Basic Business Statistics

A Casebook

Springer Science+Business Media, LLC

Basic Business Statistics

A Casebook

Dean P. Foster
Robert A. Stine
Richard P. Waterman
University of Pennsylvania
Wharton School

Springer

Dean P. Foster
Robert A. Stine
Richard P. Waterman
Department of Statistics
Wharton School
University of Pennsylvania
Philadelphia, PA 19104
USA

Library of Congress Cataloging-in-Publication Data

Foster, Dean P.
Basic business statistics: a casebook / Dean P. Foster, Robert A. Stine, Richard P. Waterman.
p. cm.
Includes index.
ISBN 978-0-387-98354-7 ISBN 978-1-4612-1696-4 (eBook)
DOI 10.1007/978-1-4612-1696-4
1. Commercial statistics. 2. Commercial statistics—Case studies. I. Stine, Robert A. II. Waterman, Richard P. III. Title.
HF1017.F67 1998
519.5–dc21 97-39225

ISBN 978-0-387-98354-7 Printed on acid-free paper.

Production managed by Timothy Taylor; manufacturing supervised by Erica Bresler.
Camera-ready copy prepared from the authors' WordPerfect files.

springeronline.com

Preface

Statistics is seldom the most eagerly anticipated course of a business student. It typically has the reputation of being a boring, complicated, and confusing mix of mathematical formulas and computers. Our goal in writing this casebook and the companion volume (*Business Analysis Using Regression*) was to change that impression by showing how statistics yields insights and answers interesting business questions. Rather than dwell on underlying formulas, we show how to use statistics to answer questions. Each case study begins with a business question and concludes with an answer to that question. Formulas appear only as needed to address the questions, and we focus on the insights into the problem provided by the mathematics. The mathematics serves a purpose.

The material in this casebook is organized into 11 "classes" of related case studies that develop a single, key idea of statistics. The analysis of data using statistics is seldom very straightforward, and each analysis has many nuances. Part of the appeal of statistics is this richness, this blending of substantive theories and mathematics. For newcomers, however, this blend is too rich, and they are easily overwhelmed and unable to sort out the important ideas from nuances. Although later cases in these notes suggest this complexity, we do not begin that way. Each class has one main idea, something big such as standard error. We begin a class by discussing an application chosen to motivate this key concept, and introduce the necessary terminology. All of the cases in that class explore and develop that idea, with perhaps a suggestion of the idea waiting in the next class. Time in the classroom is a limited commodity, and we use it to apply statistics rather than talk about statistics. We do the data analysis of these cases in class using a computer projection system. This allows us to explore tangents and gives students a chance to see the flow of data analysis. We use a supplemental textbook to fill voids in our coverage and to complete the development of underlying calculations and formulas. These casebooks remove much of the note-taking burden so that students can follow along without trying to draw plots or scribble down tables. That said, we have seen that students still seem to fill every margin with notes from class. The course seems to work. It has been very highly rated by MBA students, and some have even come to ask about majoring in statistics!

It would be easy to claim that you can use any statistics software with these casebooks, but that's not entirely true. Before we developed these notes, we needed to choose the software that we would use to do the analyses and generate the figures that appear in this book. The preferences of our students demanded that we needed something that ran on PCs and Macs; our own needs required that the software be able to handle large data sets, offer modern interactive graphics (plot linking, brushing), and include tools for the beginner up through logistic regression. JMP (whose student version is named JMP-IN) was the best fit to our criteria. Each of its analysis procedures includes a graph as part of the output, and even the student version allows an unlimited data set size (up to the capabilities of the computer). It's also very fast and was designed from the start with a graphical user interface. To top it off, the developers of JMP, led by John Sall at SAS, have been very responsive to our suggestions and have added crucial features that we needed. We're aware, though, that many will prefer to use another package, particularly Minitab. Minitab finished a close second to JMP in our comparisons of statistical software and is well-suited to accompany these notes. An appendix describes the commands from the student version of Minitab that provide the needed functionality.

