STATISTICAL TOOLS

FOR ECONOMISTS

Daniel McFadden ©2001

Department of Economics
University of California
Berkeley, CA 94720-3880
(mcfadden@econ.berkeley.edu)

REVISED VERSION, CHAP. 1-7, 1/16/2001

COMMENTS AND CORRECTIONS WELCOME

TABLE OF CONTENTS

Section		Page
1	**Economic Analysis and Econometrics**	
1.1	Introduction	1
1.2	Cab Franc's Rational Decision	1
1.3	Stock Market Efficiency	4
1.4	The Capital Asset Pricing Model	7
1.5	Conclusions	14
1.6	Exercises	14
2	**Analysis and Linear Algebra in a Nutshell**	
2.1	Some Elements of Mathematical Analysis	17
2.2	Vectors and Linear Spaces	20
2.3	Linear Transformations and Matrices	22
2.4	Eigenvalues and Eigenvectors	26
2.5	Partitioned Matrices	27
2.6	Quadratic Forms	28
2.7	LDU and Cholesky Factorizations of a Matrix	29
2.8	Singular Value Decomposition of a Matrix	32
2.9	Idempotent Matrices and Generalized Inverses	33
2.10	Projections	35
2.11	Kronecker Products	36
2.12	Shaping Operations	37
2.13	Vector and Matrix Derivatives	37
2.14	Updating and Backdating Matrix Operations	39
2.15	Notes and Comments	40
2.16	Exercises	40
3	**Probability Theory in a Nutshell**	
3.1	Sample Spaces	43
3.2	Event Fields and Information	43
3.3	Probability	46
3.4	Statistical Independence and Repeated Trials	54
3.5	Random Variables, Distribution Functions, and Expectations	58
3.6	Transformations of Random Variables	71
3.7	Special Distributions	75
3.8	Notes and Comments	79
	Exercises	84

4	**Limit Theorems in Statistics**	
4.1	Sequences of Random Variables	89
4.2	Independent and Dependent Random Sequence	98
4.2	Laws of large Numbers	101
4.3	Central Limit Theorems	105
4.4	Extensions of Limit Theorems	109
4.5	References	115
4.6	Exercises	116
5	**Experiments, Sampling, and Statistical Decisions**	
5.1	Experiments	117
5.2	Populations and Samples	119
5.3	Statistical Decisions	122
5.4	Statistical Inference	127
5.5	Exercises	127
6	**Estimation**	
6.1	Desirable Properties of Estimators	129
6.2	General Estimation Criteria	138
6.3	Estimation in Normally Distributed Populations	140
6.4	Large Sample Properties of Maximum Likelihood Estimators	143
6.5	Exercises	153
7	**Hypothesis Testing**	
7.1	The General Problem	155
7.2	The Cost of Mistakes	155
7.3	Design of the Experiment	156
7.4	Choice of Decision Procedure	158
7.5	Hypothesis Testing in Large Samples	166
7.6	Exercises	169

CHAPTER 1. ECONOMIC ANALYSIS AND ECONOMETRICS

1.1. INTRODUCTION

The study of resource allocation by the discipline of Economics is both a *pure science*, concerned with developing and validating theories of behavior of individuals, organizations, and institutions, and a *policy science*, concerned with the design of institutions and the prediction and social engineering of behavior. In both arenas, concepts from probability and statistics, and methods for analyzing and understanding economic data, play an important role. In this chapter, we give three introductory examples that illustrate the intertwining of economic behavior, statistical inference, and econometric forecasting. These examples contain some probability and statistics concepts that are explained in later chapters. What is important on first reading are the general connections between economic reasoning, probability, statistics, and economic data; details can be postponed.

1.2. CAB FRANC'S RATIONAL DECISION

Cab Franc is a typical professional economist: witty, wise, and unboundedly rational. Cab works at a university in California. Cab's life is filled with fun and excitement, the high point of course being the class he teaches in econometrics. To supplement his modest university salary, Cab operates a small vineyard in the Napa Valley, and sells his grapes to nearby wineries.

Cab faces a dilemma. He has to make a decision on whether to harvest early or late in the season. If the Fall is dry, then late-harvested fruit is improved by additional "hang time", and will fetch a premium price. On the other hand, if rains come early, then much of the late-harvested fruit will be spoiled. If Cab harvests early, he avoids the risk, but also loses the opportunity for the maximum profit. Table 1 gives Cab's profit for each possible action he can take, and each possible Event of Nature:

Table 1. Profit from Selling Grapes

		Action	
Event of Nature	Frequency	Harvest Early	Harvest Late
Wet	0.4	$30,000	$10,000
Dry	0.6	$30,000	$40,000
Expected Profit		$30,000	$28,000

Cab wants to maximize expected profit. In other words, he wants to make the probability-weighted average of the possible profit outcomes as large as possible. Cab is not adverse to risk; he figures that risks will average out over the years. From historical records, he knows that the frequency of early rain is 0.4. To calculate expected profit in this case from a specified action, Cab multiplies the profit he will receive in each event of Nature by the probability of this event, and sums. If he harvests early, the expected profit is ($30,000)·(0.4) + ($30,000)·(0.6) = $30,000. If he harvests late, the expected profit is ($10,000)·(0.4) + ($40,000)·(0.6) = $28,000. Then, in the absence of any further information, Cab will choose to harvest early and earn an expected profit of $30,000.

There is a specialized weather service, Blue Sky Forecasting, that sells long-run precipitation forecasts for the Napa Valley. Cab has to choose whether to subscribe to this service, at a cost of $1000 for the year. Table 2 gives the historical record, over the past 100 years, on the joint frequency of various forecasts and outcomes.

Table 2. Frequency of Forecasts and Outcomes

	Blue Sky Forecasts		TOTAL
Event of Nature	Early	Late	
Wet	0.3	0.1	0.4
Dry	0.2	0.4	0.6
TOTAL	0.5	0.5	1.0

The interpretation of the number 0.3 is that in 30 percent of the past 100 years, Blue Sky has forecast early rain and they do in fact occur. The column totals give the frequencies of the different Blue Sky forecasts. The row totals give the frequencies of the different Events of Nature. Thus, Blue Sky forecasts early rain half the time, and the frequency of actual early rain is 0.4. One can also form conditional probabilities from Table 2. For example, the conditional probability of dry, given the event that late rain are forecast, equals 0.4/(0.1+0.4) = 0.8.

If Cab does not subscribe to the Blue Sky forecast service, then he is in the situation already analyzed, where he will choose to harvest early and earn an expected profit of $30,000. Now suppose Cab does subscribe to Blue Sky, and has their forecast available. In this case, he can do his expected profit calculation conditioned on the forecast. To analyze his options, Cab first calculates the conditional probabilities of early rain, given the forecast:

$$\text{Prob}(\text{Wet}\,|\,\text{Forecast Early}) = 0.3/(0.3+0.2) = 0.6$$
$$\text{Prob}(\text{Dry}\,|\,\text{Forecast Late}) = 0.1/(0.1+0.4) = 0.2\,.$$

The expected profit from harvesting early is again \$30,000, no matter what the forecast. Now consider the expected profit from harvesting late. If the forecast is for early rains, the expected profit is given by weighting the outcomes by their conditional probabilities given the forecast, or

$$(\$10{,}000)\cdot(0.6) + (\$40{,}000)\cdot(0.4) = \$22{,}000.$$

This is less than \$30,000, so Cab will definitely harvest early in response to a forecast of early rain. Next suppose the forecast is for late rain. Again, the expected profit is given by weighting the outcomes by their conditional probabilities given this information,

$$(\$10{,}000)\cdot(0.2) + (\$40{,}000)\cdot(0.8) = \$34{,}000.$$

This is greater than \$30,000, so Cab will harvest late if the forecast is for late rain.

Is subscribing to Blue Sky worth while? If Cab does not, then he will always harvest early and his expected profit is \$30,000. If Cab does subscribe, then his expected profit in the event of an early rain forecast is \$30,000 and in the event of a late rain forecast is \$34,000. Since the frequency of an early rain forecast is 0.5, Cab's overall expected profit if he subscribes is

$$(\$30{,}000)\cdot(0.5) + (\$34{,}000)\cdot(0.5) = \$32{,}000.$$

This is \$2000 more than the expected profit if Cab does not subscribe, so that the value of the information provided by the subscription is \$2000. This is more than the \$1000 cost of the information, so Cab will choose to subscribe and will earn an overall expected profit, net of the subscription cost, of \$31,000.

Cab Franc's decision problem is a typical one for an economic agent facing uncertainty. He has a criterion (expected profit) to be optimized, a "model" of the probabilities of various outcomes, the possibility of collecting data (the forecast) to refine his probability model, and actions that will be based on the data collected. An econometrician facing the problem of statistical inference is in a similar situation: There is a "model" or "hypothesis" for an economic phenomenon, data that provides information that can be used to refine the model, and a criterion to be used in determining an action in response to this information. The actions of the econometrician, to declare a hypothesis "true" or "false", or to make a forecast, are similar in spirit to Cab's choice. Further, the solution to the econmetrician's inference problem will be similar to Cab's solution.

A textbook definition of econometrics is *the application of the principles of statistical inference to economic data and hypotheses*. However, Cab Franc's problem suggests a deeper connection between econometric analysis and economic behavior. The decision problems faced by rational economic agents in a world with imperfect information require statistical inference, and thus are "econometric" in nature. The solutions to these problems require the same logic and techniques that must be brought to bear more formally in scientific inference. Thus, all rational economic agents

are informal working econometricians, and the study of formal econometrics can provide at least prescriptive models for economic behavior. Turned around, econometrics is simply a codification of the "folk" techniques used by economic agents to solve their decision problems. Thus, the study of econometrics provides not only the body of tools needed in empirical and applied economics for data analysis, forecasting, and inference, but also key concepts needed to explain economic behavior.

1.3. STOCK MARKET EFFICIENCY

The hypothesis is often advanced that the stock market is *efficient*. Among the possible meanings of this term is the idea that arbitragers are sufficiently active and pervasive so that potential windows of opportunity for excess profit are immediately closed. Consider a broad-based stock market index, the New York Stock Exchange (NYSE) value-weighted index of the prices of all the stocks listed with this exchange. The gross return to be made by taking a dollar out of a "risk-free" channel (defined here to be 90-day U.S. Treasury Bills), buying a dollar's worth of the "market" stock portfolio, and selling it one day later is $g_t = \log(M_t/M_{t-1})$, where M_t equals the NYSE index on day t, and the log gives the one-day exponential growth rate in share price. It is necessary in general to account for distributions (dividends) paid by the stocks during the time they are held; this is done automatically when g_t is reported. An arbitrager on day t-1 knows the history of the market and economic variables up through that day; let $\boldsymbol{H}_{t-1}$ denote this history. In particular, $\boldsymbol{H}_{t-1}$ includes the level M_{t-1} of the index, the pattern of historical market changes, and the overnight interest rate i_{t-1} on 90-day Treasury Bills, representing the opportunity cost of not keeping a dollar in a T-Bill account. The difference $R_t = g_t - i_{t-1}$ is the profit an arbitrager makes by buying one dollar of the NYSE portfolio on t-1, and is called the *excess return* to the market on day t. (If the arbitrager sells rather than buys a dollar of the exchange index in t-1, then her profit is $-R_t$). On day t-1, conditioned on the history $\boldsymbol{H}_{t-1}$, the excess return R_t is a random variable, and the probability that it is less than any constant r is given by a cumulative distribution function $F(r|\boldsymbol{H}_{t-1})$. Then, the expected profit is $\int_{-\infty}^{+\infty} r \cdot F'(r|\boldsymbol{H}_{t-1})dr$. The argument is that if this expected profit is positive, then arbitragers will buy and drive the price M_{t-1} up until the opportunity for positive expected profit is eliminated. Conversely, if the expected profit is negative, arbitragers will sell and drive the price M_{t-1} down until the expected profit opportunity is eliminated. Then, no matter what the history, arbitrage should make expected profit zero. This argument does not take account of the possibility that there may be some trading cost, a combination of transactions charges, risk preference, the cost of acquiring information, and the opportunity cost of the time the arbitrager spends trading. These trading costs could be sufficient to allow a small positive expected excess return to persist; this is in fact observed, and is called the *equity premium*. However, even if there is an equity premium, the arbitrage argument implies that the expected excess return should be independent of history.

From the CRSP financial database and from the Federal Reserve Bank, we take observations on g_t and i_{t-1} for all days the market is open between January 2, 1968 and December 31, 1998, a total of 7806 observations. We then calculate the excess return $R_t = g_t - i_{t-1}$. The next table gives some statistics on these quantities:

Variable	Sample Average	Sample Standard Deviation
g_t	0.0514%	0.919%
i_{t-1}	0.0183%	0.007%
R_t	0.0331%	0.919%

These statistics imply that the annual return in the NYSE index over this period was 18.76 percent, and the annual rate of interest on 90-day Treasury Bills was 6.68 percent.

Now consider the efficient markets hypothesis, which we have argued should lead to excess returns that do not on average differ from one previous history to another. The table below shows the sample average excess return in the NYSE index under each of two possible conditions, a positive or a negative excess return on the previous day.

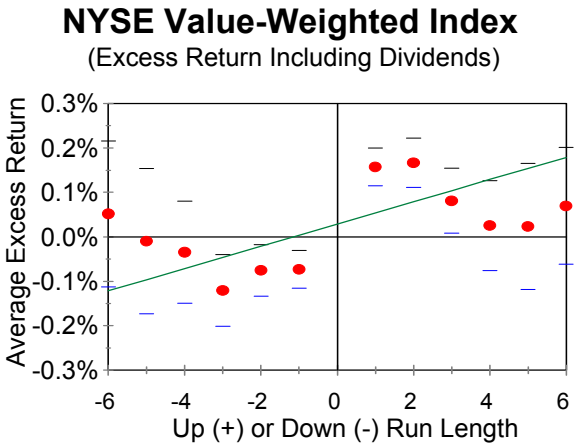
Condition	Frequency	Sample Average	Standard Error
$R_{t-1} > 0$	4082 (52.3%)	0.129%	0.013%
$R_{t-1} < 0$	3723 (47.7%)	-0.072%	0.016%

We conclude that excess return on a day following a positive excess return is on average positive, and on a day following a negative excess return is on average negative. The standard errors measure the precision of the sample averages. These averages are sufficiently precise so that we can say that the difference in sample averages did not arise by chance, and the expected excess returns under the two conditions are different.[1] We conclude that over the time period covered by our sample, the efficient markets hypothesis fails. There was a persistence in the market in which good days tended to be followed by better-than-average days, and bad days by worse-than-average days. There appears to have been a potentially profitable arbitrage strategy, to buy on up days and sell on down days, that was not fully exploited and eliminated by the arbitragers.

Define a *positive (negative) run of length* n to be a sequence of n successive days on which the excess return is positive (negative), with a day on each end on which the sign of excess return is

[1]The mean difference is 0.202% with a standard error of 0.010%, and the T-statistic for the hypothesis that the two conditional means is the same is 19.5.

reversed. The figure below plots the average excess returns conditioned on the length of an up (+) or down (-) run over the immediate past days. The horizontal lines above and below each point give an indication of the accuracy with which it is estimated; these are 95 percent confidence bounds. The figure suggests that expected returns are positive following up runs and negative following down runs. A straight line is fitted through these points, taking into account the accuracy with which they are estimated, by the least squares regression method. This is also plotted in the figure, and shows that the average returns have a positive trend. The data suggest that the trend line may overstate the impact of longer run lengths, and longer runs may be less predictive. Nevertheless, the figure supports our previous conclusion that the efficient markets hypothesis fails.



The preceding evidence suggests that arbitragers may underestimate the persistence of up and down runs, and fail to make profitable bets on persistence. The table below gives the observed counts of numbers of up and down runs of various lengths. It also includes a prediction of the numbers of runs of various lengths that you would expect to see if runs are the result of independent "coin tosses", with the probability of an "up" outcome equal to the 52.4% frequency with which the excess return was positive in our sample.[2] What one sees is that there are many fewer runs of length one and more runs of longer lengths than the "coin tossing" model predicts. There is the possibility that the differences in this table are the result of chance, but a statistical analysis using what is called a likelihood ratio test shows that the pattern we see is very unlikely to be due to chance. Then, up and down runs are indeed more persistent than one would predict if one assumed that the probability of a up or down day was independent of previous history.

[2] If P is the probability of an up day, then $P^{n-1}(1-P)$ is the probability of an up run of length n, and this multiplied by the total number of positive runs is the expected number of length n. An analogous formula applies for negative runs.

Run Length	Observed Positive	Observed Negative	Expected Positive	Expected Negative
1	1772	1772	1940.6	1938.0
2	1037	954	1015.1	924.3
3	602	495	531.0	440.8
4	313	245	277.7	210.2
5	159	120	145.3	100.3
6 or more	186	119	159.3	91.4
Total	4069	3705	4069	3705

This example shows how an economic hypothesis can be formulated as a condition on a probability model of the data generation process, and how statistical tools can be used to judge whether the economic hypothesis is true. In this case, the evidence is that either the efficient markets hypothesis does not hold, or that there is a problem with one of the assumptions that we made along the way to facilitate the analysis.[3] A more careful study of the time-series of stock market prices that was done above tends to support one aspect of the efficient markets hypothesis, that expected profit at time t-1 from an arbitrage to be completed the following period is zero. Thus, the elementary economic idea that arbitragers discipline the market is supported. However, there do appear to be longer-run time dependencies in the market, as well as heterogeneities, that are inconsistent with some stronger versions of the efficient markets hypothesis.

1.4. THE CAPITAL ASSET PRICING MODEL

The return that an investor can earn from a stock is a random variable, depending on events that impinge on the firm and on the economy. By selecting the stocks that they hold, investors can trade off between average return and risk. A basic, and influential, theory of rational portfolio selection is the Capital Asset Pricing (CAP) model. This theory concludes that if investors are concerned only with the mean and variance of the return from their portfolio, then there is a single portfolio of stocks that is optimal, and that every rational investor will hold this portfolio in some mix with a riskless asset to achieve the desired balance of mean and variance. Since every investor, no matter what her attitudes to risk, holds the same stock portfolio, this portfolio will simply be a share of the total market; that is, all investors simply hold a market index fund. This is a powerful conclusion, and

[3] For example, stock prices historically have been adjusted in units of 1/8 of a dollar, rather than to exact market-clearing levels. Some day-to-day variations reflect this institutional precularity.

one that appears to be easily refutable by examining the portfolios of individuals. This suggests that other factors, such as irrationality, transactions costs, heterogeneity in information, or preferences that take into account risk features other than mean and variance, are influencing behavior. Nevertheless, the CAP model is often useful as a normative guide to optimal investment behavior, and as a tool for understanding the benefits of diversification.

To explain the CAP model, consider a market with K stocks, indexed k = 1,2,...,K. Let P_{kt} be the price of stock k at the end of month t; this is a *random variable* when considered before the end of month t, and after that it is a number that is a *realization* of this random variable. Suppose an investor can withdraw or deposit funds in an account that holds U.S. Treasury 30-Day Bills that pay an interest rate i_{t-1} during month t. (The investor is assumed to be able to borrow money at this rate if necessary.) Conventionally, the T-Bill interest rate is assumed to be risk-free and known to the investor in advance. The profit, or *excess return*, that the investor can make from withdrawing a dollar from her T-Bill account and buying a dollar's worth of stock k is given by

$$R_{kt} = \frac{P_{kt} + d_{k,t-1} - P_{k,t-1}}{P_{k,t-1}} - i_{t-1},$$

where $d_{k,t-1}$ is the announced dividend paid by the stock at the end of month t. The excess return R_{kt} is again a random variable. Let r_k denote the *mean* of R_{kt}. Let σ_k^2 denote its variance, and let σ_{kj} denote the covariance of R_{kt} and R_{jt}. Note that σ_{kk} and σ_k^2 are two different notations for the same variance. The square root of the variance, σ_k, is called the standard deviation.

Consider an investor's portfolio of value A^* at the beginning of a month, and suppose A dollars are invested in stocks and A^*-A dollars are held in the risk-free account. Many investors will have $0 \le A \le A^*$. However, it is possible to have $A > A^*$, so that the investor has borrowed money (at the risk-free rate) and put this into stocks. In this case, the investor is said to have purchased stocks *on margin*. For the all-stock component of the portfolio, a fraction θ_k of each dollar in A is allocated to shares of stock k, for k = 1,...,K. The excess return to this portfolio is then AR_{pt}, where

$$R_{pt} = \sum_{k=1}^{K} \theta_k R_{kt}$$

is the excess return to the one dollar stock portfolio characterized by the shares $(\theta_1,...,\theta_K)$. The fractions θ_k are restricted to be non-negative. However, the list of "stocks" may also include *financial derivatives*, which can be interpreted as lottery tickets that pay off in dollars or stocks under specified conditions. For example, an investor may *short* a stock, which means that she in effect sells an IOU promising to deliver a share of the stock at the end of the month. She is then obligated to deliver a share of the stock at the end of the month, if necessary by buying a share of this stock to deliver in order to complete the transaction. Other elementary examples of financial derivatives are *futures*, which are contracts to deliver stocks at some future date, and *mutual funds*, which are institutions that sell shares and use the proceeds to buy portfolios of stocks. There are also more complex financial derivatives that require delivery under specified conditions, such as an increase

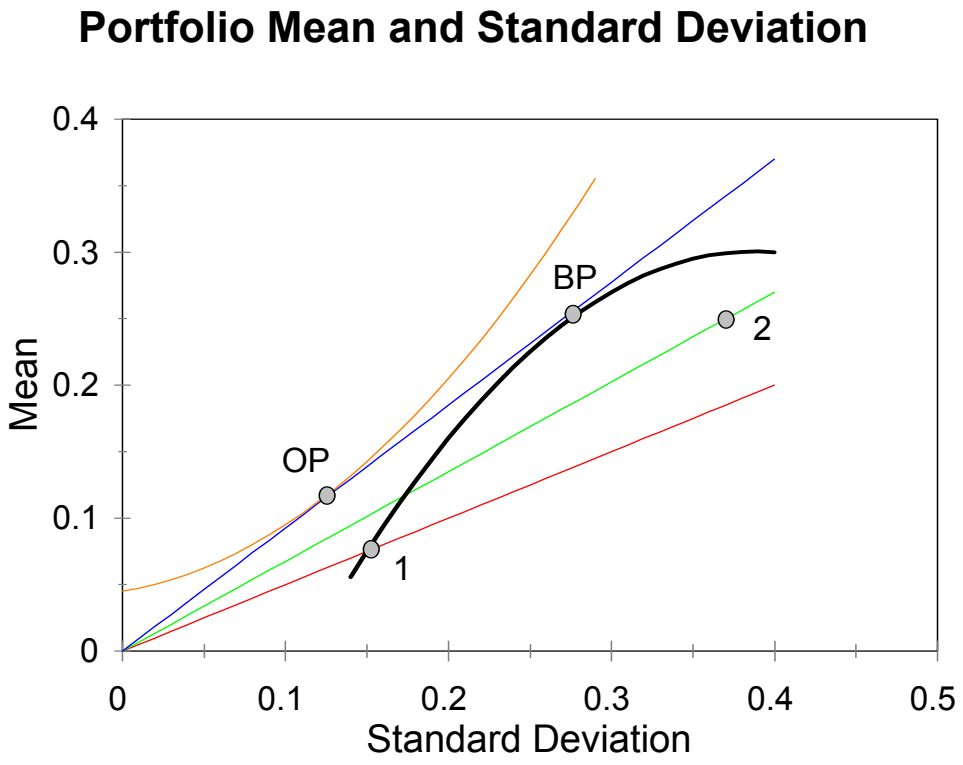
in a stock market index of more than a specified percentage. The excess return R_{pt} is again a random variable, with mean $r_p = \sum_{k=1}^{K} \theta_k r_{kt}$ and variance $\sigma_p^2 = \sum_{k=1}^{K}\sum_{j=1}^{K} \theta_k\theta_j\sigma_{kj}$. This implies that the stock portfolio with A dollars invested has an excess return with mean Ar_p and variance $A^2\sigma_p^2$ (or standard deviation $A\sigma_p$). The covariance of R_{kt} and R_{pt} is given by $\sigma_{kp} \equiv \text{cov}(R_{kt},R_{pt}) = \sum_{j=1}^{K} \theta_j\sigma_{kj}$. Define the *beta* of stock k (with respect to the stock portfolio p) by the formula $\beta_k = \sigma_{kp}/\sigma_p^2$, and note that

$$\sigma_p^2 = \sum_{k=1}^{K} \theta_k \left(\sum_{j=1}^{K} \theta_j\sigma_{kj} \right) = \sum_{k=1}^{K} \theta_k\sigma_{kp} = \sum_{k=1}^{K} \theta_k\beta_k\sigma_p^2,$$

and hence that $\sum_{k=1}^{K} \theta_k\beta_k = 1$.

Now consider the rational investor's problem of choosing the level of investment A and the portfolio mix $\theta_1,\theta_2,...,\theta_K$. Assume that the investor cares only about the mean and standard deviation of the excess return from her portfolio, and always prefers a higher mean and a lower standard deviation. The investor's tastes for risk will determine how mean and standard deviation are traded off. The figure below shows the alternatives available to the investor. The investor prefers to be as far to the northwest in this figure as possible, where mean is high and standard deviation is low, and will have indifference curves that specify her tradeoffs between mean and standard deviation. (Note that an investor with low risk aversion will have indifference curves that are almost horizontal, while an extremely risk-averse investor will have indifference curves that are almost vertical.) If preferences between mean and standard deviation are derived from a utility function that is concave in consumption, then the indifference curve will be convex. A specified mix $(\theta_1^1,...,\theta_K^1)$ of stocks determines a one-dollar all-equity portfolio with particular values for standard deviation and mean, and an all-stock portfolio of value A^* invested in this mixture will have a mean and standard deviation that are A^* times those of the one-dollar portfolio. This point is indicated in the diagram as Portfolio 1. By holding A of the all-equity portfolio and $A^* - A$ of the riskless asset in some combination, the investor can attain mean and standard deviation combinations anywhere on a straight line through the origin and the all-equity Portfolio 1 point. (Points between zero and the all-equity portfolio correspond to investing only part of the investor's assets in stocks, while points on the line to the right of the Portfolio 1 point correspond to borrowing money to take a margin position in stocks.) Another mix $(\theta_1^2,...,\theta_K^2)$ gives the point in the diagram indicated as Portfolio 2. Again, by holding the riskless asset and Portfolio 2 in some combination, the consumer could attain any point on the straight line connecting the origin and Portfolio 2. There will be a frontier envelope on the north-west boundary of all the mean and standard deviation combinations that can be attained

with all-equity portfolios with value A^*; this envelope is drawn as a heavy curve in the figure, and represents *efficient* portfolios in terms of tradeoffs between mean and standard deviation in all-equity portfolios.



Portfolio 1 is efficient, while Portfolio 2 is not because it is southwest of the all-equity portfolio frontier. Note however that the consumer can be made better off with a portfolio that is some combination of Portfolio 2 and the riskless asset than with any combination of Portfolio 1 and the riskless asset, because the line through Portfolio 2 is always northwest of the line through Portfolio 1. Consider all the lines that connect the origin and all-equity portfolios. The location of these lines reflects the operation of diversification to reduce risk; i.e., by holding a mix of stocks, some of which are likely to go up when others are going down, one may be able to reduce standard deviation for a given level of the mean. There will be an *efficient* mix $(\theta_1^*,...,\theta_K^*)$, labeled BP (for Best Portfolio) in the figure, that gives a line that is rotated as far to the northwest as possible. No matter what the specific tastes of the investor for mean versus standard deviation, she will maximize preferences

somewhere alone this tangency line, using some combination of the riskless asset and the Best Portfolio. The diagram shows for a particular indifference curve how the optimal portfolio, labeled OP in the figure, is determined. Different investors will locate at different points along the optimal line, by picking different A levels, to maximize their various preferences for mean versus standard deviation. However, all investors will choose exactly the same BP mix $(\theta_1^*,...,\theta_K^*)$ for the stocks that they hold. But if every investor holds stocks in the same proportions, then these must also be the proportions that prevail in the market as a whole. Then the Best Portfolio $(\theta_1^*,...,\theta_K^*)$ will be the shares by value of all the stocks in the market. Such a portfolio is called a *market index fund*. The CAP model then concludes that rational investors who care only about the mean and standard deviation of excess return will hold only the market index fund, with the levels of investment reflecting their heterogeneous tastes for risk. Investors may purchase individual stocks in the BP proportions; however, there are mutual funds that do precisely this, and the investor may then simply put her stock portfolio into the market index mutual fund.

The problem of determining the optimal portfolio mix $(\theta_1^*,...,\theta_K^*)$ is most easily solved by considering a closely related problem. An investor's portfolio is characterized by A and $(\theta_1,...,\theta_K)$. Given a choice among all the portfolios that achieve a specified level of mean return, the investor would want to choose the one that minimizes variance. In the figure, this corresponds to getting as far to the left as possible when constrained to the feasible portfolios on a specified horizontal line. From the previous discussion, the solution to this problem will be a portfolio with the optimal mix of stocks, and the only difference between this problem and the one of maximizing preferences will be in determining the overall investment level A. The problem of minimizing variance for a given mean is one of constrained minimization:

$$\text{Choose } A,\theta_1,...,\theta_K \geq 0 \text{ to minimize } A^2 \sum_{k=1}^{K} \sum_{j=1}^{K} \theta_k\theta_j\sigma_{kj}, \text{subject to } A \sum_{k=1}^{K} \theta_k r_k = c \text{ and } \sum_{k=1}^{K} \theta_k$$

= 1, where c is a constant that can be varied parametrically. The first-order (Kuhn-Tucker) conditions for this problem are

$$(1) \qquad 2A^2 \sum_{j=1}^{K} \theta_j^*\sigma_{kj} \geq \lambda A r_k + \mu, \text{ with equality unless } \theta_k^* = 0, \text{ for } k = 1,...,K$$

$$(2) \qquad 2A \sum_{k=1}^{K} \sum_{j=1}^{K} \theta_k^*\theta_j^*\sigma_{kj} = \lambda \sum_{k=1}^{K} \theta_k^* r_k,$$

where the scalars λ and μ are Lagrange multipliers. Multiply (1) by θ_k^* and sum over k to obtain the result

$$A\left(2A\sum_{k=1}^{K}\sum_{j=1}^{K}\theta_k^*\theta_j^*\sigma_{kj} - \lambda\sum_{k=1}^{K}\theta_k^* r_k\right) = \mu \sum_{k=1}^{K} \theta_k^*.$$

Using (2), this implies $\mu = 0$. Then, (1) implies that the optimal θ_k^* satisfy

(3) $$\sum_{j=1}^{K} \theta_j^* \sigma_{kj} \geq \gamma r_k, \text{ with equality unless } \theta_k^* = 0, \text{ for } k = 1,...,K,$$

where γ is a scalar defined so that the θ_k^* sum to one. Since by the earlier comments this mix is simply the mix in the total market, equality will hold for all stocks that are in the market and have positive value.

Assume now that $r_p = \sum_{k=1}^{K} \theta_k^* r_k$ and $\sigma_p^2 = \sum_{k=1}^{K} \sum_{j=1}^{K} \theta_k^* \theta_j^* \sigma_{kj}$ refer to the best portfolio. The left-hand-side of (3) equals $\sigma_{kp} \equiv \beta_k \sigma_p^2$, so this condition can be rewritten

(4) $$\beta_k \sigma_p^2 \geq \gamma r_k, \text{ with equality unless } \theta_k^* = 0, \text{ for } k = 1,...,K.$$

Multiplying both sides of this inequality by θ_k^* and summing yields the condition

$$\gamma r_p \equiv \gamma \sum_{k=1}^{K} \theta_k^* r_k = \sigma_p^2 \sum_{k=1}^{K} \theta_k \beta_k = \sigma_p^2,$$

or $r_p = \sigma_p^2/\gamma$. Substituting this into (4) gives us the final form of a main result of the CAP model, $r_k \leq \beta_k r_p$, with equality if the stock is held, for $k = 1,...,K$.

The mean returns are not observed directly, but the realizations of monthly returns on individual stocks and the market are observed. Write an observed return as the sum of its mean and a deviation from the mean, $R_{kt} = r_k + \varepsilon_{kt}$ and $R_{pt} = r_p + \varepsilon_{pt}$. Note that $R_{pt} = \sum_{k=1}^{K} \theta_k^* R_{kt}$, so then $\varepsilon_{pt} = \sum_{k=1}^{K} \theta_k^* \varepsilon_{kt}$.

For all stocks held in the market, the CAP model implies $r_k = \beta_k r_p$. Use the form $R_{kt} = r_k + \varepsilon_{kt}$ to rewrite the equation $r_k = \beta_k r_p$ as $R_{kt} - \varepsilon_{kt} = \beta_k (R_{pt} - \varepsilon_{pt})$. Define $\nu_{kt} = \varepsilon_{kt} - \beta_k \varepsilon_{pt}$. Then, the equation becomes

(5) $$R_{kt} = \beta_k R_{pt} + \nu_{kt}.$$

This equation can be interpreted as a relation between *market risk*, embodied in the market excess return R_{pt}, and the risk of stock k, embodied in R_{kt}. The *disturbance* ν_{kt} in this equation is sometimes called the *specific risk* in stock k, the proportion of the total risk in this stock that is not responsive to market fluctuations. This disturbance has the following properties:

(a) $$\mathbf{E}\nu_{kt} = 0;$$

(b) $$\mathbf{E}\nu_{kt}^2 = \mathbf{E}(\varepsilon_{kt} - \beta_k\varepsilon_{pt})^2 = \sigma_k^2 + \beta_k^2\sigma_p^2 - 2\beta_k\sigma_{kp} \equiv \sigma_k^2 - \beta_k^2\ \sigma_p^2$$

(c) $$\mathbf{E}\nu_{kt}R_{pt} = \mathbf{E}(\varepsilon_{kt} - \beta_k\varepsilon_{pt})\varepsilon_{pt} = \sigma_{kp} - \beta_k\sigma_p^2 = 0.$$

Equation (5) is a *linear regression* formula. Properties (a)-(c) are called the *Gauss-Markov conditions*. We will see that they imply that an estimate of β_k with desirable statistical properties can be obtained by using the *method of least squares*. Then, the CAP model's assumptions on behavior imply an econometric model that can then be fitted to provide estimates of the *market betas*, key parameters in the CAP analysis.

The market beta's of individual stocks are often used by portfolio managers to assess the merits of adding or deleting stocks to their portfolios. (Of course, the CAP model says that there is no need to consider holding portfolios different than the market index fund, and therefore no need for portfolio managers. That these things exist is itself evidence that there is some deficiency in the CAP model, perhaps due to failures of rationality, the presence of transactions cost, or the ability to mimic the excess return of the market using various subsets of all the stocks on the market because the optimal portfolio is not unique.) Further, statistical analysis of the validity of the assumptions (a)-(c) can be used to test the validity of the CAP model.

The β's in formula (5) convey information on the relationship between the excess return on an individual stock and the excess return in the market. Subtract means in (5), square both sides, and take the expectation to get the formula

$$\sigma_k^2 = \beta_k^2\sigma_p^2 + \delta_k^2,$$

where $\delta_k^2 = \mathbf{E}\nu_{kt}^2$ is the variance of the disturbance. This equation says that the risk of stock k equals the market risk, amplified by β_k^2, plus the specific risk. Stock k will have high risk if it has large specific risk, or a β_k that is large in magnitude, or both. A positive β_k implies that events that influence the market tend to influence stock k in the same way; i.e., the stock is *pro-cyclic*. A negative β_k implies that the stock tends to move in a direction opposite that of the market, so that it is *counter-cyclic*. Stocks that have small or negative β_k are *defensive*, aiding diversification and making an important contribution to reducing risk in the market portfolio.

The CAP model described in this example is a widely used tool in economics and finance. It illustrates a situation in which *economic* axioms on behavior lead to a widely used *statistical* model for the data generation process, the linear regression model with errors satisfying Gauss-Markov assumptions. An important feature of this example is that these statistical assumptions are implied by the economic axioms, not attached *ad hoc* to facilitate data analysis. It is a happy, but rare, circumstance in which an economic theory and its econometric analysis are fully integrated. To seek such harmonies, theorists need to draw out and make precise the empirical implications of their work, and econometricians need to develop models and methods that minimize the need for

facilitating assumptions that make the statistical analysis tractable but whose economic plausibility is weak or undetermined.

1.5. CONCLUSION

A traditional view of econometrics is that it is the special field that deals with economic data and with statistical methods that can be employed to use these data to test economic hypotheses and make forecasts. If this were all there was to econometrics, it would still be one of the most important parts of the training of most economists, who in their professional careers deal with economic hypotheses, policy issues, and planning that in the final analysis depend on facts and the interpretation of facts. However, the examples in this chapter are intended to show that the subject of econometrics is far more deeply intertwined with economic science, providing tools for modeling core theories of the behavior of economic agents under uncertainty, and a template for rational decision-making under incomplete information. This suggests that it is useful to understand econometrics at three levels: (1) the relatively straightforward and mechanical procedures for applied econometric data analysis and inference that are needed to understand and carry out empirical economics research, (2) a deeper knowledge of the theory of statistics in the form that it is needed to develop and adapt econometric tools for the situations frequently encountered in applied work where conventional techniques may not apply, or may not make efficient use of the data, and (3) a deeper knowledge of the concepts and formal theory of probability, statistics, and decision theory that enter models of the behavior of economic agents, and show the conceptual unity of econometrics and economic theory.

1.6. EXERCISES

Questions 1-3 refer to the decision problem of Cab Franc, from Section 1.2.

1. The Department of Agriculture offers Cab crop insurance, which costs $1000. If Cab's revenue from selling grapes falls below 90 percent of his expected net revenue of $31,000, then the insurance reimburses him for the difference between his actual revenue and 90 percent of $31,000. Is this insurance actuarially fair if Cab makes the same operating decisions that he did without insurance? Will Cab in fact make the same operating decisions if he has the insurance? Will he buy the insurance?

2. Blue Sky introduces a more detailed forecast, with the properties described below, at a cost of $1500. Will Cab buy this forecast rather than the previous one? For this problem, assume there is no crop insurance.

Frequency of Expanded Forecasts and Outcomes					
	Blue Sky Forecasts				TOTAL
Event of Nature	Bad	Poor	Fair	Good	
Wet	0.15	0.15	0.1	0.0	0.4
Dry	0.05	0.15	0.2	0.2	0.6
TOTAL	0.2	0.3	0.3	0.2	1.0

3. After reviewing his budget commitments, Cab decides he is not risk neutral. If his net income from selling grapes falls below his mortgage payment of $25,000, then he will have to borrow the difference in the local spot market, and the "vig" will double the cost. Put another way, his utility of income Y is $u = Y - 0.5 \cdot \max(25000-Y,0)$. What harvesting decision will maximize Cab's utility if he subscribes to the Blue Sky forecast? Again assume there is no crop insurance, but assume the expanded Blue Sky forecasts in question 2 are available.

Questions 4-5 refer to the market efficiency analysis in Section 1.3.

4. Formulating the expected markets hypothesis as a feature produced by large numbers of active arbitragers assumes that there is an objective probability distribution for outcomes, and this distribution is common knowledge to all arbitragers. What would you expect to happen if the probabilities are subjective and are not the same for all arbitragers? If the number of arbitragers is limited? If institutional constraints place limits on the magnitudes of the positions that arbitragers can take? If arbitragers are risk-averse rather than risk neutral? (This is a thought question, as you do not yet have the tools for a formal analysis.)

5. The narrow form of the efficient markets hypothesis states that arbitragers will buy or sell to take advantage of expected profits or losses, and thereby arbitrage away these profits and losses, so that expected returns are zero. Broader forms of the efficient markets hypothesis state that arbitragers will create *derivatives*, and trade away any expected profit for any specified position they could take in the market for these derivatives. For example, a *Call* (or Call option) is an agreement between the buyer and the seller of the option whereby the buyer obtains the right but not the obligation to *buy* an agreed amount of a stock at a pre-agreed price (the strike price), on an agreed future date (the value date). Conversely, the seller of a Call has the obligation, but not the right, to sell the agreed amount of a stock at the pre-agreed price. Obviously, the holder of a Call will exercise it if the price on the value date exceeds the strike price, and will otherwise let the option expire. The price of a Call includes a premium that reflects its value as insurance against an increase in the price of the stock, and is determined by the beliefs of buyers and sellers about the probabilities of prices above the strike price. A *Put* option is in all respect the same except that it confers the right but not the obligation to *sell* at a pre-agreed rate. Suppose the cumulative probability distribution $F(P_t|P_{t-1})$ for the price of a stock on a value date t, given the current price at t-1, is known to all investors. Derive a formula for the price of a Call option at strike price P* if all investors are risk neutral.

Questions 6-8 refer to the Capital Asset Pricing Model in Section 1.4.

6. The table below gives the probability distribution of next month's price for three stocks, each of which has a current price of $100. There are no declared dividends in this month. The risk-free interest rate for the month is 0.05. Calculate the mean excess return for each stock, and the variances and covariances of their excess returns.

Probability	Stock A	Stock B	Stock C
0.25	$120	$130	$140
0.2	$110	$110	$100
0.2	$110	$90	$110
0.25	$100	$100	$90
0.1	$90	$100	$100

7. If the market is composed of the three stocks described in Question 6, with an equal number of shares of each stock in the market, calculate the excess return and variance of the market. Calculate the beta of each stock.

8. To derive the CAP model description of the optimal portfolio, instead of minimizing variance subject to a constraint on the expected rate of return, maximize the expected rate of return subject to a constraint on the variance. Give an interpretation of the Lagrange multiplier in this formulation. Show that it leads to the same characterization of the optimal portfolio as before.

CHAPTER 2. ANALYSIS AND LINEAR ALGEBRA IN A NUTSHELL

2.1. SOME ELEMENTS OF MATHEMATICAL ANALYSIS

2.1.1. Real numbers are denoted by lower case Greek or Roman numbers; the space of real numbers is the real line, denoted by $\mathbb{R}$. The *absolute value* of a real number a is denoted by $|a|$. Complex numbers are rarely required in econometrics before the study of time series and dynamic systems. For future reference, a complex number is written $a + \iota b$, where a and b are real numbers and ι is the square root of -1, with a termed the *real* part and ιb termed the *imaginary* part. The complex number can also be written as $r(\cos\theta + \iota \sin\theta)$, where $r = (a^2+b^2)^{1/2}$ is the *modulus* of the number and $\theta = \cos^{-1}(a/r)$. The properties of complex numbers we will need in basic econometrics are the rules for sums, $(a+\iota b) + (c+\iota d) = (a+c)+\iota(b+d)$, and products, $(a+\iota b)\cdot(c+\iota d) = (ab-cd)+\iota(ad+bc)$.

2.1.2. For sets of objects **A** and **B**, the *union* $\mathbf{A}\cup\mathbf{B}$ is the set of objects in either; the *intersection* $\mathbf{A}\cap\mathbf{B}$ is the set of objects in both; and $\mathbf{A}\backslash\mathbf{B}$ is the set of objects in **A** that are not in **B**. The empty set is denoted $\boldsymbol{\varphi}$. Set inclusion is denoted $\mathbf{A} \subseteq \mathbf{B}$; we say **A** *is contained in* **B**. The complement of a set **A** (which may be relative to a set **B** that contains it) is denoted $\mathbf{A}^c$. A family of sets is *disjoint* if the intersection of each pair is empty. The symbol $a \in \mathbf{A}$ means that a is a member of **A**; and $a \notin \mathbf{A}$ means that a is not a member of **A**. The symbol $\exists$ means "there exists", the symbol $\forall$ means "for all", and the symbol $\ni$ means "such that". A proposition that A implies B is denoted "$A \Longrightarrow B$", and a proposition that A and B are equivalent is denoted "$A \Longleftrightarrow B$". The proposition that A implies B, but B does not imply A, is denoted "$A \mid\!\Longrightarrow B$". The phrase "if and only if" is often abbreviated to "iff".

2.1.3. A *function* $f:\mathbf{A} \to \mathbf{B}$ is a mapping from each object a in the *domain* **A** into an object $b = f(a)$ in the *range* **B**. The terms *function*, *mapping*, and *transformation* will be used interchangeably. The symbol $f(\mathbf{C})$, termed the *image* of **C**, is used for the set of all objects $f(a)$ for $a \in \mathbf{C}$. For $\mathbf{D} \subseteq \mathbf{B}$, the symbol $f^{-1}(\mathbf{D})$ denotes the *inverse image* of **D**: the set of all $a \in \mathbf{A}$ such that $f(a) \in \mathbf{D}$. The function f is *onto* if $\mathbf{B} = f(\mathbf{A})$; it is *one-to-one* if it is onto and if $a,c \in \mathbf{A}$ and $a \neq c$ implies $f(a) \neq f(c)$. When f is one-to-one, the mapping f^{-1} is a function from **B** onto **A**. If $\mathbf{C} \subseteq \mathbf{A}$, define the *indicator function* for **C**, denoted $\mathbf{1}_{\mathbf{C}}:\mathbf{A} \to \mathbb{R}$, by $\mathbf{1}_{\mathbf{C}}(a) = 1$ for $a \in \mathbf{C}$, and $\mathbf{1}_{\mathbf{C}}(a) = 0$ otherwise. The notation $\mathbf{1}(a\in\mathbf{C})$ is also used for the indicator function $\mathbf{1}_{\mathbf{C}}$. A function is termed *real-valued* if its range is $\mathbb{R}$.

2.1.4. The *supremum* of **A**, denoted sup **A**, is the least upper bound on **A**. A typical application has a function $f:\mathbf{C} \to \mathbb{R}$ and $\mathbf{A} = f(\mathbf{C})$; then $\sup_{c\in\mathbf{C}} f(c)$ is used to denote sup **A**. If the supremum is achieved by an object $d \in \mathbf{C}$, so $f(d) = \sup_{c\in\mathbf{C}} f(c)$, then we write $f(d) = \max_{c\in\mathbf{C}} f(c)$. When there is a unique maximizing argument, write $d = \operatorname{argmax}_{c\in\mathbf{C}} f(c)$. When there is a non-unique maximizing

argument; we will assume that $\text{argmax}_{c\in\mathbf{C}}$ f(c) is a *selection* of any one of the maximizing arguments. Analogous definitions hold for the infimum and minimum, denoted inf, min, and for argmin.

2.1.5. If a_i is a sequence of real numbers indexed by i = 1,2,..., then the sequence is said to have a *limit* (equal to a_o) if for each $\varepsilon > 0$, there exists n such that $|a_i - a_o| < \varepsilon$ for all $i \geq n$; the notation for a limit is $\lim_{i\to\infty} a_i = a_o$ or $a_i \to a_o$. The *Cauchy criterion* says that a sequence a_i has a limit if and only if, for each $\varepsilon > 0$, there exists n such that $|a_i - a_j| < \varepsilon$ for $i,j \geq n$. The notation $\text{limsup}_{i\to\infty} a_i$ means the limit of the supremum of the sets $\{a_i,a_{i+1},...\}$; because it is nonincreasing, it always exists (but may equal $+\infty$ or $-\infty$). An analogous definition holds for liminf.

2.1.6. A real-valued function $\rho(a,b)$ defined for pairs of objects in a set **A** is a *distance function* if it is non-negative, gives a positive distance between all distinct points of **A**, has $\rho(a,b) = \rho(b,a)$, and satisfies the triangle inequality $\rho(a,b) \leq \rho(a,c) + \rho(c,b)$. A set **A** with a distance function ρ is termed a *metric space*. A real-valued function ‖a‖ defined for objects in a set A is a *norm* if $\|a-b\|$ has the properties of a distance function. A typical example is the real line $\mathbb{R}$, with the absolute value of the difference of two numbers taken as the distance between them; then $\mathbb{R}$ is a metric space and a normed space. A (ε-)*neighborhood* of a point *a* in a metric space **A** is a set of the form $\{b\in\mathbf{A} | \rho(a,b) < \varepsilon\}$. A set $\mathbf{C} \subseteq \mathbf{A}$ is *open* if for each point in **C**, some neighborhood of this point is also contained in **C**. A set $\mathbf{C} \subseteq \mathbf{A}$ is *closed* if its complement is open. The *closure* of a set **C** is the intersection of all closed sets that contain **C**. The *interior* of **C** is the union of all open sets contained in **C**; it can be empty. A *covering* of a set **C** is a family of open sets whose union contains **C**. The set **C** is said to be *compact* if every covering contains a finite sub-family which is also a covering. A family of sets is said to have the *finite-intersection property* if every finite sub-family has a non-empty intersection. Another characterization of a compact set is that every family of closed subsets with the finite intersection property has a non-empty intersection. A metric space **A** is *separable* if there exists a countable subset **B** such that every neighborhood contains a member of **B**. All of the metric spaces encountered in econometrics will be separable. A sequence a_i in a separable metric space **A** is *convergent* (to a point a_o) if the sequence is eventually contained in each neighborhood of *a*; we write $a_i \to a_o$ or $\lim_{i\to\infty} a_i = a_o$ to denote a convergent sequence. A set $\mathbf{C} \subseteq \mathbf{A}$ is compact if and only if every sequence in **C** has a convergent subsequence (which converges to a *cluster point* of the original sequence).

2.1.7. Consider separable metric spaces **A** and **B**, and a function $f:\mathbf{A} \to \mathbf{B}$. The function f is *continuous* on **A** if the inverse image of every open set is open. Another characterization of continuity is that for any sequence satisfying $a_i \to a_o$, one has $f(a_i) \to f(a_o)$; the function is said to be continuous on $\mathbf{C} \subseteq \mathbf{A}$ if this property holds for each $a_o \in \mathbf{C}$. Stated another way, f is continuous on **C** if for each $\varepsilon > 0$ and $a \in \mathbf{C}$, there exists $\delta > 0$ such that for each *b* in a δ-neighborhood of *a*, f(*b*) is in a ε-neighborhood of f(*a*). For real valued functions on separable metric spaces, the concepts of supremium and limsup defined earlier for sequences have a natural extension: $\sup_{a\in\mathbf{A}} f(a)$ denotes

the least upper bound on the set $\{f(a)|a\in\mathbf{A}\}$, and $\limsup_{a\to b} f(a)$ denotes the limit as $\varepsilon \to 0$ of the suprema of $f(a)$ on ε-neighborhoods of b. Analogous definitions hold for inf and liminf. A real-valued function f is continuous at b if $\limsup_{a\to b} f(a) = \liminf_{a\to b} f(a)$. Continuity of real-valued functions f and g is preserved by the operations of absolute value $|f(a)|$, multiplication $f(a)\cdot g(a)$, addition $f(a)+g(a)$, and maximization $\max\{f(a),g(a)\}$ and minimization $\min\{f(a),g(a)\}$. The function f is *uniformly continuous* on **C** if for each $\varepsilon > 0$, there exists $\delta > 0$ such that for all $a \in \mathbf{C}$ and $b \in \mathbf{A}$ with b in a δ-neighborhood of a, one has $f(b)$ in a ε-neighborhood of $f(a)$. The distinction between continuity and uniform continuity is that for the latter a single $\delta > 0$ works for <u>all</u> $a \in \mathbf{C}$. A function that is continuous on a compact set is uniformly continuous. The function f is *Lipschitz* on **C** if there exist $L > 0$ and $\delta > 0$ such that $|f(b) - f(a)| \le L\cdot\rho(a,b)$ for all $a \in \mathbf{C}$ and $b \in \mathbf{A}$ with b in a δ-neighborhood of a.

2.1.8. Consider a real-valued function f on $\mathbb{R}$. The *derivative* of f at a_o, denoted $f'(a_o)$, $\nabla f(a_o)$, or $df(a_o)/da$, has the property if it exists that $|f(b) - f(a_o) - f'(a_o)(b-a_o)| \le \varepsilon(b-a_o)\cdot(b-a_o)$, where $\lim_{c\to 0} \varepsilon© = 0$. The function is *continuously differentiable* at a_o if f' is a continuous function at a_o. If a function is k-times continuously differentiable in a neighborhood of a point a_o, then for b in this neighborhood it has a *Taylor's expansion*

$$f(b) = \sum_{i=0}^{k} f^{(i)}(a_o)\cdot \frac{(b-a_o)^i}{i!} + \left\{f^{(k)}(\lambda b+(1-\lambda)a_o) - f^{(k)}(a_o)\right\} \cdot \frac{(b-a_o)^k}{k!},$$

where $f^{(i)}$ denotes the i-th derivative, and λ is a scalar between zero and one.

If $\lim_{i\to\infty} a_i = a_o$ and f is a continuous function at a_o, then $\lim_{i\to\infty} f(a_i) = f(a_o)$. One useful result for limits is *L'Hopital's rule*, which states that if f(1/n) and g(1/n) are functions that are continuously differentiable at zero with $f(0) = g(0) = 0$, so that f(n)/g(n) approaches the indeterminate expression 0/0, one has $\lim_{n\to\infty} f(n)/g(n) = f'(0)/g'(0)$, provided the last ratio exists.

2.1.9. If a_i for i = 0,1,2,... is a sequence of real numbers, the partial sums $s_n = \sum_{i=0}^{n} a_i$ define a *series*. We say the sequence is *summable*, or that the series is *convergent*, if $\lim_{n\to\infty} s_n$ exists and is finite. An example is the *geometric series* $a_i = r^i$, which has $s_n = (1-r^{n+1})/(1-r)$ if $r \ne 1$. When $|r| < 1$, this series is convergent, with the limit $1/(1-r)$. When $r < -1$ or $r \ge 1$, the series *diverges*. In the borderline case r = -1, the series alternates between 0 and 1, so the limit does not exist. Applying the Cauchy criterion, a summable sequence has $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} \sum_{i=n}^{\infty} a_i = 0$. A sequence satisfies a more general form of summability, called *Cesaro summability*, if $\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n} a_i$ exists. Summability implies Cesaro summability, but not vice versa. A useful result known as

Kronecker's lemma states that if a_i and b_i are positive series, b_i is monotonically increasing to $+\infty$, and $\sum_{i=0}^{n} a_i/b_i$ is bounded for all n, then $\lim_{n\to\infty} b_n^{-1} \sum_{i=0}^{n} a_i = 0$.

2.1.10. The exponential function e^a, also written $\exp(a)$, and natural logarithm $\log(a)$ appear frequently in econometrics. The exponential function is defined for both real and complex arguments, and has the properties that $e^{a+b} = e^a e^b$, $e^0 = 1$, and the Taylor's expansion $e^a = \sum_{i=0}^{\infty} \frac{a^i}{i!}$ that is valid for all a. The trigonometric functions $\cos(a)$ and $\sin(a)$ are also defined for both real and complex arguments, and have Taylor's expansions $\cos(a) = \sum_{i=0}^{\infty} \frac{(-1)^i a^{2i}}{(2i)!}$, and $\sin(a) = \sum_{i=0}^{\infty} \frac{(-1)^i a^{2i+1}}{(2i+1)!}$. These expansions combine to show that $e^{a+\iota b} = e^a(\cos(b) + \iota\cdot\sin(b))$. The logarithm is defined for positive arguments, and has the properties that $\log(1) = 0$, $\log(a\cdot b) = \log(a) + \log(b)$, and $\log(e^a) = a$. It has a Taylor's expansion $\log(1+a) = \sum_{i=1}^{\infty} a^i$, valid for $|a| < 1$. A useful bound on logarithms is that for $|a| < 1/3$ and $|b| < 1/3$, $|\mathrm{Log}(1+a+b) - a| < 4|b| + 3|a|^2$. Another useful result, obtained by applying L'Hopital's rule to the expression $\log(1+a_n/n)/(1/n)$, is that $\lim_{n\to\infty} (1+a_n/n)^n = \exp(a_0)$ when $\lim_{n\to\infty} a_n = a_0$ exists.

A few specific series appear occasionally in probability theory. The series $a_i = i^\alpha$ for i = 1,2,... is summable for $\alpha < -1$, and divergent otherwise, with $s_n = n(n+1)/2$ for $\alpha = 1$, $s_n = n(n+1)(2n+1)/6$ for $\alpha = 2$, and $s_n = n^2(n+1)^2/4$ for $\alpha = 3$. Differentiating the formula $s_n = (1-r^{n+1})/(1-r)$ for a convergent geometric series leads to the expressions $\sum_{i=1}^{\infty} i\cdot r^i = r/(1-r)^2$ and $\sum_{i=1}^{\infty} i^2\cdot r^i = r(1+r)/(1-r)^3$.

2.1.11. If a_i and b_i are real numbers and c_i are non-negative numbers for i = 1,2,..., then *Holder's Inequality* states that for p > 0, q > 0, and 1/p + 1/q = 1, one has

$$|\sum_i c_i\cdot a_i\cdot b_i| \le \sum_i c_i\cdot|a_i\cdot b_i| \le \left(\sum_i c_i\cdot|a_i|^p\right)^{1/p}\left(\sum_i c_i\cdot|b_i|^q\right)^{1/q}.$$

When p = q = 1/2, this is called the *Cauchy-Schwartz inequality*. Obviously, the inequality is useful only if the sums on the right converge. The inequality also holds in the limiting case where sums are replaced by integrals, and a(i), b(i), and c(i) are functions of a continuous index i.

