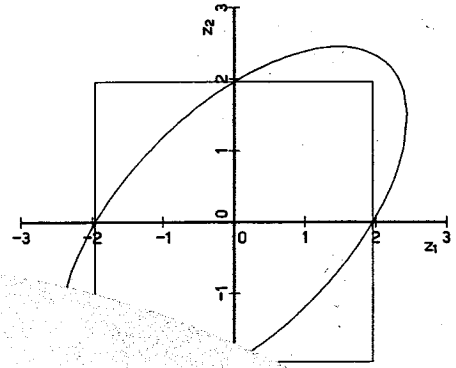
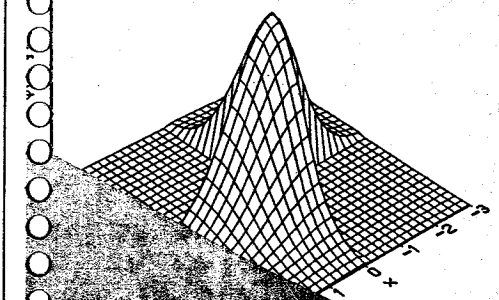
ARTHUR S. GOLDBERGER IS VILAS RESEARCH PROFESSOR OF ECONOMICS AT THE UNIVERSITY OF WISCONSIN, MADISON, AND IS A MEMBER OF THE NATIONAL ACADEMY OF SCIENCES.





COVER DESIGN BY JOYCE C. WESTON

HARVARD UNIVERSITY PRESS
CAMBRIDGE, MASSACHUSETTS
LONDON, ENGLAND

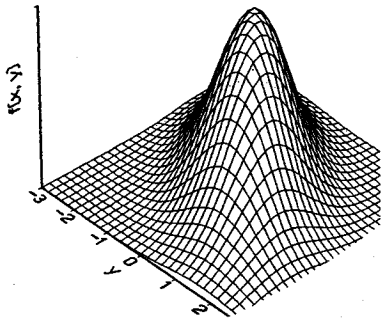
90000
9 780674 175440
ISBN 0-674-17544-1

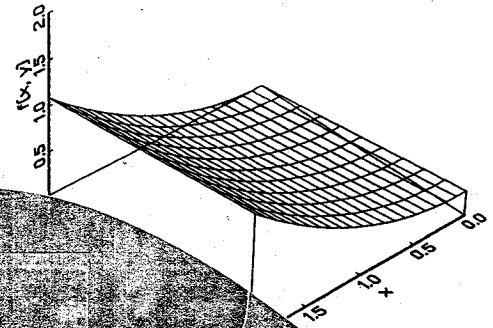
GOLDBERGER • A COURSE IN ECONOMETRICS

HARVARD

A · COURSE · IN ECONOMETRICS

ARTHUR S. GOLDBERGER





A Course in Econometrics

A Course in Econometrics

Arthur S. Goldberger

Harvard University Press
Cambridge, Massachusetts
London, England
1991

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

This book is printed on acid-free paper, and its binding materials have been chosen for strength and durability.

Library of Congress Cataloging-in-Publication Data

Goldberger, Arthur Stanley, 1930–
A course in econometrics / Arthur S. Goldberger.
p. cm.
Includes bibliographical references and index.
ISBN 0–674–17544–1 (alk. paper)
1. Econometrics. I. Title.
HB139.G634 1991
330'.01'5195—dc20 90–42284
CIP

For Guy H. Orcutt

Contents

Preface

The primary objective of this book is to prepare students for empirical research. But it also serves those who will go on to advanced study in econometric theory. Recognizing that readers will have diverse backgrounds and interests, I appeal to intuition as well as to rigor, and draw on a general acquaintance with empirical economics. I encourage readers to develop a critical sense: students ought to evaluate, rather than simply accept, what they read in journals and textbooks.

The book derives from lecture notes that I have used in the first-year graduate econometrics course at the University of Wisconsin. Students enroll from a variety of departments, including agricultural economics, finance, accounting, industrial relations, and sociology, as well as economics. All have had a year of calculus, a semester of linear algebra, and a semester of statistical inference. Some have had much more course work, including probability theory, mathematical statistics, and econometrics. Others have had substantial empirical research experience.

All of the material can be covered—indeed has been covered—in two semesters. To make that possible, I focus on a few underlying principles, rather than cataloging many potential methods. To accommodate students with varied preparation, the book begins with a review of elementary statistical concepts and methods, before proceeding to the regression model and its variants.

Although the models covered are quite standard, the approach taken is somewhat distinctive. The *conditional expectation function* (CEF) is introduced as the key feature of a multivariate population for economists who are interested in relations among economic variables. The CEF describes how the average value of one variable varies with values of the other variables in the population—a very simple concept. Another key feature of a multivariate population is the linear projection, or best

linear predictor (BLP): it provides the best linear approximation to the CEF. Alternative regression models arise according to the sampling scheme used to get drawings from the population.

The focus on CEF's and BLP's is useful. For example, whether a regression specification is "right" or "wrong," least-squares linear regression will typically estimate something in the population, namely, the BLP. Instead of emphasizing the bias (or inconsistency) of least squares, one can consider whether or not the population feature that it does consistently estimate is an interesting one. This approach also avoids visualizing empirical relations as disturbed versions of exact functions. For the most part, "disturbances" are just deviations from a mean, rather than objects that must be added to theoretical relations to produce empirical relations.

