplots and then see the true line, the least squares line, and the vertical errors that go into the residual sum of squares.

7 *Central Limit Theorem:* This lab illustrates sample means for repeated sampling from a user-specified choice of four parent populations: normal, exponential, uniform, and 0–1. This is actually two labs in one:

(a) *One-at-a-time:* One sample at a time, boxes corresponding to sample means are placed above an axis until the tallest column of boxes fills the graph.

(b) *500 Samples:* The histogram of the sample means for 500 samples is drawn with the approximating normal curve superimposed.

8 *Z, t, Chi-square, and F:* This item leads to six labs:

(a) *Critical Values:* This lab graphs rejection regions for one- and two-tailed tests for any of Z, t, χ^2, or F for user-specified α and, if necessary, degrees of freedom.

(b) *Normal Curves:* This lab starts by drawing the standard normal curve. The user can then repeatedly change the mean and/or variance, and each time the lab draws the new normal curve on the same axes.

(c) *Chi-square Curves:* This lab starts by drawing the χ^2 curve with 10 degrees of freedom. The user can then repeatedly change the degrees of freedom, and each time the lab draws the new χ^2 curve on the same axes.

(d) *F Curves:* This lab starts by drawing the F curve with 10 and 10 degrees of freedom. The user can then repeatedly change the degrees of freedom, and each time the lab draws the new F curve on the same axes.

(e) *t Converging to Z:* This lab allows the user to superimpose any part of the Z curve and the same part of the t curve for increasing degrees of freedom.

(f) *Normal Approximation to Binomial:* This is the same lab as in the *Sampling From 0–1 Populations* item.

9 *Sampling Distributions:* This lab allows the user to generate 500 samples (or pairs of samples) of user-specified size from one of three parent populations: normal, uniform, and exponential. The user can then calculate the Z, one- or two-sample t, χ^2, or F statistics and superimpose the histogram of the 500 statistics and the

theoretical normal theory curve. The lab also displays the percentiles of the 500 statistics and the theoretical curve to see the agreement (disagreement) of the two if assumptions are (are not) met.

10 *Minimum Variance Estimation:* This lab allows the user to generate 500 samples of user-specified size from one of four parent populations (Normal(0,1), Uniform(−0.5, 0.5), t with 3 degrees of freedom, and Laplace), each symmetrical about zero, and then draw the histograms of the 500 sample means and 500 sample medians. The sample mean and standard deviations of the 500 means and 500 medians are also displayed. The lab shows that the sample mean is not always the best estimator.

11 *Interpreting Confidence Intervals:* This lab allows the user to generate 50, 100, or 150 samples (or pairs of samples) of user-specified size from one of four parent populations (normal, uniform, exponential, and 0–1) and draw horizontal lines for the confidence intervals (for user-specified α) for user-specified parameters (μ, σ^2, $\mu_1 - \mu_2$, σ_1^2/σ_2^2, π, or $\pi_1 - \pi_2$), as well as a vertical line representing the true value of the parameter. This allows the user to see the effect of changing α and sample size on the width of intervals, as well as the effect of violating assumptions on the confidence interval coverage probability.

12 *Calculating Confidence Intervals:* This lab allows the user to calculate confidence intervals for the 11 different one- and two-sample inference situations for means, variances, and proportion problems usually covered in an introductory course.

13 *Tests of Significance:* This lab draws a graph of a user-specified Z, t, χ^2, or F curve and shades in the area corresponding to the p value for either a user-specified or lab-generated value of a test statistic.

14 *Level of Significance of a Test:* This lab allows the user to generate 500 samples (or pairs of samples) of user-specified size from one of three parent populations (normal, uniform, and exponential); calculate the Z, one- or two-sample t, χ^2, or F statistics; and then superimpose the histogram of the 500 statistics and the theoretical normal theory curve. It shades in the area under the curve for the rejection region (for the user-specified α and one- or two-tailed test) and displays the proportion of times the null hypothesis is rejected, thus showing the agreement (disagreement) with the value of α if assumptions are (are not) met.

15 *Calculating Tests of Hypotheses:* This lab allows the user to calculate test statistics and p values for the same 11 situations as in the *Calculating Confidence Intervals* Lab. It draws a graph with the test statistic marked and the tail areas corresponding to the p value shaded in.

16 *Power of a Test:* This lab is the same as the *Level of Significance of a Test* Lab except that now the user can specify the degree to which the null hypothesis is actually false (including actually being true). This allows the user to see the effect of sample size and degree of falseness on the power of the test.

17 *Calculating One-way ANOVA:* This lab allows the user to enter sample sizes, means, and variances (or standard deviations) for specified number of samples and then displays the ANOVA table.

18 *Between and Within Variation:* This lab starts by generating and graphing two sets of four samples, each of user-specified size. In the first set of four samples, the populations have differing means; in the second set, the population means are all the same. All the populations have the same variance. Then the lab allows the user to repeatedly change the population variance, each time redrawing the plot and displaying the p value of the one-way ANOVA test of equality of means. This allows the user to see how the test is comparing the between-sample variability to the within-sample variability.

19 *Chi-square Goodness of Fit:* This lab illustrates the χ^2 goodness-of-fit test by generating a user-specified number of points in a square and then placing a grid of (user-specified number of) boxes on the points, counting how many are in each box, and then displaying the p value of the resulting χ^2 test. This allows users to see how nonuniform the placement of points can look even though they are, in fact, being placed according to a uniform distribution.

BRIEF CONTENTS