2.2. VECTORS AND LINEAR SPACES

2.2.1. A finite-dimensional linear space is a set such that (a) linear combinations of points in the set are defined and are again in the set, and (b) there is a finite number of points in the set (a *basis*)

such that every point in the set is a linear combination of this finite number of points. The *dimension* of the space is the minimum number of points needed to form a basis. A point **x** in a linear space of dimension n has a *ordinate representation* $\mathbf{x} = (x_1,x_2,...,x_n)$, given a *basis* for the space $\{\mathbf{b}_1,...,\mathbf{b}_n\}$, where $x_1,...,x_n$ are real numbers such that $\mathbf{x} = x_1\mathbf{b}_1 + ... + x_n\mathbf{b}_n$. The point **x** is called a *vector*, and $x_1,...,x_n$ are called its *components*. The notation $(\mathbf{x})_i$ will sometimes also be used for component i of a vector **x**. In econometrics, we work mostly with *finite-dimensional real space*. When this space is of dimension n, it is denoted $\mathbb{R}^n$. Points in this space are vectors of real numbers $(x_1,...,x_n)$; this corresponds to the previous terminology with the *basis* for $\mathbb{R}^n$ being the *unit vectors* (1,0,..,0), (0,1,0,..,0),..., (0,..,0,1). Usually, we assume this representation without being explicit about the basis for the space. However, it is worth noting that the coordinate representation of a vector depends on the particular basis chosen for a space. Sometimes this fact can be used to choose bases in which vectors and transformations have particularly simple coordinate representations.

The *Euclidean norm* of a vector **x** is $\|\mathbf{x}\|_2 = (x_1^2+...+x_n^2)^{1/2}$. This norm can be used to define the distance between vectors, or neighborhoods of a vector. Other possible norms include $\|\mathbf{x}\|_1 = |x_1|+...+|x_n|$, $\|\mathbf{x}\|_\infty = \max\{|x_1|,...,|x_n|\}$, or for $1 \le p < +\infty$, $\|\mathbf{x}\|_p = \left[|x_1|^p+...+|x_n|^p\right]^{1/p}$. Each norm defines a *topology* on the linear space, based on neighborhoods of a vector that are less than each positive distance away. The space $\mathbb{R}^n$ with the norm $\|\mathbf{x}\|_2$ and associated topology is called *Euclidean* n-*space*.

The *vector product* of **x** and **y** in $\mathbb{R}^n$ is defined as $\mathbf{x}\cdot\mathbf{y} = x_1y_1+...+x_ny_n$. Other notations for vector products are $\langle\mathbf{x},\mathbf{y}\rangle$ or (when **x** and **y** are interpreted as *row* vectors) $\mathbf{x}\mathbf{y}'$ or (when **x** and **y** are interpreted as column vectors) $\mathbf{x}'\mathbf{y}$.

2.2.2. A *linear subspace* of a linear space such as $\mathbb{R}^n$ is a subset that has the property that all linear combinations of its members remain in the subset. Examples of linear subspaces in $\mathbb{R}^3$ are the plane $\{(a,b,c)|b = 0\}$ and the line $\{(a,b,c)|a = b = 2\cdot c\}$. The linear subspace *spanned* by a set of vectors $\{x_1,...,x_J\}$ is the set of all linear combinations of these vectors, $\mathbf{L} = \{x_1\alpha_1+...+x_J\alpha_J|(\alpha_1,...,\alpha_J) \in \mathbb{R}^J\}$. The vectors $\{\mathbf{x}_1,...,\mathbf{x}_J\}$ are *linearly independent* if and only if one cannot be written as a linear combination of the remainder. The linear subspace that is spanned by a set of J linearly independent vectors is said to be *of dimension* J. Conversely, each linear space of dimension J can be represented as the set of linear combinations of J linearly independent vectors, which are in fact a basis for the subspace. A linear subspace of dimension one is a *line* (through the origin), and a linear subspace of dimension (n-1) is a *hyperplane* (through the origin). If **L** is a subspace, then $\mathbf{L}^\perp = \{\mathbf{x}\in\mathbb{R}^n|\ \mathbf{x}\cdot\mathbf{y} = 0$ for all $\mathbf{y}\in\mathbf{L}\}$ is termed the *complementary subspace.* Subspaces **L** and **M** with the property that $\mathbf{x}\cdot\mathbf{y} = 0$ for all $\mathbf{y} \in \mathbf{L}$ and $\mathbf{x} \in \mathbf{M}$ are termed *orthogonal*, and denoted $\mathbf{L}\perp\mathbf{M}$. The *angle* θ between subspaces **L** and **M** is defined by $\cos\theta = \text{Min}\{\mathbf{x}\cdot\mathbf{y}|\ \mathbf{y} \in L, \|\mathbf{y}\|_2 = 1, \mathbf{x} \in M, \|\mathbf{x}\|_2 = 1\}$. Then, the angle between orthogonal subspaces is π/2, and the angle between subspaces that have a nonzero point in common is zero. A subspace that is translated by adding a nonzero vector **c** to all points in the subspace is termed an *affine subspace*.

2.2.3. The concept of a finite-dimensional space can be generalized. An example, for $1 \le p < +\infty$, is the family $\mathbf{L}_p(\mathbb{R}^n)$ of real-valued functions f on $\mathbb{R}^n$ such that the integral $\|f\|_p = \left[\int_{\mathbb{R}^n} |f(x)|^p dx\right]^{1/p}$ is well-defined and finite. This is a linear space with norm $\|f\|_p$ since linear combinations of functions that satisfy this property also satisfy (using convexity of the norm function) this property. One can think of the function f as a vector in $L_p(\mathbb{R}^n)$, and f(x) for a particular value of x as a component of this vector. Many, but not all, of the properties of finite-dimensional space extend to infinite dimensions. In basic econometrics, we will not need the infinite-dimensional generalization. It appears in more advanced econometrics, in stochastic processes in time series, and in nonlinear and nonparametric problems.

2.3. LINEAR TRANSFORMATIONS AND MATRICES

2.3.1. A mapping A from one linear space (its *domain)* into another (its *range*) is a *linear transformation* if it satisfies $A(\mathbf{x}+\mathbf{z}) = A(\mathbf{x}) + A(\mathbf{z})$ for any $\mathbf{x}$ and $\mathbf{z}$ in the domain. When the domain and range are finite-dimensional linear spaces, a linear transformation can be represented as a *matrix*. Specifically, a linear transformation A from $\mathbb{R}^n$ into $\mathbb{R}^m$ can be represented by a m×n array $\mathbf{A}$ with elements a_{ij} for $1 \le i \le m$ and $1 \le j \le n$, with $\mathbf{y} = A(\mathbf{x})$ having components $y_i = \sum_{j=1}^{n} a_{ij}x_j$ for $1 \le i \le m$. In matrix notation, this is written $\mathbf{y} = \mathbf{A}\mathbf{x}$. A matrix $\mathbf{A}$ is *real* if all its elements are real numbers, *complex* if some of its elements are complex numbers. Throughout, matrices are assumed to be real unless explicitly assumed otherwise. The set $\mathbb{N} = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{A}\mathbf{x} = \mathbf{0}\}$ is termed the *null space* of the transformation $\mathbf{A}$. The subspace $\mathbb{N}^{\perp}$ containing all linear combinations of the column vectors of $\mathbf{A}$ is termed the *column space* of $\mathbf{A}$; it is the complementary subspace to $\mathbb{N}$.

If $\mathbf{A}$ denotes a m×n matrix, then $\mathbf{A}'$ denotes its n×m *transpose* (rows become columns and vice versa). The *identity matrix* of dimension n is n×n with one's down the diagonal, zero's elsewhere, and is denoted $\mathbf{I}_n$, or $\mathbf{I}$ if the dimension is clear from the context. A matrix of zeros is denoted $\mathbf{0}$, and a n×1 vector of ones is denoted $\mathbf{1}_n$. A *permutation matrix* is obtained by permuting the columns of an identity matrix. If $\mathbf{A}$ is a m×n matrix and $\mathbf{B}$ is a n×p matrix, then the *matrix product* $\mathbf{C} = \mathbf{A}\mathbf{B}$ is of dimension m×p with elements $c_{ik} \equiv \sum_{j=1}^{n} a_{ij}b_{jk}$ for $1 \le i \le m$ and $1 \le k \le p$. For the matrix product to be defined, the number of columns in $\mathbf{A}$ must equal the number of rows in $\mathbf{B}$ (i.e., the matrices must be *commensurate*). A matrix $\mathbf{A}$ is *square* if it has the same number of rows and columns. A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}'$, *diagonal* if all off-diagonal elements are zero, *upper (lower)*

triangular if all its elements below (above) the diagonal are zero, and *idempotent* if it is symmetric and $\mathbf{A}^2 = \mathbf{A}$. A matrix $\mathbf{A}$ is *column orthonormal* if $\mathbf{A}'\mathbf{A} = \mathbf{I}$; simply *orthonormal* if it is both square and column orthonormal.

A set of linearly independent vectors in $\mathbb{R}^n$ can be recursively orthonormalized; i.e., transformed so they are orthogonal and scaled to have unit length: Suppose vectors $\mathbf{x}_1,...,\mathbf{x}_{J-1}$ have previously been orthonormalized, and $\mathbf{z}$ is the next vector in the set. Then, $\mathbf{z} - \sum_{j=1}^{J-1} (\boldsymbol{x}_j'\boldsymbol{z})\boldsymbol{x}_j$ is orthogonal to $\mathbf{x}_1,...,\mathbf{x}_{J-1}$, and is non-zero since it is linearly independent. Scale it to unit length; this defines $\mathbf{x}_J$. Each column of a n×m matrix $\mathbf{A}$ is a vector in $\mathbb{R}^n$. The *rank* of $\mathbf{A}$, denoted $r = \rho(\mathbf{A})$, is the largest number of columns that are *linearly independent*. Then $\mathbf{A}$ is of rank m if and only if $\mathbf{x} = \mathbf{0}$ is the only solution to $\mathbf{Ax} = \mathbf{0}$. If A is of rank r, then orthonormalization applied to the linearly independent columns of $\mathbf{A}$ can be interpreted as defining a r×m lower triangluar matrix $\mathbf{U}$ such that $\mathbf{AU}'$ is column orthonormal. A n×m matrix $\mathbf{A}$ is of *full rank* if $\rho(\mathbf{A}) = \min(n,m)$. A n×n matrix $\mathbf{A}$ of full rank is termed *nonsingular*. A nonsingular n×n matrix $\mathbf{A}$ has an *inverse* matrix $\mathbf{A}^{-1}$ such that both $\mathbf{AA}^{-1}$ and $\mathbf{A}^{-1}\mathbf{A}$ equal the identity matrix $\mathbf{I}_n$. An orthonormal matrix $\mathbf{A}$ satisfies $\mathbf{A}'\mathbf{A} = \mathbf{I}_n$, implying that $\mathbf{A}' = \mathbf{A}^{-1}$, and hence $\mathbf{A}'\mathbf{A} = \mathbf{AA}' = \mathbf{I}_n$. The trace $\mathrm{tr}(\mathbf{A})$ of a square matrix $\mathbf{A}$ is the sum of its diagonal elements.

2.3.2. The tables in this section summarize useful matrix and vector operations. In addition to the operations in these tables, there are statistical operations that can be performed on a matrix when its columns are vectors of observations on various variables. Discussion of these operations is postponed until later. Most of the operations in Tables 2.1-2.3 are available as part of the matrix programming languages in econometrics computer packages such as SST, TSP, GAUSS, or MATLAB. The notation in these tables is close to the notation for the corresponding matrix commands in SST and GAUSS.

TABLE 2.1. BASIC OPERATIONS

	Name	Notation	Definition
1.	Matrix Product	$\mathbf{C} = \mathbf{AB}$	For m×n $\mathbf{A}$ and n×p $\mathbf{B}$: $c_{ik} = \sum_{j=1}^{n} a_{ij}b_{jk}$
2.	Scalar Multiplication	$\mathbf{C} = b\mathbf{A}$	For a scalar b: $c_{ij} = ba_{ij}$
3.	Matrix Sum	$\mathbf{C} = \mathbf{A}+\mathbf{B}$	For $\mathbf{A}$ and $\mathbf{B}$ m×n: $c_{ij} = a_{ij} + b_{ij}$
4.	Transpose	$\mathbf{C} = \mathbf{A}'$	For m×n $\mathbf{A}$: $c_{ij} = a_{ji}$
5.	Matrix Inverse	$\mathbf{C} = \mathbf{A}^{-1}$	For $\mathbf{A}$ n×n nonsingular: $\mathbf{AA}^{-1} = \mathbf{I}_m$
6.	Trace	$c = \mathrm{tr}(\mathbf{A})$	For n×n $\mathbf{A}$: $c = \sum_{i=1}^{n} a_{ii}$

TABLE 2.2. OPERATIONS ON ELEMENTS

	Name	Notation	Definition
1.	Element Product	$\mathbf{C} = \mathbf{A}.*\mathbf{B}$	For **A**, **B** m×n: $c_{ij} = a_{ij}\, b_{ij}$
2.	Element Division	$\mathbf{C} = \mathbf{A}.\div\mathbf{B}$	For **A**, **B** m×n: $c_{ij} = a_{ij}/b_{ij}$
3.	Logical Condition	$\mathbf{C} = \mathbf{A}.\leq\mathbf{B}$	For **A**, **B** m×n: $c_{ij} = \mathbf{1}(a_{ij}\leq b_{ij})$ (Note 1)
4.	Row Minimum	$\mathbf{c}$ = vmin(**A**)	For m×n **A**: $c_i = \min_{1\leq k\leq m} a_{ik}$ (Note 2)
5.	Row Min Replace	$\mathbf{C}$ = rmin(**A**)	For m×n **A**: $c_{ij} = \min_{1\leq k\leq m} a_{ik}$ (Note 3)
6.	Column Min Replace	$\mathbf{C}$ = cmin(**A**)	For m×n **A**: $c_{ij} = \min_{1\leq k\leq n} a_{ik}$ (Note 4)
7.	Cumulative Sum	$\mathbf{C}$ = cumsum(**A**)	For m×n **A**: $c_{ij} = \sum_{k=1}^{i} a_{kj}$

NOTES:
1. **1**(P) is one of P is true, zero otherwise. The condition is also defined for the logical operations "<", ">", "≥", "=", and "≠".
2. **c** is a m×1 vector. The operation is also defined for "max".
3. **C** is a m×n matrix, with all columns the same. The operation is also defined for "max"
4. **C** is a m×n matrix, with all rows the same. The operation is also defined for "max".

TABLE 2.3. SHAPING OPERATIONS

	Name	Notation	Definition
1.	Kronecker Product	$\mathbf{C} = \mathbf{A}\otimes\mathbf{B}$	Note 1
2.	Direct Sum	$\mathbf{C} = \mathbf{A}\oplus\mathbf{B}$	Note 2
3.	diag	$\mathbf{C}$ = diag(**x**)	**C** a diagonal matrix with vector **x** down the diagonal
4.	vec *or* vecr	$\mathbf{c}$ = vecr(**A**)	vector **c** contains rows of **A**, stacked
5.	vecc	$\mathbf{c}$ = vecc(**A**)	vector **c** contains columns of **A**, stacked
6.	vech	$\mathbf{c}$ = vech(**A**)	vector **c** contains upper triangle of **A**, row by row, stacked
7.	vecd	$\mathbf{c}$ = vecd(**A**)	vector **c** contains diagonal of **A**
8.	horizontal contatination	$\mathbf{C} = \{\mathbf{A},\mathbf{B}\}$	Partitioned matrix $\mathbf{C} = [\,\mathbf{A}\ \mathbf{B}\,]$
9.	vertical contatination	$\mathbf{C} = \{\mathbf{A};\mathbf{B}\}$	Partitioned matrix $\mathbf{C}' = [\,\mathbf{A}'\ \mathbf{B}'\,]$
10.	reshape	$\mathbf{C}$ = rsh(**A**,k)	Note 3

NOTES:
1. Also termed the *direct product*, the Kronecker product creates an array made up of blocks, with each block the product of an element of **A** and the matrix **B**; see Section 2.11.

2. The *direct sum* is defined for a m×n matrix **A** and a p×q matrix **B** by the (m+p)×(n+q) partitioned array $\mathbf{A}\oplus\mathbf{B} = \begin{bmatrix} \boldsymbol{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} \end{bmatrix}$.

3. If **A** is m×n, then k must be a divisor of m·n. The operation takes the elements of **A** row by row, and rewrites the successive elements as rows of a matrix **C** that has k rows and m·n/k columns.

2.3.3. The *determinant* of a n×n matrix **A** is denoted $|\mathbf{A}|$ or det(**A**), and has a geometric interpretation as the volume of the parallelepiped formed by the column vectors of **A**. The matrix **A** is *nonsingular* if and only if det(**A**) ≠ 0. A *minor* of a matrix **A** (of order r) is the determinant of a submatrix formed by striking out n-r rows and columns. A *principal minor* is formed by striking out symmetric rows and columns of **A**. A *leading principal minor* (of order r) is formed by striking out the last n-r rows and columns. The *minor* of an element a_{ij} of **A** is the determinant of the submatrix $\mathbf{A}^{ij}$ formed by striking out row i and column j of **A**. Determinants satisfy the recursion relation

$$\det(\mathbf{A}) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}^{ij}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}^{ij}),$$

with the first equality holding for any j and the second holding for any i. This formula can be used as a recursive definition of determinants, starting from the result that the determinant of a scalar is the scalar. A useful related formula is

$$\sum_{i=1}^{n} (-1)^{i+j} a_{ik} \det(\mathbf{A}^{ij})/\det(\mathbf{A}) = \delta_{kj},$$

where δ_{kj} is one if k = j and zero otherwise.

2.3.4. We list without proof a number of useful elementary properties of matrices:

(1) $(\mathbf{A}')' = \mathbf{A}$.
(2) If $\mathbf{A}^{-1}$ exists, then $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
(3) If $\mathbf{A}^{-1}$ exists, then $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.
(4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
(5) If **A**,**B** are square, nonsingular, and commensurate, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
(6) If **A** is m×n, then Min {m,n} ≥ ρ(**A**) = ρ(**A**′) = ρ(**A**′**A**) = ρ(**AA**′).
(7) If **A** is m×n and **B** is m×r, then ρ(**AB**) ≤ min(ρ(**A**),ρ(**B**)).
(8) If **A** is m×n with ρ(**A**) = m, and **B** is m×r, then ρ(**AB**) = ρ(**B**).
(9) ρ(**A**+**B**) ≤ ρ(**A**) + ρ(**B**).
(10) If **A** is n×n, then det(**A**) ≠ 0 if and only if ρ(**A**) = n.
(11) If **B** and **C** are nonsingular and commensurate with **A**, then ρ(**BAC**) = ρ(**A**).
(12) If **A**, **B** are n×n, then ρ(**AB**) ≥ ρ(**A**) + ρ(**B**) - n.
(13) det(**AB**) = det(**A**)·det(**B**).
(14) If *c* is a scalar and **A** is n×n, then $\det(c\mathbf{A}) = c^n \det(\mathbf{A})$
(15) The determinant of a matrix is unchanged if a scalar times one column (row) is added to another column (row).
(16) If **A** is n×n and diagonal or triangular, then det(**A**) is the product of the diagonal elements.
(17) $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$.

(18) If $\mathbf{A}$ is n×n and $\mathbf{B} = \mathbf{A}^{-1}$, then $b_{ij} = (-1)^{i+j}\det(\mathbf{A}^{ij})/\det(\mathbf{A})$.
(19) The determinant of an orthonormal matrix is +1 or -1.
(20) If $\mathbf{A}$ is m×n and $\mathbf{B}$ is n×m, then $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.
(21) $\text{tr}(\mathbf{I}_n) = n$.
(22) $\text{tr}(\mathbf{A}+\mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$.
(23) A permutation matrix $\mathbf{P}$ is orthonormal; hence, $\mathbf{P}' = \mathbf{P}^{-1}$.
(24) The inverse of a (upper) triangular matrix is (upper) triangular, and the inverse of a diagonal matrix $\mathbf{D}$ is diagonal, with $(\mathbf{D}^{-1})_{ii} = 1/\mathbf{D}_{ii}$.
(25) The product of orthonormal matrices is orthonormal, and the product of permutation matrices is a permutation matrix.

2.4. EIGENVALUES AND EIGENVECTORS

An *eigenvalue* of a n×n matrix $\mathbf{A}$ is a scalar λ such that $\mathbf{Ax} = \lambda\mathbf{x}$ for some vector $\mathbf{x} \neq \mathbf{0}$. The vector $\mathbf{x}$ is called a (right) *eigenvector*. The condition $(\mathbf{A}-\lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ associated with an eigenvalue implies $\mathbf{A}-\lambda\mathbf{I}$ simgular, and hence $\det(\mathbf{A}-\lambda\mathbf{I}) = 0$. This determanental equation defines a polynomial in λ of order n, and the n roots of this polynomial are the eigenvalues. For each eigenvalue λ, the condition that $\mathbf{A}-\lambda\mathbf{I}$ is less than rank n implies the existence of one or more linearly independent eigenvectors; the number equals the multiplicity of the root λ. The following basic properties of eigenvalues and eigenvectors of a n×n matrix $\mathbf{A}$ are stated without proof:

(1) If $\mathbf{A}$ is real and symmetric, then its eigenvalues and eigenvectors are real. However, if $\mathbf{A}$ is nonsymmetric, then its eigenvalues and eigenvectors in general are complex.
(2) The number of nonzero eigenvalues of $\mathbf{A}$ equals its rank $\rho(\mathbf{A})$.
(3) If λ is an eigenvalue of $\mathbf{A}$, then λ^k is an eigenvalue of $\mathbf{A}^k$, and $1/\lambda$ is an eigenvalue of $\mathbf{A}^{-1}$ (if the inverse exists).
(4) If $\mathbf{A}$ is real and symmetric, then the eigenvalues corresponding to distinct roots are orthogonal. [$\mathbf{Ax}_i = \lambda_i\mathbf{x}_i$ implies $\mathbf{x}_i'\mathbf{Ax}_j = \lambda_i\mathbf{x}_i'\mathbf{x}_j = \lambda_j\mathbf{x}_i'\mathbf{x}_j$, which can be true for $i \neq j$ only if $\mathbf{x}_i'\mathbf{x}_j = 0$.]
(5) If $\mathbf{A}$ is real and symmetric, and $\mathbf{\Lambda}$ is a diagonal matrix with the roots of the polynomial $\det(\mathbf{A}-\lambda\mathbf{I})$ along the diagonal, then there exists an orthonormal matrix $\mathbf{C}$ such that $\mathbf{C}'\mathbf{C} = \mathbf{I}$ and $\mathbf{AC} = \mathbf{C\Lambda}$, and hence $\mathbf{C}'\mathbf{AC} = \mathbf{\Lambda}$ and $\mathbf{C\Lambda C}' = \mathbf{A}$. The transformation $\mathbf{C}$ is said to *diagonalize* $\mathbf{A}$. [Take $\mathbf{C}$ to be an array whose columns are eigenvectors of $\mathbf{A}$, scaled to unit length. In the case of a multiple root, orthonormalize the eigenvectors corresponding to this root.].
(6) If $\mathbf{A}$ is real and nonsymmetric, there exists a nonsingular complex matrix $\mathbf{Q}$ and a upper triangular complex matrix $\mathbf{T}$ with the eigenvalues of $\mathbf{A}$ on its diagonal such that $\mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{T}$.
(7) $\mathbf{A}$ real and symmetric implies $\text{tr}(\mathbf{A})$ equals the sum of the eigenvalues of $\mathbf{A}$. [Since $\mathbf{A} = \mathbf{C\Lambda C}'$, $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{C\Lambda C}') = \text{tr}(\mathbf{C}'\mathbf{C\Lambda}) = \text{tr}(\mathbf{\Lambda})$ by 2.3.20.]

(8) If $\mathbf{A}_i$ are real and symmetric for i = 1,...,p, then there exists **C** orthonormal such that $\mathbf{C}'\mathbf{A}_i\mathbf{C}$, are all diagonal if and only if $\mathbf{A}_i\mathbf{A}_j = \mathbf{A}_j\mathbf{A}_i$ for i,j = 1,...,p.

Results (5) and (6) combined with the result 2.3.13 that the determinant of a matrix product is the product of the determinants of the matrices, implies that the determinant of a matrix is the product of its eigenvalues. The transformations in (5) and (6) are called *similarity transformations*, and can be interpreted as representations of the transformation **A** when the basis of the domain is transformed by **C** (or **Q**) and the basis of the range is transformed by $\mathbf{C}^{-1}$ (or $\mathbf{Q}^{-1}$). These transformations are used extensively in econometric theory.

2.5. PARTITIONED MATRICES

It is sometimes useful to *partition* a matrix into submatrices,

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{bmatrix},$$

where **A** is m×n, $\mathbf{A}_{11}$ is $m_1 \times n_1$, $\mathbf{A}_{12}$ is $m_1 \times n_2$, $\mathbf{A}_{21}$ is $m_2 \times n_1$, $\mathbf{A}_{22}$ is $m_2 \times n_2$, and $m_1+m_2 = m$ and $n_1+n_2 = n$. Matrix products can be written for partitioned matrices, applying the usual algorithm to the partition blocks, provided the blocks are commensurate. For example, if **B** is n×p and is partitioned $\mathbf{B} = \begin{bmatrix} \boldsymbol{B_1} \\ \boldsymbol{B_2} \end{bmatrix}$ where $\mathbf{B}_1$ is $n_1 \times p$ and $\mathbf{B}_2$ is $n_2 \times p$, one has $\mathbf{AB} = \begin{bmatrix} \boldsymbol{A_{11}B_1 + A_{12}B_2} \\ \boldsymbol{A_{21}B_1 + A_{22}B_2} \end{bmatrix}$.

Partitioned matrices have the following elementary properties:

(1) **A** square and $\mathbf{A}_{11}$ square and nonsingular implies $\det(\mathbf{A}) = \det(\mathbf{A}_{11})\cdot\det(\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})$.
(2) **A** and $\mathbf{A}_{11}$ square and nonsingular implies

$$\mathbf{A}^{-1} = \begin{bmatrix} \boldsymbol{A_{11}^{-1} + A_{11}^{-1}A_{12}C^{-1}A_{21}A_{11}^{-1}} & \boldsymbol{-A_{11}^{-1}A_{12}C^{-1}} \\ \boldsymbol{-C^{-1}A_{21}A_{11}^{-1}} & \boldsymbol{C^{-1}} \end{bmatrix}$$

with $\mathbf{C} = \mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. When $\mathbf{A}_{22}$ is nonsingular, the northwest matrix in this partition can also be written as $(\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}$.

2.6. QUADRATIC FORMS

The scalar function $Q(\mathbf{x},\mathbf{A}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a n×n matrix and $\mathbf{x}$ is a n×1 vector, is termed a *quadratic form*; we call $\mathbf{x}$ the *wings* and $\mathbf{A}$ the *center* of the quadratic form. The value of a quadratic form is unchanged if $\mathbf{A}$ is replaced by its *symmetrized* version $(\mathbf{A}+\mathbf{A}')/2$. Therefore, $\mathbf{A}$ will be assumed symmetric for the discussion of quadratic forms.

A quadratic form $Q(\mathbf{x},\mathbf{A})$ may fall into one of the classes in the table below:

Class	Defining Condition
Positive Definite	$\mathbf{x}\neq\mathbf{0} \Rightarrow Q(\mathbf{x},\mathbf{A}) > 0$
Positive Semidefinite	$\mathbf{x}\neq\mathbf{0} \Rightarrow Q(\mathbf{x},\mathbf{A}) \geq 0$
Negative Definite	$\mathbf{x}\neq\mathbf{0} \Rightarrow Q(\mathbf{x},\mathbf{A}) < 0$
Negative Semidefinite	$\mathbf{x}\neq\mathbf{0} \Rightarrow Q(\mathbf{x},\mathbf{A}) \leq 0$

A quadratic form that is not in one of these four classes is termed *indefinite*. The basic properties of quadratic forms are listed below:

(1) If $\mathbf{B}$ is m×n and is of rank $\rho(\mathbf{B}) = r$, then $\mathbf{B}'\mathbf{B}$ and $\mathbf{B}\mathbf{B}'$ are both positive semidefinite; and if $r = m \leq n$, then $\mathbf{B}'\mathbf{B}$ is positive definite.

(2) If $\mathbf{A}$ is symmetric and positive semidefinite (positive definite), then the eigenvalues of $\mathbf{A}$ are nonnegative (positive). Similarly, if $\mathbf{A}$ is symmetric and negative semidefinite (negative definite), then the eigenvalues of $\mathbf{A}$ are nonpositive (negative).

(3) Every symmetric positive semidefinite matrix $\mathbf{A}$ has a symmetric positive semidefinite square root $\mathbf{A}^{1/2}$ [By 2.4.4, $\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{D}$ for some $\mathbf{C}$ orthonormal and $\mathbf{D}$ a diagonal matrix with the nonnegative eigenvalues down the diagonal. Then, $\mathbf{A} = \mathbf{C}\mathbf{D}\mathbf{C}'$ and $\mathbf{A}^{1/2} = \mathbf{C}\mathbf{D}^{1/2}\mathbf{C}'$ with $\mathbf{D}^{1/2}$ a diagonal matrix of the positive square roots of the diagonal of $\mathbf{D}$.]

(4) If $\mathbf{A}$ is positive definite, then $\mathbf{A}^{-1}$ is positive definite.

(5) If $\mathbf{A}$ and $\mathbf{B}$ are real, symmetric n×n matrices and $\mathbf{B}$ is positive definite, then there exists a n×n matrix $\mathbf{Q}$ that simultaneously diagonalizes $\mathbf{A}$ and $\mathbf{B}$: $\mathbf{Q}'\mathbf{A}\mathbf{Q} = \mathbf{\Lambda}$ diagonal and $\mathbf{Q}'\mathbf{B}\mathbf{Q} = \mathbf{I}$. [From 2.4(5), there exists a n×n orthonormal matrix $\mathbf{U}$ such that $\mathbf{U}'\mathbf{B}\mathbf{U} = \mathbf{D}$ is diagonal. Let $\mathbf{G}$ be an orthonormal matrix that diagonalizes $\mathbf{D}^{-1/2}\mathbf{U}'\mathbf{A}\mathbf{U}\mathbf{D}^{-1/2}$, and define $\mathbf{Q} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{G}$.]

(6) $\mathbf{B}$ positive definite and $\mathbf{A} - \mathbf{B}$ positive semidefinite imply $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ positive semidefinite. [For a vector $\mathbf{z}$, let $\mathbf{x} = \mathbf{Q}^{-1}\mathbf{z}$, where $\mathbf{Q}$ is the diagonalizing matrix from (5). Then $\mathbf{z}'(\mathbf{B} - \mathbf{A})\mathbf{z} = \mathbf{x}'\mathbf{Q}'(\mathbf{B} - \mathbf{A})\mathbf{Q}\mathbf{x} = \mathbf{x}'(\mathbf{\Lambda} - \mathbf{I})\mathbf{x} \geq 0$, so no diagonal element of $\mathbf{\Lambda}$ is less than one. Alternately, let $\mathbf{x} = \mathbf{Q}'\mathbf{z}$. Then $\mathbf{z}'(\mathbf{B}^{-1} - \mathbf{A}^{-1})\mathbf{z} = \mathbf{x}'\mathbf{Q}^{-1}(\mathbf{B}^{-1} - \mathbf{A}^{-1})(\mathbf{Q}')^{-1}\mathbf{x} = \mathbf{x}'(\mathbf{I} - \mathbf{\Lambda}^{-1})\mathbf{x}$ must be non-negative.]

(7) The following conditions are equivalent:

- (i) $\mathbf{A}$ is positive definite
- (ii) The principal minors of $\mathbf{A}$ are positive
- (iii) The leading principal minors of $\mathbf{A}$ are positive.

2.7. THE LDU AND CHOLESKY FACTORIZATIONS OF A MATRIX

A n×n matrix **A** has a LDU factorization if it can be written **A** = **LDU′**, where **D** is a diagonal matrix and **L** and **U** are lower triangular matrices. This factorization is useful for computation of inverses, as triangular matrices are easily inverted by recursion.

Theorem 2.1. Each n×n matrix **A** can be written as **A** = **PLDU′Q′**, where **P** and **Q** are permutation matrices, **L** and **U** are lower triangular matrices, each with ones on the diagonal, and **D** is a diagonal matrix. If the leading principal minors of **A** are all non-zero, then **P** and **Q** can be taken to be identity matrices.

Proof: First assume that the leading principal minors of **A** are all nonzero. We give a recursive construction of the required **L** and **U**. Suppose the result has been established for matrices up to order n-1. Then, write the required decomposition **A** = **LDU′** for a n×n matrix in partitioned form

$$\begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & A_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L_{11}} & \mathbf{0} \\ \boldsymbol{L_{21}} & 1 \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{D_{11}} & \mathbf{0} \\ \mathbf{0} & D_{22} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{U_{11}}' & \boldsymbol{U_{21}}' \\ 0 & 1 \end{bmatrix}',$$

where $\mathbf{A}_{11}$, $\mathbf{L}_{11}$, $\mathbf{D}_{11}$, and $\mathbf{U}_{11}'$ are (n-1)×(n-1), $\mathbf{L}_{21}$ is 1×(n-1), $\mathbf{U}_{21}$ is 1×(n-1), and A_{22} and D_{22} are 1×1. Assume that $\mathbf{L}_{11}$, $\mathbf{D}_{11}$, and $\mathbf{U}_{11}$ have been defined so that $\mathbf{A}_{11} = \mathbf{L}_{11}\mathbf{D}_{11}\mathbf{U}_{11}'$, and that $\mathbf{L}_{11}^{-1}$ and $\mathbf{U}_{11}^{-1}$ also exist and have been computed. Let $\mathbf{S} = \mathbf{L}^{-1}$ and $\mathbf{T} = \mathbf{U}^{-1}$, and partition **S** and **T** commensurately with **L** and **U**. Then, $\mathbf{A}_{11}^{-1} = \mathbf{U}_{11}'^{-1}\mathbf{D}_{11}^{-1}\mathbf{L}_{11}^{-1}$ and the remaining elements must satisfy the equations

$$\begin{aligned}
&\mathbf{A}_{21} = \mathbf{L}_{21}\mathbf{D}_{11}\mathbf{U}_{11}' \;\Rightarrow\; \mathbf{L}_{21} = \mathbf{A}_{21}\mathbf{U}_{11}'^{-1}\mathbf{D}_{11}^{-1} \equiv \mathbf{A}_{21}\mathbf{T}_{11}'\mathbf{D}_{11}^{-1} \\
&\mathbf{A}_{12} = \mathbf{L}_{11}\mathbf{D}_{11}\mathbf{U}_{21}' \;\Rightarrow\; \mathbf{U}_{21}' = \mathbf{D}_{11}^{-1}\mathbf{L}_{11}^{-1}\mathbf{A}_{12} \equiv \mathbf{D}_{11}^{-1}\mathbf{S}_{11}\mathbf{A}_{12} \\
&A_{22} = \mathbf{L}_{21}\mathbf{D}_{11}\mathbf{U}_{21}' + D_{22} \;\Rightarrow\; D_{22} = A_{22} - \mathbf{A}_{21}\mathbf{U}_{11}'^{-1}\mathbf{D}_{11}^{-1}\mathbf{L}_{11}^{-1}\mathbf{A}_{12} = A_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\
&\mathbf{S}_{21} = -\mathbf{L}_{21}\mathbf{S}_{11} \qquad \mathbf{S}_{22} = 1 \\
&\mathbf{T}_{21}' = -\mathbf{T}_{11}'\mathbf{U}_{21}' \qquad \mathbf{T}_{22} = 1
\end{aligned}$$

where $\det(\mathbf{A}) = \det(\mathbf{A}_{11})\cdot\det(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}) \neq 0$ implies $D_{22} \neq 0$. Since the decomposition is trivial for n = 1, this recursion establishes the result, and furthermore gives the triangular matrices **S** and **T** from the same recursion that can be multiplied to give $\mathbf{A}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{S}$.

Now assume that **A** is of rank r < n, and that the first r columns of **A** are linearly independent, with non-zero leading principal minors up to order r. Partition

$$\begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L_{11}} & \mathbf{0} \\ \boldsymbol{L_{21}} & \boldsymbol{I} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{D_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{U_{11}}' & \boldsymbol{U_{21}}' \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix},$$

where $\mathbf{A}_{11}$ is r×r and the remaining blocks are commensurate. Then, $\mathbf{U}_{21}' = \mathbf{D}_{11}^{-1}\mathbf{S}_{11}\mathbf{A}_{12}$ and $\mathbf{L}_{21} = \mathbf{A}_{21}\mathbf{T}_{11}'\mathbf{D}_{11}^{-1}$, and one must satisfy $\mathbf{A}_{22} = \mathbf{L}_{21}\mathbf{D}_{11}\mathbf{U}_{12}' = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. But the rank condition implies the last n-r columns of **A** can be written as a linear combination $\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} C$ of the first r columns, where **C** is some r×(n-r) matrix. But $\mathbf{A}_{12} = \mathbf{A}_{11}\mathbf{C}$ implies $\mathbf{C} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and hence $\mathbf{A}_{22} = \mathbf{A}_{21}\mathbf{C} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ as required.

Finally, consider any real matrix **A** of rank r. By column permutations, the first r columns can be made linearly independent. Then, by row permutations, the first r rows of these r columns can be made linearly independent. Repeat this process recursively on the remaining northwest principal submatrices to obtain products of permutation matrices that give nonzero leading principal minors up to order r. This defines **P** and **Q**, and completes the proof of the theorem. □

Corollary 2.1.1. If **A** is symmetric, then **L** = **U**.

Corollary 2.1.2. (LU Factorization) If **A** has nonzero leading principal minors, then **A** can be written $\mathbf{A} = \mathbf{L}\mathbf{V}'$, where $\mathbf{V}' = \mathbf{D}\mathbf{U}'$ is upper triangular with a diagonal coinciding with that of **D**.

Corollary 2.1.3. (Cholesky Factorization) If **A** is symmetric and positive definite, then **A** can be written $\mathbf{A} = \mathbf{V}\mathbf{V}'$, where $\mathbf{V} = \mathbf{L}\mathbf{D}^{1/2}$ is lower triangular with a positive diagonal.

Corollary 2.1.4. **A** symmetric positive semidefinite implies $\mathbf{A} = \mathbf{P}\mathbf{V}\mathbf{V}'\mathbf{P}'$, with **V** lower triangular with a nonnegative diagonal, **P** a permutation matrix.

Corollary 2.1.5. If **A** is m×n with m ≥ n, then there exists a factorization $\mathbf{A} = \mathbf{P}\mathbf{L}\mathbf{D}\mathbf{U}'\mathbf{Q}'$, with **D** n×n diagonal, **P** a m×m permutation matrix, **Q** a n×n permutation matrix, **U** a n×n lower triangular matrix with ones on the diagonal, and **L** a m×n lower triangular matrix with ones on the diagonal (i.e., **L** has the form $\mathbf{L}' = [\mathbf{L}_{11}' \ \mathbf{L}_{21}']$ with $\mathbf{L}_{11}$ n×n and lower triangular with ones on the diagonal, and $\mathbf{L}_{21}$ (m-n)×n. Further, if ρ(A) = n, then $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{Q}\mathbf{U}'^{-1}\mathbf{D}^{-1}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}'\mathbf{P}'$.

Corollary 2.1.6. If the system of equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ with **A** m×n of rank n has a solution, then the solution is given by $\mathbf{x} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} = \mathbf{Q}\mathbf{U}'^{-1}\mathbf{D}^{-1}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}'\mathbf{P}'\mathbf{y}$.

Proof outline: To show Corollary 3, note that a positive definite matrix has positive leading principal minors, and note from the proof of the theorem that this implies that the diagonal of **D** is positive. Take $\mathbf{V}' = \mathbf{D}^{1/2}\mathbf{U}'$, where $\mathbf{D}^{1/2}$ is the positive square root. The same construction applied to the **LDU** factorization of **A** after permutation gives Corollary 4. To show Corollary 5, note first that the rows of **A** can be permuted so that the first n rows are of maximum rank ρ(**A**). Suppose **A** =

$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ is of this form, and apply the theorem to obtain $\mathbf{A}_{11} = \mathbf{P}_{11}\mathbf{L}_{11}\mathbf{DU}'\mathbf{Q}'$. The rank condition implies that $\mathbf{A}_{21} = \mathbf{F}\mathbf{A}_{11}$ for some (m-n)×n array $\mathbf{F}$. Then, $\mathbf{A}_{21} = \mathbf{L}_{21}\mathbf{DU}'\mathbf{Q}'$, with $\mathbf{L}_{21} = \mathbf{F}\mathbf{P}_{11}\mathbf{L}_{11}$, so that

$$\mathbf{A} = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} \mathbf{DU}'\mathbf{Q}'.$$

To complete the proof, apply a left permutation if necessary to undo the initial row permutation of **A**. Corollary 6 is an implication of the last result. □

The recursion in the proof of the theorem is called *Crout's algorithm*, and is the method for matrix inversion of positive definite matrices used in many computer programs. It is unnecessary to do the permutations in advance of the factorizations; they can also be carried out recursively, bringing in rows (in what is termed a *pivot*) to make the successive elements of **D** as large in magnitude as possible. This pivot step is important for numerical accuracy.

The Cholesky factorization of a n×n positive definite matrix **A** that was obtained above as a corollary of the **LDU** decomposition states that **A** can be written as $\mathbf{A} = \mathbf{LL}'$, where **L** is lower triangular with a positive diagonal. This factorization is readily computed and widely used in econometrics. We give a direct recursive construction of **L** that forms the basis for its computation. Write the factorization in partitioned form

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ U_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}' .$$

Also, let $\mathbf{V} = \mathbf{L}^{-1}$, and partition it commensurately, so that

$$\begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} V_{11} & 0 & 0 \\ V_{21} & V_{22} & 0 \\ V_{31} & V_{32} & V_{33} \end{bmatrix} .$$

Then $\mathbf{A}_{11} = \mathbf{L}_{11}\mathbf{L}_{11}'$, $\mathbf{A}_{12} = \mathbf{L}_{11}\mathbf{L}_{21}'$, $\mathbf{A}_{22} = \mathbf{L}_{21}\mathbf{L}_{21}' + \mathbf{L}_{22}\mathbf{L}_{22}'$, $\mathbf{V}_{11} = \mathbf{L}_{11}^{-1}$, $\mathbf{V}_{22} = \mathbf{L}_{22}^{-1}$, and $\mathbf{0} = \mathbf{L}_{21}\mathbf{V}_{11} + \mathbf{L}_{22}'\mathbf{V}_{22}$. Note first that if $\mathbf{A}_{11}$ is 1×1, then $\mathbf{L}_{11} = \mathbf{A}_{11}^{1/2}$ and $\mathbf{V}_{11} = 1/\mathbf{L}_{11}$. Now suppose that one has proceeded recursively from the northwest corner of these matrices, and that $\mathbf{L}_{11}$ and $\mathbf{V}_{11}$ have already been computed up through dimension n_1. Suppose that $\mathbf{A}_{22}$ is 1×1. Then, compute in sequence $\mathbf{L}_{21}' = \mathbf{V}_{11}'\mathbf{A}_{12}$, $\mathbf{L}_{22} = (\mathbf{A}_{22} - \mathbf{L}_{12}\mathbf{L}_{12}')^{1/2}$, $\mathbf{V}_{22} = 1/\mathbf{L}_{22}$, and $\mathbf{V}_{12} = -\mathbf{V}_{11}\mathbf{L}_{21}'\mathbf{V}_{22}$. This gives the required factors

up through dimension n_1+1. Repeat this for each dimension in turn to construct the full **L** and **V** matrices.

An extension of the Cholesky decomposition holds for an n×n positive semidefinite matrix **A** of rank r, which can be written as $\mathbf{A} = \mathbf{PLL'P'}$ with **P** a permutation matrix and **L** a lower triangular matrix whose first r diagonal elements are positive. The construction proceeds recursively as before, but at each stage one may have to search among remaining columns to find one for which $\mathbf{L}_{22} > 0$, determining the **P** matrix. Once dimension r is reached, all remaining columns will have $\mathbf{L}_{22} = 0$. Now reinterpret $\mathbf{L}_{21}$ and $\mathbf{L}_{22}$ as a partition corresponding to all the remaining columns and compute $\mathbf{L}_{12}' = \mathbf{V}_{11}'\mathbf{A}_{12}$ and $\mathbf{L}_{22} = \mathbf{0}$ to complete the Cholesky factor.

2.8. THE SINGULAR VALUE DECOMPOSITION OF A MATRIX

A factorization that is useful as a tool for finding the eigenvalues and eigenvectors of a symmetric matrix, and for calculation of inverses of moment matrices of data with high multicollinearity, is the *singular value decomposition* (SVD):

Theorem 2.2. Every real m×n matrix **A** of rank r can be decomposed into a product $\mathbf{A} = \mathbf{UDV}'$, where **D** is a r×r diagonal matrix with positive nonincreasing elements down the diagonal, **U** is m×r, **V** is n×r, and **U** and **V** are column-orthonormal; i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}_r = \mathbf{V}'\mathbf{V}$.

Proof: Note that the SVD is an extension of the LDU decomposition to non-square matrices. To prove that the SVD is possible, note first that the m×m matrix $\mathbf{AA}'$ is symmetric and positive semidefinite. Then, there exists a m×m orthonormal matrix **W** whose columns are eigenvectors of $\mathbf{AA}'$ arranged in non-increasing order for the eigenvalues, partitioned $\mathbf{W} = [\mathbf{W}_1\ \mathbf{W}_2]$ with $\mathbf{W}_1$ of dimension m×r, such that $\mathbf{W}_1'(\mathbf{AA}')\mathbf{W}_1 = \mathbf{\Lambda}$ is diagonal with positive, non-increasing diagonal elements, and $\mathbf{W}_2'(\mathbf{AA}')\mathbf{W}_2 = 0$, implying $\mathbf{A}'\mathbf{W}_2 = \mathbf{0}$. Define **D** from $\mathbf{\Lambda}$ by replacing the diagonal elements of $\mathbf{\Lambda}$ by their positive square roots. Note that $\mathbf{W}'\mathbf{W} = \mathbf{I} = \mathbf{WW}' \equiv \mathbf{W}_1\mathbf{W}_1' + \mathbf{W}_2\mathbf{W}_2'$. Define $\mathbf{U} = \mathbf{W}_1$ and $\mathbf{V}' = \mathbf{D}^{-1}\mathbf{U}'\mathbf{A}$. Then, $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}'\mathbf{V} = \mathbf{D}^{-1}\mathbf{U}'\mathbf{AA}'\mathbf{UD}^{-1} = \mathbf{D}^{-1}\mathbf{\Lambda}\mathbf{D}^{-1} = \mathbf{I}_r$. Further, $\mathbf{A} = (\mathbf{I}_m - \mathbf{W}_2\mathbf{W}_2')\mathbf{A} = \mathbf{UU}'\mathbf{A} = \mathbf{UDV}'$. This establishes the decomposition. □

If **A** is symmetric, then **U** is the array of eigenvectors of **A** corresponding to the non-zero roots, so that $\mathbf{A}'\mathbf{U} = \mathbf{UD}_1$, with $\mathbf{D}_1$ the r×r diagonal matrix with the non-zero eigenvalues in descending magnitude down the diagonal. In this case, $\mathbf{V} = \mathbf{A}'\mathbf{UD}^{-1} = \mathbf{UD}_1\mathbf{D}^{-1}$. Since the elements of $\mathbf{D}_1$ and **D** are identical except possibly for sign, the columns of **U** and **V** are either equal (for positive roots) or reversed in sign (for negative roots). Then, the SVD of a square symmetric nonsingular matrix provides the pieces necessary to write down its eigenvalues and eigenvectors. For a positive definite matrix, the connection is direct.

When the m×n matrix **A** is of rank n, so that $\mathbf{A}'\mathbf{A}$ is symmetric and positive definite, the SVD provides a method of calculating $(\mathbf{A}'\mathbf{A})^{-1}$ that is particularly numerically accurate: Substituting the form $\mathbf{A} = \mathbf{UDV}'$, one obtains $(\mathbf{A}'\mathbf{A})^{-1} = \mathbf{VD}^{-2}\mathbf{V}'$. One also obtains convenient forms for a square root of $\mathbf{A}'\mathbf{A}$ and its inverse, $(\mathbf{A}'\mathbf{A})^{1/2} = \mathbf{VDV}'$ and $(\mathbf{A}'\mathbf{A})^{-1/2} = \mathbf{VD}^{-1}\mathbf{V}'$.

The numerical accuracy of the SVD is most advantageous when m is large and some of the columns of A are nearly linearly dependent. Then, roundoff errors in the matrix product $\mathbf{A}'\mathbf{A}$ can lead to quite inaccurate results when a matrix inverse of $\mathbf{A}'\mathbf{A}$ is computed directly. The SVD extracts the required information from **A** before the roundoff errors in $\mathbf{A}'\mathbf{A}$ are introduced. Computer programs for the Singular Value Decomposition can be found in Press *et al*, Numerical Recipes, Cambridge University Press, 1986.

2.9. IDEMPOTENT MATRICES AND GENERALIZED INVERSES

A symmetric n×n matrix **A** is *idempotent* if $\mathbf{A}^2 = \mathbf{A}$. Examples of idempotent matrices are **0**, **I**, and for any n×r matrix **X** of rank r, $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Idempotent matrices are intimately related to projections, discussed in the following section. Some of the properties of an n×n idempotent matrix **A** are listed below:

(1) The eigenvalues of **A** are either zero or one.
(2) The rank of **A** equals tr(**A**).
(3) The matrix **I**-**A** is idempotent.
(4) If **B** is an orthonormal matrix, then $\mathbf{B}'\mathbf{AB}$ is idempotent.
(5) If $\rho(\mathbf{A}) = r$, then there exists a n×r matrix **B** of rank r such that $\mathbf{A} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$. [Let **C** be an orthonormal matrix that diagonalizes **A**, and take **B** to be the columns of **C** corresponding to the non-zero elements in the diagonalization.]
(6) **A**, **B** idempotent implies **AB** = **0** if and only if **A**+**B** is idempotent.
(7) **A**, **B** idempotent and **AB** = **BA** implies **AB** idempotent.
(8) **A**, **B** idempotent implies **A**-**B** idempotent if and only if **BA** = **B**.

Recall that a n×n non-singular matrix **A** has an inverse $\mathbf{A}^{-1}$ that satisfies $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. It is useful to extend the concept of an inverse to matrices that are not necessarily non-singular, or even square. For an m×k matrix **A** (of rank r), define its *Moore-Penrose generalized inverse* $\mathbf{A}^-$ to be a k×m matrix with the following three properties:

(i) $\mathbf{AA}^-\mathbf{A} = \mathbf{A}$,
(ii) $\mathbf{A}^-\mathbf{AA}^- = \mathbf{A}^-$
(iii) $\mathbf{AA}^-$ and $\mathbf{A}^-\mathbf{A}$ are symmetric

The next theorem shows that the Moore-Penrose generalized inverse always exists, and is unique. Conditions (i) and (ii) imply that the matrices $\mathbf{A}\mathbf{A}^-$ and $\mathbf{A}^-\mathbf{A}$ are idempotent. There are other generalized inverse definitions that have some, but not all, of the properties (i)-(iii); in particular $\mathbf{A}^+$ will denote any matrix that satisfies (i), or $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$.

Theorem 2.3. *The Moore-Penrose generalized inverse of a* m×k matrix **A** of rank r (which has a SVD $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$, where **U** is m×r**,** **V** is k×r, **U** and **V** are column-orthogonal, and **D** is r×r diagonal with positive diagonal elements) is the matrix $\mathbf{A}^- = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}'$. Let $\mathbf{A}^+$ denote any matrix, including $\mathbf{A}^-$, that satisfies $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$. These matrices satisfy:

(1) $\mathbf{A}^- = \mathbf{A}^+ = \mathbf{A}^{-1}$ if **A** is square and non-singular.
(2) The system of equations $\mathbf{A}\mathbf{x} = \mathbf{y}$ has a solution if and only if $\mathbf{y} = \mathbf{A}\mathbf{A}^+\mathbf{y}$, and the linear subspace of all solutions is the set of vectors $\mathbf{x} = \mathbf{A}^+\mathbf{y} + [\mathbf{I} - \mathbf{A}^+\mathbf{A}]\mathbf{z}$ for $\mathbf{z} \in \mathbb{R}^k$.
(3) $\mathbf{A}\mathbf{A}^+$ and $\mathbf{A}^+\mathbf{A}$ are idempotent.
(4) If **A** is idempotent, then $\mathbf{A} = \mathbf{A}^-$.
(5) If $\mathbf{A} = \mathbf{B}\mathbf{C}\mathbf{D}$ with **B** and **D** nonsingular, then $\mathbf{A}^- = \mathbf{D}^{-1}\,\mathbf{C}^-\,\mathbf{B}^{-1}$, and any matrix $\mathbf{A}^+ = \mathbf{D}^{-1}\mathbf{C}^+\mathbf{B}^{-1}$ satisfies $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$.
(6) $(\mathbf{A}')^- = (\mathbf{A}^-)'$
(7) $(\mathbf{A}'\mathbf{A})^- = \mathbf{A}^-(\mathbf{A}^-)'$
(8) $(\mathbf{A}^-)^- = \mathbf{A} = \mathbf{A}\mathbf{A}'(\mathbf{A}^-)' = (\mathbf{A}^-)'\mathbf{A}'\mathbf{A}$.

(9) If $\mathbf{A} = \sum_i \mathbf{A}_i$ with $\mathbf{A}_i'\mathbf{A}_j = \mathbf{0}$ and $\mathbf{A}_i\mathbf{A}_j' = \mathbf{0}$ for $i \neq j$, then $\mathbf{A}^- = \sum_i \mathbf{A}_i^-$.

Theorem 2.4. If **A** is m×m, symmetric, and positive semidefinite of rank r, then
(1) There exist **Q** positive definite and **R** idempotent of rank r such that $\mathbf{A} = \mathbf{Q}\mathbf{R}\mathbf{Q}$ and $\mathbf{A}^- = \mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1}$.
(2) There exists an m×r column-orthonormal matrix **U** such that $\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{D}$ is positive diagonal, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}'$, $\mathbf{A}^- = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}' = \mathbf{U}(\mathbf{U}'\mathbf{A}\mathbf{U})^{-1}\mathbf{U}'$, and any matrix $\mathbf{A}^+$ satisfying condition (i) for a generalized inverse, $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, has $\mathbf{U}'\mathbf{A}^+\mathbf{U} = \mathbf{D}^{-1}$.
(3) **A** has a symmetric square root $\mathbf{B} = \mathbf{A}^{1/2}$, and $\mathbf{A}^- = \mathbf{B}^-\mathbf{B}^-$.

Proof: Let **U** be an m×r column-orthonormal matrix of eigenvectors of **A** corresponding to the positive characteristic roots, and **W** be a m×(m-r) column-orthonormal matrix of eigenvectors corresponding to the zero characteristic roots. Then [**U W**] is an orthonormal matrix diagonalizing **A**, with $\begin{bmatrix} \boldsymbol{U}' \\ \boldsymbol{W}' \end{bmatrix} \boldsymbol{A} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{W} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ and **D** positive diagonal. Define $\mathbf{Q} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}^{1/2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{m-r} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}' \\ \boldsymbol{W}' \end{bmatrix}$, and $\mathbf{R} = \mathbf{U}\mathbf{U}'$. The diagonalizing transformation implies $\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{D}$ and $\mathbf{A}\mathbf{W} = \mathbf{0}$. One has $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$,

$\mathbf{W}'\mathbf{W} = \mathbf{I}_{m-r}$, and $\mathbf{U}\mathbf{U}' + \mathbf{W}\mathbf{W}' = \mathbf{I}_m$. Since $\mathbf{A}\mathbf{W} = \mathbf{0}$, $\mathbf{A} = \mathbf{A}[\mathbf{U}\mathbf{U}' + \mathbf{W}\mathbf{W}'] = \mathbf{A}\mathbf{U}\mathbf{U}'$ and $\mathbf{D} = \mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{U}'\mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{A}\mathbf{U}\mathbf{U}'\mathbf{A}^+\mathbf{U}\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{D}\mathbf{U}'\mathbf{A}^+\mathbf{U}\mathbf{D}$, implying $\mathbf{U}'\mathbf{A}^+\mathbf{U} = \mathbf{D}^{-1}$. Define $\mathbf{B} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}'$. □

2.10. PROJECTIONS

Consider a Euclidean space $\mathbb{R}^n$ of dimension n, and suppose $\mathbf{X}$ is a n×p array with columns that are vectors in this space. Let $\boldsymbol{X}$ denote the linear subspace of $\mathbb{R}^n$ that is *spanned* or *generated* by $\mathbf{X}$; i.e., the space formed by all linear combinations of the vectors in $\mathbf{X}$. Every linear subspace can be identified with an array such as $\mathbf{X}$. The dimension of the subspace is the rank of $\mathbf{X}$. (The array $\mathbf{X}$ need not be of full rank, although if it is not, then a subarray of linearly independent columns also generates $\boldsymbol{X}$.) A given $\mathbf{X}$ determines a unique subspace, so that $\mathbf{X}$ characterizes the subspace. However, any set of vectors contained in the subspace that form an array with the rank of the subspace, in particular any array $\mathbf{X}\mathbf{A}$ with rank equal to the dimension of $\mathbf{X}$, also generates $\boldsymbol{X}$. Then, $\mathbf{X}$ is not a unique characterization of the subspace it generates.