The *analogy principle* is relied on to suggest estimators, which are then evaluated according to conventional criteria. Thus least-squares, instrumental-variable, and maximum-likelihood estimators are made plausible by analogy before their sampling properties are studied.

A pedagogical feature of the book is the introduction of technical ideas in simple settings. Many advanced items are covered in the context of simple regression. These include asymptotics, the effect of alternative sampling schemes, and heteroskedasticity-corrected standard errors. The asymptotic theory for the ratio of sample means in sampling from a bivariate population, derived in Chapter 10, serves as a prototype for much more elaborate problems.

From Chapter 16 on, the exercises include real micro-data analyses. These are keyed to the GAUSS programming language, but can readily be adapted to other languages or packages. Virtually all of the exercises have been used as homework assignments or exam questions.

I thank three cohorts of students at Wisconsin, and one class at Stanford (where a portion of the material was used in 1990), for pressing me on details as well as on exposition. Over the years, I have had the benefit of guidance and instruction by several past and present colleagues at Wisconsin, including Guy Orcutt, Harold Watts, Glen Cain, Laurits Christensen, Gary Chamberlain, Charles Manski, and James Powell. I am particularly grateful to Gary Chamberlain for his close critical reading of an early version of the manuscript. Frank Wolak of Stanford provided helpful comments on a later version. They all will recognize their ideas here despite my attempts at camouflage.

I am fortunate to have had the expert editorial advice of Elizabeth Gretz at Harvard University Press, and the proofreading assistance of Donghul Cho and Sangyong Joo. For permission to quote or reproduce their work, I thank Thad W. Mirer, John J. Johnston, and Aptech Systems, Inc. Passages from *Econometric Methods*, 3d ed., by John J. Johnston, copyright © 1984 by McGraw-Hill, are reproduced with permission of McGraw-Hill, Inc.; Table 1.1 is adapted with permission of the Institute for Social Research from *Consumer Behavior of Individual Families over Two and Three Years*, edited by R. Kosobud and J. N. Morgan (Ann Arbor: Institute for Social Research, The University of Michigan, 1964); Table A.6 is reprinted by permission of John Wiley & Sons, Inc., from *Principles of Econometrics* by Henri Theil, copyright © 1971 by John Wiley & Sons; Tables A.3 and A.5 are reprinted by permission of Macmillan Publishing Company from *Economic Statistics and Econometrics*, 2d ed., by Thad W. Mirer, copyright © 1988 by Macmillan Publishing Company.

Madison, Wisconsin
November 1990

A Course in Econometrics

1 *Empirical Relations*

1.1. Theoretical and Empirical Relations

Most of economics is concerned with relations among variables. For example, economists might consider how

- the output of a firm is related to its inputs of labor, capital, and raw materials;
- the inflation rate is related to unemployment, change in the money supply, and change in the wage rate;
- the quantity demanded of a product depends on household income, price of the product, and prices of substitute products;
- the proportion of income saved varies with the level of family income;
- the earnings of a worker are related to her age, education, race, region of residence, and years of work experience.

In theoretical economics, the relations are characteristically treated as exact relations, that is, as deterministic relations, that is, as (single-valued) functions. For example, consider the relation between savings and income. Let

Y = savings rate = savings/income = proportion of income saved,
X = income.

In theoretical economics, one might consider $Y = g(X)$, where $g(\cdot)$ is a function in the mathematical sense, that is, a single-valued function. Henceforth we will always use the word "function" in this strict sense. Corresponding to each value of X, there is a unique value of Y. An economist might ask such questions as: Is $g(X)$ constant with respect to X? Is $g(X)$ increasing in X? Is $g(X)$ linear in X?

The same applies when there are several explanatory variables $X_1, \ldots, X_k$, as when a firm's output is related to its inputs of labor, capital, and raw materials. In theory one considers $Y = g(X_1, \ldots, X_k)$, where corresponding to each set of values for $X_1, \ldots, X_k$, there is a unique value of Y. So g is again a (single-valued) function. The relation of Y to the X's is an exact one, that is, a deterministic one.

This is what relations look like in theory. What happens when we look at *empirical relations*, that is, at real-world data on economic variables?

Table 1.1 refers to 1027 U.S. "consumer units" (roughly, families) interviewed by the University of Michigan's Survey Research Center in 1960, 1961, and 1962. Income is averaged over the two years 1960 and 1961; the savings rate is the ratio of two-year savings to two-year income. In the source, the data were presented in grouped form, with ten brackets for income and nine brackets for the savings rate. For convenience, we have assumed that all observations in a bracket were located at a single point (the approximate midpoint of the bracket) and have labeled the values of X and Y accordingly. Across the top of the table are the ten distinct values of X = income (in thousands of dollars), which we refer to as x_i ($i = 1, \ldots, 10$). Down the left-hand side of the table are the nine distinct values of Y = savings rate, which we refer to as y_j

Table 1.1 Joint frequency distribution of X = income and Y = savings rate.