A mixture of real and simulated data are used in the cases. We often use simulated or artificial data when we want to focus on a particular technique or issue. Real data are seldom so simple and typically have numerous nuances that distract from the point being made. The resulting analysis can become so convoluted that everyone tends to get lost. The data for the included assignments were used recently when we taught this course (we have to change the assignments annually). You may want to require these or modify them to suit your emphasis. All of the data sets used in these examples (in both JMP and ASCII formats) are available over the Internet from either StatLib (http://lib.stat.cmu.edu) or our departmental web site (http://www-stat.wharton.upenn.edu).

A Brief Overview of the Cases

We have used the material in this book as the lecture notes for an intensive 3-week "pre-term" course in the MBA program at Wharton. This course is taken by virtually all incoming MBA students prior to the start of the usual semester. In many ways, the course is intended to be a leveling experience, bringing students to a common level of preparation so that all are prepared to begin the fall semester with regression analysis. The pace (3 or 4 classes a week) of the Wharton pre-term and the fact that this course is not graded require that we make the material as interesting and engaging as possible. The first 7 lectures of our course are each 2 hours long, and the remaining 4 last 1.5 hours. Generally, though, there is more material in these notes than can be adequately covered in even the 2-hour classes. At a minimum, you will want to cover the first example for each class since this example generally introduces the methods with more discussion; later examples for each class offer repetition and explore tangential (important, but still tangential) ideas. The material is inherently cumulative, though our timing for a discussion of the issues in sampling can be altered. We have found it better to cover sampling once we have introduced confidence intervals and tests so that students can appreciate the effects of sampling on these methods. The final two lectures move toward regression and can be omitted if one is pressed for time. The data in these final two lectures are, perhaps, the most interesting in the book (at least for those interested in finance).

The remainder of this preface discusses the material in the classes, with particular emphasis on the points that we try to make in each. We typically begin each class by reviewing the overview material that introduces the lecture, covering the various concepts and terminology. This "blackboard time" lays out the key ideas for the class so that students have a road map for the day. Without this introduction, we have found that some will lose sight of the concepts while focusing instead on the use of the software. Once we have the ideas down, we turn to the examples for the day, emphasizing the use of statistics to answer important business questions. Each example begins with a question, offers some relevant data, and applies statistics to answer the question.

Class 1

We use our first class as an overview and as an opportunity to set expectations. Many students have not had a good experience with statistics as undergraduates, and we find it useful to point out how this course is different from those they probably took as undergraduates. This class is also a good chance to suggest how statistics appears in their other courses, most likely finance and marketing in the MBA program.

An activity we have successfully used in this first class is an exercise in counting the number of chips in Chips Ahoy! brand chocolate chip cookies. The particular appeal of this exercise is the claim made on the package: "1,000 chips in every bag!" We have the students try to see if this claim is true. We begin by letting the students group themselves into teams of 6 to 10 members, and then give each team a sealed plastic "pouch" of cookies (about 25 cookies). Each of the packages that display the advertising claim holds two of these pouches. We leave it to the students to figure out how to verify the claim, dealing with questions such as "How many cookies should we use?" and "What's a chip?" Some have figured out that a good way to count the chips is to dissolve the cookie in water (though it can get really messy). The dough dissolves, leaving whole chips behind. Even if the students do not see to do this, the experiment works very well and makes for an intriguing, hands-on first day of class. We have also done this experiment with students working alone, but it seems to run better with teams. At the end of this exercise, we have the students report their data and we enter the data from several groups into a JMP spreadsheet with a team name and count data for, say, 10 cookies from each group. As time allows, we plot this data set to show the variation and later return to it in Class 3.

During this exercise, our focus is on variation. We want the students to recognize the presence of natural variation in the cookie-making process. This leads to some ideas from quality control. For example, it would be very expensive to have someone count out 20 chips for each cookie, and then make sure that each package had 50 cookies (for a total of exactly 1000 chips in each bag). The mass production method approximates this ideal, but in a more cost effective manner.