The *projection* of a vector y in $\mathbb{R}^n$ into the subspace $\boldsymbol{X}$ is defined as the point $\mathbf{v}$ in $\boldsymbol{X}$ that is the minimum Euclidean distance from $\mathbf{y}$. Since each vector $\mathbf{v}$ in $\boldsymbol{X}$ can be represented as a linear combination $\mathbf{X}\boldsymbol{\alpha}$ of an array $\mathbf{X}$ that generates $\boldsymbol{X}$, the projection is characterized by the value of $\boldsymbol{\alpha}$ that minimizes the squared Euclidean distance of $\mathbf{y}$ from $\boldsymbol{X}$, $(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})'(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})$. The solution to this problem is the vector $\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$ giving $\mathbf{v} = \mathbf{X}\hat{\boldsymbol{\alpha}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$. In these formulas, we use the Moore-Penrose generalized inverse $(\mathbf{X}'\mathbf{X})^{-}$ rather than $(\mathbf{X}'\mathbf{X})^{-1}$ so that the solution is defined even if $\mathbf{X}$ is not of full rank. The array $\boldsymbol{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is termed the *projection matrix* for the subspace $\boldsymbol{X}$; it is the linear transformation in $\mathbb{R}^n$ that maps any vector in the space into its projection $\mathbf{v}$ in $\boldsymbol{X}$. The matrix $\boldsymbol{P}_\mathbf{X}$ is *idempotent* (i.e., $\boldsymbol{P}_\mathbf{X}\boldsymbol{P}_\mathbf{X} = \boldsymbol{P}_\mathbf{X}$ and $\boldsymbol{P}_\mathbf{X} = \boldsymbol{P}_\mathbf{X}'$), and every idempotent matrix can be interpreted as a projection matrix. These observations have two important implications: First, the projection matrix is uniquely determined by $\mathbf{X}$, so that starting from a different array that generates $\boldsymbol{X}$, say an array $\mathbf{S} = \mathbf{X}\mathbf{A}$, implies $\boldsymbol{P}_\mathbf{X} = \boldsymbol{P}_\mathbf{S}$. (One could use the notation $\boldsymbol{P}_{\boldsymbol{X}}$ rather than $\boldsymbol{P}_\mathbf{X}$ to emphasize that the projection matrix depends only on the subspace, and not on any particular set of vectors that generate $\boldsymbol{X}$.) Second, if a vector $\mathbf{y}$ is contained in $\boldsymbol{X}$, then the projection into $\boldsymbol{X}$ leaves it unchanged, $\boldsymbol{P}_\mathbf{X}\mathbf{y} = \mathbf{y}$.

Define $\boldsymbol{Q}_\mathbf{X} = \mathbf{I} - \boldsymbol{P}_\mathbf{X} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$; it is the projection to the subspace orthogonal to that spanned by $\mathbf{X}$. Every vector $\mathbf{y}$ in $\mathbb{R}^n$ is uniquely decomposed into the sum of its projection $\boldsymbol{P}_\mathbf{X}\mathbf{y}$ onto $\boldsymbol{X}$ and its projection $\boldsymbol{Q}_\mathbf{X}\mathbf{y}$ onto the subspace orthogonal to $\boldsymbol{X}$. Note that $\boldsymbol{P}_\mathbf{X}\boldsymbol{Q}_\mathbf{X} = \mathbf{0}$, a property that holds in general for two projections onto orthogonal subspaces.

If $\boldsymbol{X}$ is a subspace generated by an array $\mathbf{X}$ and $\boldsymbol{W}$ is a subspace generated by a more inclusive array $\mathbf{W} = [\mathbf{X}\ \mathbf{Z}]$, then $\boldsymbol{X} \subseteq \boldsymbol{W}$. This implies that $\boldsymbol{P}_\mathbf{X}\boldsymbol{P}_\mathbf{W} = \mathbf{P}_{\boldsymbol{W}}\mathbf{P}_{\boldsymbol{X}} = \boldsymbol{P}_\mathbf{X}$; i.e., a projection onto a subspace is left invariant by a further projection onto a larger subspace, and a two-stage projection onto a large subspace followed by a projection onto a smaller one is the same as projecting directly onto the smaller one. The subspace of $\boldsymbol{W}$ that is orthogonal to $\boldsymbol{X}$ is generated by $\boldsymbol{Q}_\mathbf{X}\mathbf{W}$; i.e., it is the set of

linear combinations of the residuals, orthogonal to $\boldsymbol{X}$, obtained by the difference of $\mathbf{W}$ and its projection onto $\boldsymbol{X}$. Note that any $\mathbf{y}$ in $\mathbb{R}^n$ has a unique decomposition $\boldsymbol{P}_X\mathbf{y} + \boldsymbol{Q}_X\boldsymbol{P}_W\mathbf{y} + \boldsymbol{Q}_W\mathbf{y}$ into the sum of projections onto three mutually orthogonal subspaces, $\boldsymbol{X}$, the subspace of $\boldsymbol{W}$ orthogonal to $\boldsymbol{X}$, and the subspace orthogonal to $\boldsymbol{W}$. The projection $\boldsymbol{Q}_X\boldsymbol{P}_W$ can be rewritten $\boldsymbol{Q}_X\boldsymbol{P}_W = \boldsymbol{P}_W - \boldsymbol{P}_X = \boldsymbol{P}_W\boldsymbol{Q}_X = \boldsymbol{Q}_X\boldsymbol{P}_W\boldsymbol{Q}_X$, or since $\boldsymbol{Q}_X\mathbf{W} = \boldsymbol{Q}_X[\mathbf{X}\ \ \mathbf{Z}] = [\mathbf{0}\ \ \boldsymbol{Q}_X\mathbf{Z}]$, $\boldsymbol{Q}_X\boldsymbol{P}_W = \boldsymbol{P}_{\boldsymbol{Q}_X\boldsymbol{W}} = \boldsymbol{P}_{\boldsymbol{Q}_X\boldsymbol{Z}} = \boldsymbol{Q}_X\mathbf{Z}(\mathbf{Z}'\boldsymbol{Q}_X\mathbf{Z})^{-}\ \mathbf{Z}'\boldsymbol{Q}_X$.

This establishes that $\boldsymbol{P}_W$ and $\boldsymbol{Q}_X$ commute. This condition is necessary and sufficient for the product of two projections to be a projection; equivalently, it implies that $\boldsymbol{Q}_X\mathbf{P}_W$ is idempotent since $(\boldsymbol{Q}_X\boldsymbol{P}_W)(\boldsymbol{Q}_X\boldsymbol{P}_W) = \boldsymbol{Q}_X(\boldsymbol{P}_W\boldsymbol{Q}_X)\boldsymbol{P}_W = \boldsymbol{Q}_X(\boldsymbol{Q}_X\boldsymbol{P}_W)\boldsymbol{P}_W = \boldsymbol{Q}_X\boldsymbol{P}_W$.

2.11. KRONECKER PRODUCTS

If $\mathbf{A}$ is a m×n matrix and $\mathbf{B}$ is a p×q matrix, then the *Kronecker (direct) product* of $\mathbf{A}$ and $\mathbf{B}$ is the (mp)×(nq) partitioned array

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1m}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2m}\boldsymbol{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & \cdots & a_{nm}\boldsymbol{B} \end{bmatrix}.$$

In general, $\mathbf{A}\otimes\mathbf{B} \neq \mathbf{B}\otimes\mathbf{A}$. The Kronecker product has the following properties:

(1) For a scalar c, $(c\mathbf{A})\otimes\mathbf{B} = \mathbf{A}\otimes(c\mathbf{B}) = c(\mathbf{A}\otimes\mathbf{B})$.
(2) $(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C} = \mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$.
(3) $(\mathbf{A}\otimes\mathbf{B})' = (\mathbf{A}')\otimes(\mathbf{B}')$.
(4) $\mathrm{tr}(\mathbf{A}\otimes\mathbf{B}) = (\mathrm{tr}(\mathbf{A}))\cdot(\mathrm{tr}(\mathbf{B}))$ when $\mathbf{A}$ and $\mathbf{B}$ are square.
(5) If the matrix products $\mathbf{AC}$ and $\mathbf{BF}$ are defined, then $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{F}) = (\mathbf{AC})\otimes(\mathbf{BF})$.
(6) If $\mathbf{A}$ and $\mathbf{B}$ are square and nonsingular, then $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$.
(7) If $\mathbf{A}$ and $\mathbf{B}$ are orthonormal, then $\mathbf{A}\otimes\mathbf{B}$ is orthonormal.
(8) If $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite, then $\mathbf{A}\otimes\mathbf{B}$ is positive semidefinite.
(9) If $\mathbf{A}$ is k×k and $\mathbf{B}$ is n×n, then $\det(\mathbf{A}\otimes\mathbf{B}) = \det(\mathbf{A})^n\cdot\det(\mathbf{B})^k$.
(10) $\rho(\mathbf{A}\otimes\mathbf{B}) = \rho(\mathbf{A})\cdot\rho(\mathbf{B})$.
(11) $(\mathbf{A}+\mathbf{B})\otimes\mathbf{C} = \mathbf{A}\otimes\mathbf{C} + \mathbf{B}\otimes\mathbf{C}$.

2.12. SHAPING OPERATIONS

The most common operations used to reshape vectors and matrices are (1) $\mathbf{C}$ = diag($\mathbf{x}$) which creates a diagonal matrix with the elements of the vector $\mathbf{x}$ down the diagonal; (2) $\mathbf{c}$ = vecc($\mathbf{A}$) which creates a vector by stacking the columns of $\mathbf{A}$, and vecr($\mathbf{A}$) = vecc($\mathbf{A}'$); (3) $\mathbf{c}$ = vech($\mathbf{A}$) which creates a vector by stacking the portions of the rows of $\mathbf{A}$ that are in the upper triangle of the matrix; and (4) $\mathbf{c}$ = vecd($\mathbf{A}$) which creates a vector containing the diagonal of $\mathbf{A}$. (In some computer matrix languages, vec($\mathbf{A}$) stacks by row rather than by column.) There are a few rules that can be used to manipulate these operations:

(1) If $\mathbf{x}$ and $\mathbf{y}$ are commensurate vectors, diag($\mathbf{x}$+$\mathbf{y}$) = diag($\mathbf{x}$) + diag($\mathbf{y}$).
(2) vecc($\mathbf{A}$+$\mathbf{B}$) = vecc($\mathbf{A}$) + vecc($\mathbf{B}$).
(3) If $\mathbf{A}$ is m×k and $\mathbf{B}$ is k×n, then vecr($\mathbf{AB}$) = ($\mathbf{I}_n\otimes\mathbf{A}$)vecr($\mathbf{B}$) = ($\mathbf{B}'\otimes\mathbf{I}_m$)vecr($\mathbf{A}$).
(4) If $\mathbf{A}$ is m×k, $\mathbf{B}$ is k×n, $\mathbf{C}$ is n×p, then vecr($\mathbf{ABC}$) = ($\mathbf{I}_p\otimes$($\mathbf{AB}$))vecr© = ($\mathbf{C}'\otimes\mathbf{A}$)vecr($\mathbf{B}$) = (($\mathbf{C}'\mathbf{B}'$)$\otimes\mathbf{I}_m$)vecr($\mathbf{A}$).
(5) If $\mathbf{A}$ is n×n, then vech($\mathbf{A}$) is of length n(n+1)/2.
(6) vecd(diag($\mathbf{x}$)) = $\mathbf{x}$.

2.13. VECTOR AND MATRIX DERIVATIVES

The derivatives of functions with respect to the elements of vectors or matrices can sometimes be expressed in a convenient matrix form. First, a scalar function of a n×1 vector of variables, f($\mathbf{x}$), has partial derivatives that are usually written as the arrays

$$\partial f/\partial \mathbf{x} = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ | \\ \partial f/\partial x_n \end{bmatrix}, \quad \partial f^2/\partial \mathbf{x}\partial \mathbf{x}' = \begin{bmatrix} \partial f^2/\partial x_1 & \partial f^2/\partial x_1\partial x_2 & \dots & \partial f^2/\partial x_1\partial x_n \\ \partial f^2/\partial x_2\partial x_1 & \partial f^2/\partial x_2 & \dots & \partial f^2/\partial x_2\partial x_n \\ | & | & & | \\ \partial f^2/\partial x_n\partial x_1 & \partial f^2/\partial x_n\partial x_2 & \dots & \partial f^2/\partial x_n \end{bmatrix}.$$

Other common notation is $f_{\mathbf{x}}(\mathbf{x})$ or $\nabla_{\mathbf{x}}f(\mathbf{x})$ for the vector of first derivatives, and $f_{\mathbf{xx}}(\mathbf{x})$ or $\nabla_{\mathbf{xx}}f(\mathbf{x})$ for the matrix of second derivatives. Sometimes, the vector of first derivatives will be interpreted as a row vector rather than a column vector. Some examples of scalar functions of a vector are the linear function $f(\mathbf{x}) = \mathbf{a}'\mathbf{x}$, which has $\nabla_{\mathbf{x}}f = \mathbf{a}$, and the quadratic function $f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, which has $\nabla_{\mathbf{x}}f = 2\mathbf{A}\mathbf{x}$.

When $\mathbf{f}$ is a column vector of scalar functions, $\mathbf{f}(\mathbf{x}) = [f^1(\mathbf{x})\ f^2(\mathbf{x})\ \dots\ f^k(\mathbf{x})]'$, then the array of first partial derivatives is called the *Jacobean matrix* and is written

$$\boldsymbol{J}(\mathbf{x}) = \begin{bmatrix} \partial f^1/\partial x_1 & \partial f^1/\partial x_2 & \dots & \partial f^1/\partial x_n \\ \partial f^2/\partial x_1 & \partial f^2/\partial x_2 & \dots & \partial f^2/\partial x_n \\ \partial f^k/\partial x_1 & \partial f^k/\partial x_2 & \dots & \partial f^k/\partial x_n \end{bmatrix}.$$

When calculating multivariate integrals of the form $\int_A g(\mathbf{y})d\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \subseteq \mathbb{R}^n$, and g is a scalar or vector function of **y**, one may want to make a nonlinear one-to-one transformation of variables $\mathbf{y} = \mathbf{f}(\mathbf{x})$. In terms of the transformed variables, the integral becomes

$$\int_A g(\mathbf{y})d\mathbf{y} = \int_{f^{-1}(A)} g(\mathbf{f}(\mathbf{x}))\cdot|\det(\boldsymbol{J}(\mathbf{x}))|\,d\mathbf{x},$$

where $\mathbf{f}^{-1}(\mathbf{A})$ is the set of **x** vectors that map onto **A**, and the Jacobean matrix is square and nonsingular for well-behaved one-to-one transformations. The intuition for the presence of the Jacobean determinant in the transformed integral is that "d**y**" is the volume of a small rectangle in **y**-space, and because determinants give the volume of the parallelepiped formed by the columns of a linear transformation, "$\det(\boldsymbol{J}(\mathbf{x}))d\mathbf{x}$" gives the volume (with a plus or minus sign) of the image in **x**-space of the "d**y**" rectangle in **y**-space.

It is useful to define the derivative of a scalar function with respect to a matrix as an array of commensurate dimensions. Consider the bilinear form $f(\mathbf{A}) = \mathbf{x}'\mathbf{A}\mathbf{y}$, where **x** is n×1, **y** is m×1, and **A** is n×m. By collecting the individual terms $\partial f/\partial A_{ij} = x_i y_j$, one obtains the result $\partial f/\partial \mathbf{A} = \mathbf{x}\mathbf{y}'$. Another example for a n×n matrix **A** is $f(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$, which has $\partial f/\partial \mathbf{A} = \mathbf{I}_n$. There are a few other derivatives that are particularly useful for statistical applications. In these formulas, **A** is a square nonsingular matrix. We do <u>not</u> require that **A** be symmetric, and the derivatives do <u>not</u> impose symmetry. One will still get valid calculations involving derivatives when these expressions are evaluated at matrices that happen to be symmetric. There are alternative, and somewhat more complicated, derivative formulas that hold when symmetry is imposed. For analysis, it is unnecessary to introduce this complication.

(1) If $\det(\mathbf{A}) > 0$, then $\partial\log(\det(\mathbf{A}))/\partial\mathbf{A} = \mathbf{A}^{-1}$.
(2) If **A** is nonsingular, then $\partial(\mathbf{x}'\mathbf{A}^{-1}\mathbf{y})/\partial\mathbf{A} = -\mathbf{A}^{-1}\mathbf{x}\mathbf{y}'\mathbf{A}^{-1}$.
(3) If $\mathbf{A} = \mathbf{T}\mathbf{T}'$, with **T** square and nonsingular, then $\partial(\mathbf{x}'\mathbf{A}^{-1}\mathbf{y})/\partial\mathbf{T} = -2\mathbf{A}^{-1}\mathbf{x}\mathbf{y}'\mathbf{A}^{-1}\mathbf{T}$.

We prove the formulas in order. For (1), recall that $\det(\mathbf{A}) = \sum_k (-1)^{i+k}a_{ik}\det(\mathbf{A}^{ik})$, where $\mathbf{A}^{ik}$ is the minor of a_{ik}. Then, $\partial\det(\mathbf{A})/\partial\mathbf{A}_{ij} = (-1)^{i+j}\det(\mathbf{A}^{ij})$. From 2.3.17, the ij element of $\mathbf{A}^{-1}$ is $(-1)^{i+j}\det(\mathbf{A}^{ij})/\det(\mathbf{A})$. For (2), apply the chain rule to the identity $\mathbf{A}\mathbf{A}^{-1} \equiv \mathrm{I}$ to get $\boldsymbol{\Delta}_{ij}\mathbf{A}^{-1} + \mathbf{A}\cdot\partial\mathbf{A}^{-1}/\partial\mathbf{A}_{ij}$

$\equiv \mathbf{0}$, where $\mathbf{\Delta}_{ij}$ denotes a matrix with a one in row i and column j, zeros elsewhere. Then, $\partial \mathbf{x}'\mathbf{A}^{-1}\mathbf{y}/\partial \mathbf{A}_{ij} = -\mathbf{x}'\mathbf{A}^{-1}\mathbf{\Delta}_{ij}\mathbf{A}^{-1}\mathbf{y} = (\mathbf{A}^{-1}\mathbf{x})_i(\mathbf{A}^{-1}\mathbf{y})_j$. For (3), first note that $\partial \mathbf{A}_{ij}/\partial \mathbf{T}_{rs} = \delta_{ir}\mathbf{T}_{js}+\delta_{jr}\mathbf{T}_{is}$. Combine this with (2) to get

$$\partial \mathbf{x}'\mathbf{A}^{-1}\mathbf{y}/\partial \mathbf{T}_{rs} = \sum_j (\mathbf{A}^{-1}\mathbf{x})_i(\mathbf{A}^{-1}\mathbf{y})_j(\delta_{ir}\mathbf{T}_{js}+\delta_{jr}\mathbf{T}_{is})$$

$$= \sum_j (\mathbf{A}^{-1}\mathbf{x})_r(\mathbf{A}^{-1}\mathbf{y})_j\mathbf{T}_{js} + \sum_i (\mathbf{A}^{-1}\mathbf{x})_i(\mathbf{A}^{-1}\mathbf{y})_r\mathbf{T}_{is} = 2(\mathbf{A}^{-1}\mathbf{x}\mathbf{y}'\mathbf{A}^{-1}\mathbf{T})_{rs}.$$

2.14. UPDATING AND BACKDATING MATRIX OPERATIONS

Often in statistical applications, one needs to modify the calculation of a matrix inverse or other matrix operation to accommodate the addition of data, or deletion of data in bootstrap methods. It is convenient to have quick methods for these calculations. Some of the useful formulas are given below:

(1) If $\mathbf{A}$ is n×n and nonsingular, and $\mathbf{A}^{-1}$ has been calculated, and if $\mathbf{B}$ and $\mathbf{C}$ are arrays that are n×k of rank k, then $(\mathbf{A}+\mathbf{B}\mathbf{C}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I}_k+\mathbf{C}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}'\mathbf{A}^{-1}$, provided $\mathbf{I}_k+\mathbf{C}'\mathbf{A}^{-1}\mathbf{B}$ is nonsingular. No matrix inversion is required if k = 1.
(2) If $\mathbf{A}$ is m×n with m ≥ n and $\rho(\mathbf{A}) = n$, so that it has a LDU factorization $\mathbf{A} = \mathbf{PLDU}'\mathbf{Q}'$ with $\mathbf{D}$ n×n diagonal, $\mathbf{P}$ and $\mathbf{Q}$ permutation matrices, $\mathbf{L}$ and $\mathbf{U}$ lower triangular, then the array $\begin{bmatrix} A \\ B \end{bmatrix}$, with $\mathbf{B}$ k×n, has the LDU factorization $\begin{bmatrix} P & 0 \\ 0 & I_k \end{bmatrix}\begin{bmatrix} L \\ C \end{bmatrix} DU'Q'$, where $\mathbf{C} = \mathbf{BQU}'^{-1}\mathbf{D}^{-1}$.

(3) Suppose $\mathbf{A}$ is m×n of rank n, and $\mathbf{b} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}$. Suppose $\mathbf{A}^* = \begin{bmatrix} A \\ C \end{bmatrix}$ and $y^* = \begin{bmatrix} y \\ w \end{bmatrix}$ with $\mathbf{C}$ k×n and $\mathbf{w}$ k×1, and $\mathbf{b}^* = (\mathbf{A}^{*\prime}\mathbf{A}^*)^{-1}\mathbf{A}^{*\prime}\mathbf{y}^*$. Then,

$$\mathbf{b}^* - \mathbf{b} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{C}'[\mathbf{I}_k+\mathbf{C}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{C}']^{-1}(\mathbf{w}-\mathbf{C}\mathbf{b}) = (\mathbf{A}^{*\prime}\mathbf{A}^*)^{-1}\mathbf{C}'[\mathbf{I}_k-\mathbf{C}(\mathbf{A}^{*\prime}\mathbf{A}^*)^{-1}\mathbf{C}']^{-1}(\mathbf{w}-\mathbf{C}\mathbf{b}^*).$$

One can verify (1) by multiplication. To show (2), use Corollary 5 of Theorem 2.1. To show (3), apply (1) to $\mathbf{A}^{*\prime}\mathbf{A}^* = \mathbf{A}'\mathbf{A} + \mathbf{C}'\mathbf{C}$, or to $\mathbf{A}'\mathbf{A} = \mathbf{A}^{*\prime}\mathbf{A}^* - \mathbf{C}'\mathbf{C}$, and use $\mathbf{A}^{*\prime}\mathbf{y}^* = \mathbf{A}\mathbf{y} + \mathbf{C}\mathbf{w}$.

2.15. NOTES AND COMMENTS

The basic results of linear algebra, including the results stated without proof in this summary, can be found in standard linear algebra texts, such as G. Hadley (1961) Linear Algebra, Addison-Wesley or F. Graybill (1983) Matrices with Applications in Statistics, Wadsworth. The organization of this summary is based on the admirable synopsis of matrix theory in the first chapter of F. Graybill (1961) An Introduction to Linear Statistical Models, McGraw-Hill. For computations involving matrices, W. Press *et al* (1986) Numerical Recipes, Cambridge Univ. Press, provides a good discussion of algorithms and accompanying computer code. For numerical issues in statistical computation, see R. Thisted (1988) Elements of Statistical Computing, Chapman and Hall.

2.16. Exercises

1. The *conjugate* of a complex number $z = a+\iota b$ is the complex number $z^* = a - \iota b$. The square of the *modulus* of a complex number is the product of the number and its complex conjugate. Show that this definition is the same as the definition of the modulus r of a complex number written in polar representation as $z = r\cdot e^{\iota\theta} = r(\cos\theta + \iota\cdot\sin\theta)$.

2. Show that $A \setminus B = A\cap B^c$, $A\cap B = A\cup B \setminus (A\setminus B) \setminus (B\setminus A)$, $A\cup B = A\cap B \cup (A\setminus B) \cup (B\setminus A)$, and if $A\cap B = \varnothing$, then $A\setminus B = A$

3. Consider the real-valued function $y = f(x) \equiv x^2$ on the real line. Find the image of sets of the form [a,b]. Find the inverse image of sets of the form [c,d]. Is the mapping f^{-1} a real-valued function?

4. Use the Cauchy criterion to show that the sequence $a_n = 1 + ... + r^n$ has a limit if $|r| < 1$, but not if $|r| \geq 1$.

5. Show that the real line is a separable metric space for each of the following distance functions: $\rho(x,y) = |x-y|$, $\rho(x,y) = |x-y|^{1/2}$, $\rho(x,y) = \min(|x-y|,1)$. Show that $\rho(x,y) = (x - y)^2$ fails to satisfy the triangle inequality for distance functions.

6. Show that the function $f(x) = \sin(1/x)$ is continuous, but not uniformly continuous, on the set (0,1]. Prove that a continuous function on a compact set is uniformly continuous.

7. What are the differentiability properties of the real-valued function $f(x) = |x|^{7/2}$ at $x = 0$? At $x \neq 0$? Does this function have a second-order Taylor's expansion at $x = 0$?

8. Find the limit of the function $x^{\alpha}\cdot\log(x)$ for positive x as x goes to zero, where α is a positive constant. What about $(1+\alpha x)^{1/x}$?

9. Show that the series $a_n = (-1)^n$ is Cesaro summable, but not summable.

10. Use Kronecker's lemma to show $a_n = \log(1+n)$ and $b_n = n^{1/\alpha}$ for any positive constant α imply $n^{-1/\alpha}\cdot\log((n+1)!) \to 0$.

11. State Holder's inequality in the limiting case p = 1 and q = +∞.

12. Consider the matrix A = $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.25 \end{bmatrix}$. Is it symmetric? Idempotent? What is its rank?

13. What is the rank of the matrix A = $\begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{bmatrix}$?

14. For the matrices A = $\begin{bmatrix} 1 & 0.5 \\ 1 & 0.5 \end{bmatrix}$ and B = $\begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{bmatrix}$, determine which of the operations in Tables 2.1-2.3 can be applied, and calculate the result if the operations do apply.

15. The determanent of a 2×2 matrix A = $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is det(A) = ad - bc. Show that this formula satisfies the determanental identity in Section 2.3.3.

16. Prove Section 2.4 (3) and (4).

17. Prove, by multiplying out the formula, the result in Section 2.5 (2) for the inverse of partitioned matrices.

18. Prove Section 2.6 (1).

19. Calculate the Cholesky factorization and the Singular Value decomposition of the matrix A = $\begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}$.

20. The Singular Value Decomposition of a matrix A of dimension m×k and rank r was defined in Section 2.8 as a product A = UDV′, where U was m×r, D was r×r, V was k×r, the matrices U and V were both column orthogonal (i.e., U′U = I = V′V) and D was diagonal with positive diagonal elements. An alternative definition, which is equivalent, is to write A = [U W_2][D 0]′V′, where U, D, and V are the same as before, the array of 0's in [D 0] is r×(m-r), W_2 is m×(m-r), and [U W_2] is m*m and column orthogonal, and therefore orthonormal. Some computer programs give SVD's in this alternative form. Define B = V[D^{-1} 0][U W_2]′ and C = V[D^{-1} G][U W_2]′, where G is any non-zero r×(m-r) array. Show that ABA = ACA = A and that BAB = CAC = B, but then CAC ≠ C.

21. Calculate the Moore-Penrose generalized inverse of A = $\begin{bmatrix} 1 & 0.5 \\ 1 & 0.5 \end{bmatrix}$, and show that A^-A and AA^- are idempotent.

22. Consider the matrices A = $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, B = $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, C = $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$, E = $\begin{bmatrix} 0.5 & 0.5 \\ 0 & 0 \end{bmatrix}$, and F = $\begin{bmatrix} 0.5 & 0.5 \\ \alpha & -\alpha \end{bmatrix}$.
Which of the matrices B, C, E, F meet which of the conditions in Section 2.9 to be a generalized inverse of A?

23. Prove Theorem 2.3 (2). Show that if one writes the matrix in terms of its SVD, A = UDV′, then the equations have a solution iff UU′y = y, and if there is a solution, then it satisfies $x = VD^{-1}U'y + [I - UU']z$.

24. In $\mathbb{R}^3$, consider the subspace **A** generated by the vectors (1,1,1), (1,-1,1), and (1,3,1), the subspace **B** generated by the vectors (2,1,1) and (-4,-2,-2), and the subspace **C** generated by the vector (1,1,1). What are the dimensions of these subspaces? What is the projection of **B** and **C** on **A**? Of **A** and **C** on **B**? Of **A** and **B** on **C**?

25. Prove a linear transformation **A** in $\mathbb{R}^n$ is a projection iff **A** is idempotent. Show that if **A** and **B** are projections, then **A** + **B** is a projection iff **AB** = **BA** = 0, and **A** - **B** is a projection iff **AB** = **BA** = **B**.

26. Prove 2.11 (6) and (8).

27. Verify Section 2.12 (3) and (4) for the matrices **A** = $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, **B** = $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, **C** = $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$.

28. Consider the function $g(x_1,x_2) = \exp(-x_1/2 - x_2/2)$, and the transformation of variables $x_1 = r\cdot\cos\theta$ and $x_2 = r\cdot\sin\theta$ for $r \geq 0$ and $0 \leq \theta \leq 2\pi$. What is the Jacobean of this transformation? Evaluate the integral $\int_{-\infty}^{+\infty} g(x_1,x_2)dx_1dx_2$.

29. Prove 2.14 (1) and (2).

30. Suppose real-valued functions F(x) and G(x) are continuously differentiable on [a,b] with $f(x) = \nabla_x F(x)$ and $g(x) = \nabla_x G(x)$. Then, $\nabla_x(F(x)\cdot G(x)) = F(x)\cdot g(x) + f(x)\cdot G(x)$. Integrate this formula over [a,b] to establish the integration by parts formula $\int_a^b f(x)\cdot G(x)dx = F(b)G(b) - F(a)G(a) - \int_a^b F(x)\cdot g(x)dx$. Evaluate the integral $\int_0^{+\infty} x\cdot e^{-x}dx$.

CHAPTER 3. A REVIEW OF PROBABILITY THEORY

3.1. SAMPLE SPACE

The starting point for probability theory is the concept of a *state of Nature*, which is a description of everything that has happened and will happen in the universe. In particular, this description includes the outcomes of all probability and sampling experiments. The set of all possible states of Nature is called the *sample space*. Let s denote a state of Nature, and $\mathbf{S}$ the sample space. These are abstract objects that play a conceptual rather than a practical role in the development of probability theory. Consequently, there can be considerable flexibility in thinking about what goes into the description of a state of Nature and into the specification of the sample space; the only critical restriction is that there be enough states of Nature so that distinct observations are always associated with distinct states of Nature. In elementary probability theory, it is often convenient to think of the states of Nature as corresponding to the outcomes of a particular experiment, such as flipping coins or tossing dice, and to suppress the description of everything else in the universe. Sections 3.2-3.4 in this Chapter contain a few crucial definitions, for events, probabilities, conditional probabilities, and statistical independence. They also contain a treatment of measurability, the theory of integration, and probability on product spaces that is needed mostly for more advanced topics in econometrics. Therefore, readers who do not have a good background in mathematical analysis may find it useful to concentrate on the definitions and examples in these sections, and postpone study of the more mathematical material until it is needed.

3.2. EVENT FIELDS AND INFORMATION

3.2.1. An *event* is a set of states of Nature with the property that one can in principle determine whether the event occurs or not. If states of Nature describe all happenings, including the outcome of a particular coin toss, then one event might be the set of states of Nature in which this coin toss comes up heads. The family of potentially observable events is denoted by $\boldsymbol{F}$. This family is assumed to have the following properties:

(i) The "anything can happen" event $\mathbf{S}$ is in $\boldsymbol{F}$.

(ii) If event $\mathbf{A}$ is in $\boldsymbol{F}$, then the event "not $\mathbf{A}$", denoted $\mathbf{A}^c$ or $\mathbf{S}\backslash\mathbf{A}$, is in $\boldsymbol{F}$.

(iii) If $\mathbf{A}$ and $\mathbf{B}$ are events in $\boldsymbol{F}$, then the event "both $\mathbf{A}$ and $\mathbf{B}$", denoted $\mathbf{A}\cap\mathbf{B}$, is in $\boldsymbol{F}$.

(iv) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a finite or countable sequence of events in $\boldsymbol{F}$, then the event "one or more of $\mathbf{A}_1$ or $\mathbf{A}_2$ or ...", denoted $\bigcup_{i=1}^{\infty} \mathbf{A}_i$, is in $\boldsymbol{F}$.

A family $\boldsymbol{F}$ with these properties is called a σ-*field* (or *Boolean* σ-*algebra*) of subsets of $\mathbf{S}$. The pair $(\mathbf{S},\boldsymbol{F})$ consisting of an abstract set $\mathbf{S}$ and a σ-field $\boldsymbol{F}$ of subsets of $\mathbf{S}$ is called a *measurable space*, and the sets in $\boldsymbol{F}$ are called the *measurable* subsets of $\mathbf{S}$. Implications of the definition of a σ-field are

(v) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a finite or countable sequence of events in $\boldsymbol{F}$, then $\bigcap_{i=1}^{\infty}\boldsymbol{A}_i$ is also in $\boldsymbol{F}$.

(vi) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a countable sequence of events in $\boldsymbol{F}$ that is *monotone decreasing* (i.e., $\mathbf{A}_1 \supseteq \mathbf{A}_2 \supseteq ...$), then its limit, also denoted $A_i \searrow A_0$, is also in $\boldsymbol{F}$. Similarly, if a sequence in $\boldsymbol{F}$ is *monotone increasing* (i.e., $\mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq ...$), then its limit $\mathbf{A}_0 = \bigcup_{i=1}^{\infty}\boldsymbol{A}_i$, is also in $\boldsymbol{F}$

(vii) The empty event $\boldsymbol{\varphi}$ is in $\boldsymbol{F}$.

We will use a few concrete examples of sample spaces and σ-fields:

Example 1. [Two coin tosses] A coin is tossed twice, and for each toss a head or tail appears. Let HT denote the state of Nature in which the first toss yields a head and the second toss yields a tail. Then $\mathbf{S}$ = {HH,HT,TH,TT}. Let $\boldsymbol{F}$ be the class of all possible subsets of $\mathbf{S}$; $\boldsymbol{F}$ has 2^4 members.

Example 2. [Coin toss until a tail] A coin is tossed until a tail appears. The sample space is $\mathbf{S}$ = {T, HT, HHT, HHHT,...}. In this example, the sample space is infinite, but countable. Let $\boldsymbol{F}$ be the σ-field generated by the finite subsets of $\mathbf{S}$. This σ-field contains events such as "At most ten heads", and also, using the monotone closure property (vi) above, events such as "Ten or more tosses without a tail", and "an even number of heads before a tail". A set that is not in $\boldsymbol{F}$ will have the property that both the set and its complement are infinite. It is difficult to describe such a set, primarily because the language that we normally use to construct sets tends to correspond to elements in the σ-field. However, mathematical analysis shows that such sets must exist, because the cardinality of the class of all possible subsets of $\mathbf{S}$ is greater than the cardinality of $\boldsymbol{F}$.

Example 3. [S&P stock index] The stock index is a number in the positive real line $\mathbb{R}_+$, so $\mathbf{S} \equiv \mathbb{R}_+$. Take the σ-field of events to be the *Borel* σ*-field* $\boldsymbol{B}(\mathbb{R}_+)$, which is defined as the smallest family of subsets of the real line that contains all the open intervals in $\mathbb{R}_+$ and satisfies the properties (i)-(iv) of a σ-field. The subsets of $\mathbb{R}_+$ that are in $\boldsymbol{B}$ are said to be *measurable*, and those not in $\boldsymbol{B}$ are said to be non-measurable.

Example 4. [S&P stock index on successive days] The set of states of Nature is the Cartesian product of the set of calues on day one and the set of values on day 2, $\mathbf{S} = \mathbb{R}_+ \times \mathbb{R}_+$ (also denoted $\mathbb{R}_+^2$). Take the σ-field of events to be the product of the one-dimensional σ-fields, $\boldsymbol{F} = \boldsymbol{B}_1 \otimes \boldsymbol{B}_2$, where "⊗" denotes an operation that forms the smallest σ-field containing all sets of the form $\mathbf{A}\times\mathbf{C}$ with $\mathbf{A} \in \boldsymbol{B}_1$ and $\mathbf{C} \in \boldsymbol{B}_2$. In this example, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are identical copies of the Borel σ-field on $\mathbb{R}_+$. Assume that the index was normalized to be one at the beginning of the previous year. Examples of events in $\boldsymbol{F}$ are "below 1 on day 1", "at least 2 on both days", and "higher on the second day than the first day". The operation "⊗" is different than the cartesian product "×", where $\boldsymbol{B}_1 \times \boldsymbol{B}_2$ is the family of all

rectangles $\mathbf{A}\times\mathbf{C}$ formed from $\mathbf{A} \in \boldsymbol{B}_1$ and $\mathbf{C} \in \boldsymbol{B}_2$. This family is not itself a σ-field, but the σ-field that it generates is $\boldsymbol{B}_1 \otimes \boldsymbol{B}_2$. For example, the event "higher on the second day than the first day" is not a rectangle, but is obtained as a monotone limit of rectangles.

In the first example, the σ-field consisted of all possible subsets of the sample space. This was not the case in the last two examples, because the Borel σ-field does not contain all subsets of the real line. There are two reasons to introduce the complication of dealing with σ-fields that do not contain all the subsets of the sample space, one substantive and one technical. The substantive reason is that the σ-field can be interpreted as the potential information that is available by observation. If an observer is incapable of making observations that distinguish two states of Nature, then the σ-field cannot contain sets that include one of these states and excludes the other. Then, the specification of the σ-field will depend on what is observable in an application. The technical reason is that when the sample space contains an infinite number of states, it may be mathematically impossible to define probabilities with sensible properties on all subsets of the sample space. Restricting the definition of probabilities to appropriately chosen σ-fields solves this problem.

3.2.2. It is possible that more than one σ-field of subsets is defined for a particular sample space **S**. If $\boldsymbol{A}$ is an arbitrary collection of subsets of **S**, then the smallest σ-field that contains $\boldsymbol{A}$ is said to be the σ-field *generated* by $\boldsymbol{A}$. It is sometimes denoted $\sigma(\boldsymbol{A})$. If $\boldsymbol{F}$ and $\boldsymbol{G}$ are both σ-fields, and $\boldsymbol{G} \subseteq \boldsymbol{F}$, then $\boldsymbol{G}$ is said to be a *sub-field* of $\boldsymbol{F}$, and $\boldsymbol{F}$ is said to *contain more information* or *refine* $\boldsymbol{G}$. It is possible that neither $\boldsymbol{F} \subseteq \boldsymbol{G}$ nor $\boldsymbol{G} \subseteq \boldsymbol{F}$. The intersection $\boldsymbol{F}\cap\boldsymbol{G}$ of two σ-fields is again a σ-field that contains the *common information* in $\boldsymbol{F}$ and $\boldsymbol{G}$. Further, the intersection of an arbitrary countable or uncountable collection of σ-fields is again a σ-field. The union $\boldsymbol{F}\cup\boldsymbol{G}$ of two σ-fields is not necessarily a σ-field, but there is always a smallest σ-field that refines both $\boldsymbol{F}$ and $\boldsymbol{G}$, which is simply the σ-field $\sigma(\boldsymbol{F}\cup\boldsymbol{G})$ generated by the sets in the union of $\boldsymbol{F}$ and $\boldsymbol{G}$, or put another way, the intersection of all σ-fields that contain both $\boldsymbol{F}$ and $\boldsymbol{G}$.

Example 1. (continued) Let $\boldsymbol{F}$ denote the σ-field of all subsets of **S**. Another σ-field is $\boldsymbol{G}$ = {φ,**S**,{HT,HH},{TT,TH}}, containing events with information only on the outcome of the first coin toss. Yet another σ-field contains the events with information only on the number of heads, but not their order, $\boldsymbol{H}$ = {φ,**S**,{HH},{TT},{HT,TH},{HH,TT},{HT,TH,TT},{HH,HT,TH}}. Then, $\boldsymbol{F}$ contains more information than $\boldsymbol{G}$ or $\boldsymbol{H}$. The intersection $\boldsymbol{G}\cap\boldsymbol{H}$ is the "no information" σ-field {φ,**S**}. The union $\boldsymbol{G}\cup\boldsymbol{H}$ is not a σ-field, and the σ-field $\sigma(\boldsymbol{G}\cup\boldsymbol{H})$ that it generates is $\boldsymbol{F}$. This can be verified constructively (in this finite S case) by building up $\sigma(\boldsymbol{G}\cup\boldsymbol{H})$ by forming intersections and unions of members of $\boldsymbol{G}\cup\boldsymbol{H}$, but is also obvious since knowing the outcome of the first toss and knowing the total number of heads reveals full information on both tosses.

Example 3. (continued) Let $\boldsymbol{F}$ denote the Borel σ-field. Then $\boldsymbol{G}$ = {φ,**S**,(1,∞),(-∞,1]} and $\boldsymbol{D}$ = {φ,**S**,{-∞,2),[2,∞)} are both σ-fields, the first corresponding to the ability to observe whether the

index is above 1, the second corresponding to the ability to tell whether it is above 2. For shorthand, let a = (-∞,1], b = (-∞,2], c = (1,+∞), d = (2,+∞), and e = (1,2]. Neither $\boldsymbol{G}$ or $\boldsymbol{D}$ contains the other, both are contained in $\boldsymbol{F}$, and their intersection is the "no information" σ-field {φ,**S**}. The σ-field generated by their union, corresponding to the ability to tell if the index is in a,e, or d, is σ($\boldsymbol{G}\cup\boldsymbol{D}$) = {φ,**S**,a,b,c,d,e,a∪d}.

An element **B** in a σ-field $\boldsymbol{G}$ of subsets of **S** is an *atom* if the only set in $\boldsymbol{G}$ that is a proper subset of **B** is the empty set φ. In the last example, $\boldsymbol{D}$ has atoms b and d, and the atoms of σ($\boldsymbol{G}\cup\boldsymbol{D}$) are a, d, and e, but not b = a∪e or c = e∪d. The atoms of the Borel σ-field are the individual real numbers. An economic interpretation of this concept that if the σ-field defining the common information of two economic agents contains an atom, then a contingent contract between them must have the same realization no matter what state of Nature within this atom occurs.

3.3. PROBABILITY

3.3.1. Given a sample space **S** and σ-field of subsets $\boldsymbol{F}$, a *probability* (or *probability measure*) is defined as a function P from $\boldsymbol{F}$ into the real line with the following properties:

(i) P(**A**) ≥ 0 for all **A** ∈ $\boldsymbol{F}$.
(ii) P(**S**) = 1.
(iii) [Countable Additivity] If $\mathbf{A}_1$, $\mathbf{A}_2$,... is a finite or countable sequence of events in $\boldsymbol{F}$ that are mutually exclusive (i.e., $\mathbf{A}_i\cap\mathbf{A}_j = \boldsymbol{\varphi}$ for all i ≠ j), then $P(\bigcup_{i=1}^{\infty} \mathbf{A}_i) = \sum_{i=1}^{\infty} P(\mathbf{A}_i)$.

With conditions (i)-(iii), P has the following additional intuitive properties of a probability when **A** and **B** are events in $\boldsymbol{F}$:

(iv) P(**A**) + P($\mathbf{A}^c$) = 1.
(v) P(φ) = 0.
(vi) P(**A**∪**B**) = P(**A**) + P(**B**) - P(**A**∩**B**).
(vii) P(**A**) ≥ P(**B**) when **B** ⊆ **A**.
(viii) If $\mathbf{A}_i$ in $\boldsymbol{F}$ is monotone decreasing to φ (denoted $\mathbf{A}_i \searrow \boldsymbol{\varphi}$), then $P(\mathbf{A}_i) \to 0$.
(ix) If $\mathbf{A}_i \in \boldsymbol{F}$, not necessarily disjoint, then $P(\bigcup_{i=1}^{\infty} \mathbf{A}_i) \le \sum_{i=1}^{\infty} P(\mathbf{A}_i)$.
(x) If $\{\mathbf{A}_i\}$ is a finite or countable *partition* of **S** (i.e., the events $\mathbf{A}_i \in \boldsymbol{F}$ are mutually exclusive and exhaustive, or $\mathbf{A}_i\cap\mathbf{A}_j = \boldsymbol{\varphi}$ for all i ≠ j and $\bigcup_{i=1}^{\infty} \mathbf{A}_i = \mathbf{S}$), then $P(\mathbf{B}) = \sum_{i=1}^{\infty} P(\mathbf{B}\cap\mathbf{A}_i)$.

The triplet (**S**,***F***,P) consisting of a measurable space (**S**,***F***) and a probability measure P is called a *probability space*.

Example 1 (continued). Consider the σ-field ***H*** containing information on the number of heads, but not their order. The table below gives three functions P_1, P_2, P_3 defined on ***H***. All satisfy properties (i) and (ii) for a probability. Functions P_2 and P_3 also satisfy (iii), and are probabilities, but P_1 violates (iii) since $P_1(\{HH\}\cup\{TT\}) \neq P_1(\{HH\}) + P_1(\{TT\})$. The probability P_2 is generated by fair coins, and the probability P_3 by one fair coin and one biased coin.

	φ	**S**	HH	TT	HT,TH	HH,TT	HT,TH,TT	HH,HT,TH
P_1	0	1	1/3	1/3	1/2	1/2	2/3	2/3
P_2	0	1	1/4	1/4	1/2	1/2	3/4	3/4
P_3	0	1	1/3	1/6	1/2	1/2	2/3	5/6

3.3.2. If $\mathbf{A} \in \boldsymbol{F}$ has P(**A**) = 1, then **A** is said to occur *almost surely* (a.s.), or *with probability one* (w.p.1). If $\mathbf{A} \in \boldsymbol{F}$ has P(**A**) = 0, then **A** is said to occur with *probability zero* (w.p.0). Finite or countable intersections of events that occur almost surely again occur almost surely, and finite or countable unions of events that occur with probability zero again occur with probability zero.

Example 2. (continued) If the coin is fair, then the probability of k-1 heads followed by a tail is $1/2^k$. Use the geometric series formulas in 2.1.10 to verify that the probability of "At most 3 heads" is 15/16, of "Ten or more heads" is $1/2^{10}$, and of "an even number of heads" is 2/3.

Example 3. (continued) Consider the function P defined on open sets $(s,\infty) \in \mathbb{R}_+$ by $P((s,\infty)) = e^{-s/2}$. This function maps into the unit interval. It is then easy to show that P satisfies properties (i)-(iii) of a probability on the restricted family of open intervals, and a little work to show that when a probability is determined on this family of open intervals, then it is uniquely determined on the σ-field generated by these intervals. Each single point, such as {1}, is in ***F***. Taking intervals that shrink to this point, each single point occurs with probability zero. Then, a countable set of points occurs w.p.0.

3.3.3. Often a measurable space (**S**,***F***) will have an associated *measure* ν that is a countably additive function from ***F*** into the nonnegative real line; i.e., $\nu(\bigcup_{i=1}^{\infty} \mathbf{A}_i) = \sum_{i=1}^{\infty} \nu(\mathbf{A}_i)$ for any sequence of disjoint $\mathbf{A}_i \in \boldsymbol{F}$. The measure is *positive* if $\nu(\mathbf{A}) \geq 0$ for all $\mathbf{A} \in \boldsymbol{F}$; we will consider only positive measures. The measure ν is *finite* if $|\nu(\mathbf{A})| \leq M$ for some constant M and all $\mathbf{A} \in \boldsymbol{F}$, and

σ-*finite* if $\boldsymbol{F}$ contains a countable partition $\{\mathbf{A}_i\}$ of **S** such that the measure of each partition set is finite; i.e., $\nu(\mathbf{A}_i) < +\infty$. The measure ν may be a probability, but more commonly it is a measure of "length" or "volume". For example, it is common when the sample space **S** is the countable set of positive integers to define ν to be *counting measure* with ν(**A**) equal to the number of points in **A**. When the sample space **S** is the real line, with the Borel σ-field $\boldsymbol{B}$, it is common to define ν to be *Lebesgue measure*, with ν((a,b)) = b - a for any open interval (a,b). Both of these examples are positive σ-finite measures. A set **A** is said to be of *ν-measure zero* if ν(**A**) = 0. A property that holds except on a set of measure zero is said to hold *almost everywhere* (a.e.). It will sometimes be useful to talk about a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$ where μ is positive and σ-finite and may either be a probability measure or a more general counting or length measure such as Lebesgue measure.

3.3.4. Suppose f is a real-valued function on a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$. This function is *measurable* if $f^{-1}(\mathbf{C}) \in \boldsymbol{F}$ for each open set **C** in the real line. A measurable function has the property that its contour sets of the form $\{s\in\mathbf{S}|a\le f(s)\le c\}$ are contained in $\boldsymbol{F}$. This implies that if $\mathbf{B} \in \boldsymbol{F}$ is an atom, then f(s) must be constant for all $s \in \mathbf{B}$.

The integral of measurable f on a set $\mathbf{A} \in \boldsymbol{F}$, denoted $\int_{\boldsymbol{A}} f(s)\cdot\mu(ds)$, is defined for $\mu(\mathbf{A}) < +\infty$ as the limit as $n \to \infty$ of sums of the form $\sum_{k=-\infty}^{\infty} (k/n)\cdot\mu(\mathbf{C}_{kn})$, where $\mathbf{C}_{kn}$ is the set of states of Nature in **A** for which f(s) is contained in the interval (k/n,(k+1)/n]. A finite limit exists if $\sum_{k=\infty}^{\infty} |k/n|\cdot\mu(\mathbf{C}_{kn}) < +\infty$, in which case f is said to be *integrable* on **A**. Let $\{\mathbf{A}_i\} \in \boldsymbol{F}$ be a countable partition of **S** with $\mu(\mathbf{A}_i) < +\infty$, guaranteed by the σ-finite property of μ. The function f is integrable on a general set $\mathbf{A} \in \boldsymbol{F}$ if it is integrable on $\mathbf{A}\cap\mathbf{A}_i$ for each i and if $\int_{\boldsymbol{A}} |f(s)|\cdot\mu(ds) = \lim_{n\to\infty} \sum_{i=1}^{n} \int_{\boldsymbol{A}\cap\boldsymbol{A}_i} |f(s)|\cdot\mu(ds)$ exists, and simply *integrable* if it is integrable for **A** = **S**. In general, the measure μ can have point masses (at atoms), or continuous measure, or both, so that the notation for integration with respect to μ includes sums and mixed cases. The integral $\int_{\boldsymbol{A}} f(s)\mu(ds)$ will sometimes be denoted $\int_{\boldsymbol{A}} f(s)d\mu$, or in the case of Lebesgue measure, $\int_{\boldsymbol{A}} f(s)ds$.

3.3.5. For a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$, define $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $1 \le q < +\infty$ to be the set of measurable real-valued functions on **S** with the property that $|f|^q$ is integrable, and define $\|f\|_q =$

$[\int |f(s)|^q \mu(ds)]^{1/q}$ to be the *norm* of f. Then, $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ is a linear space, since linear combinations of integrable functions are again integrable. This space has many, but not all, of familiar properties of finite-dimensional Euclidean space. The set of all linear functions on the space $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $q > 1$ is the space $\mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu)$, where $1/r = 1 - 1/q$. This follows from an application of Holder's inequality, which generalizes from finite vector spaces to the condition

$$f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu) \text{ and } g \in \mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu) \text{ with } q^{-1} + r^{-1} = 1 \text{ imply } \int |f(s)\cdot g(s)| \mu(ds) \leq \|f\|_q \cdot \|g\|_r.$$

The case $q = r = 2$ gives the Cauchy-Schwartz inequality in general form. This case arises often in statistics, with the functions f interpreted as random variables and the norm $\|f\|_2$ interpreted as a quadratic mean or variance.

3.3.6. There are three important concepts for the limit of a sequence of functions $f_n \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$. First, there is *convergence in norm*, or strong convergence: f is a limit of f_n if $\|f_n - f\|_q \to 0$. Second, there is *convergence in μ-measure*: f is a limit of f_n if $\mu(\{s\in S| \, |f_n(s) - f(s)| > \varepsilon\}) \to 0$ for each $\varepsilon > 0$. Third, there is *weak convergence*: f is a limit of f_n if $\int (f_n(s) - f(s))\cdot g(s) \mu(ds) \to 0$ for each $g \in \mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu)$ with $1/r = 1 - 1/q$. The following relationship holds between these modes of convergence:

$$\text{Strong Convergence} \Longrightarrow \text{Weak Convergence} \Longrightarrow \text{Convergence in } \mu\text{-measure}$$

An example shows that convergence in μ-measure does not in general imply weak convergence: Consider $\mathbf{L}_2(\mathbf{[0,1]},\boldsymbol{B},\mu)$ where $\boldsymbol{B}$ is the Borel σ-field and μ is Lebesgue measure. Consider the sequence $f_n(s) = n\cdot\mathbf{1}(s \leq 1/n)$. Then $\mu(\{s\in S| \, |f_n(s)| > \varepsilon\}) = 1/n$, so that f_n converges in μ-measure to zero, but for $g(s) = s^{-1/3}$, one has $\|g\|_2 = 3^{1/2}$ and $\int f_n(s)g(s) \mu(ds) = 3n^{1/3}/2$ divergent. Another example shows that weak convergence does not in general imply strong convergence: Consider $\mathbf{S} = \{1,2,...\}$ endowed with the σ-field generated by the family of finite sets and the measure μ that gives weight $k^{-1/2}$ to point k. Consider $f_n(k) = n^{1/4}\cdot\mathbf{1}(k = n)$. Then $\|f_n\|_2 = 1$. If g is a function for which $\sum_{k=1}^{\infty} f_n(k)g(k)\mu(\{k\}) = g(n)\cdot n^{1/4}$ does not converge to zero, then $g(k)^2 \mu(\{k\})$ is bounded away from zero infinitely often, implying $\|g\|_2 = \sum_{k=1}^{\infty} g(k)^2 \mu(\{k\}) = +\infty$. Then, f_n converges weakly, but not strongly, to zero. The following theorem, which is of great importance in advanced econometrics, gives a uniformity condition under which these modes of convergence coincide.

Theorem 3.1. (Lebesgue Dominated Convergence) If g and f_n for n = 1,2,... are in $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $1 \le q < +\infty$ and a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$, and if $|f_n(s)| \le g(s)$ almost everywhere, then f_n converges in μ-measure to a function f if and only if $f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ and $\|f_n - f\|_q \to 0$.

One application of this theorem is a result for interchange of the order of integration and differentiation. Suppose $f(\cdot,t) \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for t in an open set $\mathbf{T} \subseteq \mathbb{R}^n$. Suppose f is *differentiable*, meaning that there exists a function $\nabla_t f(\cdot,t) \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $t \in \mathbf{T}$ such that if $t+h \in \mathbf{T}$ and $h \neq 0$, then the remainder function $r(s,t,h) = [f(s,t+h) - f(s,t) - \nabla_t f(\cdot,t)\cdot h]/|h| \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ converges in μ-measure to zero as $h \to 0$. Define $F(t) = \int f(s,t)\mu(ds)$. If there exists $g \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ which dominates the remainder function (i.e., $|r(s,t,h)| \le g(s)$ a.e.), then Theorem 3.1 implies $\lim_{h\to 0}\|r(\cdot,t,h)\|_q = 0$, and F(t) is differentiable and satisfies $\nabla_t F(t) = \int \nabla_t f(s,t)\mu(ds)$.

A finite measure P on $(\mathbf{S},\boldsymbol{F})$ is *absolutely continuous* with respect to a measure ν if $\mathbf{A} \in \boldsymbol{F}$ and $\nu(\mathbf{A}) = 0$ imply $P(\mathbf{A}) = 0$. If P is a probability measure that is absolutely continuous with respect to the measure ν, then an event of measure zero occurs w.p.0, and an event that is true almost everywhere occurs almost surely. A fundamental result from analysis is the theorem:

Theorem 3.2. (Radon-Nikodym) If a finite measure P on a measurable space $(\mathbf{S},\boldsymbol{F})$ is absolutely continuous with respect to a positive σ-finite measure ν on $(\mathbf{S},\boldsymbol{F})$, then there exists an integrable real-valued function $p \in \mathbf{L}_1(\mathbf{S},\boldsymbol{F},\nu)$ such that

$$\int_{\boldsymbol{A}} p(s)\nu(ds) = P(\mathbf{A}) \text{ for each } \mathbf{A} \in \boldsymbol{F}.$$

When P is a probability, the function p given by the theorem is nonnegative, and is called the *probability density*. An implication of the Radon-Nikodym theorem is that if a measurable space $(\mathbf{S},\boldsymbol{F})$ has a positive σ-finite measure ν and a probability measure P that is absolutely continuous with respect to ν, then there exists a density p such that for every $f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},P)$ for some $1 \le q < +\infty$, one has $\int_{\boldsymbol{S}} f(s)P(ds) = \int_{\boldsymbol{S}} f(s)\cdot p(s)\nu(ds)$.

3.3.7. In applications where the probability space is the real line with the Borel σ-field, with a probability P such that $P((-\infty,s]) = F(s)$ is continuously differentiable, the fundamental theorem of integral calculus states that $p(s) = F'(s)$ satisfies $F(\mathbf{A}) = \int_{\boldsymbol{A}} p(s)ds$. What the Radon-Nikodym theorem does is extend this result to σ-finite measure spaces and weaken the assumption from

continuous differentiability to absolute continuity. In basic econometrics, we will often characterize probabilities both in terms of the probability measure (or distribution) and the density, and will usually need only the elementary calculus version of the Radon-Nikodym result. However, it is useful in theoretical discussions to remember that the Radon-Nikodym theorem makes the connection between probabilities and densities. We give two examples that illustrate practical use of the calculus version of the Radon-Nikodym theorem.