	X									
Y	0.5	1.5	2.5	3.5	4.5	5.5	6.7	8.8	12.5	17.5
.50	.001	.011	.007	.006	.005	.005	.008	.009	.014	.004
.40	.001	.002	.006	.007	.010	.007	.008	.009	.008	.007
.25	.002	.006	.004	.007	.010	.011	.020	.019	.013	.006
.15	.002	.009	.009	.012	.016	.020	.042	.054	.024	.020
.05	.010	.023	.033	.031	.041	.029	.047	.039	.042	.007
0	.013	.013	.000	.002	.001	.000	.000	.000	.000	.000
−.05	.001	.012	.011	.005	.012	.016	.017	.014	.004	.003
−.18	.002	.008	.013	.006	.009	.008	.008	.008	.006	.002
−.25	.009	.009	.010	.006	.009	.007	.005	.003	.002	.003
$p(x)$	.041	.093	.093	.082	.113	.103	.155	.155	.113	.052

Source: Adapted from R. Kosobud and J. N. Morgan, eds., *Consumer Behavior of Individual Families over Two and Three Years* (Ann Arbor: Institute for Social Research, The University of Michigan, 1964), Table 5–5.

($j = 1, \ldots, 9$). So there are $90 = 10 \times 9$ cells in the cross-tabulation. In the i, j cell one finds

$$p(x_i, y_j) = \text{the proportion of the 1027 families who reported the combination } (X = x_i \text{ and } Y = y_j).$$

This table gives the *joint frequency distribution* of Y and X for this data set.

Here is some general notation for a joint frequency distribution of variables X and Y, where X takes on distinct values x_i ($i = 1, \ldots, I$) and Y takes on distinct values y_j ($j = 1, \ldots, J$). The joint frequencies $p(x_i, y_j)$ are defined for each of the $I \times J$ cells. Clearly $\Sigma_i \Sigma_j p(x_i, y_j) = 1$, where $\Sigma_i = \Sigma_{i=1}^{I}$, $\Sigma_j = \Sigma_{j=1}^{J}$. From the joint frequency distribution it is easy to calculate the *marginal frequency distribution* of X:

$$p(x_i) = \sum_j p(x_i, y_j)$$

$$= \text{proportion of observations having } X = x_i \quad (i = 1, \ldots, I).$$

Then $\Sigma_i p(x_i) = \Sigma_i [\Sigma_j p(x_i, y_j)] = 1$.

Return to the joint frequency distribution of Table 1.1. For each of the ten columns, add down the rows to get the marginal frequency distribution $p(x)$ in the last row—the bottom margin—and observe that the entries in the last row do add up to 1.

Evidently, the empirical relation between Y and X is not a deterministic one. For if it were, then in any column of the body of the table, there would be only a single nonzero entry. But in every column, there are several nonzero entries. Indeed, in most columns, all nine entries are nonzero. Corresponding to each value of X, there is a whole set of values of Y rather than a single value of Y. What we see is a distribution rather than a function. This is characteristic of the real world: empirical relations are not deterministic, not exact, not functional relations.

Now focus attention on the distribution of Y corresponding to a particular value of X. Take $X = x_i$, say, and ask what proportion of the observations that have $X = x_i$, also have the values $Y = y_1, \ldots, y_J$. The answers give the *conditional frequency distribution* of Y given $X = x_i$:

$$p(y_j | x_i) = \frac{p(x_i, y_j)}{p(x_i)} \qquad (j = 1, \ldots, J).$$

It follows for each $i = 1, \ldots, I$ that

$$\sum_j p(y_j|x_i) = \sum_j \frac{p(x_i, y_j)}{p(x_i)} = \frac{\Sigma_j p(x_i, y_j)}{p(x_i)} = \frac{p(x_i)}{p(x_i)} = 1.$$

Divide the entries in each column of Table 1.1 by the column sum. The resulting Table 1.2 gives the conditional frequency distributions of Y given X, one such distribution for each distinct value of X. Observe that each column sum in this table is equal to 1. The nondeterministic character of empirical relations is again apparent. If $Y = g(X)$ as in theoretical economics, each column in the body of Table 1.2 would have a single unity, all other entries being zero. But Table 1.2 does not look like that.

So we face a dilemma. We would like to use economic theory to guide our analysis of data, and to use data to implement the theory. But the savings-income relation in economic theory is deterministic, while in the empirical data it is not deterministic. How shall we resolve the dilemma?

The theory seems to say that all families with the same value of X should have the same Y. If so, the data seem to indicate that these families did not do what they should. Perhaps they tried to, but made mistakes? If so, the conditional distributions are all due to error—there

Table 1.2 Conditional frequency distributions of Y = savings rate for given values of X = income.

	X									
Y	0.5	1.5	2.5	3.5	4.5	5.5	6.7	8.8	12.5	17.5
.50	.024	.118	.075	.073	.044	.049	.052	.058	.124	.077
.40	.024	.022	.064	.086	.088	.068	.052	.058	.071	.135
.25	.049	.064	.043	.086	.088	.107	.129	.123	.115	.115
.15	.049	.097	.097	.146	.142	.194	.271	.348	.212	.384
.05	.244	.247	.355	.378	.363	.281	.303	.252	.372	.135
0	.317	.140	.000	.024	.009	.000	.000	.000	.000	.000
−.05	.024	.129	.118	.061	.106	.155	.109	.090	.035	.058
−.18	.049	.086	.140	.073	.080	.078	.052	.052	.053	.038
−.25	.220	.097	.108	.073	.080	.068	.032	.019	.018	.058
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$m_{Y\|X}$	−.012	.065	.048	.099	.079	.083	.112	.129	.154	.161

is a true value of Y for each value of X, but the families erred in their savings behavior, or perhaps in reporting savings to the interviewer. This is surely possible, but to rely on errors alone is unappealing.