Class 2

This class introduces the key numerical and graphical measures of variability. As the list of topics suggests, there's a lot in this class. Students who have not seen histograms will have a hard time keeping up. However, since we do not dwell on the calculation of these summaries, but rather on their interpretation, they seem to be able to hang in there with the others. It's easy to get bogged down trying to explain the details of how to draw a boxplot, for example. We try to avoid these and keep the discussion focused on the relative merits of numerical and graphical summaries.

Some underlying themes that need to be stressed are issues of robustness (e.g., mean versus median in the presence of outliers), the use of transformations to simplify problems, and the fact that graphs are more informative than numerical summaries. Many students will not have seen a kernel density estimate, so this is a good opportunity to introduce some of the more recent developments in statistics that involve nonparametric smoothing. The availability of a slider makes this a nice showcase for what you can do with interactive software as well. Smoothing reappears in the sequel in the context of smoothing scatterplots to identify nonlinear patterns in regression analysis.

Another underlying theme of this class is the role of assumptions. For our purposes, assumptions enter indirectly (as in the choice of bin size in a histogram) or more explicitly (as in the use of the normal distribution/empirical rule). In either case, we try to comment in class on the importance of checking such assumptions via graphical diagnostics. For example, JMP's "hand tool" provides the ability to do an animated sensitivity analysis of the choices made when constructing a histogram. Though less interactive, the normal quantile plot is a graphical diagnostic for normality.

Supplemental material in this class describes graphically how one constructs a kernel density estimate since this procedure has not yet made its way into many introductory texts.

Class 3

This class has three objectives

- Use the graphical tools introduced in Class 2, giving students a second opportunity to become familiar with these techniques;
- Introduce the idea of explaining variation (albeit graphically rather than via an explicit model); and
- Do a bit of probability and quality control, which are topics covered more thoroughly in Class 4.

The quality control case is very important, and this example is the foundation for our development of standard error. The data for these shafts come from a colleague who has done considerable consulting on quality control for a major domestic auto manufacturer. We also make considerable use of plot linking in this lecture, particularly when selecting categories from histograms.

We have found that the class discussion often returns to the "cookie example" of the first day. It has an inherent grouping by the teams that collected the data. The groups may have comparable mean values, but very different variances, conveying the important message that collections of data can differ by more than their mean value (a lesson often lost by the time ANOVA comes along in subsequent courses).

Class 4

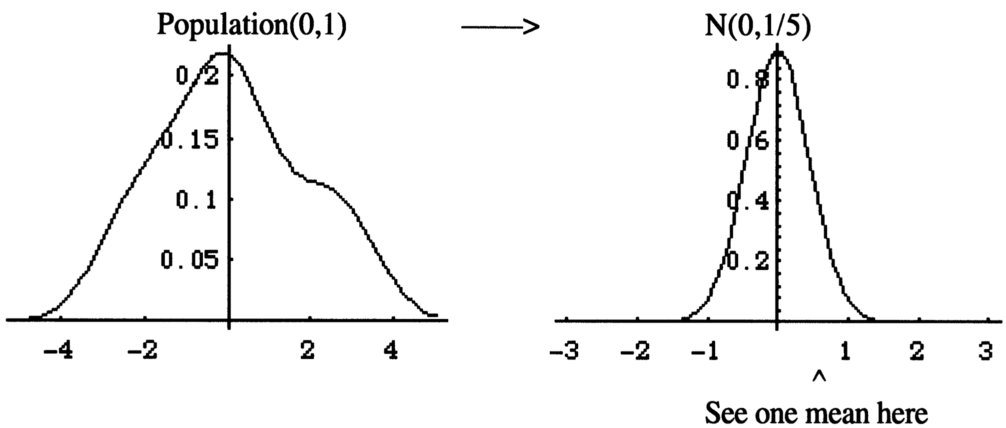
This is perhaps the most important, yet most difficult class. The difficulty arises because this class introduces standard error (SE). Standard error is hard to communicate without resorting to mathematics because it measures variation that is typically never seen, providing an estimate of the sampling variation without requiring repeated sampling.

Our method for motivating sampling variation is to use the repeated sampling associated with quality control. Rather than introduce artificially repeated samples (leading to the question "Should I take several samples?"), it makes sense in quality control that one would observe many samples. Also, the need for SE is apparent: you need SE to set limits for monitoring a process. A supplement to the first case describes why it is beneficial to look at averages rather than individual measurements.