Example 3. (continued) Given $P((s,\infty)) = e^{-s/2}$, one can use the differentiability of the function in s to argue that it is absolutely continuous with respect to Lebesgue measure on the line. Verify by integration that the density implied by the Radon-Nikodym theorem is $p(s) = e^{-s/2}/2$.

Example 5. A probability that appears frequently in statistics is the *normal*, which is defined on $(\mathbb{R},\boldsymbol{B})$, where $\mathbb{R}$ is the real line and $\boldsymbol{B}$ the Borel σ-field, by the density $n(s\text{-}\mu,\sigma) \equiv (2\pi\sigma^2)^{-1/2}\cdot e^{-(s-\mu)^2/2\sigma^2}$, so that $P(\mathbf{A}) = \int_{\boldsymbol{A}} (2\pi\sigma^2)^{-1/2}\cdot e^{-(s-\mu)^2/2\sigma^2}\,ds$. In this probability, μ and σ are parameters that are interpreted as determining the *location* and *scale* of the probability, respectively. When μ = 0 and σ = 1, this probability is called the *standard normal*.

3.3.8. Consider a probability space $(\mathbf{S},\boldsymbol{F},P)$, and a σ-field $\boldsymbol{G} \subseteq \boldsymbol{F}$. If the event $\mathbf{B} \in \boldsymbol{G}$ has $P(\mathbf{B}) > 0$, then the *conditional probability* of **A** given **B** is defined as $P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A}\cap\mathbf{B})/P(\mathbf{B})$. Stated another way, $P(\mathbf{A}|\mathbf{B})$ is a real-valued function on $\boldsymbol{F}\times\boldsymbol{G}$ with the property that $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$ for all $\mathbf{A} \in \boldsymbol{F}$ and $\mathbf{B} \in \boldsymbol{G}$. When **B** is a finite set, the conditional probability of **A** given **B** is the ratio of sums

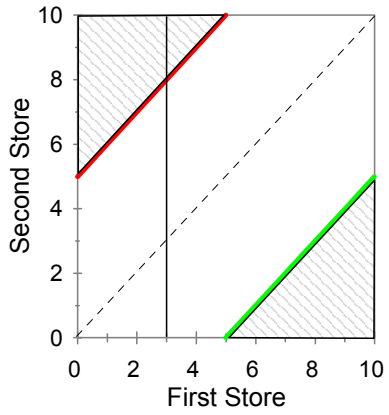
$$P(\mathbf{A}|\mathbf{B}) = \frac{\sum_{s\in\boldsymbol{A}\cap\boldsymbol{B}} P(\{s\})}{\sum_{s\in\boldsymbol{B}} P(\{s\})} .$$

Example 6. On a quiz show, a contestant is shown three doors, one of which conceals a prize, and is asked to select one. Before it is opened, the host opens one of the remaining doors which he knows does not contain the prize, and asks the contestant whether she wants to keep her original selection or switch to the other remaining unopened door. Should the contestant switch? Designate the contestant's initial selection as door 1. The sample space consists of pairs of numbers ab, where a = 1,2,3 is the number of the door containing the prize and b = 2,3 is the number of the door opened by the host, with $b \neq a$: S = {12,13,23,32}. The probability is 1/3 that the prize is behind each door. The conditional probability of b = 2, given a = 1, is 1/2, since in this case the host opens door 2 or door 3 at random. However, the conditional probability of b = 2, given a = 2 is zero and the conditional probability of b = 2 given a = 3 is one. Hence, P(12) = P(13) = (1/3)·(1/2), and P(23) = P(32) = 1/3. Let A = {12,13} be the event that door 1 contains the prize and B = {12,32} be the

event that the host opens door 2. Then the conditional probability of A given B is P(12)/(P(12)+P(32)) = (1/6)/((1/6)+(1/3)) = 1/3. Hence, the probability of receiving the prize is 1/3 if the contestant stays with her original selection, 2/3 if she switches to the other unopened door.

Example 7. Two fast food stores are sited at random points along a street that is ten miles long. What is the probability that they are less than five miles apart? Given that the first store is located at the three mile marker, what is the probability that the second store is less than five miles away? The answers are obvious from the diagram below, in which the sample space is depicted as a rectangle of dimension 10 by 10, with the horizontal axis giving the location of the first store and the vertical axis giving the location of the second store. The shaded areas correspond to the event that the two are more than five miles apart, and the proportion of the rectangle in these areas is 1/4. Conditioned on the first store being at point 3 on the horizontal axis, the second store is located at random on a vertical line through this point, and the proportion of this line that lies in the shaded area is 1/5. Let x be the location of the first store, y the location of the second. The conditional probability of the event that $|x - y| > 5$, given x, is $|x-5|/10$. This could have been derived by forming the probability of the event $|x - y| > 5$ and $c < x < c+\delta$ for a small positive δ, taking the ratio of this probability to the probability of the event $c < x < c+\delta$ to obtain the conditional probability of the event $|x - y| > 5$ given $c < x < c+\delta$, and taking the limit $\delta \to 0$.

Location of Fast Food Stores



The idea behind conditional probabilities is that one has partial information on what the state of Nature may be, and one wants to calculate the probability of events using this partial information. One way to represent partial information is in terms of a subfield; e.g., $\boldsymbol{F}$ is the field of events which distinguish outcomes in both the past and the future, and a subfield $\boldsymbol{G}$ contains events which distinguish only past outcomes. A conditional probability $P(\mathbf{A}|\mathbf{B})$ defined for $\mathbf{B} \subseteq \boldsymbol{G}$ can be interpreted for fixed **A** as a function from $\boldsymbol{G}$ into [0,1]. To emphasize this, conditional probabilities

are sometimes written $P(\mathbf{A}|\boldsymbol{G})$, and $\boldsymbol{G}$ is termed the *information set*, or a family of events with the property that you know whether or not they happened at the time you are forming the conditional probability.

Example 1. (continued) If $\boldsymbol{G} = \{\boldsymbol{\varphi},\mathbf{S},\{HT,HH\},\{TT,TH\}\}$, so that events in $\boldsymbol{G}$ describe the outcome of the first coin toss, then $P(HH|\{HH,HT\}) = P(HH)/(P(HH)+P(HT)) = ½$ is the probability of heads on the second toss, given heads on the first toss. In this example, the conditional probability of a head on the second toss equals the unconditional probability of this event. In this case, the outcome of the first coin toss provides no information on the probabilities of heads from the second coin, and the two tosses are said to be *statistically independent*. If $\boldsymbol{G} = \{\boldsymbol{\varphi},\mathbf{S},\{HT,TH\},\{HH\},\{TT\},\{HH\}^c,\{TT\}^c\}$, the family of events that determine the number of heads that occur in two tosses without regard for order, then the conditional probability of heads on the first toss, given at least one head, is $P(\{HT,HH\}|\{TT\}^c) = (P(HT)+P(HH))/(1-P(TT)) = 2/3$. Then, the conditional probability of heads on the first toss given at least one head is not equal to the unconditional probability of heads on the first toss.

Example 3. (continued) Suppose $\boldsymbol{G} = \{\boldsymbol{\varphi},\mathbf{S},(1,\infty),(-\infty,1]\}$ is the σ-field corresponding to the event that the index exceeds 1, and let $\boldsymbol{B}$ denote the Borel σ-field containing all the open intervals. The unconditional probability $P((s,\infty)) = e^{-s/2}$ implies $P((1,\infty)) = e^{-1/2} = 0.6065$. The conditional probability of $(2,\infty)$ given $(1,\infty)$ satisfies $P((2,\infty)|(1,\infty)) = P((1,\infty)\cap(2,\infty))/P((1,\infty)) = e^{-1}/e^{-1/2} = 0.6065 > P((2,\infty)) = 0.3679$. The conditional and unconditional probabilities are not the same, so that the conditioning event provides information on the probability of $(2,\infty)$.

For a probability space $(\mathbf{S},\boldsymbol{F},P)$, suppose $\mathbf{A}_1,...,\mathbf{A}_k$ is a finite partition of $\mathbf{S}$; i.e., $\mathbf{A}_i\cap\mathbf{A}_j = \boldsymbol{\varphi}$ and $\bigcup_{i=1}^{k} \mathbf{A}_i = \mathbf{S}$. The partition generates a finite field $\boldsymbol{G} \subseteq \boldsymbol{F}$. From the formula $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$ satisfied by conditional probabilities, one has for an event $\mathbf{C} \in \boldsymbol{F}$ the formula

$$P(\mathbf{C}) = \sum_{i=1}^{k} P(\mathbf{C}|\mathbf{A}_i)\cdot P(\mathbf{A}_i).$$

This is often useful in calculating probabilities in applications where the conditional probabilities are available.

3.3.9. In a probability space $(\mathbf{S},\boldsymbol{F},P)$, the concept of a conditional probability $P(\mathbf{A}|\mathbf{B})$ of $\mathbf{A} \in \boldsymbol{F}$ given an event $\mathbf{B}$ in a σ-field $\boldsymbol{G} \subseteq \boldsymbol{F}$ can be extended to cases where $P(\mathbf{B}) = 0$ by defining $P(\mathbf{A}|\mathbf{B})$ as the limit of $P(\mathbf{A}|\mathbf{B}_i)$ for sequences $\mathbf{B}_i \in \boldsymbol{G}$ that satisfy $P(\mathbf{B}_i) > 0$ and $\mathbf{B}_i \to \mathbf{B}$, provided the limit exists. If we fix $\mathbf{A}$, and consider $P(\mathbf{A}\cap\mathbf{B})$ as a measure defined for $\mathbf{B} \in \boldsymbol{G}$, this measure obviously satisfies $P(\mathbf{A}\cap\mathbf{B}) \le P(\mathbf{B})$, so that it is absolutely continuous with respect to $P(B)$. Then, Theorem 3.2

implies that there exists a function $P(\mathbf{A}|\cdot) \in \mathbf{L}_1(\mathbf{S},\boldsymbol{G},P)$ such that $P(\mathbf{A}\cap\mathbf{B}) = \int_{\boldsymbol{B}} P(\mathbf{A}|s)\cdot P(ds)$. We have written this function as if it were a conditional probability of A given the "event" {s}, and it can be given this interpretation. If $\mathbf{B} \in \boldsymbol{G}$ is an atom, then the measurability of $P(\mathbf{A}|\cdot)$ with respect to $\boldsymbol{G}$ requires that it be constant for $s \in \mathbf{B}$, so that $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|s)\cdot P(\mathbf{B})$ for any $s \in \mathbf{B}$, and we can instead write $P(\mathbf{A}\cap\mathbf{B}) = P(\mathbf{A}|\mathbf{B})\cdot P(\mathbf{B})$, satisfying the definition of conditional probability even if $P(\mathbf{B}) = 0$.

Example 4. (continued) Consider $\boldsymbol{F} = \boldsymbol{B}\otimes\boldsymbol{B}$, the product Borel σ-field on $\mathbb{R}_+$, and $\boldsymbol{G} = \boldsymbol{B}\otimes\{\boldsymbol{\varphi}.\mathbb{R}_+\}$, the σ-field corresponding to having complete information on the level of the index on the first day and no information on the second day. Suppose $P((s,\infty)\times(t,\infty)) = 2/(1+e^{s+t})$. This is a probability on these open intervals that extends to $\boldsymbol{F}$; verifying this takes some work. The conditional probability of $(s,\infty)\times(t,\infty)$ given the event $(r,\infty)\times(0,\infty) \in \boldsymbol{G}$ and $s \le r$ equals $P((r,\infty)\times(t,\infty))$ divided by $P((r,\infty)\times(0,\infty))$, or $(1+e^r)/(1+e^{r+t})$. The conditional probability of $(s,\infty)\times(t,\infty)$ given the event $(r,r+\delta)\times(0,\infty) \in \boldsymbol{G}$ and $s \le r$ is $[1/(1+e^{r+t}) - 1/(1+e^{r+\delta+t})]/[1/(1+e^r) - 1/(1+e^{r+\delta})]$. The limit of this expression as $\delta \to 0$ is $e^r\cdot(1+e^r)^2/(1+e^{r+t})^2 = P((s,\infty)\times(t,\infty)|\{r\}\times(0,\infty))$; this function of r is also the integrand that satisfies Theorem 3.2. Note that $P((s,\infty)\times(t,\infty)|\{r\}\times(0,\infty)) \neq P((s,\infty)\times(0,\infty)) = 1/(1+e^s)$, so that the conditioning event conveys information about the probability of $(s,\infty)\times(t,\infty)$.

3.4. STATISTICAL INDEPENDENCE AND REPEATED TRIALS

3.4.1. Consider a probability space $(\mathbf{S},\boldsymbol{F},P)$. Events **A** and **C** in $\boldsymbol{F}$ are *statistically independent* if $P(\mathbf{A}\cap\mathbf{C}) = P(\mathbf{A})\cdot P(\mathbf{C})$. From the definition of conditional probability, if **A** and **C** are statistically independent and $P(\mathbf{A}) > 0$, then $P(\mathbf{C}|\mathbf{A}) = P(\mathbf{A}\cap\mathbf{C})/P(\mathbf{A}) = P(\mathbf{C})$. Thus, when **A** and **C** are statistically independent, knowing that **A** occurs is unhelpful in calculating the probability that **C** occurs. The idea of statistical independence of events has an exact analogue in a concept of statistical independence of subfields. Let $\boldsymbol{A} = \{\boldsymbol{\varphi},\mathbf{A},\mathbf{A}^c,\mathbf{S}\}$ and $\boldsymbol{C} = \{\boldsymbol{\varphi},\mathbf{C},\mathbf{C}^c,\mathbf{S}\}$ be the subfields of $\boldsymbol{F}$ generated by **A** and **C**, respectively. Verify as an exercise that if **A** and **C** are statistically independent, then so are any pair of events $\mathbf{A}' \in \boldsymbol{A}$ and $\mathbf{C}' \in \boldsymbol{C}$. Then, one can say that the subfields $\boldsymbol{A}$ and $\boldsymbol{C}$ are statistically independent. One can extend this idea and talk about statistical independence in a collection of subfields. Let **N** denote an index set, which may be finite, countable, or non-countable. Let $\boldsymbol{F}_i$ denote a σ-subfield of $\boldsymbol{F}$ ($\boldsymbol{F}_i \subseteq \boldsymbol{F}$) for each $i \in \mathbf{N}$. The subfields $\boldsymbol{F}_i$ are *mutually statistically independence* (MSI) if and only if $P(\bigcap_{j\in\boldsymbol{K}} \mathbf{A}_j) = \prod_{j\in\boldsymbol{K}} P(\mathbf{A}_j)$ for all finite $\mathbf{K} \subseteq \mathbf{N}$ and $\mathbf{A}_j \in \boldsymbol{F}_j$ for $j \in \mathbf{K}$. As in the case of statistical independence between two events (subfields), the concept of MSI can be stated in terms of conditional probabilities: $\boldsymbol{F}_i$ for $i \in \mathbf{N}$ are mutually

statistically independent (MSI) if, for all i ∈ **N**, finite **K** ⊆ **N**\{i} and $\mathbf{A}_j \in \boldsymbol{F}_j$ for j ∈ {i}∪**K**, one has $P(\mathbf{A}_i \mid \bigcap_{j\in\boldsymbol{K}} \mathbf{A}_j) = P(\mathbf{A}_i)$, so the conditional and unconditional probabilities are the same.

Example 1. (continued) Let **A** = {HH,HT} denote the event of a head for the first coin, **C** = {HH,TH} denote the event of a head for the second coin, **D** = {HH,TT} denote the event of a match, **G** = {HH} the event of two heads. The table below gives the probabilities of various events.

Event	**A**	**C**	**D**	**G**	**A∩C**	**A∩D**	**C∩D**	**A∩C∩D**	**A∩G**
Prob.	½	½	½	1/4	1/4	1/4	1/4	1/4	1/4

The result P(**A**∩**C**) = P(**A**)P(**C**) = 1/4 establishes that **A** and **C** are statistically independent. Verify that **A** and **D** are statistically independent, and that **C** and **D** are statistically independent, but that **P(A∩C∩D)** ≠ P(**A**)P(**C**)P(**D**), so that **A, C**, and **D** are <u>not</u> MSI. Verify that **A** and **G** are not statistically independent.

Example 4. (continued) Recall that **S** = $\mathbb{R}^2$ with $\boldsymbol{F} = \boldsymbol{B}\otimes\boldsymbol{B}$, the *product* Borel σ-field. Define $\boldsymbol{N}$ = {φ,$\mathbb{R}$} and the subfields $\boldsymbol{F}_1 = \boldsymbol{B}\times\boldsymbol{N}$ and $\boldsymbol{F}_2 = \boldsymbol{N}\times\boldsymbol{B}$, containing information on the index levels on the first and second day, respectively. Define $\boldsymbol{G}$ to be the σ-field generated by the rectangles (0,1]×(0,1], (0,1]×(1,∞),(1,∞)×(0,1], and(1,∞)×(1,∞). Then $\boldsymbol{G}$ is the subfield of $\boldsymbol{B}$ containing information on whether the indices on the two days are above one. Define $\boldsymbol{F}_3$ to be the σ-subfield of $\boldsymbol{B}\otimes\boldsymbol{B}$ generated by sets of the form $\mathbf{A}_1\times\mathbf{A}_2$ with $\mathbf{A}_1 \in \boldsymbol{G}$ and $\mathbf{A}_2 \in \boldsymbol{B}$; then $\boldsymbol{F}_3$ contains full information on the second day index, but only the qualitative information on whether the first day index is above one. Suppose $P((s,\infty)\times(t,\infty)) = e^{-s-t}$. Then $\{\boldsymbol{F}_1,\boldsymbol{F}_2\}$ are MSI. However, $\{\boldsymbol{F}_1,\boldsymbol{F}_3\}$ are not independent.

Example 8. Consider **S** = {0, 1, 2, 3, 4, 5, 6, 7}, with $\boldsymbol{F}$ equal to all subsets of **S**. As a shorthand, let 0123 denote {0,1,2,3}, etc. Define the subfields

$\boldsymbol{F}_1$ = {φ,0123,4567,**S**}, $\boldsymbol{F}_2$ = {φ,2345,0167,**S**}, $\boldsymbol{F}_3$ = {φ,0246,1357,**S**},
$\boldsymbol{F}_4$ = {φ,01,23,4567,0123,234567,014567,**S**},
$\boldsymbol{F}_5$ = {φ,01,23,45,67,0123,0145,0167,2345,2367,4567,012345,012367,014567,234567,**S**},
$\boldsymbol{F}_6$ = {φ,06,17,24,35,0167,0246,0356,1247,1357,2345,123457,023456,013567,012467,**S**}.

The field $\boldsymbol{F}_4$ is a *refinement* of the field $\boldsymbol{F}_1$ (i.e., $\boldsymbol{F}_1 \subseteq \boldsymbol{F}_4$), and can be said to contain more information than $\boldsymbol{F}_1$. The field $\boldsymbol{F}_5$ is a *mutual refinement* of $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ (i.e., $\boldsymbol{F}_1\cup\boldsymbol{F}_2 \subseteq \boldsymbol{F}_5$), and is in fact the smallest mutual refinement. It contains all the information available in either $\boldsymbol{F}_1$ or $\boldsymbol{F}_2$. Similarly, $\boldsymbol{F}_6$ is a

mutual refinement of $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$. The intersection of $\boldsymbol{F}_5$ and $\boldsymbol{F}_6$ is the field $\boldsymbol{F}_2$; it is the common information available in $\boldsymbol{F}_5$ and $\boldsymbol{F}_6$. If, for example, $\boldsymbol{F}_5$ characterized the information available to one economic agent, and $\boldsymbol{F}_6$ characterized the information available to a second agent, then $\boldsymbol{F}_2$ would characterize the common information upon which they could base contingent contracts. Suppose P(i) = 1/8. Then $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$ are MSI. E.g., $P(0123|2345) = P(0123|0246) = P(0123|2345\cap 0246) = P(0123) = 1/2$. However, $\{\boldsymbol{F}_1, \boldsymbol{F}_4\}$ are not independent; e.g., $1 = P(0123|01) \neq P(0123) = 1/2$.

For $\mathbf{M} \subseteq \mathbf{N}$, let $\boldsymbol{F}_{\mathbf{M}}$ denote the smallest σ-field containing $\boldsymbol{F}_i$ for all $i \in \mathbf{M}$. Then MSI satisfies the following theorem, which provides a useful criterion for determining whether a collection of subfields is MSI:

Theorem 3.3. If $\boldsymbol{F}_i$ are MSI for $i \in \mathbf{N}$, and $\mathbf{M} \subseteq \mathbf{N}\backslash\{i\}$, then $\{\boldsymbol{F}_i, \boldsymbol{F}_{\mathbf{M}}\}$ are MSI. Further, $\boldsymbol{F}_i$ for $i\in\mathbf{N}$ are MSI if and only if $\{\boldsymbol{F}_i, \boldsymbol{F}_{\mathbf{N}\backslash i}\}$ are MSI for all $i\in\mathbf{N}$.

Example 5. (continued) If $\mathbf{M} = \{2,3\}$, then $\boldsymbol{F}_M \equiv \boldsymbol{F}_6$, and $P(0123|\mathbf{A}) = ½$ for each $\mathbf{A} \in \boldsymbol{F}_{\mathbf{M}}$.

3.4.2. The idea of *repeated trials* is that an experiment, such as a coin toss, is replicated over and over. It is convenient to have common probability space in which to describe the outcomes of larger and larger experiments with more and more replications. The notation for repeated trials will be similar to that introduced in the definition of mutual statistical independence. Let $\mathbf{N}$ denote a finite or countable index set of trials, $\mathbf{S}_i$ a sample space for trial i, and $\boldsymbol{G}_i$ a σ-field of subsets of $\mathbf{S}_i$. Note that $(\mathbf{S}_i, \boldsymbol{G}_i)$ may be the same for all i. Assume that $(\mathbf{S}_i, \boldsymbol{G}_i)$ is the real line with the Borel σ-field, or a countable set with the field of all subsets, or a pair with comparable mathematical properties (i.e., $\mathbf{S}_i$ is a complete separable metric space and $\boldsymbol{G}_i$ is its Borel field). Let $t = (s_1,s_2,...) = (s_i : i\in\mathbf{N})$ denote an ordered sequence of outcomes of trials, and $\mathbf{S}_{\mathbf{N}} = \bigtimes_{i\in\mathbf{N}} \mathbf{S}_i$ denote the sample space of these sequences. Let $\boldsymbol{F}_{\mathbf{N}} = \bigotimes_{i\in\mathbf{N}}\boldsymbol{G}_i$ denote the σ-field of subsets of $\mathbf{S}_{\mathbf{N}}$ generated by the *finite rectangles* which are sets of the form $\left(\bigtimes_{i\in\mathbf{K}} \mathbf{A}_i\right)\times\left(\bigtimes_{i\in\mathbf{N}\backslash\mathbf{K}} \mathbf{S}_i\right)$ with $\mathbf{K}$ a finite subset of $\mathbf{N}$ and $\mathbf{A}_i \in \boldsymbol{G}_i$ for $i \in \mathbf{K}$. The collection $\boldsymbol{F}_{\mathbf{N}}$ is called the *product σ-field* of subsets of $\mathbf{S}_{\mathbf{N}}$.

Example 9. $\mathbf{N} = \{1,2,3\}$, $\mathbf{S}_i = \{0,1\}$, $\boldsymbol{G}_i = \{\boldsymbol{\varphi}, \{0\}, \{1\}, \mathbf{S}\}$ is a sample space for a coin toss, coded "1" if heads and "0" if tails. Then $\mathbf{S}_{\mathbf{N}} = \{s_1s_2s_3 \,|\, s_i \in \mathbf{S}_i\} = \{000, 001, 010, 011, 100, 101, 110, 111\}$, where 000 is shorthand for the event $\{0\}\times\{0\}\times\{0\}$, and so forth, is the sample space for three coin tosses. The field $\boldsymbol{F}_{\mathbf{N}}$ is the family of all subsets of $\mathbf{S}_{\mathbf{N}}$.

For any subset $\mathbf{K}$ of $\mathbf{N}$, define $\mathbf{S}_{\mathbf{K}} = \bigtimes_{i\in\mathbf{K}} \mathbf{S}_i$ and $\boldsymbol{G}_{\mathbf{K}} = \bigotimes_{i\in\mathbf{K}}\boldsymbol{G}_i$. Then, $\boldsymbol{G}_{\mathbf{K}}$ is the product σ-field on $\mathbf{S}_{\mathbf{K}}$. Define $\boldsymbol{F}_{\mathbf{K}}$ to be the σ-field on $\mathbf{S}_{\mathbf{N}}$ generated by sets of the form $\mathbf{A}\times\mathbf{S}_{\mathbf{N}\backslash\mathbf{K}}$ for $\mathbf{A} \in \boldsymbol{G}_{\mathbf{K}}$. Then $\boldsymbol{G}_{\mathbf{K}}$

and $\boldsymbol{F}_{\mathbf{K}}$ contain essentially the same information, but $\boldsymbol{G}_{\mathbf{K}}$ is a field of subsets of $\mathbf{S}_{\mathbf{K}}$ and $\boldsymbol{F}_{\mathbf{K}}$ is a corresponding field of subsets of $\mathbf{S}_{\mathbf{N}}$ which contains no information on events outside of $\mathbf{K}$. Suppose $P_{\mathbf{N}}$ is a probability on $(\mathbf{S}_{\mathbf{N}}, \boldsymbol{F}_{\mathbf{N}})$. The *restriction* of $P_{\mathbf{N}}$ to $(\mathbf{S}_{\mathbf{K}},\boldsymbol{G}_{\mathbf{K}})$ is a probability $P_{\mathbf{K}}$ defined for $\mathbf{A} \in \boldsymbol{G}_{\mathbf{K}}$ by $P_{\mathbf{K}}(\mathbf{A}) = P_{\mathbf{N}}(\mathbf{A}\times\mathbf{S}_{\mathbf{N}\setminus\mathbf{K}})$. The following result establishes a link between different restrictions:

> **Theorem 3.4.** If $\mathbf{M} \subseteq \mathbf{K}$ and $P_{\mathbf{M}}$, $P_{\mathbf{K}}$ are restrictions of $P_{\mathbf{N}}$, then $P_{\mathbf{M}}$ and $P_{\mathbf{K}}$ satisfy the *compatibility condition* that $P_{\mathbf{M}}(\mathbf{A}) = P_{\mathbf{K}}(\mathbf{A}\times\mathbf{S}_{\mathbf{K}\setminus\mathbf{M}})$ for all $\mathbf{A} \in \boldsymbol{F}_{\mathbf{M}}$.

There is then a fundamental result that establishes that when probabilities are defined on all finite sequences of trials and are compatible, then there exists a probability defined on the infinite sequence of trials that yields each of the probabilities for a finite sequence as a restriction.

> **Theorem 3.5.** If $P_{\mathbf{K}}$ on $(\mathbf{S}_{\mathbf{K}},\boldsymbol{G}_{\mathbf{K}})$ for all finite $\mathbf{K} \subseteq \mathbf{N}$ satisfy the compatibility condition, then there exists a unique $P_{\mathbf{N}}$ on $(\mathbf{S}_{\mathbf{N}},\boldsymbol{F}_{\mathbf{N}})$ such that each $P_{\mathbf{K}}$ is a restriction of $P_{\mathbf{N}}$.

This result guarantees that it is meaningful to make probability statements about events such as "an infinite number of heads in repeated coin tosses".

.

Suppose trials $(\mathbf{S}_i,\boldsymbol{G}_i,P_i)$ indexed by i in a countable set $\mathbf{N}$ are mutually statistically independent. For finite $\mathbf{K} \subseteq \mathbf{N}$, let $\boldsymbol{G}_{\mathbf{K}}$ denote the product σ-field on $\mathbf{S}_{\mathbf{K}}$. Then MSI implies that the probability of a set $\times_{i\in\mathbf{K}} \mathbf{A}_i \in \boldsymbol{G}_{\mathbf{K}}$ satisfies $P_{\mathbf{K}}(\times_{i\in\mathbf{K}} \mathbf{A}_i) = \prod_{j\in\boldsymbol{K}} P_j(\mathbf{A}_j)$. Then, the compatibility condition in Theorem 3.3 is satisfied, and that result implies the existence of a probability $P_{\mathbf{N}}$ on $(\mathbf{S}_{\mathbf{N}},\boldsymbol{F}_{\mathbf{N}})$ whose restrictions to $(\mathbf{S}_{\mathbf{K}},\boldsymbol{G}_{\mathbf{K}})$ for finite $\mathbf{K} \subseteq \mathbf{N}$ are the probabilities $P_{\mathbf{K}}$.

3.4.3. The assumption of statistically independent repeated trials is a natural one for many statistical and econometric applications where the data comes from random samples from the population, such as surveys of consumers or firms. This assumption has many powerful implications, and will be used to get most of the results of basic econometrics. However, it is also common in econometrics to work with aggregate time series data. In these data, each period of observation can be interpreted as a new trial. The assumption of statistical independence across these trials is unlikely in many cases, because in most cases real random effects do not conveniently limit themselves to single time periods. The question becomes whether there are weaker assumptions that time series data are likely to satisfy that are still strong enough to get some of the basic statistical theorems. It turns out that there are quite general conditions, called *mixing conditions*, that are enough to yield many of the key results. The idea behind these conditions is that usually events that are far apart in time are nearly independent, because intervening shocks overwhelm the older history in determining the later event. This idea is formalized in Chapter 4.

3.5. RANDOM VARIABLES, DISTRIBUTION FUNCTIONS, AND EXPECTATIONS

3.5.1. A *random variable* X is a measurable real-valued function on a probability space (**S**,***F***,P), or X:S $\rightarrow \mathbb{R}$. Then each state of Nature s determines a value X(s) of the random variable, termed its *realization* in state s. When the functional nature of the random variable is to be emphasized, it is denoted X(·), or simply X. When its values or realizations are used, they are denoted X(s) or x. For each set **B** $\in$ ***B***, the probability of the event that the realization of X is contained in **B** is well-defined and equals $P'(\mathbf{B}) \equiv P(X^{-1}(\mathbf{B}))$, where *P'* is termed the probability *induced* on $\mathbb{R}$ by the random variable X. One can have many random variables defined on the same probability space; another measurable function y = Y(s) defines a second random variable. It is important in working with random variables to keep in mind that the random variable itself is a function of states of Nature, and that observations are of realizations of the random variable. Thus, when one talks about convergence of a sequence of random variables, one is actually talking about convergence of a sequence of functions, and notions of distance and closeness need to be formulated as distance and closeness of functions. Multiplying a random variable by a scalar, or adding random variables, results in another random variable. Then, the family of random variables forms a *linear vector space*. In addition, products of random variables are again random variables, so that the family of random variables forms an *Abelian group under multiplication*. The family of random variables is also closed under *majorization*, so that Z:S $\rightarrow \mathbb{R}$ defined by Z(s) = max(X(s),Y(s)) for random variables X and Y is again a random variable. Then, the family of random variables forms a *lattice* with respect to the partial order X $\leq$ Y(i.e., X(s) $\leq$ Y(s) almost surely).

3.5.2. The term *measurable* in the definition of a random variable means that for each set **A** in the Borel σ-field ***B*** of subsets of the real line, the inverse image $X^{-1}(\mathbf{A}) \equiv \{s \in \mathbf{S} \,|\, X(s) \in \mathbf{A}\}$ is in the σ-field ***F*** of subsets of the sample space **S**. The assumption of measurability is a mathematical technicality that ensures that probability statements about the random variable are meaningful. We shall not make any explicit reference to measurability in basic econometrics, and shall always assume implicitly that the random variables we are dealing with are measurable.

3.5.3. The probability that a random variable X has a realization in a set **A** $\in$ ***B*** is given by

$$F(\mathbf{A}) \equiv P(X^{-1}(\mathbf{A})) \equiv P(\{s \in \mathbf{S} \,|\, X(s) \in \mathbf{A}\}).$$

The function F is a probability on ***B***; it is defined in particular for half-open intervals of the form **A** = $(-\infty,x]$, in which case F($(-\infty,x]$) is abbreviated to F(x) and is called the *distribution function* (or, *cumulative distribution function, CDF)* of X. From the properties of a probability, the distribution function has the properties

(i) $F(-\infty) = 0$ and $F(+\infty) = 1$.
(ii) $F(x)$ is non-decreasing in x, and continuous from the right.
(iii) $F(x)$ has at most a countable number of jumps, and is continuous except at these jumps. (Points without jumps are called *continuity* points.)

Conversely, any function F that satisfies (i) and (ii) determines uniquely a probability F on $\boldsymbol{B}$. The *support* of the distribution F is the smallest closed set $\mathbf{A} \in \boldsymbol{B}$ such that $F(\mathbf{A}) = 1$.

Example 5. (continued) The standard normal CDF is $\Phi(\mathbf{x}) = \int_{-\infty}^{x} (2\pi)^{-1/2} \cdot e^{-s^2/2} ds$, obtained by integrating the density $\varphi(s) = (2\pi)^{-1/2} \cdot e^{-s^2/2}$. Other examples are the CDF for the standard exponential distribution, $F(x) = 1 - e^{-x}$ for $x > 0$, and the CDF for the logistic distribution, $F(x) = 1/(1+e^{-x})$. An example of a CDF that has jumps is $F(x) = 1 - e^{-x}/2 - \sum_{k=1}^{\infty} \mathbf{1}(k \geq x)/2^{k+1}$ for $x > 0$.

3.5.4. If F is absolutely continuous with respect to a σ-finite measure ν on $\mathbb{R}$; i.e., F gives probability zero to any set that has ν-measure zero, then (by the Radon-Nikodym theorem) there exists a real-valued function f on $\mathbb{R}$, called the *density* (or *probability density function, pdf)* of X, such that

$$F(\mathbf{A}) = \int_{\boldsymbol{A}} f(x)\nu(dx)$$

for every $A \in \boldsymbol{B}$. With the possible exception of a set of ν-measure zero, F is differentiable and the derivative of the distribution gives the density, $f(x) = F'(x)$. When the measure ν is *Lebesgue measure*, so that the measure of an interval is its length, it is customary to simplify the notation and write $F(\mathbf{A}) = \int_{\boldsymbol{A}} f(x)dx$.

If F is absolutely continuous with respect to counting measure on a countable subset **C** of $\mathbb{R}$, then it is called a *discrete* distribution, and there is a real-valued function f on **C** such that

$$F(\mathbf{A}) = \sum_{x \in \boldsymbol{A}} f(x).$$

Recall that the probability is itself a measure. This suggests a notation $F(\mathbf{A}) = \int_{\boldsymbol{A}} F(dx)$ that covers both continuous and counting cases. This is called a *Lebesgue-Stieltjes* integral.

3.5.5. If $(\mathbb{R},\boldsymbol{B},F)$ is the probability space associated with a random variable X, and $g:\mathbb{R} \to \mathbb{R}$ is a measurable function, then Y = g(X) is another random variable. The random variable Y is *integrable* with respect to the probability F if $\int_{\mathbb{R}} |g(x)|F(dx) < +\infty$;

if it is integrable, then the integral $\int_{\mathbb{R}} g(x)F(dx) \equiv \int_{\mathbb{R}} g \cdot dF$ exists, is denoted **E** g(X), and is called *the expectation of* g(X). When necessary, this expectation will also be denoted $\mathbf{E}_X g(X)$ to identify the distribution used to form the expectation. When F is absolutely continuous with respect to Lebesgue measure, so that F has a density f, the expectation is written $\mathbf{E}\, g(X) = \int_{\mathbb{R}} g(x)f(x)dx$.

Alternately, for counting measure on the integers with density f(k), $\mathbf{E}\, g(X) = \sum_{k=-\infty}^{+\infty} g(k)f(k)$.

The expectation of X, if it exists, is called the *mean* of X. The expectation of $(X - \mathbf{E}X)^2$, if it exists, is called the *variance* of X. Define $\mathbf{1}(X \le a)$ to be an indicator function that is one if $X(s) \le a$, and zero otherwise. Then, $\mathbf{E}\, \mathbf{1}(X \le a) = F(a)$, and the distribution function can be recovered from the expectations of the indicator functions. Most econometric applications deal with random variables that have finite variances. The space of these random variables is $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$, the space of random variables X for which $\mathbf{E}\, X^2 = \int_{\mathbf{S}} X(x)^2 P(ds) < +\infty$. The space $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$ is also termed the space of *square-integrable functions*. The norm in this space is root-mean-square, $\|X\|_2 = [\int_{\mathbf{S}} X(s)^2 P(ds)]^{1/2}$. Implications of $X \in \mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$ are $\mathbf{E}\,|X| \le \int_{\mathbf{S}} \max(X(s),1)P(ds) \le \int_{\mathbf{S}} (X(s)^2+1)P(ds) = \|X\|_2^2 + 1 < +\infty$ and $\mathbf{E}\,(X - \mathbf{E}X)^2 = \|X\|_2^2 - (\mathbf{E}\,|X|)^2 \le \|X\|_2^2 < +\infty$, so that X has a well-defined, finite mean and variance.

Example 1. (continued) Define a random variable X by

$$X(s) = \begin{cases} 0 & \text{if } s = TT \\ 1 & \text{if } s = TH \text{ or } HT \\ 2 & \text{if } s = HH \end{cases}$$

Then, X is the number of heads in two coin tosses. For a fair coin, $\mathbf{E}\, X = 1$.

Example 2. (continued) Let X be a random variable defined to equal the number of heads that appear before a tail occurs. Then, possible values of X are the integers $\mathbf{C} = \{0,1,2,...\}$. Then **C** is the support of X. For x real, define [x] to be the largest integer k satisfying $k \le x$. A distribution

function for X, defined on the real line, is $F(x) = \begin{cases} 1 - 2^{-[x+1]} & for\ 0 \le x \\ 0 & for\ 0 > x \end{cases}$; the associated density defined on **C** is $f(k) = 2^{-k-1}$. The expectation of X, obtained using evaluation of a special series from 2.1.10, is $\mathbf{E}\,X = \sum_{k=0}^{\infty} k \cdot 2^{-k-1} = 1$.

Example 3. (continued) Define a random variable X by $X(s) = |s - 1|$. Then, X is the magnitude of the deviation of the index from one. The inverse image of an interval (a,b) is $(1-b,1-a) \cup (1+a,1+b) \in \boldsymbol{F}$, so that X is measurable. Other examples of measurable random variables are Y defined by $Y(s) = \text{Max}\ \{1,s\}$ and Z defined by $Z(s) = s^3$.

3.5.6. Consider a random variable Y on $(\mathbb{R},\boldsymbol{B})$. The expectation $\mathbf{E}Y^k$ is the k-th *moment* of Y, and $\mathbf{E}(Y-\mathbf{E}Y)^k$ is the k-th *central moment*. Sometimes moments fail to exist. However, if g(Y) is continuous and bounded, then $\mathbf{E}g(Y)$ always exists. The expectation $m(t) = \mathbf{E}e^{tY}$ is termed the *moment generating function* (mgf) of Y; it sometimes fails to exist. Call a mgf *proper* if it is finite for t in an interval around 0. When a proper mgf exists, the random variable has finite moments of all orders. The expectation $\psi(t) = \mathbf{E}e^{\iota tY}$, where ι is the square root of -1, is termed the *characteristic function* (cf) of Y. The characteristic function always exists.

Example 5. (continued) A density f(x) that is symmetric about zero, such as the standard normal, has $\mathbf{E}X^k = \int_{-\infty}^{+\infty} x^k f(x)dx = \int_{-\infty}^{0} x^k f(-x)dx + \int_{0}^{+\infty} x^k f(x)dx = \int_{0}^{+\infty} [1 + (-1)^k] x^k f(x)dx = 0$ for k odd. Integration by parts yields the formula $\mathbf{E}X^k = 2k \int_{0}^{+\infty} x^{k-1}[1-F(x)]dx$ for k even. For the standard normal, $\mathbf{E}X^{2k} = 2 \cdot \int_0^{+\infty} (2\pi)^{-1/2} x^{2k-1} \cdot e^{-x^2/2} x dx = (2k-1) \cdot \mathbf{E}X^{2k-2}$ for $k > 2$ using integration by parts, and $\mathbf{E}X^2 = 2 \cdot \int_0^{+\infty} (2\pi)^{-1/2} \cdot e^{-x^2/2} x dx = 2 \cdot \Phi(0) = 1$. Then, $\mathbf{E}X^4 = 3$ and $\mathbf{E}X^6 = 15$. The moment generating function of the standard normal is $m(t) = \int_{-\infty}^{+\infty} (2\pi)^{-1/2} \cdot e^{tx} \cdot e^{-x^2/2} dx$. Completing the square in the exponent gives $m(t) = e^{t^2/2} \cdot \int_{-\infty}^{+\infty} (2\pi)^{-1/2} \cdot e^{-(x-t)^2/2} dx = e^{t^2/2}$.

3.5.7. If T random variables are formed into a vector, $X(\cdot) = (X(\cdot,1),...,X(\cdot,T))$, the result is termed a *random vector*. For each $s \in \mathbf{S}$, the realization of the random vector is a point $(X(s,1),...,X(s,T))$ in $\mathbb{R}^T$, and the random vector has an induced probability on $\mathbb{R}^T$ which is characterized by its multivariate CDF, $F_X(x_1,...,x_T) = P(\{s \in \mathbf{S} | X(s,1) \le x_1,...,X(s,T) \le x_T\})$. Note that all the components of a random vector are functions of the *same* state of Nature s, and the random vector can be written as a measurable function X from the probability space $(\mathbf{S},\boldsymbol{F},P)$ into $(\mathbb{R}^T,\boldsymbol{B}^{nT})$. (The notation $\boldsymbol{B}^T$ means $\boldsymbol{B}\otimes\boldsymbol{B}\otimes...\otimes\boldsymbol{B}$ T times, where $\boldsymbol{B}$ is the Borel σ-field on the real line. This is also called the *product* σ-field, and is sometimes written $\boldsymbol{B}^T = \bigotimes_{i=1,...,T} \boldsymbol{B}_i$, where the $\boldsymbol{B}_i$ are identical copies of $\boldsymbol{B}$.) The measurability of X requires $X^{-1}(\mathbf{C}) \in \mathbf{S}$ for each open rectangle $\mathbf{C}$ in $\mathbb{R}^T$. The independence or dependence of the components of X is determined by the fine structure of *P* on **S**.

A useful insight comes from considering different representations of vectors in finite-dimensional spaces, and extending these ideas to infinite-dimensional situations. To be specific, consider $\mathbb{R}^2$. When we express a function X on $T = \{1,2\}$ as a point $(X(1),X(2))$ in this space, what we are really doing is defining two functions $Z_1 = (1,0)$ and $Z_2 = (0,1)$ with the property that Z_1 and Z_2 span the space, and then writing X as the linear combination $X = X(1)\cdot Z_1 + X(2)\cdot Z_2$. The pair of functions (points) Z_1 and Z_2 is called a *Hamel basis* for $\mathbb{R}^2$, and every point in the space has a unique representation in terms of this basis. However, there may be many different Hamel bases. For example, the unit function (1,1) and the function $\cos(\pi t)$ or (-1,1) also form a Hamel basis, and in terms of this basis X has the representation $X = \frac{1}{2}(X(1)+X(2))\cdot(1,1) + \frac{1}{2}(X(2)-X(1))\cdot(-1,1)$.

Another way to write a random vector X is to define an index set $\mathbf{T} = \{1,...,T\}$, and then define X as a real-valued function on **S** and **T**, $X:\mathbf{S}\times\mathbf{T}\to\mathbb{R}$. Then, $X(\cdot,t)$ is a simple random variable for each $t \in \mathbf{T}$, and $X(s,\cdot)$ is a real vector that is a realization of X for each $s \in \mathbf{S}$. A function defined in this way is also called a *stochastic process*, particularly when **T** is not finite. The measurability requirement on X is the same as before, but can be written in a different form as requiring that the inverse image of each open interval in $\mathbb{R}$ be contained in $\boldsymbol{F}\otimes\boldsymbol{T}$, where $\boldsymbol{T}$ is a σ-field of subsets of **T** that can be taken to be the family of all subsets of **T** and "⊗" denotes the operation that forms the smallest σ-field containing all sets $\mathbf{A}\times\mathbf{B}$ with $\mathbf{A} \in \boldsymbol{F}$ and $\mathbf{B} \in \boldsymbol{T}$. There is then a complete duality between random vectors in a T-dimensional linear space and random functions on a T-dimensional index set. This duality between vectors and functions will generalize and provide useful insights into statistical applications in which **T** is a more general set indexing time. The *distribution function* (CDF) of X is

$$F(x_1,...,x_T) = P(\{s \varepsilon S \,|\, X_i(s) \le x_i \text{ for } i = 1,...,T\}).$$

If $\mathbf{A} \in \boldsymbol{B}^T$, define $F(\mathbf{A}) = P(\{s \varepsilon \mathbf{S} \,|\, X(s) \in \mathbf{A}\})$. If $F(\mathbf{A}) = 0$ for every set **A** of Lebesque measure zero, then there exists a *probability density function* (pdf) $f(x_1,...,x_T)$ such that

(1) $F(x_1,...,x_T) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_T} f(y_1,...,y_T)\, dy_1...dy_T.$

F and f are termed the *joint* or *multivariate* CDF and pdf, respectively, of X. The random variable X_1 has a distribution that satisfies

$$F_1(x_1) \equiv P(\{s\in\mathbf{S}\,|\,X_1(s) \le x_1\}) = F(x_1,+\infty,...,+\infty).$$

This random variable is measurable with respect to the σ-subfield $\boldsymbol{G}_1$ containing the events whose occurrence is determined by X_1 alone; i.e., $\boldsymbol{G}_1$ is the family generated by sets of the form $\mathbf{A}\times\mathbb{R}\times...\times\mathbb{R}$ with $\mathbf{A}\in\boldsymbol{B}$. If F is absolutely continuous with respect to Lebesque measure on $\boldsymbol{B}^T$, then there are associated densities f and f_1 satisfying

(2) $$F_1(x_1) = \int_{y_1=-\infty}^{x_1} f_1(y_1)\, dy_1$$

(3) $$f_1(x_1) = \int_{y_2=-\infty}^{+\infty}\cdots\int_{y_n=-\infty}^{+\infty} f(x_1,y_2,...,y_n)\cdot dy_2...dy_n.$$

F_1 and f_1 are termed the *marginal* CDF and pdf, respectively, of X_1.

3.5.8. Corresponding to the concept of a conditional probability, we can define a *conditional distribution*: Suppose **C** is an event in $\boldsymbol{G}_1$ with $P(\mathbf{C}) > 0$. Then, define $F_{(2)}(x_2,...,x_n|\mathbf{C}) = F(\{y\in\mathbb{R}^n|\,y_1\in\mathbf{C},y_2\le x_2,...,y_n\le x_n\})/F_1(\mathbf{C})$ to be the conditional distribution of $(X_2,...,X_n)$ given $X_1\in\mathbf{C}$. When F is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$, the conditional distribution can be written in terms of the joint density,

$$F_{(2)}(x_2,...,x_n|\mathbf{C}) = \frac{\int_{y_1\in\boldsymbol{C}}\int_{y_2=-\infty}^{x_2}\int_{y_n=-\infty}^{x_n} f(y_1,y_2,...,y_n)\cdot dy_1dy_2...dy_n}{\int_{y_1\in\boldsymbol{C}}\int_{y_2=-\infty}^{+\infty}\int_{y_n=-\infty}^{+\infty} f(y_1,y_2,...,y_n)\cdot dy_1dy_2...dy_n}.$$

Taking the limit as **C** shrinks to a point $X_1 = x_1$, one obtains the conditional distribution of $(X_2,...,X_n)$ given $X_1 = x_1$,

$$F_{(2)}(x_2,...,x_n|X_1=x_1) = \frac{\int_{y_2=-\infty}^{x_2}\int_{y_n=-\infty}^{x_n} f(x_1,y_2,...,y_n)\cdot dy_1dy_2...dy_n}{f_1(x_1)},$$

provided $f_1(x_1) > 0$. Finally, associated with this conditional distribution is the conditional density $f_{(2)}(x_2,...,x_n|X_1=x_1) = f(x_1,x_2,...,x_n)/f_1(x_1)$. More generally, one could consider the marginal distributions of any subset, say $X_1,...X_k$, of the vector X, with $X_{k+1},...X_n$ integrated out; and the

conditional distributions of one or more of the variables $X_{k+1},...X_n$ given one or more of the conditions $X_1 = x_1,...,X_k = x_k$.

3.5.9. Just as expectations are defined for a single random variable, it is possible to define expectations for a vector of random variables. For example, $\mathbf{E}(X_1 - \mathbf{E}X_1)(X_2-\mathbf{E}X_2)$ is called the *covariance* of X_1 and X_2, and $\mathbf{E}e^{t'X}$, where $t' = (t_1,...,t_n)$ is a vector of constants, is a (multivariate) moment generating function for the random vector X. Here are some useful properties of expectations of vectors:

(a) If g(X) is a function of a random vector, then $\mathbf{E}g(X)$ is the integral of g with respect to the distribution of X. When g depends on a subvector of X, then $\mathbf{E}g(X)$ is the integral of g(y) with respect to the marginal distribution of this subvector.
(b) If X and Z are random vectors of length n, and *a* and *b* are scalars, then $\mathbf{E}(aX + bZ) = a\mathbf{E}X + b\mathbf{E}Z$.
(c) [Cauchy-Schwartz inequality] If X and Z are random vectors of length n, then $(\mathbf{E}X'Z)^2 \leq (\mathbf{E}X'X)(\mathbf{E}Z'Z)$.
(d) [Minkowski Inequality] If X is a random vector of length n and $r \geq 1$ is a scalar, then $(\mathbf{E}|\sum_{i=1}^n X_i|^r)^{1/r} \leq \sum_{i=1}^n (\mathbf{E}|X_i|^r)^{1/r}$.

(e) [Loeve Inequality] If X is a random vector of length n and $r > 0$, then $\mathbf{E}|\sum_{i=1}^n X_i|^r \leq \max(1,n^{r-1}) \sum_{i=1}^n \mathbf{E}|X_i|^r$.

(f) [Jensen Inequality] If X is a random vector and g(x) is a convex function, then $\mathbf{E}\, g(X) \geq g(\mathbf{E}X)$. If g(x) is a concave function, the inequality is reversed.

When expectations exist, they can be used to bound the probability that a random variable takes on extreme values.

Theorem 3.6. Suppose X is a n×1 random vector and ε is a positive scalar.
a. [Markov bound] If $\max_i \mathbf{E}|X_i| < +\infty$, then $\max_i Pr(|X_i| > \varepsilon) < \max_i \mathbf{E}|X_i|/\varepsilon$.
b. [Chebyshev bound] If $\mathbf{E}X'X < +\infty$, then $Pr(\|X\|_2 > \varepsilon) < \mathbf{E}X'X/\varepsilon^2$.
c. [Chernoff bound] If $\mathbf{E}e^{t'X}$ exists for all vectors t in some neighborhood of zero, then for some positive scalars α and M, $Pr(\|X\|_2 > \varepsilon) < Me^{-\alpha\varepsilon}$.

Proof: All these inequalities are established by the same technique: If r(y) is a positive non-decreasing function of $y > 0$, and $\mathbf{E}r(\|X\|) < +\infty$, then

$$\Pr(\|X\| > \varepsilon) = \int_{\|x\|>\varepsilon} F(dx) \le \int_{\|x\|>\varepsilon} [r(\|x\|)/r(\varepsilon)]F(dx) \le \mathbf{E}r(\|X\|)/r(\varepsilon).$$

Taking $r(y) = y^2$ gives the result directly for the Chebyshev bound. In the remaining cases, first get a component-by-component inequality. For the Markov bound, $\Pr(|X_i| > \varepsilon) < E|X_i|/\varepsilon$ for each i gives the result. For the Chernoff bound,

$$\Pr(\|X\|_2 > \varepsilon) \le \sum_{i=1}^{n} [\Pr(X_i > \varepsilon \cdot n^{-1/2}) + \Pr(X_i < -\varepsilon \cdot n^{-1/2})]$$

since if the event on the left occurs, one of the events on the right must occur. Then apply the inequality $\Pr(|X_i| > \varepsilon) \le \mathbf{E}r(|X_i|)/r(\varepsilon)$ with $r(y) = n^{-1/2} \cdot e^{y\alpha}$ to each term in the right-hand-side sum. The inequality for vectors is built up from a corresponding inequality for each component. □

3.5.10. When the expectation of a random variable is taken with respect to a conditional distribution, it is called a *conditional expectation*. If $F(x|\mathbf{C})$ is the conditional distribution of a random vector X given the event **C**, then the conditional expectation of a function g(X) given **C** is defined as

$$\mathbf{E}_{X|\mathbf{C}}g(X) = \int g(y)F(dy|\mathbf{C}).$$

Another notation for this expectation is $\mathbf{E}(g(X)|\mathbf{C})$. When the distribution of the random variable X is absolutely continuous with respect to Lebesgue measure, so that it has a density f(x), the conditional density can be written as $f(x|\boldsymbol{C}) = f(x) \cdot \mathbf{1}(x \in \boldsymbol{C}) / \int_{\boldsymbol{C}} f(s)ds$, and the conditional expectation can then be written

$$\mathbf{E}_{X|\mathbf{C}}g(X) = \int_{\boldsymbol{C}} g(x) \cdot f(x|\boldsymbol{C})dx = \frac{\int_{\boldsymbol{C}} g(x) \cdot f(x)dx}{\int_{\boldsymbol{C}} f(x)dx}.$$

When the distribution of X is discrete, this formula becomes

$$\mathbf{E}_{X|\mathbf{C}}g(X) = \frac{\sum_{k \in \boldsymbol{C}} g(k) \cdot f(k)}{\sum_{k \in \boldsymbol{C}} f(k)}.$$

The conditional expectation is actually a <u>function</u> on the σ-field $\boldsymbol{C}$ of conditioning events, and is sometimes written $\mathbf{E}_{X|\boldsymbol{C}}\, g(X)$ or $\mathbf{E}(g(X)|\boldsymbol{C})$ to emphasize this dependence.

Suppose $\mathbf{A}_1,...,\mathbf{A}_k$ *partition* the domain of X. Then the distribution satisfies

$$F(\mathbf{x}) = \sum_{i=1}^{k} F(x|\mathbf{A}_i)\cdot F(\mathbf{A}_i),$$

implying

$$\mathbf{E}g(X) = \int g(x)F(dx) = \sum_{i=1}^{k} \int g(x)F(dx\,|\,\mathbf{A}_i)\cdot F(\mathbf{A}_i) = \sum_{i=1}^{k} \mathbf{E}\{g(X)|\mathbf{A}_i\}\cdot F(\mathbf{A}_i).$$

This is called the *law of iterated expectations*, and is heavily used in econometrics.

Example 2. (continued) Recall that X is the number of heads that appear before a tail in a sequence of coin tosses, and that the probability of X = k is 2^{-k-1} for k = 0,1,... . Let **C** be the event of an even number of heads. Then,

$$\mathbf{E}_{X|\mathbf{C}}X = \frac{\sum_{k=0,2,4,\ldots} k\cdot 2^{-k-1}}{\sum_{k=0,2,4,\ldots} 2^{-k-1}} = \frac{\sum_{j=0,1,2,\ldots} j\cdot 4^{-j}}{\sum_{j=0,1,2,\ldots} 4^{-j}/2} = 2/3,$$

where the second ratio is obtained by substituting k = 2j, and the value is obtained using the summation formulas for a geometric series from 2.1.10. A similar calculation for the event A of an odd number of heads yields $\mathbf{E}_{X|\mathbf{A}}X = 5/3$. The probability of an even number of heads is $\sum_{k=0,2,4,\ldots} 2^{-k-1} = 2/3$. The law of iterated expectations then gives

$$\mathbf{E}\,X = \mathbf{E}\{X|\mathbf{C})\cdot P(\mathbf{C}) + \mathbf{E}\{X|\mathbf{A})\cdot P(\mathbf{A}) = (2/3)(2/3) + (5/3)(1/3) = 1,$$

which confirms the direct calculation of **E** X.