We know that the families differ in characteristics other than income that may be relevant to their savings behavior. The gap between theory and reality might diminish if the theory introduced more explanatory variables, $X_2, \ldots, X_k$. Then instead of looking at $p(y_j|x_i)$ we would be looking at $p(y_j|x_{1i}, \ldots, x_{kh})$. Presumably there will be less dispersion of Y within those narrowly defined cells than there is in the coarsely defined cells of our tables. But even then the empirical relation would not be deterministic. For example, consider all households who have the same income, family size, and race. We would still see differences in their Y values. Because a gap would remain in any case, for present purposes we may as well continue with the single-X case.

1.2. Sample Means and Population Means

To resolve the dilemma, we first reinterpret the economic theory. When the theorist speaks of Y being a function of X, let us say that she means that the *average* value of Y is a function of X. If so, when she says that $g(X)$ increases with X, she means that on average, the value of Y increases with X. Or, when she says that $g(X)$ is constant, she means that the average value of Y is the same for all values of X. With that interpretation in mind, let us re-examine our data set, seeking the empirical counterpart of the theorist's average value.

Here is some algebra that shows how to calculate the average of a frequency distribution. First, for the variable X = income: if the marginal frequency distribution of X is given by $p(x_i)$ $(i = 1, \ldots, I)$, then the *marginal mean* of X is

$$m_X = \sum_i x_i p(x_i).$$

Similarly, the marginal mean of Y is $m_Y = \Sigma_j y_j p(y_j)$. Further, if the conditional frequency distribution of Y given $X = x_i$ is $p(y_j|x_i)$ $(j = 1, \ldots, J)$, then the *conditional mean* of Y given $X = x_i$ is

$$m_{Y|x_i} = \sum_j y_j p(y_j|x_i).$$

There are I such conditional means, one for each distinct value of X.

Observe that the average of the conditional means equals the marginal mean:

$$\sum_i m_{Y|x_i} p(x_i) = \sum_i \left[\sum_j y_j p(y_j \mid x_i) \right] p(x_i)$$
$$= \sum_i \sum_j y_j p(x_i, y_j)$$
$$= \sum_j y_j \left[\sum_i p(x_i, y_j) \right] = \sum_j y_j p(y_j) = m_Y.$$

Return to Table 1.2. The conditional means of Y have been calculated, one for each of the ten values of X, and are presented in the last row of the table. If we extract the top row (the x_i) and the bottom row (the $m_{Y|x_i}$), we have the *conditional mean function*, or cmf, for Y given X, which we will refer to as $m_{Y|X}$.

The cmf is a deterministic relation—that is, a function—in our data. The cmf specifies how the average value of Y is functionally related to X in the data set. For an economist who is concerned with the relation of the savings rate to income, this cmf $m_{Y|X}$ is the most interesting feature of the joint frequency distribution. We can plot it, and study it in terms of the economic theorist's concerns: Does $m_{Y|X}$ vary with X? Does it vary linearly with X, that is, is $\Delta m/\Delta X$ constant?

In Figure 1.1, the ten points that make up the cmf are plotted and, for convenience, are connected by line segments. Looking at the plot, we see a cmf that is too ragged and erratic to be taken seriously by a theorist. So a gap remains between theory and reality.

To proceed, we recognize that the theorist who discussed the relation between the savings rate and income was not talking about $m_{Y|X}$ for these particular 1027 families in 1960–1961. If the Survey Research Center had happened to interview a different 1027 families, or even a 1028th family, or even the same 1027 families in a different year, we would have had a different $p(x, y)$ table, different $p(y \mid x)$ columns, and no doubt a different $m_{Y|X}$ function.

The next step is obvious. We suppose that what we observe is only a *sample* from a *population*. Our cmf displays sample means, not population means. Presumably the theorist was referring to population means, not sample means. It will be adequate for present purposes to think of the population itself as represented by a joint frequency distribution, one that refers to millions of families rather than to our 1027. (For convenience, we continue to suppose that the X, Y pairs are confined to the same 90 combinations.) In the population the joint frequencies are given by $\pi(x_i, y_j)$, say, with $\Sigma_i \Sigma_j \pi(x_i, y_j) = 1$. So in the population, the marginal frequencies of X are