Regarding the limits for the SD of a process, we attempt to dodge the question of how these are determined, though such questions inevitably arise in class. The JMP manual has a brief description of the procedure.

Class 5

The examples for this class are shorter than those used in the previous classes. Generally, we find that we need more "blackboard time" at this point to develop standard error. Plots such as those on the next page suggest the ideas better than more mathematics. On the left we see the population (draw something that is not normal for the population, to make the point that the CLT implies that we don't need to assume normality.) On the right, we see what happens for means of samples of 5 from this population. The means are much less variable, packed in more about μ, and their distribution is pretty close to normal. The trick, or magic, of standard error is that given one sample, we can describe how close we expect that sample mean to come to μ.



We have also chosen to emphasize CIs at the expense of hypothesis testing. A confidence interval is a "positive" statement about the set of plausible values for the parameter μ, whereas a test is a "negative" statement, indicating whether a value is plausible. The supplement to the first example attempts to communicate the message that tests are closely tied to confidence intervals – you reject H_0 for values outside the confidence interval. Since subsequent output will often only show the t statistic, it's good to introduce the equivalence early. We return to hypothesis tests and this correspondence in Class 7.

Class 6

One class cannot do justice to survey sampling, so we have chosen to focus on several problems that can occur. Some are obvious, such as self-selection and poor question formulation. Others are less obvious, such as the bias arising from length-biased sampling illustrated in the hotel satisfaction example of this section. We generally supplement this class with various stories from the current press. For example, students will have ample questions about polls near the time of national or important local elections. We also try to find stories from the recent press (such as clippings from the *New York Times* or *Wall Street Journal*) that illustrate the use/misuse of a survey.

This class uses two examples, one with data that you can let students find on the Internet and a second data set that we constructed to parallel a study done by a commercial hotel chain that was seriously flawed.

Class 7

This class presents two-sample comparisons, featuring both hypothesis tests and confidence intervals. In keeping with our preference for intervals over tests, we generally stick to two-sided intervals and the associated p-values. We argue that statistical tests are not terribly meaningful without a status quo that defines a null hypotheses; the examples are a bit contrived to feature such a baseline.

The first example defines the key methods. The second example, using waiting times, discusses a tempting mistake. Confidence intervals for each mean are appealing, but can be confusing when thinking about differences of means. It is tempting to work with the simpler one-sample confidence intervals and let the overlap of intervals decide comparisons. This procedure can be very conservative, and the second case shows how.

A limitation of these examples is that they only begin to address the deeper issue of decision analysis. Each decision has a cost: the cost of keeping a poor current method versus, for example, the potential gain of switching to a newer technology. Without some discussion of these issues, hypothesis testing and the associated statistics are answering the wrong question. Unfortunately, we do not have the chance to delve into more realistic decision-making which would require a more complete assessment of the costs of the alternative decisions.

The final example is supplemental; it considers the impact of non-normality upon tests and tries to implant the idea that comparisons based on means are not always the right method (in fact, they are seldom the "best" method). We generally mention the Van der Waerden method, but lack the time to discuss it more fully, appealing to the CLT. When we deliver this material at Wharton, our time in class slips from 2 hours down to 1.5 hours at this lecture. Thus, you are likely to find that you have less material for this and subsequent classes.

Class 8

Aside from the obvious paired t-test, this class introduces both dependence and correlation as well as experimental design. Dependence is the subject of Classes 10 and 11, as well as most of regression analysis. Experimental design returns in our sequel when we consider methods for the analysis of variance. Time permitting, we have on occasion done the "Coke vs. Pepsi challenge" with this lecture. It works best if you can manage to do both two-sample and paired comparisons, but that's hard to manage. The data set for the second example is constructed to resemble some marketing data received from a drug firm that was analyzing the performance of its sales field force.