The concept of a conditional expectation is very important in econometrics and in economic theory, so we will work out its properties in some detail for the case of two variables. Suppose random variables (U,X) have a joint density f(u,x). The marginal density of X is defined by

$$g(x) = \int_{u=-\infty}^{+\infty} f(u,x)du,$$

and the conditional density of U given X = x is defined by $f(u\,|\,x) = f(u,x)/g(x)$, provided g(x) > 0. The conditional expectation of a function h(U,X) satisfies $\mathbf{E}(h(U,X)\,|\,X=x) = \int h(u,x)f(u\,|\,x)du$, and is a function of x. The unconditional expectation of h(U,X) satisfies

$$\mathbf{E}h(U,X) = \int\int h(u,x)f(u,x)dudx = \int_{x=-\infty}^{+\infty}\left(\int_{u=-\infty}^{\infty} h(u,x)f(u\,|\,x)du\right) g(x)dx = \mathbf{E}_X\mathbf{E}_{U|X}h(U,X);$$

another example of the law of iterated expectations. The *conditional mean* of U given X=x is $\mathbf{M}_{U|X}(x) \equiv \mathbf{E}_{U|X=x}U$; by the law of iterated expectations, the conditional and unconditional mean are

related by $\mathbf{E}_U U = \mathbf{E}_X \mathbf{E}_{U|X} U \equiv \mathbf{E}_X \mathbf{M}_{U|X}(X)$. The *conditional variance* of U is defined by $\mathbf{V}(U|x) = \mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(x))^2$. It is related to the unconditional variance by the formula

$$\begin{aligned}
\mathbf{E}_U(U - \mathbf{E}_U U)^2 &= \mathbf{E}_X \mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X) + \mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2 \\
&= \mathbf{E}_X \mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X))^2 + \mathbf{E}_X \mathbf{E}_{U|X}(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2 + 2\mathbf{E}_X \mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X))(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U) \\
&= \mathbf{E}_X \mathbf{V}(U|X) + \mathbf{E}_X(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2 + 2\mathbf{E}_X(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)\mathbf{E}_{U|X}(U - \mathbf{M}_{U|X}(X)) \\
&= \mathbf{E}_X \mathbf{V}(U|X) + \mathbf{E}_X(\mathbf{M}_{U|X}(X) - \mathbf{E}_U U)^2
\end{aligned}$$

Then, the unconditional variance equals the expectation of the conditional variance plus the variance of the conditional expectation.

Example 10: Suppose (U,X) are bivariate normal with means $\mathbf{E}U = \mu_u$ and $\mathbf{E}X = \mu_x$, and second moments $\mathbf{E}(U-\mu_u)^2 = \sigma_u^2$, $\mathbf{E}(X-\mu_x)^2 = \sigma_x^2$, and $\mathbf{E}(U-\mu_u)(X-\mu_x) = \sigma_{ux} \equiv \rho\sigma_u\sigma_x$. Define

$$Q = \frac{1}{1-\rho^2}\cdot\left[\left(\frac{u-\mu_u}{\sigma_u}\right)^2 + \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\cdot\rho\cdot\left(\frac{u-\mu_u}{\sigma_u}\right)\cdot\left(\frac{x-\mu_x}{\sigma_x}\right)\right],$$

and observe that

$$Q - \left(\frac{x-\mu_x}{\sigma_x}\right)^2 = \frac{1}{1-\rho^2}\cdot\left[\left(\frac{u-\mu_u}{\sigma_u}\right) - \rho\cdot\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^2 .$$

The bivariate normal density is $f(u,x) = [2\pi\sigma_u\sigma_x(1-\rho^2)^{1/2}]^{-1}\cdot\exp(-Q/2)$. The marginal density of X is normal with mean μ_x and variance σ_x^2: $n(x-\mu_x,\sigma_x) = (2\pi\sigma_x^2)^{-1}\cdot\exp(-(x-\mu_x)^2/2\sigma_x^2)$. This can be derived from the bivariate density by completing the square for u in Q and integrating over u. The conditional density of U given X then satisfies

$$f(u|x) = [2\pi\sigma_u\sigma_x(1-\rho^2)^{1/2}]^{-1}\cdot\exp(-Q/2)/(2\pi\sigma_x^2)^{-1}\cdot\exp(-(x-\mu_x)^2/2\sigma_x^2).$$

$$= [2\pi\sigma_u^2(1-\rho^2)]^{-1/2}\cdot \exp\left(\frac{-1}{2\cdot(1-\rho^2)}\cdot\left[\left(\frac{u-\mu_u}{\sigma_u}\right) - \rho\cdot\left(\frac{x-\mu_x}{\sigma_x}\right)\right]^2\right) .$$

Hence the conditional distribution of U, given X = x, is normal with conditional mean $\mathbf{E}(U|X=x) = \mu_u + \rho\sigma_u(x-\mu_x)/\sigma_x \equiv \mu_u + \sigma_{ux}(x-\mu_x)/\sigma_x^2$ and variance $\mathbf{V}(U|X=x) \equiv \mathbf{E}((U-\mathbf{E}(U|X=x))^2|X=x) = \sigma_u^2(1-\rho^2) \equiv \sigma_u^2 - \sigma_{ux}^2/\sigma_x^2$. When U and X are joint normal random <u>vectors</u> with $\mathbf{E}U = \mu_u$, $\mathbf{E}X = \mu_x$, $\mathbf{E}(U-\mu_u)(U-\mu_u)' = \Omega_{uu}$, $\mathbf{E}(X-\mu_x)(X-\mu_x)' = \Omega_{xx}$, and $\mathbf{E}(U-\mu_u)(X-\mu_x)' = \Omega_{ux}$, then $(U|X=x)$ is normal with $\mathbf{E}(U|X=x) = \mu_u + \Omega_{ux}\Omega_{xx}^{-1}(x-\mu_x)$ and $\mathbf{V}(U|X=x) = \Omega_{uu} - \Omega_{ux}\Omega_{xx}^{-1}\Omega_{xu}$.

3.5.11. Conditional densities satisfy $f(u,x) = f(u|x)g(x) = f(x|u)h(u)$, where $h(u)$ is the marginal density of U, and hence $f(u|x) = f(x|u)\, h(u)/g(x)$. This is called *Bayes Law*. When U and X are independent, $f(u,x) = h(u)\cdot g(x)$, or $f(u|x) = h(u)$ and $f(x|u) = g(x)$. For U and X independent, and $r(\cdot)$ and $s(\cdot)$ any functions, one has $E(r(U)|X=x) = \int r(u)f(u|x)du \equiv \int r(u)h(u)du = \mathbf{E}r(U)$, and $\mathbf{E}(r(U)s(X)) = \int r(u)s(x)f(u,x)dudx = \int s(x)g(x)\int r(u)f(u|x)du\, dx = \int s(x)g(x)\mathbf{E}r(U|x)dx = [\mathbf{E}s(X)][\mathbf{E}r(U)]$, or $\mathrm{cov}(r(U),s(X)) = 0$, provided $\mathbf{E}r(U)$ and $\mathbf{E}s(X)$ exist. If $r(u) = u - \mathbf{E}U$, then $\mathbf{E}(r(U)|X=x) = 0$ and $\mathrm{cov}(U,X) = \mathbf{E}(U-\mathbf{E}U)X = 0$. Conversely, suppose U and X are jointly distributed. If $\mathrm{cov}(r(U),s(X)) = 0$ for all functions $r(\cdot)$, $s(\cdot)$ such that $\mathbf{E}r(U)$ and $\mathbf{E}s(X)$ exist, then X and U are independent. To see this, choose $r(u) = 1$ for $u \le u^*$, $r(u) = 0$ otherwise; choose $s(x) = 1$ for $x \le x^*$, $s(x) = 0$ otherwise. Then $\mathbf{E}r(U) = H(u^*)$ and $\mathbf{E}s(X) = G(x^*)$, where H and G are the marginal cumulative distribution functions, and $0 = \mathrm{cov} = F(u^*,x^*) - H(u^*)\cdot G(x^*)$, where F is the joint cumulative distribution function. Hence, $F(u,x) = H(u)\cdot G(x)$, and X, U are independent.

Note that $\mathrm{cov}(U,X) = 0$ is not sufficient to imply U,X independent. For example, $g(x) = ½$ for $-1 \le x \le 1$ and $f(u|x) = ½$ for $-1 \le u-x^2 \le 1$ is nonindependent with $\mathbf{E}(U|X=x) = x^2$, but $\mathrm{cov}(U,X) = \mathbf{E}X^3 = 0$. Furthermore, $\mathbf{E}(U|X=x) \equiv 0$ is not sufficient to imply U,X independent. For example, $g(x) = ½$ for $-1 \le x \le 1$ and $f(u|x) = 1/2(1 + x^2)$ for $-(1+ x^2) \le u \le (1 + x^2)$ is nonindependent with $\mathbf{E}(U^2|x) = (1 + x^2)^2 \ne \mathbf{E}\, U^2 = 28/15$, but $\mathbf{E}(U|X=x) \equiv 0$.

Example 11. Suppose monthly family income (in thousands of dollars) is a random variable Y with a CDF $F(y) = 1 - y^{-2}$ for $y > 1$. Suppose a random variable Z is one for home owners and zero otherwise, and that the conditional probability of the event $Z = 1$, given Y, is $(Y-1)/Y$. The unconditional expectation of Y is 2. The joint density of Y and Z is $f(y)\cdot g(z|y) = (2y^{-3})\,(1 - y^{-1})$ for $z = 1$. The unconditional probability of $Z = 1$ is then $\int_{y=1}^{+\infty} f(y)\cdot g(z|y)dy = 1/3$. Bayes Law gives the conditional density of Y given $z = 1$, $f(y|z) = f(y)\cdot g(z|y)/\int_{y=1}^{+\infty} f(y)\cdot g(z|y)dy = (6y^{-3})\,(1 - y^{-1})$, so that the conditional expectation of Y given $z = 1$ is $E(Y|Z=1) = \int_{y=1}^{+\infty} y\, f(y|z)dy = 3$.

Example 12. The problem of interpreting the results of medical tests illustrates Bayes Law. A blood test for prostate cancer is known to yield a "positive" with probability 0.9 if cancer is present, and a false "positive" with probability of 0.2 if cancer is not present. The prevalence of the cancer in the population of males is 0.05. Then, the conditional probability of cancer, given a "positive" test result, equals the joint probability of cancer and a positive test result, (0.05)(0.9), divided by the probability of a positive test result, (0.05)(0.9)+(0.95)(0.2), or 0.235. Thus, a "positive" test has a low probability of identifying a case of cancer, and if all "positive" tests were followed by surgery, about 75 percent of these surgeries would prove unnecessary.

3.5.12. The discussion of expectations will be concluded with a list of detailed properties of characteristic functions and moment generating functions:

a. $\psi(t) = \mathbf{E}e^{\iota tY} \equiv \mathbf{E}\cos(tY) + \iota\mathbf{E}\sin(tY)$.
b. $Z = a + bY$ has the cf $e^{\iota ta}\psi(bt)$ and $Z = f(Y)$ has the cf $\mathbf{E}e^{\iota t\cdot f(Y)}$.
c. If $\mathbf{E}Y^k$ exists, then $\psi^{(k)}(t) \equiv d^k\psi(t)/dt^k$ exists, satisfies the bound $|d^k\psi(t)/dt^k| \leq \mathbf{E}|Y|^k$, and is uniformly continuous, and $\mathbf{E}Y^k = (\iota)^k\psi^{(k)}(0)$. If $\psi^{(k)}(t)$ exists, then $\mathbf{E}Y^k$ exists.
d. If Y has finite moments through order k, then $\psi(t)$ has a Taylor's expansion

$$\psi(t) = \sum_{j=0}^{k} \iota^j(\mathbf{E}Y^j)t^j/j! + [\psi^{(k)}(\lambda t) - \psi^{(k)}(0)]t^k/k!$$

where λ is a scalar with $0 < \lambda < 1$; the Taylor's expansion satisfies the bounds

$$|\psi(t) - \sum_{j=0}^{k-1} \iota^j(\mathbf{E}Y^j)t^j/j!| \leq |t|^k\mathbf{E}|Y|^k/k!$$

and

$$|\psi(t) - \sum_{j=0}^{k} \iota^j(\mathbf{E}Y^j)t^j/j!| \leq 2|t|^k\mathbf{E}|Y|^k/k!$$

If $\mathbf{E}Y^k$ exists, then the expression $\zeta(t) = \text{Ln } \psi(t)$, called the *second characteristic function* or *cumulant generating function*, has a Taylor's expansion

$$\zeta(t) = \sum_{j=1}^{k} \kappa_j\iota^jt^j/j! + [\zeta^{(k)}(\lambda t) - \zeta^{(k)}(t)],$$

where $\zeta^{(k)} \equiv d^k\zeta/dt^k$, and λ is a scalar with $0 < \lambda < 1$. The expressions κ_j are called the *cumulants* of the distribution, and satisfy $\kappa_1 = \mathbf{E}Y$ and $\kappa_2 = \text{Var}(Y)$. The expression $\kappa_3/\kappa_2^{3/2}$ is called the *skewness*, and the expression $\kappa_4/\kappa_2^2 - 3$ is called the *kurtosis* (i.e., thickness of tails relative to center), of the distribution.
e. If Y is normally distributed with mean μ and variance σ^2, then its characteristic function is $\exp(\iota\mu t - \sigma^2t^2/2)$. The normal has cumulants $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, $\kappa_3 = \kappa_4 = 0$.
f. Random variables X and Y have identical distribution functions if and only if they have identical characteristic functions.
g. If $Y_n \rightarrow_p Y$ (see Chap. 4.1), then the associated characteristic functions satisfy $\psi_n(t) \rightarrow \psi(t)$ for each t. Conversely, if Y_n has characteristic function $\psi_n(t)$ converging pointwise to a function $\psi(t)$ that is continuous at $t = 0$, then there exists Y such that $\psi(t)$ is the characteristic function of Y and $Y_n \rightarrow_p Y$.
h. The characteristic function of a sum of independent random variables equals the product of the characteristic functions of these random variables, and the second characteristic function of a sum of independent random variables is the sum of the second characteristic functions of these variables; the characteristic function of a mean of n independently identically distributed random variables, with characteristic function $\psi(t)$, is $\psi(t/n)^n$.

Similar properties hold for proper moment generating functions, with obvious modifications: Suppose a random variable Y has a proper mgf m(t), finite for $|t| < \tau$, where τ is a positive constant. Then, the following properties hold:

a. $m(t) = \mathbf{E}e^{tY}$ for $|t| < \tau$.
b. $Z = a + bY$ has the mgf $e^{ta}m(bt)$.
c. $\mathbf{E}Y^k$ exists for all $k > 0$, and $m \equiv d^k m(t)/dt^k$ exists and is uniformly continuous for $|t| < \tau$, with $\mathbf{E}Y^k = m_Y(0)$.
d. m(t) has a Taylor's expansion (for any k) $m_Y(t) = (\mathbf{E}Y^j)t^j/j! + [m(\lambda t) - m(0)]t^k/k!$, where λ is a scalar with $0 < \lambda < 1$.
e. If Y is normally distributed with mean μ and variance σ^2, then it has mgf $\exp(\mu t+\sigma 2t2)$.
f. Random variables X and Y with proper mgf have identical distribution functions if and only if their mgf are identical.
g. If $Y_n \rightarrow_p Y$ and the associated mgf are finite for $|t| < \tau$, then the mgf of Y_n converges pointwise to the MGF of Y. Conversely, if Y_n have proper MGF which converges pointwise to a function m(t) that is finite for $|t| < \tau$, then there exists Y such that m(t) is the mgf of Y and $Y_n \rightarrow_p Y$.
h. The mgf of a sum of independent random variables equals the product of the mgf of these random variables; the mgf of the mean of n independently identically distributed random variables, each with proper mgf m(t), is $m(t/n)^n$.

The definitions of characteristic and moment generating functions can be extended to vectors of random variables. Suppose Y is a n×1 random vector, and let $\mathbf{t}$ be a n×1 vector of constants. Then $\psi(\mathbf{t}) = \mathbf{E}e^{\mathbf{t}'Y}$ is the characteristic function and $m(\mathbf{t}) = \mathbf{E}e^{\mathbf{t}'Y}$ is the moment generating function. The properties of cf and mgf listed above also hold in their multivariate versions, with obvious modifications. For characteristic functions, two of the important properties translate to

(b') $Z = \mathbf{a} + \mathbf{B}Y$, where $\mathbf{a}$ is a m×1 vector and $\mathbf{B}$ is a m×n matrix, has cf $e^{\mathbf{t}'\mathbf{a}}\psi(\mathbf{Bt})$.
(e') if Y is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then its characteristic function is $\exp(\iota\boldsymbol{\mu}'\mathbf{t} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$.

A useful implication of (b') and (e') is that a linear transformation of a multivariate normal vector is again multivariate normal. Conditions (c) and (d) relating Taylor's expansions and moments for univariate cf have multivariate versions where the expansions are in terms of partial derivatives of various orders. Conditions (f) through (h) are unchanged in the multivariate version.

The properties of characteristic functions and moment generating functions are discussed and established in C. R. Rao Linear Statistical Inference, 2b.4, and W. Feller An Introduction to Probability Theory, II, Chap. 13 and 15.

3.6. TRANSFORMATIONS OF RANDOM VARIABLES

6.1. Suppose X is a measurable random variable on $(\mathbb{R},\boldsymbol{B})$ with a distribution F(x) that is absolutely continuous with respect to Lebesgue measure, so that X has a density f(x). Consider an increasing transformation Y = H(X); then Y is another random variable. Let h denote the inverse function of H; i.e., y = H(x) implies x = h(y). The distribution function of Y is given by

$$G(y) = Pr(Y \leq y) = Pr(H(X) \leq y) = Pr(X \leq h(y)) = F(h(y)).$$

When h(y) is differentiable, with a derivative $h'(y) = dh(y)/dy$, the density of Y is obtained by differentiating, and satisfies $g(y) = f(h(y))h'(y)$. Since $y \equiv H(h(y))$, one obtains by differentiation the formula $1 \equiv H'(h(y))h'(y)$, or $h'(y) = 1/H'(h(y))$. Substituting this formula gives $g(y) = f(h(y))/H'(h(y))$.

Example 13. Suppose X has the distribution function $F(x) = 1-e^{-x}$ for $x > 0$, with F(x) = 0 for $x \leq 0$; then X is said to have an exponential distribution. Suppose $Y = H(X) \equiv \log X$, so that $X = h(Y) \equiv e^{Y}$. Then, $G(y) = 1-\exp(-e^{y})$ and $G(y) = \exp(-e^{y})e^{y} = \exp(y-e^{y})$ for $-\infty < y < +\infty$. This is called an extreme value distribution. A third example is X with some distribution function F and density f, and Y = F(X), so that for any value of X, the corresponding value of Y is the proportion of all X that are below this value. Let x_p denote the solution to F(x) = p. The distribution function of Y is $G(y) = F(x_y) = y$. Hence, Y has the uniform density on the unit interval.

The rule for an increasing transformation of a random variable X can be extended in several ways. If the transformation Y = H(X) is decreasing rather than increasing, then

$$G(y) = Pr(Y \leq y) = Pr(H(X) \leq y) = Pr(X \geq h(y)) = 1-F(h(y)),$$

where h is the inverse function of H. Differentiating,

$$g(y) = f(h(y))(-h'(y)).$$

Then, combining cases, one has the result that *for any one-to-one transformation* Y = H(X) *with inverse* X = h(Y)*, the density of* Y *is*

$$g(y) = f(h(y))|h'(y)| \equiv f(h(y))/|H'(h(y)|.$$

An example of a decreasing transformation is X with the exponential density e^{-x} for $x > 0$, and $Y = 1/X$. Show as an exercise that $G(y) = e^{-1/y}$ and $g(y) = e^{-1/y}/y^2$.

Consider a transformation Y = H(X) that is not one-to-one. The interval $(-\infty,y)$ is the image of a set A_y of x values that may have a complicated structure. One can write

$$G(y) = Pr(Y \leq y) = Pr(H(X) \leq y) = Pr(X \in A_y) = F(A_y).$$

If this expression is differentiable, then its derivative gives the density.

Example 14. If X has a distribution F and density f, and $Y = |X|$, then $A_y = [-y,y]$, implying $G(y) = F(y) - F(-y)$ and $f(y) = f(y) + f(-y)$.

Example 15. If $Y = X^2$, then $A_y = [-y^{1/2},y^{1/2}]$, $G(y) = F(y^{1/2}) - F(-y^{1/2})$. Differentiating for $y \neq 0$, $g(y) = (f(y^{1/2}) + f(-y^{1/2}))/2y^{1/2}$. Applying this to the standard normal with $F(x) = \Phi(x)$, the density of Y is $g(y) = \varphi(y^{1/2})/y^{1/2} = (2\pi y)^{-½}\cdot e^{-y/2}$, called the chi-square with one degree of freedom.

3.6.2. Next consider transformations of random vectors. These transformations will permit us to analyze sums or other functions of random variables. Suppose X is a n×1 random vector. Consider first the transformation Y = AX, where A is a nonsingular n×n matrix. The following result from multivariate calculus relates the densities of X and Y:

Theorem 3.8. If X has density f(x), and Y = **A**X, with **A** nonsingular, then the density of Y is

$$g(y) = f(\mathbf{A}^{-1}y)/|\det(\mathbf{A})|\,.$$

Proof: We will prove the result in two dimensions, leaving the general case to the reader. First, consider the case $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{22} > 0$. One has $G(y_1,y_2) \equiv F(y_1/a_{11},y_2/a_{22})$. Differentiating with respect to y_1 and y_2, $g(y_1,y_2) \equiv f(y_1/a_{11},y_2/a_{22})/a_{11}a_{22}$. This establishes the result for diagonal transformations. Second, consider $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{22} > 0$. Then

$G(y_1,y_2) \equiv \int_{x_1=-\infty}^{y_1/a_{11}} \int_{x_2=-\infty}^{(y_2-a_{21})/a_{22}} f(x_1,x_2)dx_2dx_1$. Differentiating with respect to y_1 and y_2 yields

$$\partial^2 G(y_1,y_2)/\partial y_1 \partial y_2 \equiv g(y_1,y_2) = (a_{11}a_{22})^{-1} f(y_1/a_{11},(y_2-y_1a_{21}/a_{11})/a_{22}).$$

This establishes the result for triangular transformations. Finally, consider the general transformation $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with $a_{11} > 0$ and $a_{11}a_{22}-a_{12}a_{21} > 0$. Apply the result for triangular transformations first to $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & a_{12}/a_{11} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, and second to $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22}-a_{12}a_{21}/a_{11} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$. This gives the general transformation, as $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22}-a_{12}a_{21}/a_{11} \end{bmatrix} \begin{bmatrix} 1 & a_{12}/a_{11} \\ 0 & 1 \end{bmatrix}$. The density of Z is $h(z_1,z_2) = f(z_1-z_2a_{12}/a_{11},z_2)$, and of Y is $g(y_1,y_2) = h(y_1/a_{11},(y_2-y_1a_{21}/a_{11})/(a_{22}-a_{12}a_{21}/a_{11}))$. Substituting for h in the last expression and simplifying gives

$$g(y_1,y_2) = f((a_{22}y_1-a_{12}y_2)/D,(a_{11}y_2-a_{21}y_1)/D)/D,$$

where $D = a_{11}a_{22}-a_{12}a_{21}$ is the determinant of the transformation.

We leave as an exercise the proof of the theorem for the density of Y = AX in the general case with A n×n and nonsingular. First, recall that A can be factored so that A = PLDU′Q′, where P and Q are permutation matrices, L and U are lower triangular with ones down the diagonal, and D is a nonsingular diagonal matrix. Write Y = PLDUQ′X. Then consider the series of intermediate transformations obtained by applying each matrix in turn, constructing the densities as was done previously. □

3.6.3. The extension from linear transformations to one-to-one nonlinear transformations of vectors is straightforward. Consider Y = H(X), with an inverse transformation X = h(Y). At a point y^o and $x^o = h(y^o)$, a first-order Taylor's expansion gives

$$y - y^o = \mathbf{A}(x - x^o) + o(x - x^o),$$

where **A** is the *Jacobean* matrix

$$\mathbf{A} = \begin{bmatrix} \partial H^1(x^o)/\partial x_1 & \dots & \partial H^1(x^o)/\partial x_n \\ | & & | \\ \partial H^n(x^o)/\partial x_1 & \dots & \partial H^n(x^o)/\partial x_n \end{bmatrix}$$

and the notation $o(z)$ means an expression that is small relative to z. Alternately, one has

$$\mathbf{B} = \mathbf{A}^{-1} = \begin{bmatrix} \partial h^1(y^o)/\partial y_1 & \dots & \partial h^1(y^o)/\partial y_n \\ | & & | \\ \partial h^n(x^o)/\partial y_1 & \dots & \partial h^n(y^o)/\partial y_n \end{bmatrix}.$$

The probability of Y in the little rectangle $[y^o, y^o+\Delta y]$ is approximately equal to the probability of X in the little rectangle $[x^o, x^o+A^{-1}\Delta y]$. This is the same situation as in the linear case, except there the equality was exact. Then, the formulas for the linear case carry over directly, with the Jacobean matrix of the transformation replacing the linear transformation matrix A. If f(x) is the density of X, then $g(y) = f(h(y))\cdot|\det(\mathbf{B})| = f(h(y))/|\det(\mathbf{A})|$ is the density of Y.

Example 16. Suppose a random vector (X,Z) has a density f(x,z) for x,z > 0, and consider the nonlinear transformation W = X·Z and Y = X/Z, which has the inverse transformation $X = (WY)^{1/2}$ and $Z = (W/Y)^{1/2}$. The Jacobean matrix is $B = \begin{bmatrix} W^{-1/2}Y^{1/2}/2 & W^{1/2}Y^{-1/2}/2 \\ W^{-1/2}Y^{-1/2}/2 & -W^{1/2}Y^{-3/2}/2 \end{bmatrix}$, and det(**B**) = 1/2y. Hence, the density of (w,y) is $f((wy)^{1/2},(w/y)^{1/2})/2y$.

In principle, it is possible to analyze n-dimensional nonlinear transformations that are <u>not</u> one-to-one in the same manner as the one-dimensional case, by working with the one-to-many inverse transformation. There are no general formulas, and each case needs to be treated separately.

Often in applications, one is interested in a transformation from a n×1 vector of random variables X to a lower dimension. For example, one may be interested in the scalar random variable $S = X_1 + \dots + X_n$. If one "fills out" the transformation in a one-to-one way, so that the random variables of interest are components of the complete transformation, then Theorem 3.6 can be applied. In the case of S, the transformation $Y_1 \equiv S$ filled out by $Y_i = X_i$ for i = 2,...,n is one-to-one, with

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ | \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ | & | & | & & | \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ | \\ X_n \end{bmatrix}.$$

Example 17. Consider a random vector (X,Z) with a density f(x,z), and the transformation S = X + Z and T = Z, or $\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}$. The Jacobean of this transformation is one, and its inverse is $\begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S \\ T \end{bmatrix}$, so the density of (S,T) is g(s,t) = f(s-t,t). The marginal density of S is then $g_1(s) = \int_{t=-\infty}^{+\infty} f(s-t,t)dt$. If X and Z are statistically independent, so that their density is $f(x,z) = f_1(x) \cdot f_2(z)$, then this becomes $g_1(s) = \int_{t=-\infty}^{+\infty} f_1(s-t) \cdot f_2(t)dt$. This is termed a *convolution* formula.

3.7. SPECIAL DISTRIBUTIONS

3.7.1. A number of special probability distributions appear frequently in statistics and econometrics, because they are convenient for applications or illustrations, because they are useful for approximations, or because they crop up in limiting arguments. The tables at the end of this Chapter list many of these distributions.

3.7.2. Table 3.1 lists discrete distributions. The binomial and geometric distributions are particularly simple, and are associated with statistical experiments such as coin tosses. The Poisson distribution is often used to model the occurrence of rare events. The hypergeometric distribution is associated with classical probability experiments of drawing red and white balls from urns, and is also used to approximate many other distributions.

3.7.3. Table 3.2 list a number of continuous distributions, including some basic distributions such as the gamma and beta from which other distributions are constructed. The extreme value and logistic distributions are used in the economic theory of discrete choice, and are also of statistical interest because they have simple closed form CDF's.

3.7.4. The normal distribution and its related distributions play a central role in econometrics, both because they provide the foundation for finite-sample distribution results for regression models with normally distributed disturbances, and because they appear as limiting approximations in large samples even when the finite sample distributions are unknown or intractable. Table 3.3 lists the normal distribution, and a number of other distributions that are related to it. The t and F distributions appear in the theory of hypothesis testing, and the chi-square distribution appears in

large-sample approximations. The non-central versions of these distributions appear in calculations of the power of hypothesis tests.

It is a standard exercise in mathematical statistics to establish the relationships between normal, chi-square, F, and t distributions. For completeness, we state the most important result:

Theorem 3.9. Normal and chi-square random variables have the following properties:
(I) If $S = Y_1^2 + ... + Y_k^2$, where the Y_k are independent normal random variables with means μ_k and unit variances, then S has a non-central chi-square distribution with degrees of freedom parameter k and non-centrality parameter $\delta = \mu_1^2 + ... + \mu_k^2$, denoted $\chi'^2(k,\delta)$. If $\delta = 0$, this is a (central) chi-square distribution with degrees of freedom parameter k, denoted $\chi^2(k)$.
(ii) If Y and S are independent, Y is normal with mean λ and unit variance, and S is chi-square with k degrees of freedom, then $T = Y/(S/k)^{1/2}$ is non-central t-distributed with degrees of freedom parameter k and non-centrality parameter λ, denoted $t'(k,\lambda)$. If $\lambda = 0$, this is a (central) t-distribution with degrees of freedom parameter k, denoted t(k).
(iii) If R and S are independent, R is non-central chi-square with degrees of freedom parameter k and non-centrality parameter δ, and S is central chi-square with degrees of freedom parameter n, then $F = nR/kS$ is non-central F-distributed with degrees of freedom parameters (k,n) and non-centrality parameter δ, denoted $F'(k,n,\delta)$. If $\delta = 0$, this distribution is F-distributed with degrees of freedom parameters (k,n), and is denoted F(k,n).
(iv) T is non-central t-distributed with degrees of freedom parameter k and non-centrality parameter λ if and only if $F = T^2$ is non-central F-distributed with degrees of freedom parameters (1,k) and non-centrality parameter $\delta = \lambda^2$.

Proof: These results can be found in most classical texts in mathematical statistics; see particularly Rao (1973), pp. 166-167, 170-172, 181-182, Johnson & Kotz (1970), Chap. 26-31, and Graybill (1961), Chap. 4.. □

In applied statistics, it is important to be able to calculate values $x = G^{-1}(p)$, where G is the CDF of the central chi-square, F, or t, distribution, and values $p = G(x)$ where G is the CDF of the non-central chi-square, F, or t distribution. Selected points of these distributions are tabled in many books of mathematical and statistical tables, but it is more convenient and accurate to calculate these values within a statistical or econometrics software package. Most current packages, including TSP, STATA, and SST, can provide these values.

3.7.5. One of the most heavily used distributions in econometrics is the multivariate normal. We describe this distribution and summarize some of its properties. A n×1 random vector **Y** is multivariate normal with a vector of means **μ** and a positive definite covariance matrix **Σ** if it has the density

$$n(\mathbf{y} - \boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-n/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp(-((\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})/2).$$

This density is also sometimes denoted $n(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$, and the CDF denoted $N(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$. Its characteristic function is $\exp(\imath\boldsymbol{\mu}'\mathbf{t} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$, and it has the moments E Y = $\boldsymbol{\mu}$ and E $(Y-\boldsymbol{\mu})(Y-\boldsymbol{\mu})' = \boldsymbol{\Sigma}$. From the characteristic function and the rule for linear transformations, one has immediately the property that a linear transformations of a multivariate normal vector is again multivariate normal. Specifically, if Y is distributed $N(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$, then the linear transformation $Z = \mathbf{a} + \mathbf{B}Y$, which has mean $\mathbf{a} + \mathbf{B}\boldsymbol{\mu}$ and covariance matrix $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}$, is distributed $N(\mathbf{z};\mathbf{a} + \mathbf{B}\boldsymbol{\mu},\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B})$. The dimension of Z need not be the same as the dimension of Y, nor does B have to be of maximum rank; if $\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}$ is less than full rank, then the distribution of Z is concentrated on an affine linear subspace of dimension n through the point $\mathbf{a} + \mathbf{B}\boldsymbol{\mu}$. Let $\sigma_k = (\Sigma_{kk})^{1/2}$ denote the standard deviation of Y_k, and let $\rho_{kj} = \Sigma_{kj}/\sigma_k\,\sigma_j$ denote the correlation of Y_k and Y_j. Then the covariance matrix $\boldsymbol{\Sigma}$ can be written

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} \begin{bmatrix} 1 & \varrho_{12} & .. & \varrho_{1n} \\ \varrho_{21} & 1 & \dots & \varrho_{2n} \\ \vdots & \vdots & & \vdots \\ \varrho_{n1} & \varrho_{n2} & .. & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} = \mathbf{DRD},$$

where $\mathbf{D} = \mathrm{diag}(\sigma_1,\dots,\sigma_n)$ and **R** is the array of correlation coefficients.

Theorem 3.10. Suppose Y is partitioned $\mathbf{Y}' = (\mathbf{Y}_1'\ \mathbf{Y}_2')$, where $\mathbf{Y}_1$ is m×1, and let $\boldsymbol{\mu}' = (\boldsymbol{\mu}_1'\ \boldsymbol{\mu}_2')$ and $\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$ be commensurate partitions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then the marginal density of $\mathbf{Y}_1$ is multivariate normal with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{11}$. The conditional density of $\mathbf{Y}_2$, given $\mathbf{Y}_1 = \mathbf{y}_1$, is multivariate normal with mean $\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$ and covariance matrix $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\,\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Then, the conditional mean of a multivariate normal is linear in the conditioning variables.

Proof: The easiest way to demonstrate the theorem is to recall from Chapter 2 that the positive definite matrix $\boldsymbol{\Sigma}$ has a Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$, where **L** is lower triangular, and that **L** has an inverse **K** that is again lower triangular. If **Z** is a n×1 vector of independent standard normal random variables (e.g., each $\mathbf{Z}_i$ has mean zero and variance 1), then $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z}$ is normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Conversely, if Y has density $n(\mathbf{y} - \boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\mathbf{Z} = \mathbf{K}(Y - \boldsymbol{\mu})$ is a vector of i.i.d. standard normal random variables. These statement use the important property of normal random vectors that a linear transformation is again normal. This can be shown directly by using the formulas in Section 3.6 for densities of linear transformations, or by observing that the (multivariate) characteristic function of Y with density $n(\mathbf{y} - \boldsymbol{\mu},\boldsymbol{\Sigma})$ is $\exp(\imath\mathbf{t}'\boldsymbol{\mu} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2)$, and the form of this characteristic function is unchanged by linear transformations.

The Cholesky construction $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{LZ}$ provides an easy demonstration for the densities of marginal or conditional subvectors of $\mathbf{Y}$. Partition $\mathbf{L}$ and $\mathbf{Z}$ commensurately with $(\mathbf{Y}_1' \; \mathbf{Y}_2')$, so that $\mathbf{L} = \begin{bmatrix} \boldsymbol{L}_{11} & \mathbf{0} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{bmatrix}$ and $Z' = (\mathbf{Z}_1' \; \mathbf{Z}_2')$. Then $\boldsymbol{\Sigma}_{11} = \mathbf{L}_{11}\mathbf{L}_{11}'$, $\boldsymbol{\Sigma}_{21} = \mathbf{L}_{21}\mathrm{L}_{11}'$, $\boldsymbol{\Sigma}_{22} = \mathbf{L}_{22}\mathbf{L}_{22}' + \mathbf{L}_{21}\mathbf{L}_{21}'$, and hence $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \mathbf{L}_{21}\mathbf{L}_{11}^{-1}$, implying $\mathbf{L}_{22}\mathbf{L}_{22}' = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Then, $\mathbf{Y}_1 = \boldsymbol{\mu}_1 + \mathbf{L}_{11}\mathbf{Z}_1$ has a marginal multivariate normal density with mean $\boldsymbol{\mu}_1$ and covariance matrix $\mathbf{L}_{11}\mathrm{L}_{11}' = \boldsymbol{\Sigma}_{11}$. Also, $\mathbf{Y}_2 = \boldsymbol{\mu}_2 + \mathbf{L}_{21}\mathbf{Z}_1 + \mathbf{L}_{22}\mathrm{Z}_2$, implying $\mathbf{Y}_2 = \boldsymbol{\mu}_2 + \mathbf{L}_{21}\mathrm{L}_{11}^{-1}(\mathbf{Y}_1 - \boldsymbol{\mu}_1) + \mathbf{L}_{22}\mathbf{Z}_2$. Conditioned on $\mathbf{Y}_1 = \mathbf{y}_1$, this implies $\mathbf{Y}_2 = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1) + \mathbf{L}_{22}\mathrm{Z}_2$ is multivariate normal with mean $\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. □

The next theorem gives some additional useful properties of the multivariate normal and of quadratic forms in normal vectors.

Theorem 3.11. Let Y be a n×1 random vector. Then,

(i) If $\mathbf{Y}' = (\mathbf{Y}_1' \; \mathbf{Y}_2')$ is multivariate normal, then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if they are uncorrelated. However, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ can be uncorrelated and each have a marginal normal distribution without necessarily being independent.

(ii) If every linear combination $\mathbf{c}'\mathrm{Y}$ is normal, then Y is multivariate normal.

(iii) If Y is i.i.d. standard normal and $\mathbf{A}$ is an idempotent n×n matrix of rank k, then $\mathrm{Y}'\mathbf{A}\mathrm{Y}$ is distributed $\chi^2(\mathrm{k})$.

(iv) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$ and $\mathbf{A}$ is an idempotent n×n matrix of rank k, then $\mathrm{Y}'\mathbf{A}\mathrm{Y}$ is distributed $\chi'^2(\mathrm{k},\delta)$ with $\delta = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

(v) If Y is i.i.d. standard normal and $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite n×n matrices, then $\mathrm{Y}'\mathbf{A}\mathrm{Y}$ and $\mathrm{Y}'\mathbf{B}\mathrm{Y}$ are independent if and only if $\mathbf{AB} = \mathbf{0}$.

(vi) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$, and $\mathbf{A}_\mathrm{i}$ is an idempotent n×n matrix of rank k_i for I = 1,...K, then the $\mathrm{Y}'\mathbf{A}_\mathrm{i}\,\mathrm{Y}$ are mutually independent and distributed $\chi'^2(\mathrm{k}_\mathrm{i},\delta_\mathrm{i})$ with $\delta_\mathrm{i} = \boldsymbol{\mu}'\mathbf{A}_\mathrm{i}\boldsymbol{\mu}$ if and only if either (a) $\mathbf{A}_\mathrm{i}\mathbf{A}_\mathrm{j} = 0$ for $\mathrm{I} \neq \mathrm{j}$ or (b) $\mathbf{A}_1 + ... + \mathbf{A}_\mathrm{K}$ is idempotent.

(vii) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$, $\mathbf{A}$ is a positive semidefinite n×n matrix, $\mathbf{B}$ is a k×n matrix, and $\mathbf{BA} = \mathbf{0}$, then $\mathbf{B}\mathrm{Y}$ and $\mathrm{Y}'\mathbf{A}\mathrm{Y}$ are independent.

(viii) If Y is distributed $N(\boldsymbol{\mu},\mathbf{I})$ and $\mathbf{A}$ is a positive semidefinite n×n matrix, then $\mathrm{E}\; \mathrm{Y}'\mathbf{A}\mathrm{Y} = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A})$.

Proof: Results (i) and (ii) are proved in Anderson (1958), Thm. 2.4.2 and 2.6.2. For (iii) and (iv), write $\mathbf{A} = \mathbf{UU}'$, where this is its singular value decomposition with $\mathbf{U}$ a n×k column orthogonal matrix. Then $\mathbf{U}'\mathrm{Y}$ is distributed $N(\mathbf{U}'\boldsymbol{\mu},\mathbf{I}_\mathrm{k})$, and the result follows from Theorem 3.8. For (v), let k be the rank of $\mathbf{A}$ and m the rank of $\mathbf{B}$. There exists a n×k matrix $\mathbf{U}$ of rank k and a n×m matrix $\mathbf{V}$ of rank m such that $\mathbf{A} = \mathbf{UU}'$ and $\mathbf{B} = \mathbf{VV}'$. The vectors $\mathbf{U}'\mathrm{Y}$ and $\mathbf{V}'\mathrm{Y}$ are uncorrelated, hence

independent, if and only if $\mathbf{U'V} = \mathbf{0}$. But $\mathbf{AB} = \mathbf{U}(\mathbf{U}'\mathbf{V})\mathbf{V}'$ is zero if and only if $\mathbf{U'V} = \mathbf{0}$ since $\mathbf{U}$ and $\mathbf{V}'$ are of maximum rank. For (vi), use the SVD decomposition as in (iv). For (vii), write $\mathbf{A} = \mathbf{U}\mathbf{U}'$ with $\mathbf{U}$ of maximum rank as in (v). Then $\mathbf{BA} = (\mathbf{BU})\mathbf{U}' = \mathbf{0}$ implies $\mathbf{BU} = \mathbf{0}$, so that $\mathbf{B}Y$ and $\mathbf{U}'Y$ are independent by (i). For (vii), $E\ Y'\mathbf{A}Y = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + E\ (Y-\boldsymbol{\mu})'\mathbf{A}(Y-\boldsymbol{\mu}) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + tr(E\ (Y-\boldsymbol{\mu})'\mathbf{A}(Y-\boldsymbol{\mu})) = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + tr(\mathbf{A})$. □

3.8. NOTES AND COMMENTS

The purpose of this chapter has been to collect the key results from probability theory that are used in econometrics. While the chapter is reasonably self-contained, it is expected that the reader will already be familiar with most of the concepts, and can if necessary refer to one of the excellent texts in basic probability theory and mathematical statistics, such as P. Billingsley, *Probability and Measure*, Wiley, 1986; or Y. Chow and H. Teicher, *Probability Theory*, 1997. A classic that provides an accessible treatment of fields of subsets, measure, and statistical independence is J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, 1965. Another classic that contains many results from mathematical statistics is C. R. Rao (1973) *Linear Statistical Inference and Its Applications*, Wiley. A comprehensive classical text with treatment of many topics, including characteristic functions, is W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1&2, Wiley, 1957. For special distributions, properties of distributions, and computation, a four-volume compendium by N. Johnson and S. Kotz, Distributions in Statistics, Houghton-Mifflin, 1970, is a good source. For the multivariate normal distribution, T. Anderson (1958) *An Introduction to Multivariate Statistical Analysis*, Wiley, and F. Graybill (1961) *An Introduction to Linear Statistical Models*, McGraw-Hill, are good sources. Readers who find some sections of this chapter unfamiliar or too dense may find it useful to first review an introductory text at the undergraduate level, such as K. Chung, *A Course in Probability Theory*, Academic Press, New York, or R. Larsen and M. Marx, *Probability Theory*, Prentice-Hall.

TABLE 3.1. SPECIAL DISCRETE DISTRIBUTIONS

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Binomial k = 0,1,...,n	$\binom{n}{k} p^k (1-p)^{n-k}$ 0 < p < 1	μ = np $\sigma^2 = np(1-p)$	$(1-p+pe^{tt})^n$ Note 1
2. Hypergeometric k an integer max{0,n-w} ≤ k & k ≤ min{r,n}	$\binom{r}{k}\binom{w}{n-k} \div \binom{r+w}{n}$ r+w > n r,w,n positive integers	μ = nr/(r+w) $\sigma^2 = \frac{nrw}{(r+w)^2} \cdot \frac{r+w-n}{r+w-1}$	Note 2
3. Geometric k = 0,1,2,...	$p(1-p)^k$ 0 < p < 1	μ = (1-p)/p $\sigma^2 = (1-p)/p^2$	Note 3
4. Poisson k = 0,1,2,...	$e^{-\lambda}\lambda^k/k!$ λ > 0	μ = λ $\sigma^2 = \lambda^2$	$\exp[\lambda(e^{tt}-1)]$ Note 4
5. Negative Binomial k = 0,1,2,...	$\binom{r+k-1}{k} p^r(1-p)^k$ r integer, r > 0 & 0 < p < 1	μ = r(1-p)/p $\sigma^2 = r(1-p)/p^2$	 Note 5

NOTES

1. μ ≡ EX (the mean), and $\sigma^2 = E(X-\mu)^2$ (the variance). The density is often denoted b(k;n,p). The moment generating function is $(1-p+pe^t)^n$.
2. The characteristic and moment generating functions are complicated.
3. The characteristic function is $p/(1-(1-p)e^{tt})$ and the moment generating function is $p/(1-(1-p)e^t)$, defined for $t < -\ln(1-p)$.
4. The moment generating function is $\exp(\lambda(e^t-1))$, defined for all t.
5. The characteristic function is $p^r/(1-(1-p)e^{tt})^r$, and the moment generating function is $p^r/(1-(1-p)e^t)^r$, defined for $t < -\ln(1-p)$.

TABLE 3.2. SPECIAL CONTINUOUS DISTRIBUTIONS

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Uniform $a \le x \le b$	$1/(b-a)$	$\mu = (a+b)/2$ $\sigma^2 = (b-a)^2/12$	$\frac{e^{\iota bt} - e^{\iota at}}{\iota t(b-a)}$ Note 1
2. Triangular $\lvert x\rvert < a$	$(1-\lvert x\rvert/a)/a$	$\mu = 0$ $\sigma^2 = a^2/6$	$2\frac{1-\cos at}{a^2t^2}$
3. Cauchy $-\infty < x < +\infty$	$a/\pi(a^2 + (x-\mu)^2)$	none	$e^{\iota t\mu - \lvert t\delta\rvert}$
4. Exponential $x \ge 0$	$e^{-x/\lambda}/\lambda$	$\mu = \lambda$ $\sigma^2 = \lambda^2$	$1/(1-\iota\lambda t)$ Note 2
5. Pareto $x \ge a$	$ba^b x^{-b-1}$	$\mu = ab/(b-1)$ $\sigma^2 = ba^2/(b-1)^2(b-2)$	Note 3
6. Gamma $x > 0$	$\frac{x^{a-1}e^{x/b}}{\Gamma(a)b^a}$	$\mu = ab$ $\sigma^2 = ab^2$	$(1-\iota bt)^{-a}$ Note 4
7. Beta $0 < x < 1$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\ x^{a-1}(1-x)^{b-1}$	$\mu = a/(a+b)$ $\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$	Note 5
8. Extreme Value $-\infty < x < +\infty$	$\frac{1}{b}\ \exp\left(-\frac{x-a}{b} - e^{-(x-a)/b}\right)$	$\mu = a + 0.57721 \cdot b$ $\sigma^2 = (\pi b)^2/12$	Note 6
9. Logistic $-\infty < x < +\infty$	$\frac{1}{b} \cdot \frac{\exp((a-x)/b)}{(1+\exp((a-x)/b))^2}$	$\mu = a$ $\sigma^2 = (\pi b)^2/6$	Note 7

NOTES

1. The moment generating function is $(e^{bt} - e^{at})/(b-a)t$, defined for all t.
2. The moment generating function is $1/(1 - \lambda t)$, defined for $t < 1/\lambda$.
3. The moment generating function does not exist. The mean exists for $b > 1$, the variance exists for $b > 2$.
4. For $a > 0$, $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$ is the gamma function. If a is an integer, $\Gamma(a) = (a-1)!$.
5. For the characteristic function, see C. R. Rao, Linear Statistical Inference, Wiley, 1973, p. 151.
6. The moment generating function is $e^{at}\Gamma(1 - tb)$ for $t < 1/b$.
7. The moment generating function is $e^{at}\pi bt/\sin(\pi bt)$ for $\lvert t\rvert < 1/2b$.

TABLE 3.3. THE NORMAL DISTRIBUTION AND ITS RELATIVES

NAME & DOMAIN	DENSITY	MOMENTS	CHAR. FN.
1. Normal $n(x-\mu,\sigma)$ $-\infty < x < +\infty$, $\sigma > 0$	$(2\pi\sigma^2)^{-1/2}\cdot\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$	μ = mean σ^2 = variance	$\exp(\imath t\mu-\sigma^2t^2/2)$ Note 1
2. Standard Normal $-\infty < x < +\infty$	$\varphi(x) = (2\pi)^{-1/2}\cdot\exp(-x^2/2)$	$\mu = 0$ $\sigma^2 = 1$	$\exp(-t^2/2)$
3. Chi-Square $0 < x < +\infty$	$\chi^2(x;k) = \frac{x^{(k/2)-1}\cdot e^{-x/2}}{\Gamma(k/2)2^{k/2}}$	$\mu = k$ $\sigma^2 = 2k$ k = 1,2,...	$(1-2\imath t)^{-k/2}$ Note 2
4. F-distribution $0 < x < +\infty$	F(x;k,n) k,n positive integers	μ = if n > 2 $\sigma^2 = \frac{2n^2(k+n-2)}{k(n-2)^2(n-4)}$ if n > 4	Note 3
5. t-distribution $-\infty < x < +\infty$	$\frac{\Gamma(\frac{k+1}{2})(1+x^2/k)^{-(k+1)/2}}{\sqrt{k}\ \Gamma(\frac{1}{2})\Gamma(\frac{1+2k}{2})}$	$\mu = 0$ if k > 1 σ^2 = k/(k-2) if k > 2	Note 4
1. Noncentral Chi-Squared $x > 0$	$\chi^2(x;k,\delta)$ k pos. integer $\delta \geq 0$	$\mu = k+\delta$ $\sigma^2 = 2(k+2\delta)$	Note 5
2. Noncentral F-distribution $x > 0$	F(x;k,n,δ) k,n positive integers $\delta \geq 0$	if n > 2, μ = n(k+δ)/k(n-2) if n > 4, σ^2 = $\frac{2(n/k)^2(k+\delta)^2+(k+2\delta)(n-2)}{(n-2)^2(n-4)}$	Note 6
3. Noncentral t-distribution	t(x;k,λ) k pos. integer	$\mu = \frac{\Gamma((k-1)/2)\lambda}{\Gamma(k/2)}$ if k > 1 $\sigma^2 = (1+\lambda^2)k/(k-2) - \mu^2$ if k > 2	Note 7

NOTES TO TABLE 3.3

1. The density is often denoted $n(x\text{-}\mu,\sigma^2)$, and the cumulative distribution referred to as $N(x\text{-}\mu,\sigma^2)$, or simply $N(\mu,\sigma^2)$. The moment generating function is $\exp(\mu t+\sigma^2t^2/2)$, defined for all t. The standard normal density is often denoted $\varphi(x)$, and the standard normal CDF is denoted $\Phi(x)$. The general normal and standard normal formulas are related by $n(x\text{-}\mu,\sigma^2) = \varphi((x\text{-}\mu)/\sigma)/\sigma$ and $N(x\text{-}\mu,\sigma^2) = \Phi((x\text{-}\mu)/\sigma)$.

2. The moment generating function is $(1\text{-}t/2)^{-k/2}$ for $t < 2$. The Chi-Square distribution with parameter k (≡ degrees of freedom) is the distribution of the sum of squares of k independent standard normal random variables. The Chi-Square density is the same as the gamma density with $b = 2$ and $a = k/2$.

3. The F-distribution is the distribution of the expression nU/kV, where U is a random variable with a Chi-square distribution with parameter k, and V is an independent random variable with a Chi-square distribution with parameter n.

The density is $\dfrac{\Gamma(\frac{k+n}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{n}{2})}\cdot\dfrac{k^{k/2}n^{n/2}x^{k/2-1}}{(n+kx)^{(k+n)/2}}$. For $n \le 2$, the mean does not exist, and for $n \le 4$, the variance does not exist.

The characteristic and moment generating functions are complicated.

4. If Y is standard normal and Z is independently Chi-squared distributed with parameter k, then $Y/\sqrt{Z/k}$ has a T-Distribution with parameter k (≡ degrees of freedom). The characteristic function is complicated; the moment generating function does not exist.

5. The Noncentral Chi-square is the distribution of the sum of squares of k independent normal random variables, each with variance one, and with means whose squares sum to δ. The Noncentral Chi-Square density is a Poisson mixture of (central) Chi-square densities, $\sum_{j=0}^{\infty}\ [e^{-\delta/2}(\delta/2)^j/j!]\chi^2(x;k+2j)$.

6. The Non-central F-distribution has a density that is a Poisson mixture of rescaled (central) F-distributed densities, $\sum_{j=0}^{\infty}\ [e^{-\delta/2}(\delta/2)^j/j!]\ \dfrac{k}{k+2j}\ F(\ \dfrac{kx}{k+2j}\ ;k+2j,n)$. It is the distribution of the expression nU'/kV, where U' is a Noncentral Chi-Squared random variable with parameters k and δ, and V is an independent central Chi-Squared distribution with parameter n.

7. If Y is standard normal and Z is independently Chi-squared distributed with parameter k, then $(Y+\lambda)/\sqrt{(Z/k)}$ has a Noncentral T-Distribution with parameters k and λ. The density is a Poisson mixture of scaled Beta distributed densities,

$$\sum_{j=0}^{\infty}\ \left[\ e^{-\lambda^2/2}\ (\lambda^2/2)^j/j!\right]\ \frac{xk}{(k+x^2)^2}\ B(\ \frac{k}{k+x^2},\frac{k}{2},\frac{1+2j}{2}\ ).$$

The square of a Noncentral T-Distributed random variable has a Noncentral F-Distribution with parameters 1, k, and $\delta = \lambda^2$.

3.9 EXERCISES

1. In Example 1, write out all the members of $\boldsymbol{F}$.

2. Prove that a σ-field of events contains *countable* intersections of its members.

3. Example 2 claims that the class of all subsets of countable **S** has greater cardinality than **S** itself. Mathematically, this means that it is not possible to associate a unique element of **S** with each member of the class. Use the following device to convince yourself this is true: Write each number in the unit interval as a fraction in binary notation, $0.b_1b_2....$ Associate with each number the class member that contains the sequence with j heads if and only if bj = 1. Then, the real numbers, which are uncountable, map into unique members of the class, so the class is also uncountable.

4. In Example 4, show that the event "the change is the same on successive days" is not in $\boldsymbol{B}_1\times\boldsymbol{B}_2$, but is a monotone limit of sets in $\boldsymbol{B}_1\times\boldsymbol{B}_2$.

5. Economic agents can make contingent trades only if it is common knowledge if the contingency is realized. In Example 1, Agent 1 knows $\boldsymbol{F}$, Agent 2 knows $\boldsymbol{G}$, Agent 3 knows $\boldsymbol{H}$ = {∅,{HH,TT},{HT,TH},**S**}. What is the common knowledge of Agents 1 and 2? Of Agents 1 and 3?

6. Suppose, in Example 1, that instead of H and T being equally likely, the probability measure satisfies

HH	HT	TH	TT
0.2	0.4	0.1	0.3

What is the probability that the first coin is heads? That the second coin is heads? That the two coins give the same result?

7. Consider the sequence of functions $f_n(x) = x^{1/n}$ for $0 < x < 1$. These are square integrable. Do they converge to a limit, and if so, what is the convergence strong, in measure, or weak?

8. Consider the probability measure P([0,x]) = x1/2 on $0 \le x \le 1$. Does it meet the Radon-Nikodym conditions for the existence of a probability density?

9. It is known that 0.2 percent of the population is HIV-positive. It is known that a screening test for HIV has a 10 percent chance of incorrectly showing positive when the subject is negative, and a 2 percent chance of incorrectly showing negative when the subject is positive. What proportion of the population that tests positive has HIV?

10. John and Kate are 80 years old. The probability that John will die in the next year is 0.08, and the probability that Kate will die in the next year is 0.05. The probability that John will die, given that Kate dies, is 0.2. What is the probability that both will die? That at least one will die? That Kate will die, given that John dies?

11. The probability that a driver will have an accident next year if she has a Ph.D. is 0.2. The probability she will have an accident if she does not have a Ph.D. is 0.25. The probability the driver has a Ph.D. and an accident is 0.01. What is the probability the driver has a Ph.D.? What is the probability of a Ph.D. given an accident?

12. A quiz show offers you the opportunity to become a millionaire if you answer nine questions correctly. Questions can be easy (E), moderate (M), or hard (H). The respective probabilities that you will answer an E, M, or H question correctly are 2/3, 1/2, and 1/3. If you get an E question, your next question will be E, M, or H with probabilities 1/4, 1/2, and 1/4 respectively. If you get a M question, your next question will be E, M, or H with probabilities 1/3, 1/3, and 1/3 respectively. If you get a H question, your next question will be E, M, or H with probabilities 1/2, 0, and 1/2 respectively. The first question is always an E question. What is the probability that you will become a millionaire? [Hint: Show that the probability of winning if you reach question 9 is independent of whether this question is E, M, or H. Then use backward recursion.]

13. Show that if $\mathbf{A} \subseteq \mathbf{B}$ and $P(\mathbf{A}) > 0$, then $P(\mathbf{C}|\mathbf{A})$ can be either larger or smaller than $P(\mathbf{C}|\mathbf{B})$.

14. An airplane has 100 seats. The probability that a ticketed passenger shows up for the flight is 0.9, and the events that any two different passengers show up is statistically independent. If the airline sells 105 seats, what is the probability that the plane will be overbooked? How many seats can the airline sell, and keep the probability of overbooking to 5 percent or less?

15. Prove that the expectation $\mathbf{E}(X - c)^2$ is minimized when $c = \mathbf{E}X$.

16. Prove that the expectation $\mathbf{E}|X - c|$ is minimized when c = median(X).

17. What value of c minimizes $\mathbf{E}\{\alpha\cdot\max(X-c,0) + (1-\alpha)\cdot\max(c-X,0)\}$? (Hint: describe the solution in terms of the distribution F of X.]

18. A sealed bid auction has a tract of land for sale to the highest of n bidders. You are bidder 1. Your experience is that the bids of each other bidder is distributed with a Power distribution $F(X) = X^{\alpha}$ for $0 \le X \le 1$. Your profit if you are successful in buying tract at price y is 1 - y. What should you bid to maximize your expected profit? What is your probability of winning the auction?