$$\pi(x_i) = \sum_j \pi(x_i, y_j),$$

and the conditional frequencies of Y given X are

$$\pi(y_j \mid x_i) = \pi(x_i, y_j)/\pi(x_i).$$

Further, the population mean of X is

$$\mu_X = \sum_i x_i \pi(x_i),$$

and the population conditional means of Y given X are

$$\mu_{Y|x_i} = \sum_j y_j \pi(y_j \mid x_i).$$

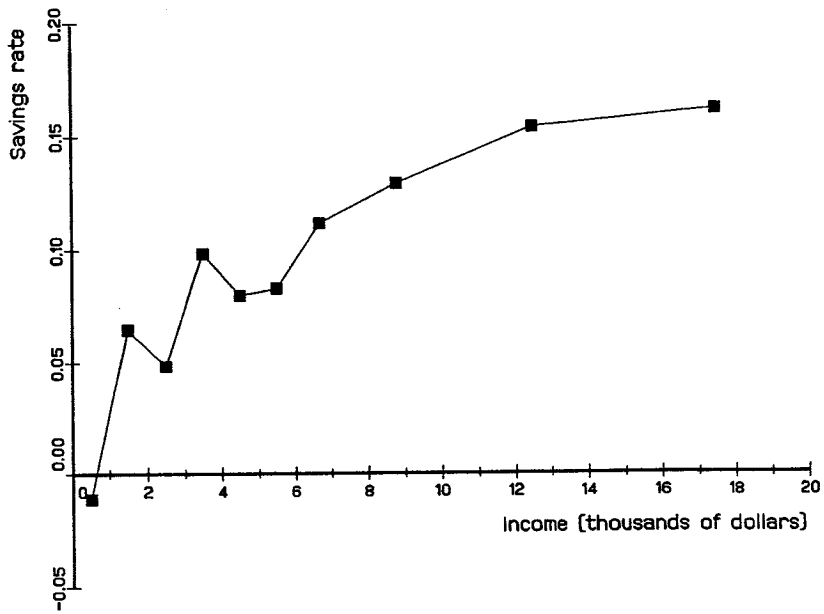


Figure 1.1 Conditional mean function: savings rate on income.

We have arrived at the following position. When a theorist talks about $Y = g(X)$, she is really referring to the population conditional mean function $\mu_{Y|X} = g(X)$, which is indeed a function of X. We now have the theorist referring to the population features $\pi(x, y)$, $\pi(y|x)$, $\mu_{Y|X}$, while the empirical material refers to the sample features $p(x, y)$, $p(y|x)$, $m_{Y|X}$. This leaves us with the gap between the hypothetical population π's and μ's and the observed sample p's and m's.

1.3. Sampling

Imagine this physical representation of the population and sample. Each family in the population is represented by a chip on which its (X, Y) pair is printed. The millions of chips are in a barrel. The joint frequency distribution in the barrel is $\pi(x, y)$. We draw 1027 chips with replacement, record the x, y combinations, and tabulate the (relative) frequencies as $p(x, y)$. Our $p(x, y)$ table is just one of the possible $p(x, y)$ tables that might have been obtained in this manner. Our data set is just one sample from the population. In general none of the possible sample $p(x, y)$ tables will be identical to the population $\pi(x, y)$ table, and none of the possible sample $m_{Y|X}$ functions will coincide with the population $\mu_{Y|X}$ function.

The dilemma has been substantially resolved. The questions that remain include: What sort of samples come from a population? How do sample joint frequency distributions, conditional frequency distributions, and cmf's depart from the population joint frequency distribution, conditional frequency distribution, and cmf? How can we best use a sample to learn about the population from which it came? How confident can we be in our conclusions? These are precisely the questions that are addressed in classical statistical theory.

1.4. Estimation

A large part of empirical econometrics is concerned with estimating population conditional mean functions from a sample. That is, economists very often want to learn how the average value of one variable varies in a population with one, or several, other variables.

If so, what remains to be discussed? After all, in introductory statistics courses, we have learned all about estimating population means. In particular we have learned that the sample mean is an attractive estimator of a population mean—perhaps even that it is the best estimator. That attractiveness should carry over to the present situation, where we are concerned with a population conditional mean *function*. A population cmf is just a set of population means, and our joint sample can be viewed as a collection of conditional subsamples. So it is natural to use the sample conditional means as estimates of the population conditional means. That is, it is natural to take $m_{Y|x_i}$ as the estimate of $\mu_{Y|x_i}$, thus taking $m_{Y|X}$ as the estimate of $\mu_{Y|X}$, bearing in mind that $m \neq \mu$.

But is that always the right way to proceed? Suppose that an economic theory says that the population cmf for the savings rate on income is linear: $\mu_{Y|X} = \alpha + \beta X$, with α and β unknown. As Figure 1.1 shows, our sample $m_{Y|X}$ is not linear in income—the ten $m_{Y|x_i}$'s do not fall on a straight line. As empirical economists who wish to be guided by economic theory, shall we retain the ten sample $m_{Y|x_i}$'s as they stand? Or shall we smooth the sample $m_{Y|x_i}$'s by fitting a straight line to the ten points, obtaining $m^*_{Y|X} = a + bX$, and use those $m^*_{Y|x_i} = a + bx_i$ $(i = 1, \ldots, 10)$ as the estimates of the $\mu_{Y|x_i}$, thus using a and b as the estimates of α and β? If we decide to smooth, how shall we fit the line? And do we know that the smoothed estimates are better than the sample means as estimates of the population means? Or suppose a theory said that the population cmf is exponential: $\mu_{Y|X} = \alpha X^{\beta}$. How should we fit that *curve*? And does the smoothed sample line $m^*_{Y|X}$ still tell us anything about the population curve $\mu_{Y|X}$?

These are typical of the issues that arise in this book. To address them seriously, we turn to a review of the framework provided by the random variable–probability distribution model of classical statistics.

Exercises

The following all refer to the empirical joint frequency distribution of Tables 1.1 and 1.2.

1.1 Calculate the marginal frequency distribution of Y. Then calculate the mean of the conditional means of Y, verifying that it equals the marginal mean of Y (up to round-off error).

1.2 Calculate the conditional frequency distributions of X given Y, and the conditional mean function of X given Y.

1.3 Plot the two conditional mean functions $m_{Y|X}$ and $m_{X|Y}$ on a single diagram, using the horizontal axis for x-values and the vertical axis for y-values. Comment on the differences between those two functions.