Unfortunately, JMP can get in the way of the success of this class. When trying to compare a paired test to the corresponding two-sample test, the manipulations of the data (dividing the rows of a single column into several) can confuse students. If possible, have students learn how to "split" a JMP data sheet at some point prior to this class. An exercise would likely do the trick. Otherwise, the substantive content of the class (design, dependence) can be obscured by concerns over the procedural aspects of the *Tables* command.

Class 9

This class introduces the issue of confounding in a study: the groups that you want to compare are different in other aspects aside from the factor that determines group membership.

This class, with one example, provides a nice opportunity to apply some of the concepts from previous classes and review difficult ideas, such as the meaning of a p-value. The data used in this example are patterned on a large study of salary equity done at the University of Pennsylvania several years ago. Obviously, we have not used that data set, but rather constructed one with similar patterns and issues.

This data set is also used in the sequel. When we return to this problem, we show how regression analysis (rather than conditioning via subsetting) brings all of the data to bear on the question at hand, rather than just a subset.

Class 10

This class is one of the more challenging yet interesting since it mixes quite a bit of finance with the introduction of covariance. Going through this case carefully, elaborating on the needed finance, can easily consume two lecture periods. Covariance is the fundamental statistical component used in the formation of portfolios, so this example will get those who are interested in finance very engaged in the classroom discussion. Unfortunately, we also have students who, though interested in business, are not familiar with terms such as "short selling" a stock. They'll find this class more challenging since they will be trying to learn two things at once: elementary finance as well as covariance.

We have found that scatterplot matrices can consume quite a bit of classroom discussion, even though our use of them here is quite limited – we treat them as visual counterparts of a correlation matrix. It might be useful to have students work with these as part of a prior assignment so that this plot will be familiar to them.

Finally, the added note discussing the covariance of pairs of means may be of interest to those who are following the underlying details. It explains why the paired t-test works better than the two-sample t-test when the samples are correlated. The note is also important because it implies that, like data, statistics (here two sample means) can be correlated. Correlation is not just a property of data, it's also a property of the things that we compute from data. Students will need some sense of this idea to appreciate collinearity's impact on slope estimates in regression analysis.

Class 11

This class both reviews the material of the first 10 lectures as well as advertises what is to come with regression analysis. The data set is inherently interesting to explore graphically, identifying the mutual funds by name. Some of the big outliers have rather distinctive names (such as funds investing in gold) that can lead to considerable discussion in class. This case also makes a very subtle point about covariance and independence. The sample sizes for this example are quite large, some 1533 different mutual funds. These large counts lead to spuriously precise claims about the year-to-year correlation in mutual fund returns and an apparent paradox. For example, the returns in 1992 and 1993 are negatively correlated (significantly so from the regression output), whereas those for 1992 and 1991 are significantly negatively correlated.

A particularly simple explanation for this flip-flop is that the 1533 mutual funds are not independent observations. All of them invest in the same financial market and suffer common rises and falls. The 1533 funds are "worth" quite a bit fewer independent observations, to the extent that the observed changes in correlation are not surprising. Large data sets do not always imply statistical significance when correlation is present among the observations. This sort of "hidden" correlation provides fair warning to students that independence is both important and yet hard to verify. Unlike the assumption of normality or common variance across groups, there is no simple graphical tool for judging the independence of the mutual funds.

Acknowledgments

We have benefited from the help of many colleagues in preparing this material. Two have been very involved and we want to thank them here. Paul Shaman kept us honest, shared his remarkable editorial skills, and as the chairman of our department provided valuable resources. Dave Hildebrand offered numerous suggestions, and is the source of the data for many of our examples, including the car seam and computer chip data in Class 4 and the primer and food processing data in Class 7. We thank him for his generosity and encouragement along the way.

Dean P. Foster
Robert A. Stine
Richard P. Waterman

Department of Statistics
Wharton School
University of Pennsylvania
Philadelphia, PA

LECTURE TEMPLATE

Quick recap of previous class

Overview and/or key application

Definitions

These won't always make sense until you have seen some data, but at least you have them written down.

Concepts

A brief overview of the new ideas we will see in each class.

Heuristics

Colloquial language, rules of thumb, etc.

Potential Confusers

Nip these in the bud.

Contents