19. A random variable X has a normal distribution if its density is $f(x) = (2\pi\sigma^2)^{-½}\cdot\exp(-(x-\mu)^2/2\sigma^2)$, where μ and σ^2 are parameters. Prove that X has mean μ and variance σ^2. Prove that $\mathbf{E}(X-\mu)^3 = 0$ and $\mathbf{E}(X-\mu)^4 = 3\sigma^4$. [Hint: First show that $\int x\cdot\exp(-x^2/2)dx = -\exp(-x^2/2)$ and for $k > 1$, the integration by parts formula $\int x^k\cdot\exp(-x^2/2)dx = -x^{k-1}\cdot\exp(-x^2/2) + (k-1)\int x^{k-2}\cdot\exp(-x^2/2)dx$.]

20. Suppose the stock market has two regimes, Up and Down. In an Up regime, the probability that the market index will rise on any given day is P. In a Down regime, the probability that the market index will rise on any given day is Q, with $Q < P$. Within a regime, the probability that the market rises on a given day is independent of its history. The probability of being in a Up regime is 1/2, so that if you do not know which regime you are in, then all you can say is that the probability that the market will rise on any given day is R = (P+Q)/2. Assume that regimes persist far longer than runs of rises, so that when analyzing runs the regime can be treated as persisting indefinitely. Show that when you are in the Up regime, the probability of a run of k or more successive days in which the market rises is P^{k-1}, and that the probability of a run of exactly k days in which the market rises is $P^{k-1}(1-P)$. A similar formula with Q instead of P holds when you are in a Down regime. Show that expected length in an Up regime of a run of rises is 1/(1-P). Show that $½/(1-P) + ½/(1-Q) \ge 1/(1-R)$.

21. The random vector (X_1,X_2) has the distribution function $\exp(-(\exp(-2x_1)+\exp(-2x_2))^{1/2})$. What is the marginal distribution of X_1? What is the conditional distribution of X_1 given $X_2 \le c$? Given $X_2 = c$?

22. The expectation $E(X+aZ)^2 \geq 0$ for random variables X, Z and any scalar a.. Use this property to prove the Cauchy-Schwartz inequality.

23. Prove Jensen's inequality for a probability concentrated at two points.

24. In Example 2, use the law of iterated expectations to calculate the expectation of the number of heads, given that the number exceeds one.

25. If X and Z are bivariate normal with means 1 and 2, and variances 1 and 4, respectively, and covariance , what is the density of X given Z = z? Use Bayes law to deduce the conditional density of Z given X = x?

26. Prove the formula for the characteristic function of a standard normal random variable.

27. What is the domain of the moment generating function of an exponentially distributed random variable with density $f(x) = \exp(-3x)$ for $x > 0$?

28. If (X,Z) is a random vector with density f(x,z) and $z > 0$, and $S = X/Z$, $T = Z$, what is the Jacobean of the transformation?

29. If X and Y are multivariate normal with zero means, $EXX' = A$, $EYY' = B$, and $EXY' = C$, show that X and $Z = Y - XB^{-1}C$ are independent.

30. For the binomial distribution b(k;n,p), what is the variance of the frequency f = k/n?

31. The hypergeometric distribution describes the probability that k of n balls drawn from an urn will be red, where the urn contains r red and w white balls, and sampling is without replacement. Calculate the same probability if sampling is with replacement. Calculate the probabilities, with and without replacement, when r = 10, w = 90, n = 5, k = 1.

32. In a Poisson distribution, what is the expected count conditioned on the count being positive?

33. Under what conditions is the characteristic function of a uniform distribution of [-a,b] real?

34. Show that if X and Y are independent identically distributed extreme value, then X - Y is logistic distributed.

35. Suppose that the duration of a spell of unemployment (in days) can be described by a geometric distribution, $Prob(k) = p^k(1-p)$, where $0 < p < 1$ is a parameter and k is a non-negative integer. What is the expected duration of unemployment? What is the probability of a spell of unemployment lasting longer than K days? What is the conditional expectation of the duration of unemployment, given the event that K > m, where m is a positive integer? [Hint: Use formulas for geometric series, see 2.1.10.]

36. Use the moment generating function to find $\mathbf{E}X^3$ when X has density $e^{-x/\lambda}/\lambda$, $x > 0$.

37. A log normal random variable Y is one that has log(Y) normal. If log(Y) has mean μ and variance σ^2, find the mean and variance of Y. [Hint: It is useful to find the moment generating function of Z = log(Y).]

38. If X and Y are independent normal, then X+Y is again normal, so that one can say that *the normal family is closed under addition*. (Addition of random variables is also called convolution, from the formula for the density of the sum.) Now suppose X and Y are independent and have extreme value distributions, Prob($X \leq x$) = $\exp(-e^{a-x})$ and Prob($Y \leq y$) = $\exp(-e^{b-y})$, where a and b are location parameters. Show that max(X,Y) once again has an extreme value distribution (with location parameter $c = \log(e^a+e^b)$), so that *the extreme value family is closed under maximization*.

39. If X is standard normal, derive the density and characteristic function of $Y = X^2$, and confirm that this is the same as the tabled density of a chi-square random variable with one degree of freedom. If X is normal with variance one and a mean μ that is not zero, derive the density of Y, which is non-central chi-square distributed with one degree of freedom and noncentrality parameter μ^2.

40. Random Variables X and Y are bivariate normal, with **E**X = 1, **E**Y = 3, and Var(X) = 4, Var(Y) = 9, Covariance(X,Y) = 5.
(a) What is the mean of Z = 2X - Y?
(b) What is the variance of Z = 2X - Y?
© What is the conditional mean of Z given X = 5?
(d) What is the conditional variance of Z given X = 5?

41. What is the probability that the larger of two random observations from any continuous distribution will exceed the population median?

42. If random variables X and Y are independent, with EX = 1, EY = 2, $EX^2 = 4$, $EY^2 = 9$, what is the unconditional mean and variance of $3 \cdot X \cdot Y$? What is the conditional mean and variance of $3 \cdot X \cdot Y$ given Y = 5?

43. Jobs are characterized by a wage rate W and a duration of employment X, and (W,X) can be interpreted as a random vector. The duration of employment has an exponential density $\lambda \cdot e^{-\lambda x}$, and the wage rate W has a exponential density, conditioned on X = x, equal to $(\alpha+\beta x)e^{-(\alpha+\beta x)w}$, where λ, α, and β are positive parameters. What is the marginal density of W? The conditional density of X given W?

44. Random Variables X and Y are bivariate normal, with **E**X = 1, **E**Y = 3, and Var(X) = 4, Var(Y) = 9, Covariance(X,Y) = 5.
(a) What is the mean of Z = 2X - Y?
(b) What is the variance of Z = 2X - Y?
© What is the conditional mean of Z given X = 5?
(d) What is the conditional variance of Z given X = 5?

45. The data set nyse.txt in the class data area of the class home page contains daily observations on stock market returns from Jan. 2, 1968 through Dec. 31, 1998, a total of 7806 observations corresponding to days the market was open. There are four variables, in columns delimited by spaces. The first variable (DAT) is the date in *yymmdd* format, the second variable (RNYSE) is the daily return to the NYSE market index, defined as the log of the ratio of the closing value of the index today to the closing index on the previous day the market was open, with distributions (dividends) factored in. The third variable (SP500) is the S&P500 market index, an index of a majority of the high market value stocks in the New York stock exchange. The fourth variable (RTB90) is the rate of interest in the secondary market for 90-day Treasury Bills, converted to a daily rate commensurate with RNYSE..

a. Let $\mathbf{E}_n$ denote a sample average (empirical expectation). Find the sample mean $\mu = \mathbf{E}_n X$, variance $\sigma^2 = \mathbf{E}_n(X - \mu)$, skewness $\mathbf{E}_n(X - \mu)^3/\sigma^3$, and kurtosis $\mathbf{E}_n(X - \mu)^4/\sigma^4 - 3$, for the variables RNYSE and RTB90. Normally distributed

random variables have zero skewness and kurtosis in the population. Making an "eyeball" comparison, do the sample moments appear to be consistent with the proposition that RNYSE and RTB90 are normally distributed?

b. For RNYSE, form the *standardized* variable $Z = (RNYSE - \mu)/\sigma$, by subtracting this variable's sample mean and then dividing by the square root of its variance (or standard deviation). Sort the values of Z from low to high, and then construct a new variable Y that equals i/7806 for $1 \leq i \leq 7806$. The values of Z are called the *order statistics* of the sample, and Y is the empirical CDF, a CDF that puts 1/7806 probability at each observed value of RNYSE. Plot Y against $\Phi(Z)$, where Φ is the standard normal CDF. If RNYSE is normal, then these curves will differ only because of sampling noise in Y. Does it appear by eyeball comparison that they are likely to be the same? A particular issue is the theoretical question of whether the distribution of returns has fat tails, so that the variance and higher moments are hard to estimate precisely or may fail to exist. In a normal sample, one would expect that on average 99 percent of standardized observations are less than 2.575 in magnitude. Do the standardized values Z appear to be consistent with this frequency?

c. A claim in the analysis of stock market returns is that the introduction of financial derivatives and index funds through the 1980's made it easier for arbitragers to close windows of profit opportunity. The argument is made that the resulting actions of arbitragers have made the market more volatile. Compare the subsamples of NYSE excess returns (EXCESS = RNYSE - NRTB90) for the periods 1968-1978 and 1988-1998. By eyeball comparison, were there differences in mean excess return in the two decades? In the variance (or standard deviation) of excess return? Now do a 2×2 table of sample means classified by the two decades above and by whether or not the previous day's excess return was above its decade average. Does it appear that the gap between mean excess returns on days following previous rises and falls has increased or shrunk in the decade of the 90's?

CHAPTER 4. LIMIT THEOREMS IN STATISTICS

4.1. SEQUENCES OF RANDOM VARIABLES

4.1.1. A great deal of econometrics uses relatively large data sets and methods of statistical inference that are justified by their desirable properties in large samples. The probabilistic foundations for these arguments are "laws of large numbers", sometimes called the "law of averages", and "central limit theorems". This chapter presents these foundations. It concentrates on the simplest versions of these results, but goes some way in covering more complicated versions that are needed for some econometric applications. For basic econometrics, the most critical materials are the limit concepts and their relationship covered in this section, and for independent and identically distributed (i.i.d.) random variables the first Weak Law of Large Numbers in Section 4.3 and the first Central Limit Theorem in Section 4.4. The reader may want to postpone other topics, and return to them as they are needed in later chapters.

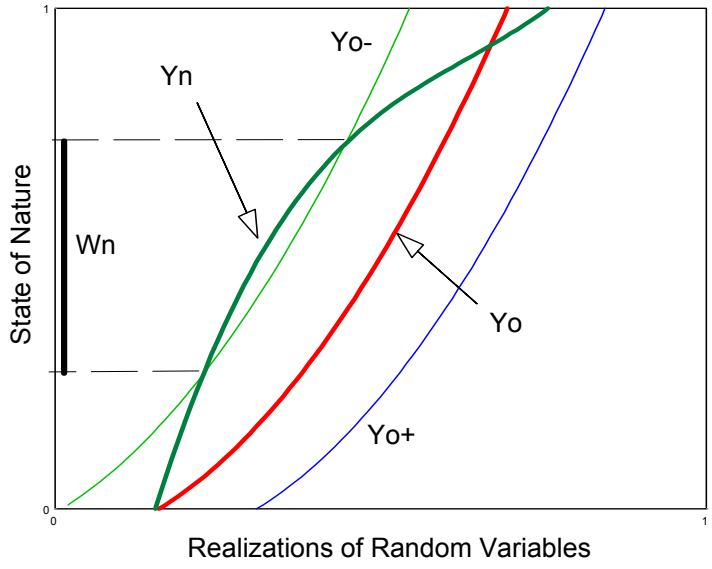
4.1.2. Consider a sequence of random variables $Y_1,Y_2,Y_3,...$. These random variables are all functions $Y_k(s)$ of the same state of Nature s, but may depend on different parts of s. There are several possible concepts for the limit Y_o of a sequence of random variables Y_n. Since the Y_n are functions of states of nature, these limit concepts will correspond to different ways of defining limits of functions. Figure 4.1 will be used to discuss limit concepts. Panel (a) graphs Y_n and Y_o as functions of the state of Nature. Also graphed are curves denoted $Y_{o\pm}$ and defined by $Y_o \pm \varepsilon$ which for each state of Nature s delineate an ε-neighborhood of $Y_o(s)$. The set of states of Nature for which $|Y_o(s) - Y_n(s)| > \varepsilon$ is denoted $\mathbf{W}_n$. Panel (b) graphs the CDF's of Y_o and Y_n. For technical completeness, note that a random variable Y is a measurable real-valued function on a probability space (S,***F***,P), where ***F*** is a σ-field of subsets of **S**, P is a probability on ***F***, and "measurable" means that F contains the inverse image of every set in the Borel σ-field of subsets of the real line. The CDF of a vector of random variables is then a measurable function with the properties given in 3.5.3.

4.1.3. Y_n *converges in probability* to Y_o, if for each $\varepsilon > 0$, $\lim_{n\to\infty} \text{Prob}(|Y_n - Y_o| > \varepsilon) = 0$. Convergence in probability is denoted $Y_n \to_p Y_o$, or $\text{plim}_{n\to\infty} Y_n = Y_o$. With $\mathbf{W}_n$ defined as in Figure 4.1, $Y_n \to_p Y_o$ iff $\lim_{n\to\infty} \text{Prob}(\mathbf{W}_n) = 0$ for each $\varepsilon > 0$.

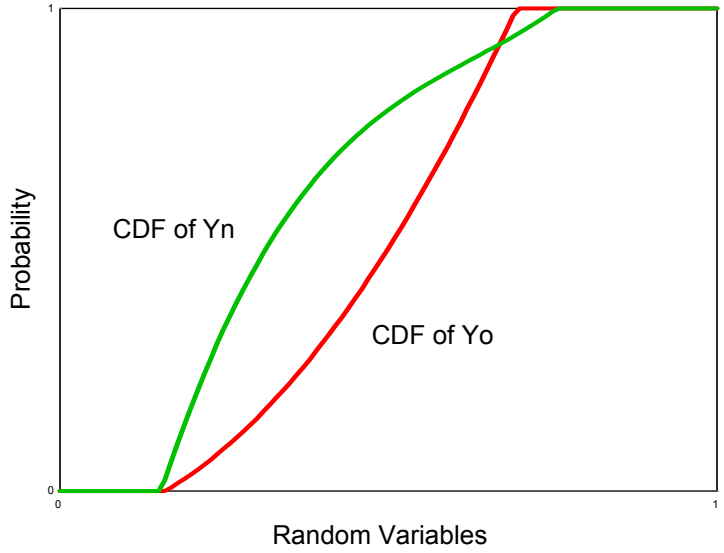
4.1.4. Y_n *converges almost surely* to Y_o, denoted $Y_n \to_{as} Y_o$, if for each $\varepsilon > 0$, $\lim_{n\to\infty} \text{Prob}(\sup_{m\geq n}|Y_m - Y_o| > \varepsilon) = 0$. For $\mathbf{W}_n$ defined in Figure 4.1, the set of states of nature for which $|Y_m(w) - Y_o(w)| > \varepsilon$ for some $m \geq n$ is $\bigcup_{m\geq n} \mathbf{W}_m$, and $Y_n \to_{as} Y_o$ iff $\text{Prob}(\bigcup_{m\geq n} \mathbf{W}_{n'}) \to 0$. An implication of almost sure convergence is $\lim_{n\to\infty} Y_n(s) = Y_o(s)$ a.s. (i.e., except for a set of states of Nature of probability zero); this is not an implication of $Y_n \to_p Y_o$.

FIGURE 4.1. CONVERGENCE CONCEPTS FOR RANDOM VARIABLES

Panel (a)



Panel (b)



4.1.5. Y_n *converges in* ρ*-mean* (also called *convergence in* $\|\cdot\|_\rho$ *norm,* or *convergence in* $\mathbf{L}_\rho$ *space*) to Y_o if $\lim_{n\to\infty} \mathbf{E}|Y_n - Y_o|^\rho = 0$. For ρ = 2, this is called *convergence in quadratic mean*. The norm is defined as $\|Y\|_\rho = [\int_S |Y(s)|^\rho \cdot P(ds)]^{1/\rho} = [\mathbf{E}|Y|^\rho]^{1/\rho}$, and can be interpreted as a probability-weighted measure of the distance of Y from zero. The norm of a random variable is a moment. There are random variables for which the ρ-mean will not exist for any ρ > 0; for example, Y with CDF F(y) = 1 - 1/(log y) for y ≥ *e* has this property. However, in many applications moments such as variances exist, and the quadratic mean is a useful measure of distance.

4.1.6. Y_n *converges in distribution* to Y_o, denoted $Y_n \to_d Y_o$, if the CDF of Y_n converges to the CDF of Y_o at each continuity point of Y_o. In Figure 4.1(b), this means that F_n converges to the function F_o point by point for each argument on the horizontal axis, except possibly for points where F_o jumps. (Recall that distribution functions are always continuous from the right, and except at jumps are continuous from the left. Since each jump contains a distinct rational number and the rationals are countable, there are at most a countable number of jumps. Then the set of jump points has Lebesgue measure zero, and there are continuity points arbitrarily close to any jump point. Because of right-continuity, distribution functions are uniquely determined by their values at their continuity points.) If **A** is an open set, then $Y_n \to_d Y_o$ implies $\liminf_{n\to\infty} F_n(\mathbf{A}) \geq F_o(\mathbf{A})$; conversely, **A** closed implies $\limsup_{n\to\infty} F_n(\mathbf{A}) \leq F_o(\mathbf{A})$ see P. Billingsley (1968), Theorem 2.1. Convergence in distribution is also called *weak convergence* in the space of distribution functions.

4.1.7. The relationships between different types of convergence are summarized in Figure 4.2. In this table, "A $\models\Rightarrow$ B" means that A implies B, but not vice versa, and "A $\Leftarrow\Rightarrow$B" means that A and B are equivalent. Explanations and examples are given in Sections 4.1.8-4.1.18. On first reading, skim these sections and skip the proofs.

4.1.8. $Y_n \to_{as} Y_o$ implies $\text{Prob}(\mathbf{W}_n) \leq \text{Prob}(\bigcup_{m\geq n} \mathbf{W}_m) \to 0$, and hence $Y_n \to_p Y_o$. However, $\text{Prob}(\mathbf{W}_n) \to 0$ does not necessarily imply that the probability of $\bigcup_{m\geq n} \mathbf{W}_m$ is small, so $Y_n \to_p Y_o$ does not imply $Y_n \to_{as} Y_o$. For example, take the universe of states of nature to be the points on the unit circle with uniform probability, take the $\mathbf{W}_n$ to be successive arcs of length $2\pi/n$, and take Y_n to be 1 on $\mathbf{W}_n$, 0 otherwise. Then $Y_n \to_p 0$ since $\Pr(Y_n \neq 0) = 1/n$, but Y_n fails to converge almost surely to zero since the successive arcs wrap around the circle an infinite number of times, and every s in the circle is in an infinite number of $\mathbf{W}_n$.

4.1.9. Suppose $Y_n \to_p Y_o$. It is a good exercise in manipulation of probabilities of events to show that $Y_n \to_d Y_o$. Given ε > 0, define $\mathbf{W}_n$ as before to be the set of states of Nature where $|Y_n(s) - Y_o(s)| > \varepsilon$. Given y, define $\mathbf{A}_n$, $\mathbf{B}_o$, and $\mathbf{C}_o$ to be, respectively, the states of Nature with $Y_n \leq y$, $Y_o \leq y - \varepsilon$,

and $Y_o \le y + \varepsilon$. Then $\mathbf{B}_o \subseteq \mathbf{A}_n \cup \mathbf{W}_n$ (i.e., $Y_o(s) \le y - \varepsilon$ implies either $Y_n(s) \le y$ or $|Y_o(s) - Y_n(s)| > \varepsilon$) and $\mathbf{A}_n \subseteq \mathbf{C}_o \cup \mathbf{W}_n$ (i.e., $Y_n(s) \le y$ implies $Y_o(s) \le y + \varepsilon$ or $|Y_o(s) - Y_n(s)| > \varepsilon$). Hence, for n large enough so $\text{Prob}(\mathbf{W}_n) < \varepsilon$, $F_o(y-\varepsilon) \equiv \text{Prob}(\mathbf{B}_o) \le \text{Prob}(\mathbf{A}_n) + \text{Prob}(\mathbf{W}_n) < F_n(y) + \varepsilon$, and $F_n(y) \equiv \text{Prob}(\mathbf{A}_n) \le \text{Prob}(\mathbf{C}_o) + \text{Prob}(\mathbf{W}_n) < F_o(Y+\varepsilon) + \varepsilon$, implying $F_o(y-\varepsilon) - \varepsilon \le \lim_{n\to\infty} F_n(y) \le F_o(y+\varepsilon) + \varepsilon$. If y is a continuity point of Y_o, then $F_o(y-\varepsilon)$ and $F_o(y+\varepsilon)$ approach $F_o(y)$ as $\varepsilon \to 0$, implying $\lim_{n\to\infty} F_n(y) = F_o(y)$. This establishes that $Y_n \to_d Y_o$.

Convergence in distribution of Y_n to Y_o does not imply that Y_n and Y_o are close to each other. For example, if Y_n and Y_o are i.i.d. standard normal, then $Y_n \to_d Y_o$ trivially, but clearly not $Y_n \to_p Y_o$ since $Y_n - Y_o$ is normal with variance 2, and $|Y_n - Y_o| > \varepsilon$ with a positive, constant probability. However, there is a useful representation that is helpful in relating convergence in distribution and almost sure convergence; see P. Billingsley (1986), p.343.

Theorem 4.1. (Skorokhod) If $Y_n \to_d Y_o$, then there exist random variables Y_n' and Y_o' such that Y_n and Y_n' have the same CDF, as do Y_o and Y_o', and $Y_n' \to_{as} Y_o'$.

4.1.10. Convergence in distribution and convergence in probability to a constant are equivalent. If $Y_n \to_p c$ constant, then $Y_n \to_d c$ as a special case of 4.1.9 above. Conversely, $Y_n \to_d c$ constant means $F_n(y) \to F_o(y)$ at continuity points, where $F_c(y) = 0$ for $y < c$ and $F_c(y) = 1$ for $y \ge c$. Hence $\varepsilon > 0$ implies $\text{Prob}(|Y_n - c| > \varepsilon) = F_n(c-\varepsilon) + 1 - F_n(c+\varepsilon) \to 0$, so $Y_n \to_p c$. This result implies particularly that the statements $Y_n - Y_o \to_p 0$ and $Y_n - Y_o \to_d 0$ are equivalent. Then, $Y_n - Y_o \to_d 0$ implies $Y_n \to_d Y_o$, but the reverse implication does not hold.

4.1.11. The condition that convergence in distribution is equivalent to convergence of expectations of all bounded continuous functions is a fundamental mathematical result called the *Helly-Bray theorem*. Intuitively, the reason the theorem holds is that bounded continuous functions can be approximated closely by sums of continuous "almost-step" functions, and the expectations of "almost step" functions closely approximate points of CDF's. A proof by J. Davidson (1994), p. 352, employs the Skorokhod representation theorem 4.1.

4.1.12. A Chebyshev-like inequality is obtained by noting for a random variable Z with density f(z) that $\mathbf{E}|Z|^\rho = \int |z|^\rho f(z)dz \ge \int_{|z|\ge\varepsilon} \varepsilon^\rho f(z)dz = \varepsilon^\rho \text{Prob}(|Z| > \varepsilon)$, or $\text{Prob}(|Z| > \varepsilon) \le \mathbf{E}|Z|^\rho/\varepsilon^\rho$.

(When $\rho = 2$, this is the conventional Chebyshev inequality. When $\rho = 1$, one has $\text{Prob}(|Z| > \varepsilon) \le \mathbf{E}|Z|/\varepsilon$.) Taking $Z = Y_n - Y_o$, one has $\lim_{n\to\infty} \text{Prob}(|Y_n - Y_o| > \varepsilon) \le \varepsilon^{-\rho} \cdot \lim_{n\to\infty} \mathbf{E}|Y_n - Y_o|^\rho$. Hence, convergence in ρ-mean (for any $\rho > 0$) implies convergence in probability. However, convergence almost surely or in probability does not necessarily imply convergence in ρ-mean. Suppose the sample space is the unit interval with uniform probability, and $Y_n(s) = e^{n'}$ for $s \le n^{-2}$, zero otherwise. Then $Y_n \to_{as} 0$ since $\text{Prob}(Y_m \ne 0 \text{ for any } m > n) \le n^{-2}$, but $\mathbf{E}|Y_n|^\rho = e^{\rho n}/n^2 \to +\infty$ for any $\rho > 0$.

FIGURE 4.2. RELATIONS BETWEEN STOCHASTIC LIMITS

(Section numbers for details are given in parentheses)

1	$Y_n \to_{as} Y_o$ $\overset{(1.8)}{\Longrightarrow}$ $Y_n \to_p Y_o$ $\overset{(1.9)}{\Longrightarrow}$ $Y_n \to_d Y_o$
	$\Uparrow$ (1.4) $\Downarrow$ $\quad$ $\Uparrow$ (1.3) $\Downarrow$ $\quad$ $\Uparrow$ (1.10) $=$
2	$Y_n - Y_o \to_{as} 0$ $\overset{(1.8)}{\Longrightarrow}$ $Y_n - Y_o \to_p 0$ $\overset{(1.10)}{\Longleftrightarrow}$ $Y_n - Y_o \to_d 0$
3	$Y_n \to_d c$ (a constant) $\Longleftrightarrow Y_n \to_p c$ (1.10)
4	$Y_n \to_d Y_o \Longleftrightarrow \mathbf{E}g(Y_n) \to \mathbf{E}g(Y_o)$ for all bounded continuous g (1.11)
5	$\|Y_n - Y_o\|_\rho \to 0$ for some $\rho > 0 \Longrightarrow Y_n \to_p Y_o$ (1.12)
6	$\|Y_n - Y_o\|_\rho \le M$ (all n) & $Y_n \to_p Y_o \Longrightarrow \|Y_n - Y_o\|_\lambda \to 0$ for $0 < \lambda < \rho$ (1.13)
7	$Y_n \to_p Y_o \Longrightarrow Y_{n_k} \to_{as} Y_o$ for some subsequence n_k, k = 1,2,... (1.14)
8	$\sum_{n=1}^{\infty} P(\|Y_n - Y_o\| > \varepsilon) < +\infty$ for each $\varepsilon > 0 \Longrightarrow Y_n \to_{as} Y_o$ (1.15)
9	$\sum_{n=1}^{\infty} \mathbf{E}\|Y_n - Y_o\|^\rho < +\infty$ (for some $\rho > 0$) $\Longrightarrow Y_n \to_{as} Y_o$ (1.15)
10	$Y_n \to_d Y_o$ & $Z_n - Y_n \to_p 0 \Longrightarrow Z_n \to_d Y_o$ (1.16)
11	$Y_n \to_p Y_o \Longrightarrow g(Y_n) \to_p g(Y_o)$ for all continuous g (1.17)
12	$Y_n \to_d Y_o \Longrightarrow g(Y_n) \to_d g(Y_o)$ for all continuous g (1.18)

4.1.13. Adding a condition of a uniformly bounded ρ-order mean $\mathbf{E}|Y_n|^\rho \le M$ to convergence in probability $Y_n \to_p Y_o$ yields the result that $\mathbf{E}|Y_o|^\lambda$ exists for $0 < \lambda \le \rho$, and $\mathbf{E}|Y_n|^\lambda \to \mathbf{E}|Y_o|^\lambda$ for $0 < \lambda < \rho$. This result can be restated as "the moments of the limit equal the limit of the moments" for moments of order λ less than ρ. Replacing Y_n by $Y_n - Y_o$ and Y_o by 0 gives the result in Figure 4.2.

To prove these results, we will find useful the property of moments that $\mathbf{E}|Y|^\lambda \le (\mathbf{E}|Y|^\rho)^{\lambda/\rho}$ for $0 < \lambda < \rho$. (This follows from Holder′s inequality (2.1.11), which states $\mathbf{E}|UV| \le (\mathbf{E}|U|^r)^{1/r}(\mathbf{E}|V|^s)^{1/s}$ for $r,s > 0$ and $r^{-1} + s^{-1} = 1$, by taking $U = |Y|^\lambda$, $V = 1$, and $r = \rho/\lambda$.) An immediate implication is $\mathbf{E}|Y_n|^\lambda \le M^{\lambda/\rho}$. Define $g(y,\lambda,k) = \min(|y|^\lambda, k^\lambda)$, and note that since it is continuous and bounded, the Healy-Bray theorem implies $\mathbf{E}g(Y_n,\lambda,k) \to \mathbf{E}g(Y_o,\lambda,k)$. Therefore,

$$M^{\lambda/\rho} \ge \mathbf{E}|Y_n|^\lambda \ge \mathbf{E}g(Y_n,\lambda,k) = \int_{-k}^{k} |y|^\lambda f_n(y)dy + k^\lambda \cdot \text{Prob}(|Y_n| > k)$$

$$\to \int_{-k}^{k} |y|^\lambda f_o(y)dy + k^\lambda \text{Prob}(|Y_o| > k).$$

Letting $k \to \infty$ establishes that $\mathbf{E}|Y_o|^\lambda$ exists for $0 < \lambda \le \rho$. Further, for $\lambda < \rho$,

$$0 \le \mathbf{E}|Y_n|^\lambda - \mathbf{E}g(Y_n,\lambda,k) \le \int_{|y|>k} |y|^\lambda f_n(y)dy \le k^{\lambda-\rho} \int_{|y|>k} |y|^\rho f_n(y)dy \le k^{\lambda-\rho}M.$$

Choose k sufficiently large so that $k^{\lambda-\rho}M < \varepsilon$. The same inequality holds for Y_o. Choose n sufficiently large so that $|\mathbf{E}g(Y_n,\lambda,k) \to \mathbf{E}g(Y_o,\lambda,k)| < \varepsilon$. Then

$$|\mathbf{E}|Y_n|^\lambda - \mathbf{E}|Y_o|^\lambda| \le |\mathbf{E}|Y_n|^\lambda - \mathbf{E}g(Y_n)| + |\mathbf{E}g(Y_n) - \mathbf{E}g(Y_o)| + |\mathbf{E}g(Y_o) - \mathbf{E}|Y_o|^\lambda| \le 3\varepsilon.$$

This proves that $\mathbf{E}|Y_n|^\lambda \to \mathbf{E}|Y_o|^\lambda$.

An example shows that $\mathbf{E}|Z_n|^\lambda \to 0$ for $\lambda < \rho$ does not imply $\mathbf{E}|Z_n|^\rho$ bounded. Take Z_n discrete with support {0,n} and probability log(n)/n at n. Then for $\lambda < 1$, $\mathbf{E}|Z_n|^\lambda = \log(n)/n^{1-\lambda} \to 0$, but $\mathbf{E}|Z_n|^1 = \log(n) \to +\infty$.

4.1.14. If $Y_n \to_p Y_o$, then $\text{Prob}(\mathbf{W}_n) \to 0$. Choose a subsequence n_k such that $Prob(\boldsymbol{W}_{n_k}) \le 2^{-k}$.

Then $\text{Prob}(\cup_{k'>k} \boldsymbol{W}_{n_{k'}}) \le \sum_{k'>k} Prob(\boldsymbol{W}_{n_{k'}}) \le \sum_{k'>k} 2^{-k'} = 2^{-k}$, implying $Y_{n_k} \to_{as} Y_o$.

4.1.15. Conditions for a.s. convergence follow from this basic probability theorem:

Theorem 4.2. (Borel-Cantelli) If $\mathbf{A}_i$ is any sequence of events in a probability space $(\mathbf{S},\boldsymbol{F},P)$, $\sum_{n=1}^{\infty} P(\mathbf{A}_i) < +\infty$ implies that almost surely only a finite number of the events $\mathbf{A}_i$ occur. If $\mathbf{A}_i$ is a sequence of independent events, then $\sum_{n=1}^{\infty} P(\mathbf{A}_i) = +\infty$ implies that almost surely an infinite number of the events $\mathbf{A}_i$ occur.

Apply the Borel-Cantelli theorem to the events $\mathbf{A}_i = \{s\in\mathbf{S}\,|\,|Y_i - Y_o| > \varepsilon\}$ to conclude that $\sum_{n=1}^{\infty} P(\mathbf{A}_i) < +\infty$ implies that almost surely only a finite number of the events $\mathbf{A}_i$ occur, and hence $|Y_i - Y_o| \le \varepsilon$ for all i sufficiently large. Thus, $Y_n - Y_o \to_{as} 0$, or $Y_n \to_{as} Y_o$. For the next result in the table, use (1.12) to get $\text{Prob}(\cup_{m\ge n}\boldsymbol{W}_m) \le \sum_{m>n} \text{Prob}(\mathbf{W}_m) \le \varepsilon^{-\rho}\sum_{m>n} \mathbf{E}|Y_m - Y_o|^{\rho}$. Apply Theorem 4.2 to conclude that if this right-hand expression is finite, then $Y_n \to_{as} Y_o$. The example at the end of (1.12) shows that almost sure convergence does not imply convergence in ρ-mean. Also, the example mentioned in 1.8 which has convergence in probability but not almost sure convergence can be constructed to have ρ-mean convergence but not almost sure convergence.

4.1.16. A result termed the *Slutsky theorem* which is very useful in applied work is that if two random variables Y_n and Z_n have a difference which converges in probability to zero, and if Y_n converges in distribution to Y_o, then $Z_n \to_d Y_o$ also. In this case, Y_n and Z_n are termed *asymptotically equivalent*. The argument demonstrating this result is similar to that for 4.1.9. Let F_n and G_n be the CDF's of Y_n and Z_n respectively. Let y be a continuity point of F_o and define the following events:

$$\mathbf{A}_n = \{s|Z_n(s) < y\},\ \mathbf{B}_n = \{s|Y_n(s) \le y - \varepsilon\},\ \mathbf{C}_n = \{s|Y_n(s) \le y + \varepsilon\},\ \mathbf{D}_n = \{s|\ |Y_n(s) - Z_n(s)| > \varepsilon\}.$$

Then $\mathbf{A}_n \subseteq \mathbf{C}_n\cup\mathbf{D}_n$ and $\mathbf{B}_n \subseteq \mathbf{A}_n\cup\mathbf{D}_n$, implying $F_n(y-\varepsilon) - \text{Prob}(\mathbf{D}_n) \le G_n(y) \le F_n(y+\varepsilon) + \text{Prob}(\mathbf{D}_n)$. Given $\delta > 0$, one can choose $\varepsilon > 0$ such that $y-\varepsilon$ and $y+\varepsilon$ are continuity points of F_n, and such that $F_o(y+\varepsilon) - F_o(y-\varepsilon) < \delta/3$. Then one can choose n sufficiently large so that $\text{Prob}(\mathbf{D}_n) < \delta/3$, $|F_n(y+\varepsilon) - F_o(y+\varepsilon)| < \delta/3$ and $|F_n(y+\varepsilon) - F_o(y+\varepsilon)| < \delta/3$. Then $|G_n(y) - F_o(y)| < \delta$.

4.1.17 A useful property of convergence in probability is the following result:

Theorem 4.3. (Continuous Mapping Theorem) If g(y) is a continuous function on an open set containing the support of Y_o, then $Y_n \to_p Y_o$ implies $g(Y_n) \to_p g(Y_o)$. The result also holds for vectors of random variables, and specializes to the rules that if $Y_{1n} \to_p Y_{10}$, $Y_{2n} \to_p Y_{20}$, and $Y_{3n} \to_p Y_{30}$ then (a) $Y_{1n} \cdot Y_{2n} + Y_{3n} \to_p Y_{10} \cdot Y_{20} + Y_{30}$, and (b) if $\text{Prob}(|Y_{20}| < \varepsilon) = 0$ for some $\varepsilon > 0$, then $Y_{1n}/Y_{2n} \to_p Y_{10}/Y_{20}$. In these limits, Y_{10}, Y_{20}, and/or Y_{30} may be constants.

Proof: Given $\varepsilon > 0$, choose M such that $P(|Y_o| > M) < \varepsilon$. Let $\mathbf{A}_o$ be the set of y in the support of Y_o that satisfy $|y| \leq M$. Then $\mathbf{A}_o$ is compact. Mathematical analysis can be used to show that there exists a nested sequence of sets $\mathbf{A}_o \subseteq \mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq \mathbf{A}_3$ with $\mathbf{A}_3$ an open neighborhood of $\mathbf{A}_o$ on which g is continuous, $\mathbf{A}_2$ compact, and $\mathbf{A}_1$ open. From 4.16, $\liminf_{n\to\infty} F_n(\mathbf{A}_1) \geq F_o(\mathbf{A}_1) \geq 1\text{-}\varepsilon$ implies there exists n_1 such that for $m > n_1$, $F_m(\mathbf{A}_1) \geq 1\text{-}2\varepsilon$. The continuity of g implies that for each $y \in \mathbf{A}_2$, there exists $\delta_y > 0$ such that $|y'\text{-}y| < \delta_y \Rightarrow |g(y') - g(y)| < \varepsilon$. These δ_y-neighborhoods cover $\mathbf{A}_2$. Then $\mathbf{A}_2$ has a finite subcover. Let δ be the smallest value of δ_y in this finite subcover. Then, g is uniformly continuous: $y \in \mathbf{A}_2$ and $|y'\text{-}y| < \delta$ imply $|g(y') - g(y)| < \varepsilon$. Choose $n > n_1$ such that for $m > n$, $P(|Y_m - Y_o| > \delta) < \varepsilon/2$. Then for $m > n$, $P(|g(Y_m) - g(Y_o)| > \varepsilon) \leq P(|Y_n - Y_o| > \delta) + P(|Y_o| > M) + 1 - F_m(\mathbf{A}_1) \leq 4\varepsilon$. □

4.1.18 The preceding result has an analog for convergence in distribution. This result establishes, for example, that if $Y_n \to_d Y_o$, with Y_o standard normal and $g(y) = y^2$, then Y_o is chi-squared, so that that Y_n^2 converges in distribution to a chi-squared random variable.

Theorem 4.4. If g(y) is a continuous function on an open set containing the support of Y_o, then $Y_n \to_d Y_o$ implies $g(Y_n) \to_d g(Y_o)$. The result also holds for vectors of random variables.

Proof: The Skorokhod representation given in Theorem 4.1 implies there exist Y_n' and Y_o' that have the same distributions as Y_n and Y_o, respectively, and satisfy $Y_n' \to_{as} Y_o'$. Then, Theorem 4.3 implies $g(Y_n') \to_{as} g(Y_o')$, and results 4.1.8 and 4.1.9 above then imply $g(Y_n') \to_d g(Y_o')$. Because of the common distributions, this is the result in Theorem 4.4. For this reason, this result is also sometimes referred to as (part of) the continuous mapping theorem. The Slutsky theorem, result 4.1.10, is a special case of the continous mapping Theorems 4.3 and 4.4. For clarity, I also give a direct proof of Theorem 4.4. Construct the sets $\mathbf{A}_o \subseteq \mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq \mathbf{A}_3$ as in the proof of Theorem 4.3. A theorem from mathematical analysis (Urysohn) states that there exists a continuous function r with values between zero and one that satisfies $r(y) = 1$ for $y \in \mathbf{A}_1$ and $r(y) = 0$ for $y \notin \mathbf{A}_3$. Then $g^*(y) = g(y) \cdot r(y)$ is continuous everywhere. From the Healy-Bray theorem, $Y_n \to_d Y_o \Longleftrightarrow \mathbf{E}\, h(Y_n) \to \mathbf{E}\, h(Y_o)$ for all continuous bounded $h \Longrightarrow \mathbf{E}\, h(g^*(Y_n)) \to \mathbf{E}\, h(g^*(Y_o))$ for all continuous bounded h, since the composition of continuous bounded functions is continuous and bounded $\Longleftrightarrow g^*(Y_n) \to_d g^*(Y_o)$.

But $P(g^*(Y_n) \neq g(Y_n)) \leq P(Y_n \notin \mathbf{A}_1) \leq 2\varepsilon$ for n sufficiently large, and $g^*(Y_o) = g(Y_o)$. Then, 4.1.16 and $g^*(Y_n) - g(Y_n) \rightarrow_p 0$ imply $g^*(Y_n) \rightarrow_d g^*(Y_o)$. □

4.1.19. Convergence properties are sometimes summarized in a notation called $O_p(\cdot)$ and $o_p(\cdot)$ which is very convenient for manipulation. (Sometimes too convenient; it is easy to get careless and make mistakes using this calculus.) The definition of $o_p(\cdot)$ is that a random sequence Y_n is $o_p(n^\alpha)$ if $n^{-\alpha}Y_n$ converges in probability to zero; and one then writes $Y_n = o_p(n^\alpha)$. Then, $Y_n \rightarrow_p Y_o$ is also written $Y_n = Y_o + o_p(1)$, and more generally $n^{-\alpha}(Y_n - Y_o) \rightarrow_p 0$ is written $Y_n - Y_o = o_p(n^\alpha)$. Thus $o_p(\cdot)$ is a notation for convergence in probability to zero of a suitably normalized sequence of random variables. When two sequences of random variables Y_n and Z_n are asymptotically equivalent, so that they satisfy $Y_n - Z_n = o_p(1)$, then they have a common limiting distribution by Slutsky's theorem, and this is sometime denoted $Y_n \sim_a Z_n$.

The notation $Y_n = O_p(1)$ is defined to mean that given $\varepsilon > 0$, there exists a large M (not depending on n) such that $\text{Prob}(|Y_n| > M) < \varepsilon$ for all n. A sequence with this property is called *stochastically bounded*. More generally, $Y_n = O_p(n^\alpha)$ means $\text{Prob}(|Y_n| > M \cdot n^\alpha) < \varepsilon$ for all n. A sequence that is convergent in distribution is stochastically bounded: If $Y_n \rightarrow_d Y_o$, then one can find M and n_o such that $\pm M$ are continuity points of Y_o, $\text{Prob}(|Y_o| \leq M) > 1-\varepsilon/2$, $|F_n(M) - F_o(M)| < \varepsilon/4$ and $|F_n(-M) - F_o(-M)| < \varepsilon/4$ for $n > n_o$. Then $\text{Prob}(|Y_n| > M) < \varepsilon$ for $n > n_o$. This implies $Y_n = O_p(1)$. On the other hand, one can have $Y_n = O_p(1)$ without having convergence to any distribution (e.g., consider $Y_n \equiv 0$ for n odd and Y_n standard normal for n even). The notation $Y_n = O_p(n^\alpha)$ means $n^{-\alpha}Y_n = O_p(1)$.

Most of the properties of $O_p(\cdot)$ and $o_p(\cdot)$ are obvious restatements of results from Figure 4.2. For example, $n^{-\alpha}Y_n = o_p(1)$, or $n^{-\alpha}Y_n \rightarrow_p 0$, immediately implies for any $\varepsilon > 0$ that there exists n_o such that for $n > n_o$, $\text{Prob}(|n^{-\alpha}Y_n| > \varepsilon) < \varepsilon$. For each $n \leq n_o$, one can find M_n such that $\text{Prob}(|n^{-\alpha}Y_n| > M_n) < \varepsilon$. Then, taking M to be the maximum of ε and the M_n for $n \leq n_o$, one has $\text{Prob}(|n^{-\alpha}Y_n| > M) < \varepsilon$ for all n, and hence $n^{-\alpha}Y_n = O_p(1)$. The results above can be summarized in the following string of inplications:

$$n^{-\alpha}Y_n \textit{ converges in probability to } 0 \iff n^{-\alpha}Y_n = o_p(1) \implies n^{-\alpha}Y_n \textit{ converges in distribution to } 0 \implies n^{-\alpha}Y_n = O_p(1)$$

An abbreviated list of rules for o_p and O_p is given in Figure 4.3. We prove the very useful rule 6 in this figure: Given $\varepsilon > 0$, $Y_n = O_p(n^\alpha) \Longrightarrow \exists\ M > 0$ such that $\text{Prob}(|n^{-\alpha}Y_n| > M) < \varepsilon/2$. Next $Z_n = o_p(n^\beta)$ implies $\exists\ n_o$ such that for $n > n_o$, $\text{Prob}(|n^{-\beta}Z_n| > \varepsilon/M) < \varepsilon/2$. Hence $\text{Prob}(|n^{-\alpha-\beta}Y_nZ_n| > \varepsilon) \leq \text{Prob}(|n^{-\alpha}Y_n| > M) + \text{Prob}(|n^{-\beta}Z_n| > \varepsilon/M) < \varepsilon$. Demonstration of the remaining rules is left as an exercise.

FIGURE 4.3. RULES FOR $O_p(\cdot)$ AND $o_p(\cdot)$

	Definition: $Y_n = o_p(n^\alpha) \Longrightarrow \text{Prob}(\lvert n^{-\alpha}Y_n\rvert > \varepsilon) \to 0$ for each $\varepsilon > 0$. **Definition: $Y_n = O_p(n^\alpha) \Longrightarrow$ for each $\varepsilon > 0$, there exists $M > 0$ such that $\text{Prob}(\lvert n^{-\alpha}Y_n\rvert > M) < \varepsilon$ for all n**
1	$Y_n = o_p(n^\alpha) \Longrightarrow Y_n = O_p(n^\alpha)$
2	$Y_n = o_p(n^\alpha)\ \&\ \beta > \alpha \Longrightarrow Y_n = o_p(n^\beta)$
3	$Y_n = O_p(n^\alpha)\ \&\ \beta > \alpha \Longrightarrow Y_n = o_p(n^\beta)$
4	$Y_n = o_p(n^\alpha)\ \&\ Z_n = o_p(n^\beta) \Longrightarrow Y_n \cdot Z_n = o_p(n^{\alpha+\beta})$
5	$Y_n = O_p(n^\alpha)\ \&\ Z_n = O_p(n^\beta) \Longrightarrow Y_n \cdot Z_n = O_p(n^{\alpha+\beta})$
6	$Y_n = O_p(n^\alpha)\ \&\ Z_n = o_p(n^\beta) \Longrightarrow Y_n \cdot Z_n = o_p(n^{\alpha+\beta})$
7	$Y_n = o_o(n^\alpha)\ \&\ Z_n = o_p(n^\beta)\ \&\ \beta \geq \alpha \Longrightarrow Y_n + Z_n = o_p(n^\beta)$
8	$Y_n = O_p(n^\alpha)\ \&\ Z_n = O_p(n^\beta)\ \&\ \beta \geq \alpha \Longrightarrow Y_n + Z_n = O_p(n^\beta)$
9	$Y_n = O_p(n^\alpha)\ \&\ Z_n = o_p(n^\beta)\ \&\ \beta > \alpha \Longrightarrow Y_n + Z_n = o_p(n^\beta)$
10	$Y_n = O_p(n^\alpha)\ \&\ Z_n = o_p(n^\beta)\ \&\ \beta < \alpha \Longrightarrow Y_n + Z_n = O_p(n^\alpha)$
11	$Y_n = O_p(n^\alpha)\ \&\ Z_n = o_p(n^\alpha) \Longrightarrow Y_n + Z_n = O_p(n^\alpha)$

4.2. INDEPENDENT AND DEPENDENT RANDOM SEQUENCES

4.2.1. Consider a sequence of random variables $Y_1, Y_2, Y_3, \ldots$. The *joint distribution* (CDF) of a finite subsequence $(Y_1,\ldots,Y_n)$, denoted $F_{1,\ldots,n}(y_1,\ldots,y_n)$, is defined as the probability of a state of Nature such that all of the inequalities $Y_1 \leq y_1,\ldots,Y_n \leq y_n$ hold. The random variables in the sequence are *mutually statistically independent* if for every finite subsequence $Y_1\ldots,Y_n$, the joint CDF factors:

$$F_{1,\ldots,n}(y_1,\ldots,y_n) \equiv F_1(y_1)\cdot\ldots\cdot F_n(y_n).$$

The variables are *independent and identically distributed* (i.i.d.) if in addition they have a common univariate CDF $F_1(y)$. The case of i.i.d. random variables leads to the simplest theory of stochastic limits, and provides the foundation needed for much of basic econometrics. However, there are

many applications, particularly in analysis of economic time series, where i.i.d. assumptions are not plausible, and a limit theory is needed for dependent random variables. We will define two types of dependence, martingale and mixing, that will cover a variety of econometric time series applications and require a modest number of tools from probability theory. We have introduced a few of the needed tools in Chapter 3, notably the idea of information contained in σ-fields of events, with the evolution of information captured by refinements of these σ-fields, and the definitions of measurable functions, product σ-fields, and compatability conditions for probabilities defined on product spaces. There are treatments of more general forms of dependence than martingale or mixing, but these require a more comprehensive development of the theory of stochastic processes.

4.2.2. Consider a sequence of random variables Y_k with k interpreted as an index of (discrete) time. One can think of k as the infinite sequence $k \in \mathbf{K} = \{1,2,...\}$, or as a doubly infinite sequence, extending back in time as well as forward, $k \in \mathbf{K} = \{...,-2,-1,0,1,2,...\}$. The set of states of Nature can be defined as the product space $\mathbf{S} = \times_{i\in\mathbf{K}} \mathbb{R}$, or $\mathbf{S} = \mathbb{R}^{\mathbf{K}}$, where $\mathbb{R}$ is the real line, and the "complete information" σ-field of subsets of **S** defined as $\boldsymbol{F}_{\mathbf{K}} = \otimes_{i\in\mathbf{K}} \boldsymbol{B}$, where $\boldsymbol{B}$ is the Borel σ-field of subsets of the real line; see 3.2. (The same apparatus, with **K** equal to the real line, can be used to consider continuous time. To avoid a variety of mathematical technicalities, we will not consider the continuous time case here.) Accumulation of information is described by a nondecreasing sequence of σ-fields $... \subseteq \boldsymbol{G}_{-1} \subseteq \boldsymbol{G}_0 \subseteq \boldsymbol{G}_1 \subseteq \boldsymbol{G}_2 \subseteq ...$, with $\boldsymbol{G}_t = (\otimes_{i\le t} \boldsymbol{B}) \otimes (\otimes_{i>t} \{\varphi, \mathbf{S}\})$ capturing the idea that at time t the future is unknown. The monotone sequence of σ-fields $\boldsymbol{G}_t$, i = ...,-1,0,1,2,... is called a *filtration*. The sequence of random variables Y_t is *adapted* to the filtration if Y_t is measurable with respect to $\boldsymbol{G}_t$ for each t. Some authors use the notation $\sigma(...,Y_{t-2},Y_{t-1},Y_t)$ for $\boldsymbol{G}_t$ to emphasize that it is the σ-field generated by the information contained in Y_s for $s \le t$. The sequence $...,Y_{-1},Y_0,Y_1,Y_2,...$ adapted to $\boldsymbol{G}_t$ for $k \in \mathbf{K}$ is termed a *stochastic process*. One way of thinking of a stochastic process is to recall that random variables are functions of states of Nature, so that the process is a function $Y: S \times K \to \mathbb{R}$. Then Y(s,k) is the *realization* of the random variable in period k, Y(s,·) a realization or *time-path* of the stochastic process, and Y(·,k) the random variable in period k. Note that there may be more than one sequence of σ-fields in operation for a particular process. These might correspond, for example, to the information available to different economic agents. We will need in particular the sequence of σ-fields $\boldsymbol{H}_t = \sigma(Y_t, Y_{t+1}, Y_{t+2},...)$ adapted to the process from time t forward; this is a nonincreasing sequence of σ-fields... $\supseteq \boldsymbol{H}_{t-1} \supseteq \boldsymbol{H}_t \supseteq \boldsymbol{H}_{t+1} \supseteq ...$. Sometimes $\boldsymbol{G}_t$ is termed the *natural upward filtration*, and $\boldsymbol{H}_t$ the *natural downward filtration*.

Each subsequence $(Y_m,...,Y_{m+n})$ of the stochastic process has a multivariate CDF $F_{m,..,m+n}(y_m,...,y_{m+n})$. It is said to be *stationary* if for each n, this CDF is the same for every m. A stationary process has the obvious property that moments such as means, variances, and covariances between random variables a fixed number of time periods apart are the same for all times m. Referring to 4.2.1, a sequence i.i.d. random variables is always stationary.

4.2.3. One circumstance that arises in some economic time series is that while the successive random variables are not independent, they have the property that their expectation, given history, is zero. Changes in stock market prices, for example, will have this property if the market is efficient, with arbitragers finding and bidding away any component of change that is predictable from history. A sequence of random variables X_t adapted to $\mathcal{G}_t$ is a *martingale* if almost surely $\mathbf{E}\{X_t|\mathcal{G}_{t-1}) = X_{t-1}$. If X_t is a martingale, then $Y_t = X_t - X_{t-1}$ satisfies $\mathbf{E}\{Y_t|\mathcal{G}_{t-1}) = 0$, and is called a *martingale difference* (m.d.) *sequence*. Thus, stock price changes in an efficient market form a m.d. sequence. It is also useful to define a *supermartingale* (resp., *submartingale*) if almost surely $\mathbf{E}\{X_t|\mathcal{G}_{t-1}) \le X_{t-1}$ (resp,, $\mathbf{E}\{X_t|\mathcal{G}_{t-1}) \ge X_{t-1}$). The following result, called the *Kolmogorov maximal inequality*, is a useful property of martingale difference sequences.

Theorem 4.5. If random variables Y_k are have the property that $E(Y_k|Y_1.,,,.Y_{k-1}) = 0$, or more technically the property that Y_k adapted to $\sigma(...,Y_{k-1},Y_k)$ is a martingale difference sequence, and if $\mathbf{E}Y_k^2 = \sigma_k^2$, then $P(\max_{1\le k\le n} | \sum_{i=1}^{k} Y_i| > \varepsilon) \le \sum_{i=1}^{n} \sigma_i^2/\varepsilon^2$.

Proof: Let $S_k = \sum_{i=1}^{k} Y_i$. Let Z_k be a random variable that is one if $S_j \le \varepsilon$ for $j < k$ and $S_k > \varepsilon$, zero otherwise. Note that $\sum_{i=1}^{n} Z_i \le 1$ and $\mathbf{E}(\sum_{i=1}^{n} Z_i) = P(\max_{1\le k\le n} | \sum_{i=1}^{k} Y_i| > \varepsilon)$. The variables S_k and Z_k depend only on Y_i for $i \le k$. Then $E(S_n - S_k|S_k,Z_k) = 0$. Hence

$$\mathbf{E}s_n^2 \ge \sum_{k=1}^{n} \mathbf{E}S_n^2\cdot Z_k = \sum_{k=1}^{n} \mathbf{E}[S_k + (S_n - S_k)]^2\cdot Z_k \ge \sum_{k=1}^{n} \mathbf{E}S_k^2\cdot Z_k \ge \varepsilon^2 \sum_{k=1}^{n} \mathbf{E}Z_k. \ \square$$

4.2.4. As a practical matter, many economic time series exhibit correlation between different time periods, but these correlations dampen away as time differences increase. Bounds on correlations by themselves are typically not enough to give a satisfactory theory of stochastic limits, but a related idea is to postulate that the degree of statistical dependence between random variables approaches negligibility as the variables get further apart in time, because the influence of ancient history is buried in an avalance of new information (*shocks*). To formalize this, we introduce the concept of *stochastic mixing*. For a stochastic process Y_t, consider events $\mathbf{A} \in \mathcal{G}_t$ and $\mathbf{B} \in \mathcal{H}_{t+s}$; then **A** draws only on information up through period t and **B** draws only on information from period t+s on. The idea is that when s is large, the information in **A** is too "stale" to be of much use in determining the probability of **B**, and these events are nearly independent. Three definitions of mixing are given in the table below; they differ only in the manner in which they are normalized, but this changes their strength in terms of how broadly they hold and what their implications are. When the process is stationary, mixing depends only on time differences, not on time location.

Form of Mixing	**Coefficient**	**Definition** (for all $\mathbf{A} \in \boldsymbol{G}_t$ and $\mathbf{B} \in \boldsymbol{H}_{t+s}$, and all t)
Strong	$\alpha(s) \to 0$	$\lvert P(\mathbf{A}\cap\mathbf{B}) - P(\mathbf{A})\cdot P(\mathbf{B})\rvert \le \alpha(s)$
Uniform	$\varphi(s) \to 0$	$\lvert P(\mathbf{A}\cap\mathbf{B}) - P(\mathbf{A})\cdot P(\mathbf{B})\rvert \le \varphi(s)P(\mathbf{A})$
Strict	$\psi(s) \to 0$	$\lvert P(\mathbf{A}\cap\mathbf{B}) - P(\mathbf{A})\cdot P(\mathbf{B})\rvert \le \psi(s)P(\mathbf{A})\cdot P(\mathbf{B})$

There are links between the mixing conditions and bounds on correlations between events that are remote in time:

(1) Strict mixing $\Longrightarrow$ Uniform mixing $\Longrightarrow$ Strong mixing.
(2) (Serfling) If the Y_i are uniform mixing with $\mathbf{E}Y_i = 0$ and $\mathbf{E}Y_t^2 = \sigma_t^2 < +\infty$, then $\lvert\mathbf{E}Y_tY_{t+s}\rvert \le 2\varphi(s)^{1/2}\sigma_t\sigma_{t+s}$.
(3) (Ibragimov) If the Y_i are strong mixing with $\mathbf{E}Y_t = 0$ and $\mathbf{E}\lvert Y_t\rvert^d < +\infty$ for some $d > 2$, then $\lvert\mathbf{E}Y_tY_{t+s}\rvert \le 8\alpha(s)^{1-2/d}\sigma_t\sigma_{t+s}$.
(4) If there exists a sequence ρ_t with $\lim_{t\to\infty}\rho_t = 0$ such that $\lvert\mathbf{E}(U-\mathbf{E}U)(W-\mathbf{E}W)\rvert \le \rho_t[(\mathbf{E}(U-\mathbf{E}U)^2)(\mathbf{E}(W-\mathbf{E}W)^2)]^{1/2}$ for <u>all</u> bounded continuous functions $U = g(Y_1,...,Y_t)$ and $W = h(Y_{t+n},...,Y_{t+n+m})$ and <u>all</u> t, n, m, then the Y_t are strict mixing.