1.4 Let Z = savings (in thousands of dollars), so $Z = XY$. The savings of a family with income x_i and savings rate y_j is $z_{ij} = x_i y_j$, so that the mean savings for the families in our sample is given by

$$m_Z = \sum_i \sum_j x_i y_j p(x_i, y_j).$$

Will this equal $m_X m_Y$? That is, can mean savings be obtained by multiplying mean savings rate by mean income? Explain.

2 *Univariate Probability Distributions*

2.1. Introduction

The general framework of probability theory involves an experiment that has various possible outcomes. Each distinct outcome is represented as a point in a set, the sample space. Probabilities are assigned to certain outcomes in accordance with certain axioms, and then the probabilities of other events, which are subsets of the sample space, are deduced. Let S denote the sample space, A denote an event, and $\Pr(\cdot)$ the probability assignment. Then the axioms are $0 \leq \Pr(A) \leq 1$, $\Pr(S) = 1$, and, where $A_1, A_2, \ldots$ are disjoint events, $\Pr(\cup_j A_j) = \Sigma_j \Pr(A_j)$.

We proceed somewhat more concretely. The distinct possible outcomes are identified, that is, distinguished, by the value of a single variable X. Each trial of the experiment produces one and only one value of X. Here X is called a *random variable*, a label that merely indicates that X is a variable whose value is determined by the outcome of an experiment. The values that X takes on are denoted by x. So we may refer to events such as $\{X = x\}$ and $\{X \leq x\}$. We distinguish two cases of probability distributions: discrete and continuous.

2.2. Discrete Case

In the *discrete case*, the number of distinct possible outcomes is either finite or countably infinite, so one can compile a list of them: x_1, $x_2, \ldots$. The convention is to list these *mass points* in increasing order: $x_1 < x_2 < \ldots$. The assignment of probabilities is done via a function $f(x)$, with these properties:

$f(x) \geq 0$ everywhere,

$f(x) = 0$ except at the mass points $x_1, x_2, \ldots,$

$\sum_i f(x_i) = 1,$

where Σ_i denotes summation over all the mass points. The function $f(\cdot)$ is called a *probability mass function*, or pmf.

The initial assignment of probabilities is $\Pr(X = x) = f(x)$. That is, the probability that the random variable capital X takes on the value lowercase x is $f(x)$. Then the probabilities of various events are deducible by rules of probability theory. For example, supposing that the list is in increasing order, $\Pr(X \leq x_5) = \Sigma_{i=1}^{5} f(x_i)$. For another example, if x_0 is not a mass point, then $\Pr(X = x_0) = f(x_0) = 0$.

Observe that the pmf $f(\cdot)$ has exactly the formal properties that $p(\cdot)$ had in univariate *frequency* distributions. (And observe the perverse notation: we used $p(\cdot)$ for *f*requency distributions, and now use $f(\cdot)$ for *p*robability distributions.) Because of the formal resemblance, it may be helpful to interpret the pmf $f(\cdot)$ as the $\pi(\cdot)$ of Chapter 1, namely the frequency distribution in a population.

Here are several examples of discrete univariate probability distributions:

(1) *Bernoulli* with parameter p $(0 \leq p \leq 1)$. Here

$$f(x) = p^x(1 - p)^{(1-x)} \quad \text{for } x = 0, 1,$$

with $f(x) = 0$ elsewhere. So $\Pr(X = 0) = f(0) = p^0(1 - p)^1 = 1 - p$, $\Pr(X = 1) = f(1) = p^1(1 - p)^0 = p$, and $\Pr(X = x) = f(x) = 0$ for all other values of x. Observe that $f(x) \geq 0$ everywhere, that $f(x) = 0$ except at the two mass points $x = 0$ and $x = 1$, and that $\Sigma_i f(x_i) = f(0) + f(1) = 1$, as required. So this is a legitimate pmf.

In what contexts might the Bernoulli distribution be relevant? That is, for what experiments might it be appropriate? A familiar example is a coin toss: $X = 1$ if heads, $X = 0$ if tails. The Bernoulli pmf says that $\Pr(X = 1) = p$, $\Pr(X = 0) = 1 - p$. Special cases are $p = 0.5$ (fair coin), and $p = 0.7$ (loaded coin). A more interesting example concerns unemployment. Let $X = 1$ if unemployed, $X = 0$ otherwise, the experiment being drawing an adult at random from the U.S. population. The Bernoulli pmf says that the probability of being unemployed is p.

(2) *Discrete Uniform* with parameter N (N positive integer). Here

$$f(x) = 1/N \quad \text{for } x = 1, 2, \ldots, N,$$

with $f(x) = 0$ elsewhere. Observe that $f(x) \geq 0$ everywhere, that $f(x) = 0$ except at the N mass points, and that $\Sigma_i f(x_i) = 1/N + \cdots + 1/N = 1$. So this is a legitimate pmf.

In what contexts might a discrete uniform distribution be relevant? A very familiar example is the roll of a fair die: X = the number on the face that comes up, and $N = 6$. This discrete uniform pmf says that $\Pr(X = 1) = \Pr(X = 2) = \cdots = \Pr(X = 6) = 1/6$.