An example gives an indication of the restrictions on a dependent stochastic process that produce strong mixing at a specified rate. First, suppose a stationary stochastic process Y_t satisfies $Y_t = \rho Y_{t-1} + Z_t$, with the Z_t imdependent standard normal. Then, $\text{var}(Y_t) = 1/(1-\rho^2)$ and $\text{cov}(Y_{t+s},Y_t) = \rho^s/(1-\rho^2)$, and one can show with a little analysis that $\lvert P(Y_{t+s} \le a, Y_t \le b) - P(Y_{t+s} \le a)\cdot P(Y_t \le b)\rvert \le \rho^s/\pi(1-\rho^{2s})^{1/2}$. Hence, this process is strong mixing with a mixing coefficient that declines at a geometric rate. This is true more generally of processes that are formed by taking stationary linear transformations of independent processes. We return to this subject in the chapter on time series analysis.

4.3. LAWS OF LARGE NUMBERS

4.3.1. Consider a sequence of random variables $Y_1,Y_2,...$ and a corresponding sequence of averages $X_n = n^{-1}\sum_{i=1}^{n} Y_i$ for n = 1,2,... . *Laws of large numbers* give conditions under which the averages X_n converge to a constant, either in probability (weak laws, or WLLN) or almost surely (strong laws, or SLLN). Laws of large numbers give formal content to the intuition that sample averages are accurate analogs of population averages when the samples are large, and are essential to establishing that statistical estimators for many problems have the sensible property that with sufficient data they are likely to be close to the population values they are trying to estimate. In

econometrics, convergence in probability provided by a WLLN suffices for most purposes. However, the stronger result of almost sure convergence is occasionally useful, and is often attainable without additional assumptions.

4.3.2 Figure 4.4 lists a sequence of laws of large numbers. The case of independent identically distributed (i.i.d.) random variables yields the strongest result (Kolmogorov I). With additional conditions it is possible to get a laws of large numbers even for correlated variable provided the correlations of distant random variables approach zero sufficiently rapidly.

FIGURE 4.4. LAWS OF LARGE NUMBERS FOR $X_n = n^{-1}\sum_{k=1}^{n} Y_k$

WEAK LAWS (WLLN)

1 (Khinchine) If the Y_k are i.i.d., and $\mathbf{E}\, Y_k = \mu$, then $X_n \rightarrow_p \mu$

2 (Chebyshev) If the Y_k are uncorrelated with $\mathbf{E}\, Y_k = \mu$ and $\mathbf{E}(Y_k - \mu)^2 = \sigma_k^2$ satisfying

$$\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty, \text{ then } X_n \rightarrow_p \mu$$

3 If the Y_k have $\mathbf{E}\, Y_k = \mu$, $\mathbf{E}(Y_k-\mu)^2 \equiv \sigma_k^2$, and $|\mathbf{E}(Y_k-\mu)(Y_m-\mu)| \leq \rho_{km}\sigma_k\sigma_m$ with

$$\sum_{k=1}^{\infty} \sigma_k^2/k^{3/2} < +\infty \text{ and } \lim_{n\rightarrow\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{m=1}^{n} \rho_{km} < +\infty, \text{ then } X_n \rightarrow_p \mu$$

STRONG LAWS (SLLN)

1 (Kolmogorov I) If the Y_k are i.i.d., and $\mathbf{E}\, Y_k = \mu$, then $X_n \rightarrow_{as} \mu$

2 (Kolmogorov II) If the Y_k are independent, with $\mathbf{E}\, Y_k = \mu$, and $\mathbf{E}(Y_k-\mu)^2 = \sigma_k^2$ satisfying

$$\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty, \text{ then } X_n \rightarrow_{as} \mu$$

3 (Martingale) Y_k adapted to $\sigma(...,Y_{k-1},Y_k)$ is a martingale difference sequence, $\mathbf{E}Y_t^2 = \sigma_t^2$, and $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$, then $X_n \rightarrow_{as} 0$

4 (Serfling) If the Y_k have $\mathbf{E}\, Y_k = \mu$, $\mathbf{E}(Y_k-\mu)^2 = \sigma_k^2$, and $|\mathbf{E}(Y_k-\mu)(Y_m-\mu)| \leq \rho_{|k-m|}\sigma_k\sigma_m$,

with $\sum_{k=1}^{\infty} (\log k)^2\sigma_k^2/k^2 < +\infty$ and $\sum_{k=1}^{\infty} \rho_{|k-m|} < +\infty$, then $X_n \rightarrow_{as} \mu$

To show why WLLN work, I outline proofs of the first three laws in Figure 4.4.

Theorem 4.6. (Khinchine) If the Y_k are i.i.d., and $\mathbf{E}\ Y_k = \mu$, then $X_n \rightarrow_p \mu$.

Proof: The argument shows that the characteristic function (c.f.) of X_n converges pointwise to the c.f. for a constant random variable μ. Let $\psi(t)$ be the c.f. of Y_1. Then X_n has c.f. $\psi(t/n)^n$. Since $\mathbf{E}Y_1$ exists, ψ has a Taylor's expansion $\psi(t) = 1 + \psi'(\lambda t)t$, where $0 < \lambda < 1$ (see 3.5.12). Then $\psi(t/n)^n = [1 + (t/n)\ \psi'(\lambda t/n)]^n$. But $\psi'(\lambda t/n) \rightarrow \psi'(0) = \imath\mu$. A result from 2.1.10 states that if a sequence of scalars α_n has a limit, then $[1+\alpha_n/n]^n \rightarrow \exp(\lim \alpha_n)$. Then $\psi(t/n)^n \rightarrow e^{\imath\mu t}$. But this is the c.f. of a constant random variable μ, implying $X_n \rightarrow_d \mu$, and hence $X_n \rightarrow_p \mu$. □

.

Theorem 4.7. (Chebyshev) If the Y_k are uncorrelated with $\mathbf{E}\ Y_k = \mu$ and $\mathbf{E}(Y_k - \mu)^2 = \sigma_k^2$ satisfying $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$, then $X_n \rightarrow_p \mu$.

Proof: One has $\mathbf{E}(X_n-\mu)^2 = \sum_{k=1}^{n} \sigma_n^2/n^2$. Kronecker's Lemma (see 2.1.9) establishes that $\sum_{k=1}^{\infty} \sigma_k^2/k^2$ bounded implies $\mathbf{E}(X_n-\mu)^2 \rightarrow 0$. Then Chebyshev's inequality implies $X_n \rightarrow_p \mu$. □

The condition $\sum_{k=1}^{\infty} \sigma_k^2/k^2$ bounded in Theorem 4.7 is obviously satisfied if σ_k^2 is uniformly bounded, but is also satisfied if σ_k^2 grows modestly with k; e.g., it is sufficient to have $\sigma_k^2(\log K)/k$ bounded.

Theorem 4.8. (WLLN 3) If the Y_k have $\mathbf{E}\ Y_k = \mu$, $\mathbf{E}(Y_k-\mu)^2 \equiv \sigma_k^2$, and $|\mathbf{E}(Y_k-\mu)(Y_m-\mu)| \leq \rho_{km}\sigma_k\sigma_m$ with $\sum_{k=1}^{\infty} \sigma_k^2/k^{3/2} < +\infty$ and $\lim_{n\rightarrow\infty} \frac{1}{n}\sum_{k=1}^{n} \sum_{m=1}^{n} \rho_{km} < +\infty$, then $X_n \rightarrow_p \mu$

Proof: Using Chebyshev's inequality, it is sufficient to show that $\mathbf{E}(X_n-\mu)^2$ converges to zero. The Cauchy-Schwartz inequality (see 2.1.11) is applied first to establish

$$\left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m\rho_{km}\right)^2 \leq \left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m^2\right)\left(\frac{1}{n}\sum_{m=1}^{n} \rho_{km}^2\right)$$

and then to establish that

$$\mathbf{E}(X_n-\mu)^2 = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n} \sigma_k\sigma_m\rho_{km} = \frac{1}{n}\sum_{k=1}^{n} \sigma_k\left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m\rho_{km}\right)$$

$$\leq \left(\frac{1}{n}\sum_{k=1}^{n}\sigma_k^2\right)^{1/2}\left[\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}\sum_{m=1}^{n}\sigma_m\rho_{km}\right)^2\right]^{1/2} \leq \left(\frac{1}{n}\sum_{k=1}^{n}\sigma_k^2\right)^{1/2}\left[\left(\frac{1}{n}\sum_{m=1}^{n}\sigma_m^2\right)\left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n}\rho_{km}^2\right)\right]^{1/2}$$

$$= \left(\frac{1}{n}\sum_{k=1}^{n}\sigma_k^2\right)\left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n}\rho_{km}^2\right)^{1/2} = \left(\frac{1}{n^{3/2}}\sum_{k=1}^{n}\sigma_k^2\right)\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{m=1}^{n}\rho_{km}^2\right)^{1/2}.$$

The last form and Kronecker's lemma (2.1.11) give the result. □

The conditions for this result are obviously met if the σ_k^2 are uniformly bounded and the correlation coefficients decline at a sufficient rate with the distance between observations; examples are geometric decline with ρ_{km} bounded by a multiple of $\lambda^{|k-m|}$ for some $\lambda < 1$ and an arithmetic decline with ρ_{km} bounded by a multiple of $|k-m|^{-1}$.

The Kolmogorov SLLN 1 is a better result than the Kinchine WLLN, yielding a stronger conclusion from the same assumptions. Similarly, the Kolmogorov SLLN 2 is a better result than the Chebyshev WLLN. Proofs of these theorems can be found in C. R. Rao (1973), p. 114-115. The Serfling SLLN 4 is broadly comparable to WLLN 3, but Serfling gets the stronger almost sure conclusion with somewhat stronger assumptions on the correlations and somewhat weaker assumptions on the variances. If variances are uniformly bounded and correlation coefficients decline at least at a rate inversely proportional to the square of the time difference, this sufficient for either the WLLN 3 or SLLN 4 assumptions.

The SLLN 3 in the table applies to martingale difference sequences, and shows that Kolmogorov II actually holds for m.d. sequences.

Theorem 4.9. If Y_t adapted to $\sigma(...,Y_{k-1},Y_k)$ is a martingale difference sequence with $\mathbf{E}Y_t^2 = \sigma_t^2$ and $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$, then $X_n \rightarrow_{as} 0$.

Proof: The theorem is stated and proved by J. Davidson (1994), p. 314. To give an idea why SLLN work, I will give a simplified proof when the assumption $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$ is strengthened to $\sum_{k=1}^{\infty} \sigma_k^2/k^{3/2} < +\infty$. Either assumption handles the case of constant variances with room to spare. Kolmogorov's maximal inequality (Theorem 4.5) with $n = (m+1)^2$ and $\varepsilon = \delta m^2$ implies that

$$P(\max_{m^2\leq k\leq (m+1)^2} |X_k| > \delta) \leq P(\max_{1\leq k\leq n} |\sum_{i=1}^{k} Y_i| > \delta m^2) \leq \sum_{i=1}^{(m+1)^2} \sigma_i^2/\delta^2 m^4.$$

The sum over m of the right-hand-side of this inequality satisfies

$$\sum_{m=1}^{\infty} \sum_{i=1}^{(m+1)^2} \sigma_i^2/\delta^2 m^4 = \sum_{i=1}^{\infty} \sum_{m \geq i^{1/2}}^{\infty} \sigma_i^2/\delta^2 m^4 \leq 36 \sum_{i=1}^{\infty} \sigma_i^2/i^{3/2}\delta^2.$$

Then $\sum_{m=1}^{\infty} P(\sup_k |X_k| > \delta) \leq 36 \sum_{i=1}^{\infty} \sigma_i^2/i^{3/2}\delta^2 < +\infty$. Theorem 4.2 gives the result. □

4.4. CENTRAL LIMIT THEOREMS

4.4.1. Consider a sequence of random variables $Y_1,...,Y_n$ with zero means, and the associated sequence of scaled averages $Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i$. Central limit theorems (CLT) are concerned with conditions under which the Z_n, or variants with more generalized scaling, converge in distribution to a normal random variable Z_o. I will present several basic CLT, prove the simplest, and discuss the remainder. These results are summarized in Figure 4.5.

The most straighforward CLT is obtained for *independent and identically distributed* (i.i.d.) random variables, and requires only that the random variables have a finite variance. Note that the finite variance assumption is an additional condition needed for the CLT that was not needed for the SLLN for i.i.d. variables.

Theorem 4.10. (Lindeberg-Levy) If random variables Y_k are i.i.d. with mean zero and finite positive variance σ^2, then $Z_n \rightarrow_d Z_o \sim N(0,\sigma^2)$.

Proof: The approach is to show that the characteristic function of Z_n converges for each argument to the characteristic function of a normal. The CLT then follows from the limit properties of characteristic functions (see 3.5.12). Let $\psi(t)$ be the cf of Y_1. Then Z_n has cf $\psi(t \cdot n^{-1/2})^n$. Since $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^2 = \sigma^2$, $\psi(t)$ has a Taylor's expansion $\psi(t) = [1 + \psi''(\lambda t)t^2/2]$, where $0 < \lambda < 1$ and ψ'' is continuous with $\psi''(0) = -\sigma^2$. Then $\psi(t \cdot n^{-1/2})^n = [1 + \psi''(\lambda t \cdot n^{-1/2})t^2/2n]^n$. Then the limit result 2.1.10 gives $\lim_{n\rightarrow\infty} [1 + \psi''(\lambda t \cdot n^{-1/2})t^2/2n]^n = \exp(-\sigma^2 t^2/2)$. Thus, the cf of Z_n converges for each t to the cf of $Z_o \sim N(0,\sigma^2)$. □

4.4.2. When the variables are independent but not identically distributed, an additional bound on the behavior of tails of the distributions of the random variables, called the *Lindeberg condition*, is needed. This condition ensures that sources of relatively large deviations are spread fairly evenly through the series, and not concentrated in a limited number of observations. The Lindeberg condition can be difficult to interpret and check, but there are a number of sufficient conditions that are useful in applications. The main result, stated next, allows more general scaling than by $n^{-1/2}$.

FIGURE 4.5. CENTRAL LIMIT THEOREMS FOR $Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i$

1 (Lindeberg-Levy) Y_k i.i.d., $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma^2$ positive and finite $\Longrightarrow Z_n \rightarrow_d Z_o \sim N(0,\sigma^2)$

2 (Lindeberg-Feller) If Y_k independent, $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma_k^2 \in (0,+\infty)$, $c_n^2 = \sum_{k=1}^{n} \sigma_k^2$, then $c_n^2 \rightarrow +\infty$, $\lim_{n\rightarrow\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, and $U_n = \sum_{k=1}^{n} Y_k/c_n \rightarrow_d U_o \sim N(0,1)$ if and only if the *Lindeberg condition* holds: for each $\varepsilon > 0$, $\sum_{k=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 \rightarrow 0$

3 If Y_k independent, $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma_k^2 \in (0,+\infty)$, $c_n^2 = \sum_{k=1}^{n} \sigma_k^2$ have $c_n^2 \rightarrow +\infty$ and $\lim_{n\rightarrow\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, then each of the following conditions is sufficient for the Lindeberg condition:

(i) For some $r > 2$, $\sum_{k=1}^{n} \mathbf{E}\,|Y_k|^r/c_n^r \rightarrow 0$.

(ii) (Liapunov) For some $r > 2$, $\mathbf{E}\,|Y_k/\sigma_k|^r$ is bounded uniformly for all n.

(iii) For some $r > 2$, $\mathbf{E}\,|Y_k|^r$ is bounded, and c_k^2/k is bounded positive, uniformly for all k.

4 Y_k a martingale difference sequence adapted to $\sigma(...,Y_{k-1},Y_k)$ with $|Y_k| < M$ for all t and $\mathbf{E}Y_k^2 = \sigma_k^2$ satisfying $n^{-1}\sum_{k=1}^{n} \sigma_k^2 \rightarrow \sigma_o^2 > 0 \Longrightarrow Z_n \rightarrow_d Z_o \sim N(0,\sigma_o^2)$

5 (Ibragimov-Linnik) Y_k stationary and strong mixing with $\mathbf{E}\,Y_k = 0$, $\mathbf{E}\,Y_k^2 = \sigma^2 \in (0,+\infty)$, $\mathbf{E}Y_{k+s}Y_k = \sigma^2\rho_s$, and for some $r > 2$, $\mathbf{E}\,|Y_n|^r < +\infty$ and $\sum_{k=1}^{\infty} \alpha(k)^{1-2/r} < +\infty \Longrightarrow \sum_{s=1}^{\infty} |\rho_s| < +\infty$ and $Z_n \rightarrow_d Z_o \sim N(0,\sigma^2(1+2\sum_{s=1}^{\infty} \rho_s))$

Theorem 4.11. (Lindeberg-Feller) Suppose random variables Y_k are independent with mean zero and positive finite variances σ_k^2. Define $c_n^2 = \sum_{k=1}^{n} \sigma_k^2$ and $U_n = \sum_{k=1}^{n} Y_k/c_n$. Then $c_n^2 \to \infty$, $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, and $U_n \to_d U_o \sim N(0,1)$ if and only if the Y_k satisfy the Lindeberg condition that for $\varepsilon > 0$, $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 = 0$.

A proof of Theorem 4.11 can be found, for example, in P. Billingsley (1986), p. 369-375. It involves an analysis of the characteristic functions, with detailed analysis of the remainder terms in their Taylor's expansion. To understand the theorem, it is useful to first specialize it to the case that the σ_k^2 are all the same. Then $c_n^2 = n\sigma_1^2$, the conditions $c_n^2 \to \infty$ and $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$ are met automatically, and in the terminology at the start of this section, $U_n = Z_n/\sigma_1$. The theorem then says $U_n \to_d U_o \sim N(0,1)$ if and only if the sample average of $\mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon n^{1/2})$ converges to zero for each $\varepsilon > 0$. The last condition limits the possibility that the deviations in a single random variable could be as large in magnitude as the sum, so that the shape of the distribution of this variable makes a significant contribution to the shape of the distribution of the sum. An example shows how the Lindeberg condition bites. Consider independent random variables Y_k that equal $\pm k^r$ with probability $1/2k^{2r}$, and zero otherwise, where r is a positive scalar. The Y_k have mean zero and variance one, and $\mathbf{1}(|Y_k| > \varepsilon n^{1/2}) = 1$ if $k^r > \varepsilon n^{1/2}$, implying $n^{-1} \sum_{i=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon n^{1/2}) = \max(0, 1-\varepsilon^{1/r} n^{(1-2r)/2r})$. This converges to zero, so the Lindeberg condition is satisfied iff $r < 1/2$. Thus, the tails of the sequence of random variables cannot "fatten" too rapidly.

The Lindeberg condition allows the variances of the Y_k to vary within limits. For example, the variables $Y_k = \pm 2^k$ with probability 1/2 have σ_n/c_n bounded positive, so that the variances grow too rapidly and the condition fails. The variables $Y_k = \pm 2^{-k}$ with probability 1/2 have c_n bounded, so that σ_1/c_n is bounded positive, the variances shrink too rapidly, and the condition fails. The next result gives some easily checked sufficient conditions for the Lindeberg condition.

Theorem 4.12. Suppose random variables Y_k are independent with mean zero and positive finite variances σ_k^2 that satisfy $c_n^2 = \sum_{k=1}^{n} \sigma_k^2 \to \infty$ and $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$. Then, each of the following conditions is sufficient for the Lindeberg condition to hold:

(i) For some $r > 2$, $\sum_{k=1}^{n} \mathbf{E}\, |Y_k|^r/c_n^r \to 0$.

(ii) (Liapunov) For some $r > 2$, $\mathbf{E}\, |Y_k/\sigma_k|^r$ is bounded uniformly for all n.

(iii) For some $r > 2$, $\mathbf{E}\, |Y_k|^r$ is bounded, and c_k^2/k is bounded positive, uniformly for all k.

Proof: To show that (i) implies the Lindeberg condition, write

$$\sum_{k=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 \le (\varepsilon c_n)^{2-r} \sum_{k=1}^{n} \mathbf{E}\, |Y_k|^r \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 \le \varepsilon^{2-r} \sum_{k=1}^{n} \mathbf{E}\, |Y_k/c_n|^r.$$

For (ii), the middle expression in the string of inequalities above satisfies

$$(\varepsilon c_n)^{2-r} \sum_{k=1}^{n} \mathbf{E}\, |Y_k|^r \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 \le \varepsilon^{2-r}(\max_{k\le n} \mathbf{E}|Y_k/\sigma_k|^r) \cdot \sum_{k=1}^{n} \sigma_k^r/c_n^r$$

$$\le \varepsilon^{2-r}(\max_{k\le n} \mathbf{E}|Y_k/\sigma_k|^r) \cdot \sum_{k=1}^{n} (\sigma_k^2/c_n^2)\cdot(\max_{k\le n} (\sigma_k/c_n)^{r-2}),$$

and the assumptions ensure that $\max_{k\le n} \mathbf{E}|Y_k/\sigma_k|^r$ is bounded and $\max_{k\le n} (\sigma_k/c_n)^{r-2} \to 0$.

Finally, if (iii), then continuing the first string of inequalities,

$$\sum_{i=1}^{n} \mathbf{E}\, |Y_k|^r/c_n^r \le c_n^{2-r} n \cdot (\sup_k \mathbf{E}\, |Y_k|^r)/n \cdot (\inf_n c_n^2/n),$$

and the right-hand-side is proportional to c_n^{2-r}, which goes to zero. □

4.4.3. The following theorem establishes a CLT for the scaled sum $Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i$ of martingale differences; or $Z_n = n^{-1/2}(X_n - X_o)$. The uniform boundedness assumption in this theorem is a strong restriction, but it can be relaxed to a Lindeberg condition or to a "uniform integratability" condition; see P. Billingsley (1984), p. 498-501, or J. Davidson (1994), p. 385. Martingale differences can display dependence that corresponds to important economic applications, such as conditional variances that depend systematically on history.

Theorem 4.13. Suppose Y_k is a martingale difference adapted to $\sigma(...,Y_{k-1},Y_k)$, and Y_k satisfies a uniform bound $|Y_k| < M$. Let $\mathbf{E}Y_k^2 = \sigma_k^2$, and assume that $n^{-1} \sum_{k=1}^{n} \sigma_k^2 \to \sigma_o^2 > 0$. Then $Z_n \to_d Z_o \sim N(0,\sigma_o^2)$.

4.4.4. Intuitively, the CLT results that hold for independent or martingale difference random variables should continue to hold if the degree of dependence between variables is negligible. The following theorem from I. Ibragimov and Y. Linnik, 1971, gives a CLT for stationary strong mixing processes. This result will cover a variety of economic applications, including stationary linear transformations of independent processes like the one given in the last example.

Theorem 4.14. (Ibragimov-Linnik) Suppose Y_k is stationary and strong mixing with mean zero, variance σ^2, and covariances $\mathbf{E}\ Y_{k+s}Y_k = \sigma^2\rho_s$. Suppose that for some $r > 2$, $\mathbf{E}|Y_n|^r < +\infty$ and $\sum_{k=1}^{\infty} \alpha(k)^{1-2/r} < +\infty$. Then, $\sum_{s=1}^{\infty} |\rho_s| < +\infty$ and $Z_n \rightarrow_d Z_o \sim N(0,\sigma^2(1+2\ \sum_{s=1}^{\infty}\ \rho_s))$.

4.5. EXTENSIONS OF LIMIT THEOREMS

4.5.1. Limit theorems can be extended in several directions: (1) obtaining results for "triangular arrays" that include weighted sums of random variables, (2) sharpening the rate of convergence to the limit for "well-behaved" random variables, and (3) establishing "uniform" laws that apply to random functions. In addition, there are a variety of alternatives to the cases given above where independence assumptions are relaxed. The first extension gives limit theorems for random variables weighted by other (non-random) variables, a situation that occurs often in econometrics. The second extension provides tools that allow us to bound the probability of large deviations of random sums. This is of direct interest as a sharper version of a Chebychev-type inequality, and also useful in obtaining further results. To introduce uniform laws, first define a *random function* (or *stochastic process*) $y = Y(\theta,s)$ that maps a state of Nature s and a real variable (or vector of variables) θ into the real line. This may also be written, suppressing the dependence on s, as $Y(\theta)$. Note that $Y(\cdot,w)$ is a *realization* of the random function, and is itself an ordinary non-random function of θ. For each value of θ, $Y(\theta,\cdot)$ is an ordinary random variable. A uniform law is one that bounds sums of random functions uniformly for all arguments θ. For example, a uniform WLLN would say $\lim_{n\rightarrow\infty} P(\sup_\theta \mid n^{-1}\sum_{i=1}^{n}\ Y_i(\theta,\cdot)| > \varepsilon) = 0$. Uniform laws play an important role in establishing the properties of statistical estimators that are nonlinear functions of the data, such as maximum likelihood estimates.

4.5.2 Consider a doubly indexed array of constants a_{in} defined for $1 \le i \le n$ and $n = 1,2,...$, and weighted sums of the form $X_n = \sum_{i=1}^{n}\ a_{in}Y_i$. If the Y_i are i.i.d., what are the limiting properties of X_n? We next give a WLLN and a CLT for weighted sums. The way arrays like a_{in} typically arise is that there are some weighting constants c_i, and either $a_{in} = c_i/\ \sum_{i=1}^{n}\ c_j$ or $a_{in} = c_i/[\ \sum_{i=1}^{n}\ c_j]^{1/2}$. If $c_i = 1$ for all i, then $a_{in} = n^{-1}$or $n^{-1/2}$, respectively, leading to the standard scaling in limit theorems.

Theorem 4.15. Assume random variables Y_i are independently identically distributed with mean zero. If an array a_{in} satisfies $\lim_{n\to\infty} \sum_{i=1}^{n} |a_{jn}| = 0$ and $\lim_{n\to\infty} \max_{j\le n} |a_{jn}| = 0$, then $X_n \to_p 0$.

Proof: This is a weighted version of Khinchine's WLLN, and is proved in the same way. Let $\zeta(t)$ be the second characteristic function of Y_1. From the properties of characteristic functions we have $\zeta'(0) = 0$ and a Taylor's expansion $\zeta(t) = t\cdot\zeta'(\lambda t)$ for some $0 < \lambda < 1$. The second characteristic function of X_n is then $\gamma(t) = \sum_{i=1}^{n} a_{in} t\cdot\zeta'(\lambda_{in} a_{in} t)$, implying $|\gamma(t)| \le \sum_{i=1}^{n} |a_{in} t\cdot\zeta'(\lambda_{in} a_{in} t)| \le |t|\cdot(\max_{i\le n} |\zeta'(\lambda_{in} a_{in} t)|)\cdot \sum_{i=1}^{n} |a_{in}|$. Then $\lim \sum_{i=1}^{n} |a_{in}| < \infty$ and $\lim (\max_{i\le n} |a_{in}|) = 0$ imply $\gamma(t) \to 0$ for each t, and hence X_n converges in distribution, hence in probability, to 0. □

Theorem 4.16. Assume random variables Y_i are i.i.d. with mean zero and variance $\sigma^2 \in (0,+\infty)$. If an array a_{in} satisfies $\lim_{n\to\infty} \max_{j\le n} |a_{jn}| = 0$ and $\lim_{n\to\infty} \sum_{i=1}^{n} a_{in}^2 = 1$, then $X_n \to_d X_o \sim N(0,\sigma^2)$.

Proof: The argument parallels the Lindeberg-Levy CLT proof. The second characteristic function of X_n has the Taylor's expansion $\gamma(t) = -(1/2)\sigma^2 t^2 a_{in} + [\zeta''(\lambda_{in} a_{in} t) + \sigma^2]\cdot a_{in}^2 t^2/2$, where $\lambda_{in} \in (0,1)$. The limit assumptions imply $\gamma(t) + (1/2)\sigma^2 t^2$ is bounded in magnitude by

$$\sum_{i=1}^{n} |\zeta''(\lambda_{in} a_{in} t) + \sigma^2|\cdot a_{in} t^2/2 \le [\sum_{i=1}^{n} a_{in}^2 t^2/2]\cdot\max_{i\le n}|\zeta''(\lambda_{in} a_{in} t) + \sigma^2|.$$

This converges to zero for each t since $\lim_{n\to\infty} \max_{i\le n}|\zeta''(\lambda_{in} a_{in} t) + \sigma^2| \to 0$. Therefore, $\gamma(t)$ converges to the characteristic function of a normal with mean 0 and variance σ^2. □

4.5.3. The limit theorems 4.13 and 4.14 are special cases of a limit theory for what are called *triangular arrays* of random variables, Y_{nt} with t = 1,2,...,n and n = 1,2,3,... . (One additional level of generality could be introduced by letting t range from 1 up to a function of n that increases to infinity, but this is not needed for most applications.) This setup will include simple cases like $Y_{nt} = Z_t/n$ or $Y_{nt} = Z_t/n^{1/2}$, and more general weightings like $Y_{nt} = a_{nt} Z_t$ with an array of constants a_{nt}, but can also cover more complicated cases. We first give limit theorems for Y_{nt} that are uncorrelated or independent within each row. These are by no means the strongest obtainable, but they have the merit of simple proofs.

Theorem 4.17. Assume random variables Y_{nt} for t = 1,2,...,n and n = 1,2,3,... are uncorrelated across t for each n, with $\mathbf{E}\, Y_{nt} = 0$, $\mathbf{E}\, Y_{nt}^2 = \sigma_{nt}^2$. Then, $\sum_{i=1}^{n} \sigma_{nt}^2 \to 0$ implies $\sum_{i=1}^{n} Y_{nt} \to_p 0$.

Proof: Apply Chebyshev's inequality. □

Theorem 4.18. Assume random variables Y_{nt} for t = 1,2,...,n and n = 1,2,3,... are independent across t for each n, with $\mathbf{E}\ Y_{nt} = 0$, $\mathbf{E}\ Y_{nt}^2 = \sigma_{nt}^2$, $\sum_{i=1}^{n} \sigma_{nt}^2 \to 1$, $\sum_{i=1}^{n} \mathbf{E}|Y_{nt}|^3 \to 0$, and $\sum_{i=1}^{n} \sigma_{nt}^4 \to 0$. Then $X_n = \sum_{i=1}^{n} Y_{nt} \to_d X_o \sim N(0,1)$.

Proof: From the properties of characteristic functions (see 3.5.12), the c.f. of Y_{nt} has a Taylor's expansion that satisfies $|\psi_{nt}(s) - 1 + s^2\sigma_{nt}^2/2| \le |s|^3\mathbf{E}|Y_{nt}|^3/6$. Therefore, the c.f. $\gamma_n(s)$ of X_n satisfies $\log \gamma_n(s) = \sum_{i=1}^{n} \log(1 - s^2\sigma_{nt}^2/2 + \lambda_{nt}|s|^3\mathbf{E}|Y_{nt}|^3/6)$, where $|\lambda_{nt}| \le 1$. From 2.1.10, we have the inequality that for $|a| < 1/3$ and $|b| < 1/3$, $|\mathrm{Log}(1+a+b) - a| < 4|b| + 3|a|^2$. Then, the assumptions guarantee that $|\log \gamma_n(s) + s^2 \sum_{i=1}^{n} \sigma_{nt}^2/2| \le 4|s|^3 \sum_{i=1}^{n} \mathbf{E}|Y_{nt}|^3/6 + 3\ s^4 \sum_{i=1}^{n} \sigma_{nt}^4/4$. The assumptions then imply that $\log \gamma_n(s) \to -s^2/2$, establishing the result. □

In the last theorem, note that if $Y_{nt} = n^{-1/2}Z_t$, then $E|Z_t|^3$ bounded is sufficient to satisfy all the assumptions. Another set of limit theorems can be stated for triangular arrays with the property that the random variables within each row form a martingale difference sequence. Formally, consider random variables Y_{nt} for t = 1,...,n and n = 1,2,3,... that are adapted to σ-fields $\boldsymbol{G}_{nt}$ that are a filtration in t for each n, with the property that $E\{Y_{nt}|\boldsymbol{G}_{n,t-1}\} = 0$; this is called a *martingale difference array*. A WLLN for this case is adapted from J. Davidson (1994), p. 299.

Theorem 4.19. If Y_{nt} and $\boldsymbol{G}_{nt}$ for t = 1,...,n and n = 1,2,3,... is an adapted martingale difference array with $|Y_{nt}| \le M$, $\mathbf{E}\ Y_{nt}^2 = \sigma_{nt}^2$, $\sum_{i=1}^{n} \sigma_{nt}$ uniformly bounded, and $\sum_{i=1}^{n} \sigma_{nt}^2 \to 0$, then $\sum_{i=1}^{n} Y_{nt} \to_p 0$.

The following CLT for martingale difference arrays is taken from D. Pollard (1984), p. 170-174.

Theorem 4.20. If Y_{nt} and $\boldsymbol{G}_{nt}$ for t = 1,...,n and n = 1,2,3,... is an adapted martingale difference array, $\lambda_{nt}^2 = E(Y_{nt}^2|\boldsymbol{G}_{n,t-1})$ is the conditional variance of Y_{nt}, $\sum_{i=1}^{n} \lambda_{nt}^2 \to_p \sigma^2 \in (0,+\infty)$, and if for each ε > 0, $\sum_{t=1}^{n} \mathbf{E}\ Y_{nt}^2 \cdot \mathbf{1}(|Y_{nt}| > \varepsilon) \to 0$, then $X_n = \sum_{i=1}^{n} Y_{nt} \to_d X_o \sim N(0,\sigma^2)$.

4.5.4. Chebyshev's inequality gives an easy, but crude, bound on the probability in the tail of a density. For random variables with well behaved tails, there are sharper bounds that can be used to get sharper limit theorems. The following inequality, due to Hoeffding, is one of a series of results called *exponential inequalities* that are stated and proved in D. Pollard (1984), p. 191-193: If Y_n are independent random variables with zero means that satisfy the bounds $-a_n \le Y_n \le b_n$, then $P(\ n^{-1}\sum_{i=1}^{n}\ Y_i \ge \varepsilon) \le \exp(-2n^2\varepsilon^2/\ \sum_{i=1}^{n}\ (b_i+a_i)^2)$. Note that in Hoeffding's inequality, if $|Y_n| \le M$, then $P(|\ n^{-1}\sum_{i=1}^{n}\ Y_i| \ge \varepsilon) \le 2\cdot\exp(-n\varepsilon^2/2M^2)$. The next theorem gets a strong law of large numbers with weaker than usual scaling:

Theorem 4.21. If Y_n are independent random variables with zero means and $|Y_n| \le M$, then $X_n = \ n^{-1}\sum_{i=1}^{n}\ Y_i$ satisfies $X_k\cdot k^{1/2}/\log(k) \to_{as} 0$.

Proof: Hoeffding's inequality implies $\text{Prob}(k^{1/2}|X_k| > \varepsilon\cdot\log k) < 2\cdot\exp(-(\log k)\varepsilon^2/2M^2)$, and hence

$$\sum_{k=n+1}^{\infty}\ \text{Prob}(k^{1/2}|X_k| > \varepsilon\cdot\log k) \le \ \int_{z=n}^{\infty}\ 2\cdot\exp(-(\log z)^2\varepsilon^2/2M^2)dz$$

$$\le (6/\varepsilon)\cdot\exp(M^2/2\varepsilon^2)\cdot\Phi(-\varepsilon\cdot(\log n)/M + M/\varepsilon),$$

with the standard normal CDF Φ resulting from direct integration. Applying Theorem 4.2, this inequality implies $n^{1/2}|X_n|/\log n \to_{as} 0$. □

If the Y_i are not necessarily bounded, but have a proper moment generating function, one can get an exponential bound from the moment generating function.

Theorem 4.22. If i.i.d. mean-zero random variables Y_i have a proper moment generating function, then $X_n = \ n^{-1}\sum_{i=1}^{n}\ Y_i$ satisfies $P(X_n > \varepsilon) < \exp(-\tau\varepsilon n^{1/2}+\kappa)$, where τ and κ are positive constants determined by the distribution of Y_i.

Proof: $P(Z > \varepsilon) = \ \int_{z>\varepsilon}\ F(dz) \le \ \int_{z>\varepsilon}\ e^{(z-\varepsilon)t}F(dz) \le e^{-\varepsilon t}\mathbf{E}e^{Zt}$ for a random variable Z. Let m(t) be the moment generating function of Y_i and τ be a constant such that m(t) is finite for $|t| < 2\tau$. Then one has $m(t) = 1 + m''(\lambda t)t^2/2$ for some $|\lambda| < 1$, for each $|t| < 2\tau$, from the properties of mgf (see 3.5.12).

The mgf of X_n is $m(t/n)^n = (1 + m''(\lambda t/n)t^2/2n^2)^n$, finite for $|t|/n \leq 2\tau$. Replace t/n by $\tau n^{-1/2}$ and observe that $m''(\lambda t/n) \leq m''(\tau n^{-1/2})$ and $(1+m''(\tau n^{-1/2})\tau^2/2n)^n \leq \exp(m''(\tau n^{-1/2})\,\tau^2/2)$. Substituting these expressions in the initial inequality gives $P(X_n > \varepsilon) \leq \exp(-\tau\varepsilon n^{1/2} + m''(\tau n^{-1/2})\,\tau^2/2\,)$, and the result holds with $\kappa = m''(\tau)\tau^2/2$. □

Using the same argument as in the proof of Theorem 4.19 and the inequality $P(X_n > \varepsilon) < \exp(-\tau\varepsilon n^{1/2}+\kappa)$ from Theorem 4.20, one can show that $X_k \cdot k^{1/2}/(\log k)^2 \rightarrow_{as} 0$, a SLLN with weak scaling.

4.5.5. This section states a uniform SLLN for random functions on compact set Θ in a Euclidean space $\mathbb{R}^k$. Let $(\mathbf{S},\boldsymbol{F},P)$ denote a probability space. Define a *random function* as a mapping Y from $\Theta\times S$ into $\mathbb{R}$ with the property that for each $\theta \in \Theta$, $Y(\theta,\cdot)$ is measurable with respect to (S,F,P). Note that $Y(\theta,\cdot)$ is simply a random variable, and that $Y(\cdot,s)$ is simply a function of $\theta \in \Theta$. Usually, the dependence of Y on the state of nature is suppressed, and we simply write $Y(\theta)$. A random function is also called a *stochastic process*, and $Y(\cdot,s)$ is termed a *realization* of this process. A random function $Y(\theta,\cdot)$ is *almost surely continuous* at $\theta_o \in \Theta$ if for s in a set that occurs with probability one, $Y(\cdot,s)$ is continuous in θ at θ_o. It is useful to spell out this definition in more detail. For each $\varepsilon > 0$, define $\mathbf{A}_k(\varepsilon,\theta_o) = \left\{ s\in\boldsymbol{S} \,\middle|\, \sup_{|\theta-\theta_o|\leq 1/k} |Y(\theta,s)-Y(\theta_o,s)| > \varepsilon \right\}$. Almost sure continuity states that these sets converge monotonically as $k\rightarrow\infty$ to a set $\mathbf{A}_o(\varepsilon,\theta_o)$ that has probability zero.

The condition of almost sure continuity allows the modulus of continuity to vary with s, so there is not necessarily a fixed neighborhood of θ_o independent of s on which the function varies by less than ε. For example, the function $Y(\theta,s) = \theta^s$ for $\theta \in [0,1]$ and s uniform on [0,1] is continuous at $\theta = 0$ for every s, but $\mathbf{A}_k(\varepsilon,0) = [0,(-\log \varepsilon)/(\log k))$ has positive probability for all k. The exceptional sets $\mathbf{A}_k(\varepsilon,\theta)$ can vary with θ, and there is no requirement that there be a set of s with probability one, or for that matter with positive probability, where $Y(\theta,s)$ is continuous for all θ. For example, assuming $\theta \in [0,1]$ and s uniform on [0,1], and defining $Y(\theta,s) = 1$ if $\theta \geq s$ and $Y(\theta,s) = 0$ otherwise gives a function that is almost surely continuous everywhere and always has a discontinuity.

Theorem 4.3 in Section 4.1 established that convergence in probability is preserved by continuous mappings. The next result extends this to almost surely continuous transformations; the result below is taken from Pollard (1984), p. 70.

Theorem 4.23. (Continuous Mapping**).** If $Y_n(\theta) \rightarrow_p Y_o(\theta)$ uniformly for θ in $\Theta \subseteq \mathbb{R}^k$, random vectors $\tau_o,\tau_n \in \Theta$ satisfy $\tau_n \rightarrow_p \tau_o$, and $Y_o(\theta)$ is almost surely continuous at τ_o, then $Y_n(\tau_n) \rightarrow_p Y_o(\tau_o)$.

Consider i.i.d. random functions $Y_i(\theta)$ that have a finite mean $\psi(\theta)$ for each θ, and consider the average $X_n(\theta) = n^{-1} \sum_{i=1}^{n} Y_i(\theta)$. Kolmogorov's SLLN I implies that pointwise, $X_n(\theta) \rightarrow_{as} \psi(\theta)$. However, we sometimes need in statistics a stronger result that $X_n(\theta)$ is uniformly close to $\psi(\theta)$ over the whole domain Θ. This is <u>not</u> guaranteed by pointwise convergence. For example, the random function $Y_n(s,\theta) = 1$ if $n^2 \cdot |s - \theta| \leq 1$, and $Y_n(s,\theta) = 0$ otherwise, where the sample space is the unit interval with uniform probability, has $P(Y_n(\cdot,\theta) > 0) \leq 2/n^2$ for each θ. This is sufficient to give $Y_n(\cdot,\theta) \rightarrow_{as} 0$ pointwise. However, $P(\sup_\theta Y_n(\theta) > 0) = 1$.

Theorem 4.24. (Uniform SLLN**).** Assume $Y_i(\theta)$ are independent identically distributed random functions with a finite mean $\psi(\theta)$ for θ in a closed bounded set $\Theta \subseteq \mathbb{R}^k$. Assume $Y_i(\cdot)$ is almost surely continuous at each $\theta \in \Theta$. Assume that $Y_i(\cdot)$ is dominated; i.e., there exists a random variable Z with a finite mean that satisfies $Z \geq \sup_{\theta\in\Theta}|Y_1(\theta)|$. Then $\psi(\theta)$ is continuous in θ and $X_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\theta)$ satisfies $\sup_{\theta\in\Theta}|X_n(\theta) - \psi(\theta)| \rightarrow_{as} 0$.

Proof: We follow an argument of Tauchen (1985). Let $(\mathbf{S},\boldsymbol{F},P)$ be the probability space, and write the random function $Y_i(\theta,s)$ to make its dependence on the state of Nature explicit. We have $\psi(\theta) = \int_S Y(\theta,s)P(ds)$. Define $u(\theta_o,s,k) = \sup_{|\theta-\theta_o|\leq 1/k} |Y(\theta,s) - Y(\theta_o,s)|$. Let $\varepsilon > 0$ be given. Let $\mathbf{A}_k(\varepsilon/2,\theta_o)$ be the measurable set given in the definition of almost sure continuity, and note that for $k = k(\varepsilon/2,\theta_o)$ sufficiently large, the probability of $\mathbf{A}_k(\varepsilon/2,\theta_o)$ is less than $\varepsilon/(4\cdot\mathbf{E}\,Z)$. Then,

$$\mathbf{E}u(\theta_o,\cdot,k) \leq \int_{A_k(\varepsilon/2,\theta_o)} u(\theta_o,s,k)P(ds) + \int_{A_k(\varepsilon/2,\theta_o)^c} u(\theta_o,s,k)P(ds)$$

$$\leq \int_{A_k(\varepsilon/2,\theta_o)} 2\cdot Z(s)\cdot P(ds) + \int_{A_k(\varepsilon/2,\theta_o)^c} (\varepsilon/2)\cdot P(ds) \leq \varepsilon\,.$$

Let $\mathbf{B}(\theta_o)$ be an open ball of radius $1/k(\varepsilon/2,\theta_o)$ about θ_o. These balls constructed for each $\theta_o \in \Theta$ cover the compact set Θ, and it is therefore possible to extract a finite subcovering of balls $\mathbf{B}(\theta_j)$ with centers at points θ_j for $j = 1,...,J$. Let $\mu_j = \mathbf{E}u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) \leq \varepsilon$. For $\theta \in \mathbf{B}(\theta_j)$, $|\psi(\theta) - \psi(\theta_j)| \leq \mu_j \leq \varepsilon$. Then

$$\sup_{\theta\in B(\theta_j)} |X_n(\theta) - \psi(\theta)| \le |X_n(\theta) - X_n(\theta_j) - \mu_j| + \mu_j + |X_n(\theta_j) - \psi(\theta_j)| + |\psi(\theta_j) - \psi(\theta)|$$

$$\le \left| \frac{1}{n} \sum_{i=1}^{n} u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) - \mu_j \right| + \varepsilon + |X_n(\theta_j) - \psi(\theta_j)| + \varepsilon .$$

Apply Kolmogorov's SLLN to each of the first and third terms to determine a sample size n_j such that

$$P\left(\sup_{n\ge n_j} \left| n^{-1} \sum_{i=1}^{n} u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) - \mu_j \right| > \varepsilon \right) < \varepsilon/2J$$

and

$$P\left(\sup_{n\ge n_j} |X_n(\theta_j) - \psi(\theta_j)| > \varepsilon \right) < \varepsilon/2J .$$

With probability at least $1 - \varepsilon/J$, $\sup_{\theta\in B(\theta_j)} |X_n(\theta) - \psi(\theta)| \le 4\varepsilon$. Then, with probability at least $1 - \varepsilon$,

$\sup_{\theta\in\Theta} |X_n(\theta) - \psi(\theta)| \le 4\varepsilon$ for $n > n_o = \max(n_j)$. □

The construction in the proof of the theorem of a finite number of approximating points can be reinterpreted as the construction of a finite family of functions, the $Y(\theta_j,\cdot)$, with the approximation property that the expectation of the absolute difference between $Y(\theta,\cdot)$ for any θ and one of the members of this finite family is less than ε. Generalizations of the uniform SLLN above can be obtained by recognizing that it is this approximation property that is critical, with a limit on how rapidly the size of the approximating family can grow with sample size for a given ε, rather than continuity per se; see D. Pollard (1984).

4.6. REFERENCES

P. Billingsley (1968) Convergence of Probability Measures, Wiley.
P. Billingsley (1986) Probability and Measure, Wiley.
J. Davidson (1994) Stochastic Limit Theory, Oxford.
W. Feller (1966) An Introduction to Probability Theory and Its Applications, Wiley.
I. Ibragimov and Y. Linnik, Independent and Stationary sequences of Random Variables, Wolters-Noordhoff, 1971.
J. Neveu (1965) Mathematical Foundations of the Calculus of Probability, Holden-Day.
D. Pollard (1984) Convergence of Stochastic Processes, Springer-Verlag.
C. R. Rao (1973) Linear Statistical Inference and Its Applications, Wiley.

R. Serfling (1970) "Convergence Properties of S_n Under Moment Restrictions," *Annals of Mathematical Statistics*, 41, 1235-1248.
R. Serfling (1980) Approximation Theorems of Mathematical Statistics, Wiley.
G. Tauchen (1985)
H. White (1984) Asymptotic Theory for Econometricians, Academic Press.

4.7 EXERCISES

1. The sequence of random variables X_n satisfy $X_n(s) = s^n$, where s is a state of Nature in the sample space S = [0,1] with uniform probability on S. Does X_n have a stochastic limit, and if so in what sense (weak, strong, quadratic mean, distribution)? What about $Y_n = n \cdot X_n$ or $Z_n = \log(X_n)$?

2. A sequence of random variables Z_n are multivariate normal with mean zero, variance $\sigma^2 n$, and covariances $\mathbf{E}\, Z_n Z_{n+m} = \sigma^2 n$ for m > n. (For an infinite sequence, this means that every finite subsequence is multivariate normal.) Let $S_n = \sum_{k=1}^{n} Z_k$. Does S_n/n converge in probability? Is there a scale factor $\alpha(n)$ such that $S_n/\alpha(n)$ converges in probability? Is there a scale factor $\beta(n)$ such that $S_n/\beta(n)$ is asymptotically normal?

3. Ignoring adjustments for family composition and location, an American family is said to be below the poverty line if its annual income is less than $14,800 per year. Let Y_i be the income level of family i, drawn randomly and independently from the American population, and let Q_i be one if Y_i is less than $14,800, zero otherwise. Family income can obviously never be larger than GDP, so that it is bounded above by a (very big) constant G, and cannot be negative. Let μ denote the population mean annual income and π denote the population proportion below the poverty line. Let m_n and p_n denote the corresponding sample means in a simple random sample of size n. Prove that sample mean annual income m_n converges in probability to population mean annual income; i.e., show the requirements for a WLLN are met. Prove that $n^{1/2}(m_n - \mu)$ converges in distribution to a normal; i.e., show the requirements for a CLT are met. Similarly, prove that p_n converges in probability to π and $n^{1/2}(p_n - \pi)$ converges in distribution to a normal with mean 0 and variance $\pi(1-\pi)$.

4. Empirical illustration of stochastic limits: On the computer, construct a simple random sample of observations X_k by drawing independent uniform random variables U_k and V_k from (0,1) and defining $X_k = 1$ if $U_k > 1/2$ and $X_k = \log(V_k)$ if $U_k \leq 1/2$. Let m_n be the sample mean of the X_k from a sample of size n for n = 10, 100, 1000. (a) Does m_n appear to converge in probability? To what limit? (b) Draw 100 samples of size 10 by the procedure described above, and keep the sample means from each of the 100 samples. Calculate what are called "studentized residuals" by subtracting the mean of the 100 sample means, and dividing these differences by their sample standard deviation (i.e., the square root of the average of the squared deviations). Sort these studentized residuals from low to high and plot them against quantiles of the standard normal, $Q_k = \Phi^{-1}((k-0.5)/n)$. This is called a normal probability plot, and deviations from the 45-degree line reflect differences in the exact distribution of the sample means from normal, plus random noise. Are there systematic deviations that suggest the normal approximation is not very good? (c) Repeat part (b) with 100 samples of size 100. Has the normal approximation become more accurate?

CHAPTER 5. EXPERIMENTS, SAMPLING, STATISTICAL DECISIONS

5.1. EXPERIMENTS

The business of economics is to explain how consumers and firms behave, and the implications of this behavior for the operation of the economy. To do this, economists need to be able to describe the features of the economy and its economic agents, to model behavior, and to test the validity of these models. For example, economists are interested in determining the effects of medicare eligibility on retirement decisions. They believe that the incentives implicit in medical insurance programs influence willingness to work, so that changes in these programs may *cause* retirement behavior to change. A first level of empirical interest is a description of the current situation, a snapshot of current retirement patterns under the current program. This description could be based on a census or sample of the current population. Statistics will play a role if a sample is used, providing tools for judging the accuracy of estimates of population parameters. At a deeper level, economists want to estimate how patterns would change if eligibility rules were altered. This interest requires that one conduct, or at least observe, an "experiment" in which different workers face different programs, and the impact of the program differences on their responses can be observed. The objective is to uncover a *causal* mapping from programs to behavioral responses. Major barriers to accomplishing this are confounding causal factors, or mutual causation by deeper hidden effects. A well-designed experiment or intelligent use of a "natural experiment" that provides a clear separation between the factor of interest and possible confounding effects will provide the most compelling empirical evidence on the economic question.

The most reliable way to try to uncover a causal relationship is through a *designed experiment*. For example, to study the effect of the medicare program on retirement, one could in principle establish several different levels of medicare eligibility, or *treatments*, and assign these treatments at random to members of the population. The measured response of employment to these treatment is the causal effect we were looking for, with the random assignment of treatments assuring that the effects we see are arising from this source alone, not from other, uncontrolled factors that might happen to be correlated with the treatment. Classical prototypes for designed experiments are those done in chemistry or biology labs, where a good procedure will be effective in eliminating potential confounding factors so the effect of the one factor of interest can be measured. Even here, there can be problems of measurement error and contaminated experiments, and statistical issues arise. Perhaps better prototypes for experiments in economics are designed field experiments in ecology or agronomy. For example, consider the classical experiment to measure the impact of fertilization on the productivity of corn plants. The agronomist prepares different test plots, and tries to keep conditions other than fertilizer, such as irrigation levels, comparable across the plots. However, there will be a variety of factors, such as wind and sunshine, that may differ from one plot to another. To isolate the effect of fertilizer from these confounding effects, the agronomist assigns the fertilizer

treatments to the different plots at random. This *randomized treatment design* is a powerful tool for measuring the causal effect of the treatments. Economists rarely have the freedom to study economic relationships by designing classical experiments with random assignment of treatments. At the scale of economic policy, it would often be invasive, time-consuming, and costly to conduct the experiments one would need. In addition, being experimented on can make economic agents and economies testy. However, there are various arenas in which designed experiments are done in economics. Field experiments have examined the impact of different marginal tax rates on employment behavior in low-income families, and the effect of different job training programs. Economics laboratory experiments have studied behavior in artificial markets. However, some areas of economic interest are beyond the reach of designed experiments because of technical and ethical barriers. No one would seriously propose, for example, to study the effect of life expectancy on savings behavior by randomly assigning execution dates, or the returns to education by randomly assigning years of schooling that students may receive. This makes economics primarily an observational or *field* science, like astronomy. Economists must search for *natural experiments* in which economic agents are subjected to varying levels of a causal factor of interest under circumstances where the effects of potential confounding factors are controlled, either by something like random assignment of treatments by Nature, or by measuring the levels of potential confounding factors and using modeling and data analysis methods that can untangle the separate effects of different factors. For example, to study the impact of schooling on income, we might try to use as a "Natural experiment" individuals such as Vietnam war draftees and non-draftees who as a result of the random draft lottery have different access to schooling. To study the impact of medical insurance eligibility on retirement decisions, we might try to study individuals in different states where laws for State medical welfare programs differ. This is not as clean as random assignment of treatments, because economic circumstances differ across States and this may influence what welfare programs are adopted. Then, one is left with the problem of determining how much of a variation in retirement patterns between States with strong and weak work incentives in their medical welfare programs is due to these incentives, and how much is due to overall demographics or income levels that induced States to adopt one welfare program or the other.

Looking for good natural experiments is an important part of econometric analysis. The most persuasive econometric studies are those where Nature has provided an experiment in which there is little possibility than anything other than the effect you are interested in could be causing the observed response. In data where many factors are at work jointly, the ability of statistical analysis to identify the separate contribution of each factor is limited. Regression analysis, which forms the core of econometric technique, is a powerful tool for separating the contributions of different factors, but even so it is rarely definitive. A good way to do econometrics is to look for good natural experiments *and* use statistical methods that can tidy up the confounding factors that Nature has not controlled for us.

5.2. POPULATIONS AND SAMPLES

5.2.1. Often, a population census is impractical, but it is possible to *sample* from the population. A core idea of statistics is that a properly drawn sample is a representation of the population, and that one can exploit the analogies between the population and the sample to draw inferences from the sample about features of the population. Thus, one can measure the average retirement age in the sample, and use it to infer the mean retirement age in the population. Statistics provides the tools necessary to develop these analogies, and assess how reliable they are.

A basic statistical concept is of a *simple random sample*. The properties of a simple random sample are that every member of the population has the same probability of being included, and the sample observations are statistically independent. A simple random sample can be defined formally in terms of independent trials from a probability space; see Chap. 3.4. However, for current purposes, it is sufficient to think of a population that is characterized by a probability distribution, and think of a random sample as a sequence of observations drawn independently from this distribution.

5.2.2. A simple random sample is representative of the underlying population in the sense that each sample observation has the population probability distribution. However, there is a more fundamental sense in which a simple random sample is an analog of the population, so that sample statistics are appealing approximations to their population analogs. Suppose one is dealing with a random variable X that is distributed in the population with a CDF F(x), and that one is interested in some feature of this distribution, such as its mean $\mu = \int_{-\infty}^{+\infty} x \cdot F(dx)$. This expectation depends on F, and we could make the dependence explicit by writing it as $\mu(x,F)$. More generally, the target may be $\mu(g,F)$, where $g = g(x)$ is some function of x, such as $g(x) = x^2$ or $g(x) = \mathbf{1}(x \leq 0)$.

Now suppose $(x_1, .,x_n)$ is a simple random sample drawn from this CDF, and define

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(x_i \leq x).$$

Then $F_n(x)$ equals the fraction of the sample values that are no greater than x. This is called the *empirical CDF* of the sample. It can be interpreted as coming from a probability measure that puts weight 1/n on each sample point. The population mean $\mu(x,F)$ has a sample analog which is usually written as $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$, but can also be written as $\mu(x,F_n) = \int_{-\infty}^{+\infty} x \cdot F_n(dx)$. This notation emphasizes that the sample mean is a function of the empirical CDF of the sample. The population mean and the sample mean are then the same function $\mu(x,\cdot)$, the only difference being that the first is evaluated at F and the second is evaluated at F_n. The following proposition, sometimes called the "fundamental theorem of statistics", establishes that as a simple random sample gets larger and

larger, its empirical CDF approximates the population CDF more and more closely. Then, intuitively, if $\mu(g,\cdot)$ is "continuous" in its second argument, an analogy principle suggests that $\mu(g,F_n)$ will converge to $\mu(g,F)$, so that $\mu(g,F_n)$ will be a good estimator of $\mu(g,F)$.