(3) *Binomial* with parameters n, p (n positive integer, $0 \leq p \leq 1$). Here

$$f(x) = \frac{n!}{x!(n - x)!} p^x(1 - p)^{(n-x)} \quad \text{for } x = 0, 1, 2, \ldots, n,$$

with $f(x) = 0$ elsewhere. (Recall factorial notation: $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$, ...) Observe that $f(x) \geq 0$ everywhere, that $f(x) = 0$ except at the mass points, and (as can be confirmed by summing from 0 to n) that $\Sigma_i f(x_i) = [p + (1 - p)]^n = 1$.

In what contexts might the binomial distribution be relevant? Suppose we toss n identical coins at once, and let X = number of heads. That is, we run the Bernoulli(p) experiment n times, independently, and record the number of 1's. Or if we observe an adult over n months, let X = number of months unemployed. The binomial distribution may be appropriate.

Special cases of the binomial include:

(a) $n = 1$: $f(0) = 1 \times p^0(1 - p)^1 = (1 - p)$,
$f(1) = 1 \times p^1(1 - p)^0 = p.$

So the binomial distribution with parameters $(1, p)$ is the same as the Bernoulli distribution with parameter p.

(b) $n = 2$: $f(0) = 1 \times p^0(1 - p)^2 = (1 - p)^2$,
$f(1) = 2 \times p^1(1 - p)^1 = 2p(1 - p)$,
$f(2) = 1 \times p^2(1 - p)^0 = p^2.$

Clearly $f(0) + f(1) + f(2) = [p + (1 - p)]^2 = 1$.

(4) *Poisson* with parameter λ $(\lambda > 0)$. Here

$$f(x) = e^{-\lambda} \lambda^x / x! \quad \text{for } x = 0, 1, 2, \ldots,$$

with $f(x) = 0$ elsewhere. Observe that $f(x) \geq 0$ everywhere, that $f(x) = 0$ except at the mass points, and (using the series expansion

$$e^{\lambda} = \sum_{x=0}^{\infty} (\lambda^x/x!) = 1 + \lambda + \lambda^2/2 + \lambda^3/6 + \ldots)$$

that $\Sigma_i f(x_i) = 1$. In the Poisson distribution, the number of distinct possible outcomes is countably infinite.

Applications of the Poisson distribution might include the number of phone calls received at a switchboard in an hour, or the number of job offers an individual receives in a year.

2.3. Continuous Case

In the *continuous case*, we again consider an experiment whose outcomes are distinguished by the value of a single real variable X. But now there is a continuum of distinct possible outcomes, so we cannot compile them in a list.

The assignment of probabilities is done via a function $f(x)$ with these properties: $f(x) \geq 0$ everywhere, $\int_{-\infty}^{\infty} f(x)\, dx = 1$. This function $f(\cdot)$ is called a *probability density function*, or pdf. The initial assignment of probabilities via $f(x)$ is as follows. For any pair of numbers a, b with $a \leq b$:

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\, dx.$$

That is, the probability that the random variable X lies in the closed interval $[a, b]$ is given by the area under the $f(x)$ curve between the points a and b.

To see what we are committed to in the continuous case, consider several specific events:

(1) $\quad A = \{-\infty \leq X \leq \infty\}.$

Here $a = -\infty$, $b = \infty$, so $\Pr(A) = \int_{-\infty}^{\infty} f(x)\, dx = 1$, as it should, since A exhausts the sample space.

(2) $\quad A = \{X \leq b\} = \{-\infty \leq X \leq b\}.$

Here $a = -\infty$, so $\Pr(A) = \int_{-\infty}^{b} f(x)\, dx$.

(3) $\quad A = \{X = a\} = \{a \leq X \leq a\}.$

Here $b = a$, so $\Pr(A) = \int_a^a f(x)\, dx = 0$.

Consider (3), which says that in the continuous case $\Pr(X = x) = 0$ for every x. This means that the probability that X takes on a particular value x is zero, for every such particular value. And yet on every run of the experiment some value of x is taken on. Is that a contradiction? No, not unless one confuses two distinct concepts, zero probability and impossibility. In the continuous case, a zero-probability event is not an impossible event. Although this seems awkward, no other assignment of probabilities to events of the form $\{X = x\}$ is possible when the distinct possible outcomes form a continuum.

Further, the following events all have the same probability, namely $\int_a^b f(x)\, dx$:

$$A_1 = \{a \leq X \leq b\}, \qquad A_2 = \{a \leq X < b\},$$
$$A_3 = \{a < X \leq b\}, \qquad A_4 = \{a < X < b\}.$$

For example, $A_1 = A_2 \cup A_0$ where $A_0 = \{X = b\}$. But A_2 and A_0 are disjoint, and $\Pr(A_0) = 0$, so $\Pr(A_2) = \Pr(A_1)$.

The *cumulative distribution function*, or cdf, is defined as

$$F(x) = \int_{-\infty}^{x} f(t)\, dt,$$

with t being a dummy argument. The cdf gives the area under the pdf from $-\infty$ up to x, so $F(x) = \Pr(X \leq x)$. Some properties of a cdf are immediate:

- $F(-\infty) = 0$, $F(\infty) = 1$.
- $F(\cdot)$ is monotonically nondecreasing (because $f(t) \geq 0$).
- Wherever differentiable, $dF(x)/dx = f(x)$, because $F = \int f(t)\, dt$, and the derivative of an integral with respect to its upper limit is just the argument (the integrand) evaluated at the upper limit.

In the continuous case the cdf is convenient because

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\, dx = \int_{-\infty}^{b} f(x)\, dx - \int_{-\infty}^{a} f(x)\, dx$$
$$= F(b) - F(a).$$

The cdf could have been introduced in the discrete case as $F(x) = \Pr(X \leq x)$, but it is not so crucial there.

Here are several examples of continuous univariate probability distributions:

(1) *Rectangular* (or continuous uniform) on the interval $[a, b]$, with parameters $a < b$. The pdf is

$$f(x) = 1/(b - a) \quad \text{for } a \le x \le b,$$

with $f(x) = 0$ elsewhere. Observe that $f(x) \ge 0$ everywhere, and that

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{a} 0\,dx + \int_{a}^{b} [1/(b-a)]\,dx + \int_{b}^{\infty} 0\,dx$$
$$= [1/(b-a)]\,x\|_{a}^{b} = 1.$$

(Note: The symbol $\|$ is used to denote an integral to be evaluated.) So this is a legitimate pdf. It plots as a rectangle, with base $b - a$ and height $1/(b - a)$; the area of the rectangle is base $\times$ height $= 1$. The cdf is

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 0 & \text{for } x < a, \\ (x-a)/(b-a) & \text{for } a \le x \le b, \\ 1 & \text{for } b < x. \end{cases}$$

(2) *Exponential* with parameter $\lambda > 0$. The pdf is

$$f(x) = \lambda\, e^{-\lambda x} \quad \text{for } x > 0,$$

with $f(x) = 0$ for $x \le 0$. The relevant indefinite integral is

$$\int \lambda\, e^{-\lambda t}\,dt = \lambda \int e^{-\lambda t}\,dt = \lambda\,(e^{-\lambda t})/(-\lambda) = -e^{-\lambda t},$$

so the cdf is

$$F(x) = \begin{cases} 0 & \text{for } x \le 0, \\ 1 - e^{-\lambda x} & \text{for } x > 0. \end{cases}$$

The exponential pdf and cdf for $\lambda = 2$ are plotted in Figure 2.1.

The exponential distribution may be appropriate for the length of time until a light bulb fails. It may also be relevant for the duration of unemployment among those who leave a job, with time being measured continuously.

(3) *Standard Normal.* The standard normal distribution plays a central role in statistical theory. The pdf is

$$f(x) = (2\pi)^{-1/2} \exp(-x^2/2),$$

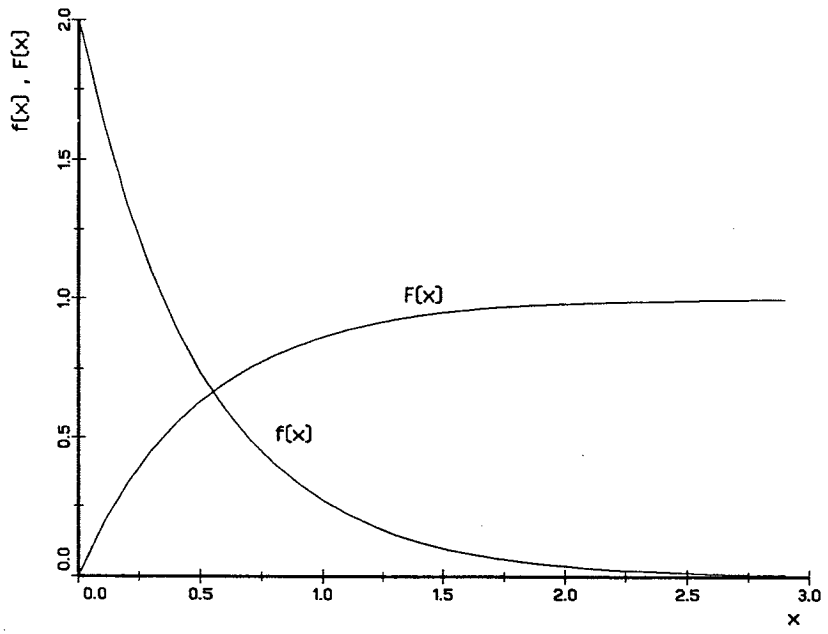


Figure 2.1 Exponential distribution: pdf and cdf, $\lambda = 2$.

which plots as a familiar bell-shaped curve, as in Figure 2.2. Some features of the curve are apparent by inspection of the pdf formula. The curve is symmetric about zero: $f(-x) = f(x)$. The ordinate at zero is $f(0) = 1/\sqrt{(2\pi)} = 0.3989$. The slope is

$$f'(x) = (2\pi)^{-1/2} \exp(-x^2/2)(-x) = -xf(x).$$

So $f'(x) > 0$ for $x < 0$, $f'(x) = 0$ for $x = 0$, $f'(x) < 0$ for $x > 0$. The second derivative is $f''(x) = -[xf'(x) + f(x)] = -[-x^2 f(x) + f(x)] = (x^2 - 1)f(x)$. So the curve has inflection points at $x = 1$ and $x = -1$.

The cdf is

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(t)\,dt.$$

No closed form is available, but the standard normal cdf is plotted in Figure 2.2 and tabulated in Table A.1. The tabulation is confined to $x > 0$, which suffices because the symmetry of $f(x)$ about 0 implies that $F(-x) = 1 - F(x)$.