Theorem 5.1. (Glivenko-Cantelli) If random variables $X_1,X_2,...$ are independent and have a common CDF F, then $\sup_x|F_n(x) - F(x)|$ converges to zero almost surely.

Proof: Given $\varepsilon, \delta > 0$, there exists a finite number of points $z_1 < ... < z_K$ such that the monotone right-continuous function F varies at most $\varepsilon/2$ between the points; i.e., $F(z_k^*) - F(z_{k-1}) < \varepsilon/2$, where z_k^* denotes the limiting value as one approaches z_k from the left. Any point where F jumps by more than $\varepsilon/4$ will be included as a z_k point. By convention, assume $z_1 = -\infty$ and $z_K = +\infty$. For every x, bracketed by $z_{k-1} \le x < z_k$, one has $F_n(z_{k-1}) - F(z_{k-1}) - \varepsilon/2 \le F_n(x) - F(x) \le F_n(z_k) - F(z_k) + \varepsilon/2$. The event $\sup_k|F_n(z_k) - F(z_k)| < \varepsilon/2$ then implies the event $\sup_x |F_n(x) - F(x)| < \varepsilon$. At each z_k, the Kolmogorov SLLN establishes that $F_n(z_k) \to_{as} F(z_k)$. Then there exists n_k such that the probability of $|F_n(z_k) - F(z_k)| > \varepsilon/3$ for any $n \ge n_k$ is less than δ/K. Let $n = \max_k n_k$. Then, with probability at least $1 - \delta$, the event $\sup_{m>n} \sup_k |F_m(z_k) - F(z_k)| < \varepsilon/2$ occurs, implying the event $\sup_{m>n} \sup_x |F_m(x) - F(x)| < \varepsilon$, occurs. □

The Glivenko-Cantelli theorem implies that F_n converges in distribution to F, but is stronger, establishing that the convergence is uniform rather than pointwise, and is not restricted to continuity points of F. It is useful to state the Kolmogorov SLLN in the terminology used here: If the population statistic $\mu(g,F)$ exists, then the sample statistic $\mu(g,F_n)$ converges almost surely to $\mu(g,F)$. This provides a fundamental justification for the use of simple random samples, and for the use of sample statistics $\mu(g,F_n)$ that are analogs of population statistics $\mu(g,F)$ that are of interest.

5.2.3. While the idea of simple random sampling is straightforward, implementation in applications may not be. The way sampling is done is to first establish a *sample frame* and a *sampling protocol*. The sample frame essentially identifies the members of the population in an operational way that makes it possible for them to be sampled, and the sampling protocol spells out precisely how the sampling is to be done and the data collected. For example, suppose your target is the population of individuals who were 55 years of age in 1980 and over the following twenty years have retired in some pattern that may have been influenced by their access to medical insurance. An ideal sample frame would be a master list containing the names and current telephone numbers of all individuals in this target population. The sampling protocol could then be to use a random number generator to select and call individuals from this list with equal probability, and collect data on their retirement age and economic circumstances. However, the required master list does not exist, so this simple sample design is infeasible. A practical sample frame might instead start from a list of all working residential telephone numbers in the U.S. The sampling protocol would be to call numbers at random from this list, ask screening questions to determine if anyone

from the target population lives at that number, and interview an eligible resident if there is one. This would yield a sample that is not exactly a simple random sample, because some members of the target population have died or do not have telephones, households with multiple telephones are over sampled relative to those with one telephone, some households may contain more than one eligible person, and there may be attrition because some telephones are not answered or the respondent declines to participate. Even this sampling plan is infeasible if there is no master list of all the working residential telephone numbers. Then one might turn instead to random digit dialing (RDD), with a random number generator on a computer making up potential telephone numbers at random until the phone is answered. At first glance, it may seem that this is guaranteed to produce at least a simple random sample of working telephones, but even here complications arise. Different prefixes correspond to different numbers of working phones, and perhaps to different mixes of residential and business phones. Further, the probability that a number is answered may depend on the economic status of the owner. An important part of econometric analysis is determining when deviations from simple random sampling matter, and developing methods for dealing with them.

There are a variety of sampling schemes that are more complex variants on simple random sampling, with protocols that produce various forms of *stratification*. An example is *cluster sampling*, which first selects geographical units (e.g., cities, census tracts, telephone prefixes), and then samples residences within each chosen unit. Generally, these schemes are used to reduce the costs of sampling. Samples produced by such protocols often come with sample weights, the idea being that when these are applied to the sample observations, sample averages will be reasonable approximations to population averages. Under some conditions, econometric analysis can be carried out on these stratified samples by treating them *as if* they were simple random samples. However, in general it is important to consider the implications of sampling frames and protocols when one is setting up a statistical analysis.

We have given a strong theoretical argument that statistical analysis of simple random samples will give reasonable approximations to target population features. On the other hand, the history of statistics is filled with horror stories where an analysis has gone wrong because the sample was not random. The classical example is the Liberty telephone poll in 1936 that predicted that Roosevelt would lose the Presidential election that he won in a landslide. The problem was that only the rich had telephones in 1936, so the sample was systematically biased. One should be very skeptical of statistical analyses that use purposive or selected samples, as the safeguards provided by random sampling no longer apply and sample statistics may be poor approximations to population statistics. Claims that a given sample frame and protocol have produced a simple random sample also deserve scrutiny,

An arena where the sampling theory is particularly obscure is in analysis of economic time-series. Here, one is observing a slice of history, and the questions are what is the population is from which this sample is drawn, and in what sense does this slice has properties of a random sample. One way statisticians have thought about this is to visualize our universe as being one draw from a population of "parallel universes".. This helps for doing formal probability theory, but is

unsatisfying for the economist whose target is a hypothesis about the one universe we are in. Another way to approach the problem is to think about the time series sample as a slice of a stochastic process that operates through time, with certain rules that regulate the relationship between behavior in a slice and behavior through all time. For example, one might postulate that the stochastic process is *stationary* and *ergodic*, which would mean that the distributions of variables depend only on their relative position in time, not their absolute position, and that long run averages converge to limits.

In this chapter and several chapters following, we will assume that the samples we are dealing with are simple random samples. Once we have a structure for statistical inference in this simplest case, we turn in later chapters to the problems that arise under alternative sampling protocols.

5.3. STATISTICAL DECISIONS

5.3.1. The process of statistical estimation can be thought of as decision-making under uncertainty. The economic problem faced by Cab Franc in Chapter 1 is an example. In decision-making under uncertainty, one has limited information, based upon observed data. There are costs to mistakes. On the basis of the available information, one wants to choose an action that minimizes cost. Let **x** denote the data, which may be a vector of observations from a simple random sample, or some more complex sample such as a slice from a time series process. These observations are governed by a *probability law*, or *data generation process* (DGP). We do not know the true DGP, but assume now that we do know that it is a member of some family of possible DGP's which we will index by a parameter θ. The true DGP will correspond to a value θ_o of this index. Let $F(\mathbf{x},\theta)$ denote the CDF for the DGP corresponding to the index θ. For the remainder of this discussion, we will assume that this CDF has a density, denoted by $f(\mathbf{x},\theta)$. The density f is called the *likelihood* function of the data. The unknown parameter θ_o might be some population feature, such as age of retirement in the example discussed at the beginning of this chapter. The statistical decision problem might then be to estimate θ_o, taking into account the cost of errors. Alternately, θ_o might be one of two possible values, say 0 and 1, corresponding to the DGP an economist would expect to see when a particular hypothesis is true or false, respectively. In this case, the decision problem is to infer whether the hypothesis is in fact true, and again there is a cost of making an error.

Where do we get values for the costs of mistakes in statistical decisions? If the client for the statistical analysis is a business person or a policy-maker, an inference about θ_o might be an input into an action that has a payoff in profits or in a measure of social welfare that is indexed in dollars. A mistake will lower this payoff. The cost of a mistake is then the opportunity cost of foregoing the higher payoff to be obtained if one could avoid mistakes. For the example of retirement behavior, making a mistake on the retirement age may cause the planned medicare budget to go out of balance, and cost may be a known function of the magnitude of the unanticipated imbalance. However, if there are multiple clients, or the analysis is being performed for the scientific audience, there may

not be precise costs, and it may be necessary to provide sufficient information from the analysis so that potential users can determine their most appropriate action based on their personal cost assessments. Before considering this situation, we will look at the case where there is a known cost function $C(\theta,\theta_o,\mathbf{x})$ that depends on the true parameter value θ_o and on the inference θ made from the data, and in general can also depend directly on $\mathbf{x}$.

5.3.2. A decision rule, or *action*, will be a mapping $T(\cdot)$ from the data $\mathbf{x}$ into the space of possible θ values. Note that while $T(\mathbf{x})$ depends on the data $\mathbf{x}$, it cannot depend directly on the unknown parameter θ_o, only indirectly through the influence of θ_o on the determination of $\mathbf{x}$. Because the data are random variables, $T(\cdot)$ is also a random variable, and it will have a density $\psi(t,\theta_o)$ that could be obtained from $f(\mathbf{x},\theta_o)$ by considering a one-to-one transformation from $\mathbf{x}$ to a vector that contains $T(\mathbf{x})$ and is filled out with some additional variables $\mathbf{Z}(\mathbf{x})$. The cost associated with the action $T(\cdot)$, given data $\mathbf{x}$, is $C(T(\mathbf{x}),\theta_o,\mathbf{x})$. One would like to choose this to be as small as possible, but the problem is that usually one cannot do this without knowing θ_o. However, the client may, prior to the observation of $\mathbf{x}$, have some beliefs about the likely values of θ_o. We will assume that these *prior beliefs* can be summarized in a density $h(\theta)$. Given prior beliefs, it is possible to calculate an expected cost for an action $T(\cdot)$. First, apply Bayes law to the *joint* density $f(x,\theta_o)\cdot h(\theta_o)$ of $\mathbf{x}$ and θ_o to obtain the conditional density of θ_o given x,

$$(1) \qquad p(\theta_o|\mathbf{x}) = f(\mathbf{x},\theta_o)\cdot h(\theta_o)/\int_{-\infty}^{+\infty} f(\mathbf{x},\theta)\cdot h(\theta)d\theta\,.$$

This is called the *posterior* density of θ_o, given the data $\mathbf{x}$. Using this posterior density, the expected cost for an action $T(\mathbf{x})$ is

$$(2) \qquad R(T(\mathbf{x}),\mathbf{x}) = \int_{-\infty}^{+\infty} C(T(\mathbf{x}),\theta_o,\mathbf{x})p(\theta_o|\mathbf{x})d\theta_o$$

$$= \int_{-\infty}^{+\infty} C(T(\mathbf{x}),\theta_o,\mathbf{x})\cdot f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o/\int_{-\infty}^{+\infty} f(\mathbf{x},\theta)\cdot h(\theta)d\theta\,.$$

This expected cost is called the *Bayes risk*. It depends on the function $T(\cdot)$. The optimal action $T^*(\cdot)$ is the function $T(\cdot)$ that minimizes the expected cost for each $\mathbf{x}$, and therefore minimizes the Bayes risk. One has $R(T^*(x),\mathbf{x}) \le R(T^*(x) + \lambda,\mathbf{x})$ for a scalar λ, implying for each x the first-order condition $0 = \int_{-\infty}^{+\infty} \nabla_t C(T(\mathbf{x}),\theta_o,\mathbf{x})p(\theta_o|\mathbf{x})d\theta_o$. A strategy $T'(\mathbf{x})$ is called *inadmissible* if there is a second strategy $T''(\mathbf{x})$ such that $C(T''(\mathbf{x}),\theta,\mathbf{x}) \le C(T'(\mathbf{x}),\theta,\mathbf{x})$ for all θ for which $f(\mathbf{x},\theta) > 0$, with the inequality strict for some θ. Clearly the search for the optimal action $T^*(\mathbf{x})$ can be confined to the set of strategies that are admissible. In general, it is not obvious what the optimal action $T^*(\mathbf{x})$ that solves this problem looks like. A few examples help to provide some intuition:

(I) Suppose $C(\theta,\theta_o,\mathbf{x}) = (\theta - \theta_o)^2$, a quadratic cost function in which cost is proportional to the square of the distance of the estimator $T(\mathbf{x})$ from the true value θ_o. For a given $\mathbf{x}$, the argument $\theta = T(\mathbf{x})$ that minimizes (2) has to satisfy the first- order condition

$$0 = \int_{-\infty}^{+\infty} (T^*(\mathbf{x}) - \theta_o)\cdot f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o, \text{ or}$$

$$(3) \qquad T^*(\mathbf{x}) = \int_{-\infty}^{+\infty} \theta_o\cdot f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o / \int_{-\infty}^{+\infty} f(\mathbf{x},\theta)\cdot h(\theta)d\theta = \int_{-\infty}^{+\infty} \theta\cdot p(\theta|\mathbf{x})\cdot d\theta\,.$$

Then, $T^*(\mathbf{x})$ equals the *mean of the posterior density*.

(ii) Suppose $C(\theta,\theta_o,\mathbf{x}) = \alpha\cdot\max(0,\theta - \theta_o) + (1-\alpha)\cdot\max(0,\theta_o - \theta)$ where α is a cost parameter satisfying $0 < \alpha < 1$. This cost function is linear in the magnitude of the error. When $\alpha = 1/2$, the cost function is symmetric; for smaller α it is non-symmetric with a unit of positive error costing less than a unit of negative error. The first-order condition for minimizing cost is

$$0 = -(1-\alpha)\cdot \int_{-\infty}^{T^*(x)} f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o + \alpha\cdot \int_{T^*(x)}^{+\infty} f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o,$$

or letting $P(\theta|\mathbf{x})$ denote the CDF of the posterior density, $P(T^*(\mathbf{x})|\mathbf{x}) = \alpha$. Then $T^*(\mathbf{x})$ equals the α-*level quantile* of the posterior distribution. In the case that $\alpha = 1/2$, so that costs are symmetric in positive and negative errors, this criterion picks out the *median of the posterior density*.

(iii) Suppose $C(\theta,\theta_o,\mathbf{x}) = -1/2\alpha$ for $|\theta-\theta_o| \le \alpha$, and $C(\theta,\theta_o,\mathbf{x}) = 0$ otherwise. This is a cost function that gives a profit of $1/2\alpha$ when the action is within a distance α of θ_o, and zero otherwise; with α a positive parameter. The criterion (2) requires that $\theta = T^*(\mathbf{x})$ be chosen to minimize the expression $(-1/2\alpha)\int_{\theta-\alpha}^{\theta+\alpha} p(\theta_o|\mathbf{x})\cdot d\theta_o$. If α is very small, then $(-1/2\alpha)\int_{\theta-\alpha}^{\theta+\alpha} p(\theta_o|\mathbf{x})\cdot d\theta_o \simeq$ $-p(\theta|\mathbf{x})$. The argument minimizing $-p(\theta|\mathbf{x})$ is called the *maximum posterior likelihood* estimator; it picks out the *mode of the posterior density*. Then for α small, the optimal estimator is approximately the maximum posterior likelihood estimator. Recall that $p(\theta|\mathbf{x})$ is proportional to $f(\mathbf{x},\theta)\cdot h(\theta)$. Then, the first-order condition for its maximization can be written

$$(4) \qquad 0 = \frac{\nabla_\theta f(\boldsymbol{x},\theta)}{f(\boldsymbol{x},\theta)} + \frac{\nabla_\theta h(\theta)}{h(\theta)}\,.$$

The first term on the right-hand-side of this condition is the derivative of the log of the likelihood function, also called the *score*. The second term is the derivative of the log of the prior density. If prior beliefs are strong and tightly concentrated, then the second term will be very important, and the maximum will be close to the mode of the prior density, irrespective of the data. On the other hand, if prior beliefs are weak and very disperse, the second term will be small and the

maximum will be close to the mode of the likelihood function. In this limiting case, the solution to the statistical decision problem will be close to a general-purpose classical estimator, the maximum likelihood estimator.

The cost function examples above were analyzed under the assumption that prior beliefs were characterized by a density with respect to Lebesgue measure. If, alternately, the prior density had a finite support, then one would have analogous criteria, with sums replacing integrals, and the criteria would pick out the best point from the support of the prior.

5.3.3. The idea that there are prior beliefs regarding the true value of θ_o, and that these beliefs can be characterized in terms of a probability density, is called the *Bayesian* approach to statistical inference. It is philosophically quite different than an approach that thinks of probabilities as being associated only with particular random devices such as coin tosses that can produce frequencies. Bayesian statistics assumes that humans have a coherent system of beliefs that can attach probabilities to events such as "the Universe will continue to expand forever" and "40 percent of workers age 65 will work another year if medicare is unavailable", and these personal probabilities satisfy the basic axioms of probability theory. One of the implications of this way of thinking is that it is meaningful to talk about the *probability* that an event occurs, even if the "event" is something like a mathematical theorem whose truth is completely determinable by logic, and not the result of some cosmic coin toss. (In this case, if you do not know if the theorem is true, your probability may reflect your opinion of the mathematical abilities of its author.) How one thinks about probabilities influences how one thinks about an economic hypothesis, such as the hypothesis that retirement age does not depend on the age of medicare eligibility. In classical statistics, a hypothesis is either true or false, and the purpose of statistical inference is to decide whether it is true. In Bayesian statistics, this would correspond to concluding that the probability that the event is true is either zero or one. For a Bayesian statistician, it is more meaningful to talk about a high or low probability of the hypothesis being true.

5.3.4. The statistical decision theory just developed assumed that the analysis had a client with precisely defined prior beliefs. As in the case of the cost of errors, there will be circumstances where the client's prior beliefs are not known, or there is not even a well-defined client. It is the lack of a clearly identified prior that is one of the primary barriers to acceptance of the Bayesian approach to statistics. (Bayesian computations can be quite difficult, and this is a second major barrier.) There are three possible options in the situation where there is not a well-defined prior.

5.3.5. The first option is to carry out the statistical decision analysis with prior beliefs that carry "zero information". For example, an analysis may use a "diffuse" prior that gives every value of θ an equal probability. There are some technical problems with this approach. If the set of possible θ values is unbounded, "diffuse" priors may not be proper probability densities that integrate to one.

This problem can be skirted by using the prior without normalization, or by forming it as a limit of proper priors. More seriously, the idea of equal probability as being equivalent to "zero information" is flawed. A one-to-one but nonlinear transformation of the index θ can change a "diffuse" prior with equal probabilities into a prior in which probabilities are not equal, without changing anything about available information or beliefs. Then, equal probability is not in fact a characterization of "zero information". The technique of using diffuse or "uninformed" priors is fairly popular, in part because it simplifies some calculations. However, one should be careful to not assume that an analysis based on a particular set of diffuse priors is "neutral" or "value-free".

5.3.6. The second option is based on the idea that you are in a game against Nature in which Nature plays θ_o and reveals information **x** about her strategy, and you know that **x** is a draw from the DGP $f(\mathbf{x},\theta_o)$. You then play $T(\mathbf{x})$. Of course, if you had a prior $h(\theta_o)$, which in this context might be interpreted as a conjecture about Nature's play, you could adopt the optimal Bayes strategy $T^*(x)$. A conservative strategy in games is to play in such a way that you minimize the maximum cost your opponent can impose on you. This strategy picks $T^*(\mathbf{x}) = \text{argmin}_t \max_h R(t,\mathbf{x},h)$, where R is the Bayes risk, now written with an argument h to emphasize that it depends on the prior h. The idea is that the worst Nature can do is draw θ_o from the prior that is the least favorable to you in terms of costs, and this strategy minimizes this maximum expected cost. This is called a *minimax* strategy. Unless the problem has some compactness properties, the minimax strategy may not exist, although there may be a sequence of strategies that come close. A minimax strategy is a sensible strategy in a zero-sum game with a clever opponent, since your cost is your opponent's gain. It is not obvious that it is a good strategy in a game against Nature, since the game is not necessarily zero-sum and it is unlikely that Nature is an aware opponent who cares about your costs. There may however "meta-Bayesian" solutions in which search for a least favorable prior is limited to a class that the analyst considers "possible".

5.3.7. The final option is to stop the analysis short of a final solution, and simply deliver sufficient information from the sample to enable each potential user to compute the action appropriate to her own cost function and prior beliefs. Suppose there is a one-to-one transformation of the data **x** into two components $(\mathbf{y},\mathbf{z})$ so that the likelihood function $f(\mathbf{x},\theta)$ factors into the product of a marginal density of **y** that depends on θ, and a conditional density of **z**, given **y**, that does not depend on θ, $f(\mathbf{x},\theta) \equiv f_1(\mathbf{y},\theta)\cdot f_2(\mathbf{z}|\mathbf{y})$. In this case, **y** is termed a *sufficient statistic* for θ. If one forms the posterior density of θ given **x** and a prior density h, one has in this case

$$(5) \qquad p(\theta|\mathbf{x}) = \frac{f_1(y,\theta)\cdot f_2(z|y)\cdot h(\theta)}{\int_{-\infty}^{+\infty} f_1(y,\theta')\cdot f_2(z|y)\cdot h(\theta')d\theta'} = \frac{f_1(y,\theta)\cdot h(\theta)}{\int_{-\infty}^{+\infty} f_1(y,\theta')\cdot h(\theta')d\theta'} = p(\theta|\mathbf{y}).$$

Then, all the information to be learned about θ from **x**, reflected in the posterior density, can be learned from the summary data **y**. Then it is unnecessary to retain all the original data in **x** for purposes of statistical inference on θ; rather it is enough to retain the sufficient statistic **y**. By reporting **y**, the econometrician leaves the user completely free to form a prior, and calculate the posterior likelihood and the action that minimizes the user's Bayes risk. The limitations of this approach are that the dimensionality of sufficient statistics can be high, in many cases the dimension of the full sample, and that a substantial computational burden is being imposed on the user.

5.4. STATISTICAL INFERENCE

Statistical decision theory provides a template for statistical analysis when it makes sense to specify prior beliefs and costs of mistakes. Its emphasis on using economic payoffs as the criterion for statistical inference is appealing to economists as a model of decision-making under uncertainty, and provides a comprehensive, but not necessarily simple, program for statistical computations. While the discussion in this chapter concentrated on estimation questions, we shall see in Chapter 7 that it is also useful for considering tests of hypotheses.

The primary limitation of the Bayesian analysis that flows from statistical decision theory is that it is difficult to rationalize and implement when costs of mistakes or prior beliefs are not fully spelled out. In particular, in most scientific work where the eventual user of the analysis is not identified, so there is no consensus on costs of mistakes or priors, there is often a preference for "purely objective" solutions rather than Bayesian ones. Since a Bayesian approach can in principle be structured so that it provides solutions for all possible costs and priors, including those of any prospective user, this preference may seem puzzling. However, there may be compelling computational reasons to turn to "classical" approaches to estimation as alternative to the Bayesian statistical decision-making framework. We will do this in the next chapter.

5.5. EXERCISES

1. Suppose you are concerned about the question of whether skilled immigrants make a positive net contribution to the U.S. economy, to the non-immigrant residents, and to the domestic workers who are competing directly for the same skilled jobs. If you could design an experiment to measure these effects, how would you do it? If you have to rely on a natural experiment, what would you look for?

2. Suppose you have a random sample of size n from a population with CDF F(x) which has mean μ and variance σ^2. You estimate μ by forming the sample average, or mean of the empirical distribution F_n. The distribution of this estimator could be determined by drawing repeated samples from F, forming the empirical distribution of the sample averages, and taking the limit. An approximation to this calculation, called the *bootstrap*, starts from the known empirical distribution of the original sample F_n rather than the unknown F. Try this computationally. Take F to be uniform on [0,1], and draw a base sample of size 10. Now estimate the CDF of the sample mean by (1) repeatedly

sampling from the uniform distribution and (2) sampling with replacement from the base sample. Do 100 draws from each, and compare their medians.

3. Discuss how you would go about drawing a simple random sample of commuters in the San Francisco Bay Area. What problems do you face in defining the universe and the sample frame. Discuss ways in which you could implement the sampling. What problems are you likely to encounter?

4. Review the decision problem of Cab Franc in Chapter 1, and put it in the terminology of Section 5.3.2. Discuss the impact on Cab's decision of the particular loss function he has.

CHAPTER 6. ESTIMATION

6.1. DESIRABLE PROPERTIES OF ESTIMATORS

6.1.1 Consider data **x** that comes from a *data generation process* (DGP) that has a density f(**x**). Suppose we do not know f(·), but do know (or assume that we know) that f(·) is a member of a family of densities ***G***. The estimation problem is to use the data **x** to select a member of ***G*** which is some appropriate sense is close to the true f(·). Suppose we index the members of ***G*** by the elements of some set **Θ**, and identify f (·) with a particular index value θ_o. Then, another way of stating the estimation problem is that in the family of densities f(x,θ) parameterized by $\theta \in \mathbf{\Theta}$, we want to use the data **x** to estimate the true parameter value θ_o. The parameterization chosen for an estimation problem is not necessarily unique; i.e., there may be more than one way to parameterize the same family of densities ***G***. Sometimes this observation can be used to our advantage, by choosing parameterizations that simplify a problem. However, a parameterization can create difficulties. For example, you might set up **Θ** in such a way that more than one value of θ picks out the true density f; e.g., for some $\theta_o \neq \theta_1$, one has $f(x,\theta_o) = f(x,\theta_1)$ for all x. Then you are said to have an *identification problem*. Viewed within the context of a particular parameterization, identification problems cause real statistical difficulties and have to be dealt with. Viewed from the standpoint of the fundamental estimation problem, they are an artificial consequence of an unfortunate choice of parameterization. Another possible difficulty is that the family of densities generated by your parametric specification f(x,θ), $\theta \in \mathbf{\Theta}$, may fail to coincide with ***G***. A particularly critical question is whether the true f(·) is in fact in your parametric family. You cannot be sure that it is unless your family contains all of ***G***. Classical statistics always assumes that the true density is in the parametric family, and we will start from that assumption too. In Chapter 28, we will ask what the statistical properties and interpretation of parameter estimates are when the true f is not in the specified parametric family. A related question is whether your parametric family contains densities that are not in ***G***. This can affect the properties of statistical inference; e.g., degrees of freedom for hypothesis tests and power calculations.

In basic statistics, the parameter θ is assumed to be a scalar, or possibly a finite-dimensional vector. This will cover many important applications, but it is also possible to consider problems where θ is infinite-dimensional. It is customary to call estimation problems where θ is finite-dimensional *parametric*, and problems where θ is infinite-dimensional *semiparametric* or *nonparametric*. (It would have been more logical to call them "finite-parametric" and "infinite-parametric", respectively, but the custom is too ingrained to change.) Several chapters in the latter half of this book, particularly Chapter 28, deal with infinite-parameter problems.

6.1.2. In most initial applications, we will think of **x** as a simple random sample of size n, $\mathbf{x} = (x_1,...,x_n)$, drawn from a population in which x has a density $f(x,\theta_o)$, so that the DGP density is $f(\mathbf{x},\theta) = f(x_1,\theta_o)\cdot...\cdot f(x_n,\theta_o)$. However, the notation $f(\mathbf{x},\theta_o)$ can also cover more complicated DGP, such as time-series data sets in which the observations are serially correlated. Suppose that θ_o is an unknown k×1 vector, but one knows that this DGP is contained in a family with densities f(**x**,θ)

indexed by $\theta \in \boldsymbol{\Theta}$. An important leading case is k = 1, so that θ_o is a scalar. For many of the topics in this Chapter, it is useful to concentrate first on this case, and postpone dealing with the additional complications introduced by having a vector of parameters. However, we will use definitions and notation that cover the vector as well as the scalar case. Let $\boldsymbol{X}$ denote the domain of $\mathbf{x}$, and Θ denote the domain of θ. In the case of a simple random sample where an observation x is a point in a space X, one has $\boldsymbol{X} = \boldsymbol{X}^n$. The statistical inference task is to estimate θ_o. In Chapter 5, we saw that an estimator $T(\mathbf{x})$ of θ_o was desirable from a Bayesian point of view if $T(\cdot)$ minimized the expected cost of mistakes. For typical cost functions where the larger the mistake, the larger the cost, Bayes estimators will try to get "close" to the true parameter value. That is, the Bayes procedure will seek estimators whose probability densities are concentrated tightly around the true θ_o. Classical statistical procedures lack the expected cost criterion for choosing estimators, but also seek estimators whose probability densities are near the true density $f(\mathbf{x},\theta_o)$.

In this Chapter, we will denote the expectation of a function $r(\mathbf{x},\gamma)$ of $\mathbf{x}$ and a vector of "parameters" γ by $\mathbf{E}\, r(\mathbf{x},\gamma)$, or when it is necessary to identify the parameter vector of the true DGP, by $\mathbf{E}_{\mathbf{x}|\theta} r(\mathbf{x},\gamma) = \int_{-\infty}^{+\infty} r(\mathbf{x},\gamma)\cdot f(\mathbf{x},\theta)d\mathbf{x}$. Sometimes, the notation $\mathbf{E}_{\mathbf{x}|\theta} r(\mathbf{x},\gamma)$ is abbreviated to $\mathbf{E}_{\theta} r(\mathbf{x},\gamma)$.

This notation also applies when the parameters γ are also in $\boldsymbol{\Theta}$. Then $\mathbf{E}_{\mathbf{x}|\theta} r(\mathbf{x},\theta)$ is the expectation of $r(\mathbf{x},\gamma)$ when γ is set equal to the true parameter vector θ, and $\mathbf{E}_{\mathbf{x}|\theta} r(\mathbf{x},\gamma)$ is the expectation when r is evaluated at an argument γ that is not necessarily equal to the true parameter vector θ. The first of these expectations can be interpreted as a function of θ, and the second as a function of γ and θ.

The class of functions $r(\mathbf{x})$ that do not depend on any unknown parameters are called *statistics*. Examples of statistics are sample means and sample variances. The expectation $\rho(\theta) \equiv \mathbf{E}_{\mathbf{x}|\theta} r(\mathbf{x})$ of a statistic $r(\mathbf{x})$ is a function of θ (if the expectation exists), and it is sometimes said that $r(\mathbf{x})$ is an *unbiased estimator* of $\rho(\theta)$, or if $\rho(\theta) \equiv \theta$, an unbiased estimator of θ. It is important to note that $\rho(\theta) \equiv \mathbf{E}_{\mathbf{x}|\theta} r(\mathbf{x})$ is an *identity* that holds for all θ, and unbiasedness is a statement about this identity, not about the expectation for a specific value of θ. Observe that the condition $\rho(\theta) \equiv \theta$ is quite special, since in general $r(\mathbf{x})$ need not be in the same space or of the same dimension as θ. The concept of unbiased estimators and their properties will be developed further later in this section. However, it is worth noting at this point that every statistic that has an expectation is *by definition* an unbiased estimator of this expectation. Colloquially, "every estimator is an unbiased estimator of what it estimates".

When $r(\mathbf{x})$ is in the same space as θ, so that the Euclidean distance between $r(\mathbf{x})$ and θ is defined, the expectation $MSE(\theta) \equiv \mathbf{E}_{\mathbf{x}|\theta}(r(\mathbf{x}) - \theta)^2$ is termed the *mean square error* (MSE) of r. Note that the MSE may fail to exist, and when it does exist, it is a function of θ. The MSE is a classical concept in statistics, but note that it is identical to the the quadratic cost function that is widely used in statistical decision theory. As a result, there will be a strong family resemblance between estimators that have optimality properties for the statistical decision problem with quadratic cost and estimators that have some classical optimality properties in terms of MSE.

It is sometimes useful to think of each possible value of θ as defining a coordinate in a space, and $MSE(\theta)$ as determining a point along this coordinate. For example, if there are two possible values of the parameter, θ_1 and θ_2, then $(MSE(\theta_1),MSE(\theta_2))$ is a point in two-dimensional space that

characterizes its MSE as a function of θ. Different estimators have different MSE points in this space, and we may use the geometry of their relationship to select among them. There is an obvious generalization from two dimensions to a finite-dimensional vector space, corresponding to **Θ** finite, but the geometry continues to be well defined even when the set of possible parameter vectors is not finite.

The MSE can always be written as

$$\text{MSE}(\theta) \equiv \mathbf{E}_{\mathbf{x}|\theta}(r(\mathbf{x}) - \theta)^2 \equiv \mathbf{E}_{\mathbf{x}|\theta}(r(\mathbf{x}) - \rho(\theta) + \rho(\theta) - \theta)^2 \equiv \mathbf{E}_{\mathbf{x}|\theta}(r(\mathbf{x}) - \rho(\theta))^2 + (\rho(\theta) - \theta)^2,$$

where $\rho(\theta)$ is the expectation of $r(\mathbf{x})$, and there is no cross-product term since it has expectation zero. The term $V(\theta) \equiv \mathbf{E}_{\mathbf{x}|\theta}(r(\mathbf{x}) - \rho(\theta))^2$ is the variance of $r(\mathbf{x})$, the term $B(\theta) = \rho(\theta) - \theta$ is the bias of $r(x)$, and we have the result that *MSE equals variance plus squared bias*.

6.1.3. Listed below are some of the properties that are deemed desirable for classical estimators. Classical statistics often proceeds by developing some candidate estimators, and then using some of these properties to choose among the candidates. It is often not possible to achieve all of these properties at the same time, and sometimes they can even be incompatible. Some of the properties are defined relative to a *class* of candidate estimators, a set of possible $T(\cdot)$ that we will denote by ***T***. The density of an estimator $T(\cdot)$ will be denoted $\psi(t,\theta_o)$, or when it is necessary to index the estimator, $\psi_T(t,\theta_o)$. Sometimes the parameter vector θ will consist of a subvector α that is of primary interest for the application and a subvector β that is not. Then, α is termed the *primary parameter vector*, β is termed a *nuisance parameter vector*, and the DGP $f(x,\alpha,\beta)$ depends on both the primary and nuisance parameters. In this case, the problem is often to estimate α, dealing with the nuisance parameters as expediently as possible. One approach with fairly wide applicability is to replace β in the DGP by some appropriate function $r(x,\alpha)$, obtaining a *concentrated* DGP $f(x,\alpha,r(x,\alpha))$ that is a function only of the α parameters. Some statistical analysis is needed to determine when this is feasible and can be used as a device to get estimates of α with reasonable statistical properties. A specific choice of $r(x,\alpha)$ that often works is the argument that solves the problem $\max_\beta f(x,\alpha,\beta)$. Keep in mind that choice of parameterization is to some extent under the control of the analyst. Then it may be possible to choose a parameterization that defines α and isolates nuisance parameters in a way that helps in estimation of the primary parameters α.

6.1.4. *Sufficiency*. Suppose there is a one-to-one transformation from the data **x** into variables $(\mathbf{y},\mathbf{z})$. Then the DGP density $f(\mathbf{x},\theta)$ can be described in terms of the density of $(\mathbf{y},\mathbf{z})$, which we might denote $g(\mathbf{y},\mathbf{z},\theta)$ and write as the product of the marginal density of **y** and the conditional density of **z** given **y**, $g(\mathbf{y},\mathbf{z},\theta) = g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$. The relationship of the density $f(\mathbf{x},\theta)$ and the density $g(\mathbf{y},\mathbf{z},\theta)$ comes from the rules for transforming random variables; see Chapter 3.8. Let $\mathbf{x} = \mathbf{x}(\mathbf{y},\mathbf{z})$ denote the inverse of the one-to-one transformation from **x** to **y** and **z**, and let $J(\mathbf{y},\mathbf{z})$ denote the *Jacobian* of this mapping; i.e., the determinant of the array of derivatives of $\mathbf{x}(\mathbf{y},\mathbf{z})$ with respect to its arguments, signed so that it is positive. Then $g(\mathbf{y},\mathbf{z},\theta) = f(\mathbf{x}(\mathbf{y},\mathbf{z}))\cdot J(\mathbf{y},\mathbf{z})$. The Jacobian $J(\mathbf{y},\mathbf{z})$ does not depend on θ, so $g(\mathbf{y},\mathbf{z},\theta)$ factors into a term depending only on **y** and θ and a term independent of θ if and only if $f(\mathbf{x}(\mathbf{y},\mathbf{z}))$ factors in the same way.

In general, both the marginal and the conditional densities depend on θ. However, if the conditional distribution of **z** given **y** is independent of θ, $g_2(\mathbf{z}|\mathbf{y},\theta) = g_2(\mathbf{z}|\mathbf{y})$, then the variables **y** are said to be *sufficient* for θ. In this case, all of the information in the sample about θ is summarized in **y**, and once you know **y**, knowing **z** tells you nothing more about θ. (In Chapter 5.4, we demonstrated this by showing that the posterior density for θ, given **y** and **z**, depended only on **y**, no matter what the prior. Sufficiency of **y** is equivalent to a factorization $g(\mathbf{y},\mathbf{z},\theta) = g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y})$ of the density into one term depending only on **y** and θ and a second term depending only on **z** and **y**, where the terms g_1 and g_2 need not be densities; i.e., if there is such a factorization, then there is always an additional normalization by a function of **y** that makes g_1 and g_2 into densities. This characterization is useful for identifying sufficient statistics. Sufficiency can also be defined with respect to a subvector of primary parameters: if $g(\mathbf{y},\mathbf{z},\alpha,\beta) = g_1(\mathbf{y},\alpha)\cdot g_2(\mathbf{z}|\mathbf{y},\beta)$, then **y** is sufficient for α. Another situation that could arise is $g(\mathbf{y},\mathbf{z},\alpha,\beta) = g_1(\mathbf{y},\alpha)\cdot g_2(\mathbf{z}|\mathbf{y},\alpha,\beta)$, so the marginal distribution of **y** does not depend on the nuisance parameters, but the conditional distribution of **z** given **y** depends on all the parameters. It may be possible in this case to circumvent estimation of the nuisance parameters by concentrating on $g_1(\mathbf{y},\alpha)$. However, **y** is *not* sufficient for α in this case, as $g_2(\mathbf{z}|\mathbf{y},\alpha,\beta)$ contains additional information on α, unfortunately entangled with the nuisance parameters β.

An implication of sufficiency is that the search for a good estimator can be restricted to estimators $T(\mathbf{y})$ that depend only on sufficient statistics **y**. In some problems, only the full sample **x** is a sufficient statistic, and you obtain no useful restriction from sufficiency. In others there may be many different transformations of **x** into (**y**,**z**) for which **y** is sufficient. Then, among the alternative sufficient statistics, you will want to choose a **y** that is a *minimal sufficient statistic*. This will be the case if there is no further one-to-one transformation of **y** into variables (**u**,**v**) such that **u** is sufficient for θ and of lower dimension than **y**. Minimal sufficient statistics will be most useful when their dimension is low, and/or they isolate nuisance parameters so that the marginal distribution of **y** depends only on the primary parameters.

An example shows how sufficiency works. Suppose one has a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from an exponential distribution with an unknown scale parameter λ. The DGP density is the product of univariate exponential densities, $f(\mathbf{x},\lambda) = (\lambda\cdot\exp(-\lambda x_1))\cdot...\cdot(\lambda\cdot\exp(-\lambda x_n)) = \lambda^n\cdot\exp(-\lambda(x_1 + ... + x_n))$. Make the one-to-one transformation $y = x_1 + ... + x_n$, $z_1 = x_1,..., z_{n-1} = x_{n-1}$, and note that the inverse transformation implies $x_n = y - z_1 - ... - z_{n-1}$. Substitute the inverse transformation into f to obtain $g(\mathbf{y},\mathbf{z}) = f(\mathbf{x}(y,\mathbf{z})) = \lambda^n\cdot e^{-\lambda y}$. Then, g factors trivially into a marginal gamma density $g_1(y,\lambda) = \lambda^n y^{n-1}\cdot e^{-\lambda y}/(n-1)!$ for y, and a conditional uniform density $g_2(\mathbf{z}|y) = (n-1)!/y^{n-1}$ on the simplex $0 \le z_1 + ... + z_{n-1} \le y$. Then, y is a sufficient statistic for λ, and one need consider only estimators for λ that are functions of the univariate statistic $y = x_1 + ... + x_n$. In this case, y is a minimal sufficient statistic since obviously no further reduction in dimension is possible.

In this exponential example, there are other sufficient statistics that are not minimal. For example, any **y** such as $\mathbf{y} = (x_1 + ... + x_{n-2},x_{n-1},x_n)$ whose components can be transformed to recover the sum of the x's is sufficient. Obviously, the lower the dimension of the sufficient statistic, the less extensive the search needed to find a satisfactory estimator among all functions of the sufficient statistic. Then, it is worth while to start from a minimal sufficient statistic.

6.1.5. *Ancillarity*. Analogously to the discussion of sufficiency, suppose there is a one-to-one transformation from the data **x** into variables (**y**,**w**,**z**). Also suppose that the parameter vector θ is composed of a vector α of primary parameters and a vector β of nuisance parameters. Suppose that (**y**,**w**) are jointly sufficient for (α,β). Then the DGP density can be written as the product of the marginal density of **(y**,**w)** and the conditional density of **z** given **y** and **w**, $g_1(\mathbf{y},\mathbf{w},\alpha,\beta)\cdot g_2(\mathbf{z}|\mathbf{y},\mathbf{w})$. Further, the marginal density g_2 is the product of the conditional density of **y** given **w** times the marginal density of **w**, so that the density of the sample is $g_3(\mathbf{y}|\mathbf{w},\alpha,\beta)\cdot g_4(\mathbf{w},\alpha,\beta)\cdot g_2(\mathbf{z}|\mathbf{y},\mathbf{w})$. Both g_3 and g_4 depend in general on α and β. However, now consider the case where the sample density can be written $g_3(\mathbf{y}|\mathbf{w},\alpha)\cdot g_4(\mathbf{w},\beta)\cdot g_2(\mathbf{z}|\mathbf{y},\mathbf{w})$, with g_3 independent of β and g_4 independent of α. In this case, all the information in the data about α is contained in the *conditional* distribution of **y** given **w**. In this situation, the statistics **w** are said to be *ancillary* to α, and **y** is said to be *conditionally sufficient* for α given **w**. The search for an estimator for α should then concentrate solely on the conditional density of **y** given **w**, for the same reasons that it is sufficient to look only at estimators that are functions of sufficient statistics, and the nuisance parameters drop out of the analysis.

In these definitions, the distinction between **w** and **z** is not essential; if one absorbs all of **z** into **w**, it is still true that (**y**,**w**) is sufficient and all the information on α is contained in the conditional density of **y** given **w**. However, in applications, it is useful to distinguish **w** and **z** in order to reduce the dimensions of the statistics in $g_3(\mathbf{y}|\mathbf{w},\alpha)$ as much as possible.

Consider an estimator $T(\mathbf{y},\mathbf{w})$ of α that depends on the sufficient statistics (**y**,**w**), and suppose we now want to examine the properties of this estimator. For example, we might want to look at its expected value, and compare this to α. When **w** is ancillary, α influences the distribution of **y** given **w**, via $g_3(\mathbf{y}|\mathbf{w},\alpha)$, but it has *no* influence on the distribution of **w**. This suggests that in assessing $T(\mathbf{y},\mathbf{w})$, we should examine its *conditional* expectation given **w**, rather than its unconditional expectation with respect to both **y** and **w**. Put another way, we should not be satisfied with the estimator $T(\mathbf{y},\mathbf{w})$ if its conditional expectations are far away from α, even if its unconditional expectation were close to α, since the latter property is an accident of the distribution of **w**. Stated more generally, this is a *principle of conditionality* (or, *principle of ancillarity*) which says that the properties of an estimator should be considered conditional on ancillary statistics.

An artificial example makes it clear why the principle of conditionality is sensible. Suppose a survey is taken to estimate the size μ of a population of cows. There are alternative sets of instructions for the data collector. Instruction A says to count and report the number of noses. Instruction B says to count and report the number of ears. A sufficient statistic for μ is (y,w), where y is the reported count and w is an indicator that takes the value 1 if instruction A is used and the value 2 if instruction B is used. Each instruction has a 50/50 chance of being selected. Consider three estimators of μ, $T(y,w) = y/w$, $T'(y,w) = 2y/3$, and $T''(y,w) = y/w + w - 3/2$. Now, the expected value of y if w = 1 is μ, and the expected value of y if w = 2 is 2μ. Then, the conditional and unconditional expectations of the three estimators are

$$\mathbf{E}_{y|w}\, T(y,w) = \mu \qquad \mathbf{E}_{y|w}\, T'(y,w) = \begin{cases} 2\mu/3 & \text{if } w=1 \\ 4\mu/3 & \text{if } w=2 \end{cases} \qquad \mathbf{E}_{y|w}\, T''(y,w) = \begin{cases} \mu-1/2 & \text{if } w=1 \\ \mu+1/2 & \text{if } w=2 \end{cases}$$

$$\mathbf{E}\, T(y,w) = \mu \qquad \mathbf{E}\, T'(y,w) = (1/2)(2\mu/3) + (1/2)(4\mu/3) = \mu \qquad \mathbf{E}\, T''(y,w) = \mu + \mathbf{E}w - 3/2 = \mu$$

All three estimators are functions of the sufficient statistics and have unconditional expectation μ, so that they are "centered" at the true value. However, only the estimator T satisfies the principle of conditionality that this centering property holds conditional on w. The estimator T′ is not a function of w, and has the property that conditional on w, it will be systematically off-center. Obviously, it is unappealing to use an estimator that has a systematic bias given the measurement method we use, even if the expectation over the choice of measuring instruments accidently averages out the bias. The estimator T″ again has a systematic bias, given w, and is an unappealing estimator even though the further expectation over the selection of measurement method averages out the bias. In application, the principle of conditionality is usually, but not universally, desirable. First, some estimators that fail to satisfy this principle in finite samples will do so approximately in large samples, so that they satisfy what might be called asymptotic conditionality. This will often be enough to assure that they have reasonable large sample statistical properties. Second, there may be a tradeoff between the principal of conditionality and tractability; e.g., it is sometimes possible by deliberately intorducing some randomization into an estimator to make it easier to compute or easier to characterize its distribution, and this gain will sometimes offset the loss of the conditionality property.

A more realistic example where ancillarity and the principle of conditionality are useful arises in data $\mathbf{x} = (x_1,...,x_n)$ where the x_i are independent observations from an exponential density and the sample size n is random with a Poisson density $\gamma^{n-1}\cdot e^{-\gamma}/(n-1)!$ for n = 1,2,.... The DGP density is then $\lambda^n\cdot\exp(-\lambda(x_1 + ... + x_n))\cdot\gamma^{n-1}\cdot e^{-\gamma}/(n-1)!$. This density factors into the density $\lambda^n y^{n-1}\cdot e^{-\lambda y}$, with $y = x_1 +...+ x_n$, that is now the conditional density of y given n, times a marginal density that is a function of n, y, and γ, but not of λ. Then, the principle of ancillarity says that to make inferences on λ, one should condition on n and not be concerned with the nuisance parameter γ that enters only the marginal density of n, and the principle of conditionality says that in evaluating an estimator of λ, one should condition on n and not rely on some averaging over the distribution of n to yield some apparently desirable property.

6.1.6. *Unbiasedness*. An estimator $T(\cdot)$ is *unbiased* for θ if $\mathbf{E}_{x|\theta}T(\mathbf{x}) \equiv \theta$ for all θ; i.e., $\theta \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})f(\mathbf{x},\theta)d\mathbf{x}$. An estimator with this property is "centered" at the true parameter value, and will not systematically be too high or too low. Unbiasedness is an intuitively appealing criterion that is often used in classical statistics to select estimators. One implication of unbiasedness is that a comparison of the MSE of two unbiased estimators reduces to a comparison of their variances, since bias makes no contribution to their respective MSE. When $T(\mathbf{y},\mathbf{w})$ is a function of statistics $(\mathbf{y},\mathbf{w})$, the estimator is *conditionally unbiased* for θ given $\mathbf{w}$ if $\mathbf{E}_{x|\theta,\mathbf{w}}T(\mathbf{y},\mathbf{w}) \equiv \theta$. If w is ancillary to θ, then conditiionally unbiased estimators, conditioned on the ancillary $\mathbf{w}$, will be appealing, while the principle of conditionality suggests that an (unconditionally) unbiased estimator falls short. However, the next concept questions whether unbiasedness in general is a desirable property.

6.1.7. Admissibility. An estimator $T(\cdot)$ from a class of estimators $\boldsymbol{T}$ for a scalar parameter θ is *admissible relative to* $\boldsymbol{T}$ if there is no other estimator $T'(\cdot)$ in $\boldsymbol{T}$ with the property that the MSE of T′ is less than or equal to the MSE of T for all $\theta \in \boldsymbol{\Theta}$, with inequality strict for at least one θ. This is

the same as the definition of admissibility in statistical decision theory when the cost of a mistake is defined as *mean squared error* (MSE), the expected value of the square of the difference between the estimate and the true value of θ. An inadmissible estimator is undesirable because there is an identified alternative estimator that is more closely clustered around the true parameter value. Geometrically, if one interprets the MSE of an estimator as a point in a space where each possible θ is a coordinate, then the estimator T is admissible relative to $\boldsymbol{T}$ if there is no estimator in the class whose MSE is southwest of the MSE of T. The class $\boldsymbol{T}$ may have many admissible estimators, arrayed northwest or southeast of each other so that there is no uniform in θ ranking of their MSE. Then, admissibility is a relatively weak criterion may when applied leave many candidate estimators to be sorted out on other grounds. Another limitation of the admissibility criterion is that one might in fact have a cost of mistakes that is inconsistent with minimizing mean squared error. Suppose, for example, you incur a cost of zero if your estimate is no greater than a distance M from the true value, and a cost of one otherwise. Then, you will prefer the estimator that gives a higher probability of being within distance M, even if it occasionally has large deviations that make its MSE large.

The admissibility criterion is usually inconsistent with the unbiasedness criterion, a conflict between two reasonable properties. An example illustrates the issue. Suppose T(·) is an unbiased estimator. Suppose θ* is an arbitrary point in $\boldsymbol{\Theta}$ and c is a small positive constant, and define T′(·) = (1-c)T(·) + cθ* ; this is called a *Stein shrinkage* estimator. Then

$$\mathbf{E}_{\mathbf{x}|\theta}(T'(\mathbf{x}) - \theta)^2 = \mathbf{E}_{\mathbf{x}|\theta}\,[(1\text{-}c)(T(x) - \theta) + c(\theta^* - \theta)]^2 = c^2(\theta^* - \theta)^2 + (1\text{-}c)^2\mathbf{E}_{\mathbf{x}|\theta}\,[T(x) - \theta]^2,$$

implying that $\partial\mathbf{E}_{\mathbf{x}|\theta}(T'(\mathbf{x}) - \theta)^2/\partial c = 2c(\theta^* - \theta)^2 - 2(1\text{-}c)\mathbf{E}_{\mathbf{x}|\theta}\,[T(x) - \theta]^2 < 0$ for c sufficiently small. Then, for a problem where $(\theta^* - \theta)^2$ and $\mathbf{E}_{\mathbf{x}|\theta}\,[T(x) - \theta]^2$ are bounded for all $\theta \in \boldsymbol{\Theta}$, one can find c for which T′(·) has lower MSE than T(·), so that T(·) is inadmissible.

The concept of admissibility can be extended to vectors of parameters by saying that an estimator is admissible if each linear combination of the estimators is admissible for the same linear combination of the parameters. Consequently, if a vector of estimators is admissible, each coordinate estimator is admissible, but the reverse is not true

6.1.8. *Efficiency*. An estimator T(·) of a scalar parameter is *efficient relative to* an estimator T′(·) if for all θ the MSE of T(·) is less than or equal to the MSE of T′(·). The estimator T(·) is efficient relative to a class of estimators $\boldsymbol{T}$ if it is efficient relative to T′(·) for all T′(·) in $\boldsymbol{T}$. An efficient estimator provides estimates that are most closely clustered around the true value of θ, by the MSE measure, among all the estimators in $\boldsymbol{T}$. In terms of the geometric interpretation of MSE as a point in a space with a coordinate for each possible θ, an efficient estimator T(·) in $\boldsymbol{T}$ has the property that every other estimator T′(·) in $\boldsymbol{T}$ has a MSE to the northeast of T(·). It is possible for two distinct estimators to map into the same MSE point, in which case they are termed *equivalent*. This does not imply they are identical, or even that they agree in terms of other criteria that may be relevant to the user. Recall that admissibility requires that there be no other estimator in $\boldsymbol{T}$ with a MSE to the southwest of T(·). If there are estimators to the northwest or southeast of T(·), then T(·) will not be efficient because it no longer has a uniformly (in θ) weakly smallest MSE; however, T(·) can still be admissible even if it is not efficient. Thus, an efficient estimator must be admissible, but in

general an admissible estimator need not be efficient. Then there can be many admissible estimators, but no efficient estimator. If $\boldsymbol{T}$ contains an efficient estimator $T(\cdot)$, then another estimator $T'(\cdot)$ in $\boldsymbol{T}$ is admissible only if it is also efficient. The concept of efficiency extends to parameter vectors by requiring that it apply to each linear combination of the parameter vector. The following theorem establishes an important efficiency result for estimators that are functions of sufficient statistics:

Theorem 6.1. (Blackwell) If $T'(\cdot)$ is any estimator of θ from data $\mathbf{x}$, and $\mathbf{y}$ is a sufficient statistic, then there exists an estimator $T(\cdot)$ that is a function solely of the sufficient statistic and that is efficient relative to $T'(\cdot)$. If $T'(\cdot)$ is unbiased, then so is $T(\cdot)$. If an unbiased estimator $T(\cdot)$ is uncorrelated with every unbiased estimator of zero, then $T(\cdot)$ has a smaller variance than any other unbiased estimator, and is the unique efficient estimator in the class of unbiased estimators.

Proof: Suppose there is a scalar parameter. Make a one-to-one transformation of the data $\mathbf{x}$ into $(\mathbf{y},\mathbf{z})$, where $\mathbf{y}$ is the sufficient statistic, and let $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y})$ denote the DGP density. Define $T(\mathbf{y}) = \mathbf{E}_{z|y}T'(\mathbf{y},\mathbf{z})$. Write $T'(\mathbf{y},\mathbf{z}) - \theta = T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}) + T(\mathbf{y}) - \theta$. Then

$$\mathbf{E}(T'(\mathbf{y},\mathbf{z}) - \theta)^2 = \mathbf{E}(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}))^2 + \mathbf{E}(T(\mathbf{y}) - \theta)^2 + 2\cdot\mathbf{E}(T(\mathbf{y}) - \theta)\cdot(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}))\,.$$

But the last term satisfies

$$2\cdot\mathbf{E}(T(\mathbf{y}) - \theta)\cdot(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) = 2\cdot\mathbf{E}_y(T(\mathbf{y}) - \theta)\cdot\mathbf{E}_{z|y}(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) = 0\,.$$

Therefore, $\mathbf{E}(T'(\mathbf{y},\mathbf{z}) - \theta)^2 \geq \mathbf{E}(T(\mathbf{y}) - \theta)^2$. If $T'(\mathbf{y},\mathbf{z})$ is unbiased, then $\mathbf{E}T(\mathbf{y}) = \mathbf{E}_y\mathbf{E}_{z|y}T'(\mathbf{y},\mathbf{z}) = \theta$, and $T(\cdot)$ is also unbiased. Finally, suppose $T(\cdot)$ is uncorrelated with any estimator $U(\cdot)$ that is an unbiased estimator of zero, i.e., $\mathbf{E}U(\mathbf{y},\mathbf{z}) = 0$ implies $\mathbf{E}U(\mathbf{y},\mathbf{z})\cdot(T(\mathbf{y}) - \theta) = 0$. Then, any unbiased $T'(\mathbf{y},\mathbf{z})$ has $U(\mathbf{y},\mathbf{z})\cdot = T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})$ an unbiased estimator of zero, implying

$$\begin{aligned}\mathbf{E}(T'(\mathbf{x}) - \theta)^2 &= \mathbf{E}(T'(\mathbf{x}) - T(\mathbf{x}) + T(\mathbf{x}) - \theta)^2 = \mathbf{E}(T'(\mathbf{x}) - T(\mathbf{x}))^2 + \mathbf{E}(T(\mathbf{x}) - \theta)^2 + 2\cdot\mathbf{E}T(\mathbf{x})\cdot(T'(\mathbf{x}) - T(\mathbf{x}))\\ &= \mathbf{E}(T'(\mathbf{x}) - T(\mathbf{x}))^2 + \mathbf{E}(T(\mathbf{x}) - \theta)^2 > \mathbf{E}(T(\mathbf{x}) - \theta)^2.\end{aligned}$$

The theorem also holds for vectors of parameters, and can be established by applying the arguments above to each linear combination of the parameter vector. □

6.1.9. *Minimum Variance Unbiased Estimator (MVUE)*. If $\boldsymbol{T}$ is a class of unbiased estimators of a scalar parameter, so that $\mathbf{E}_{x|\theta}T'(\mathbf{x}) \equiv \theta$ for every estimator $T'(\cdot)$ in this class, then an estimator is efficient in this class if its variance is no larger than the variance of any other estimator in the class, and is termed a MVUE. There are many problems for which no MVUE exists. We next give a lower bound on the variance of an unbiased estimator. If a candidate satisfies this bound, then we can be sure that it is MVUE. However, the converse is not true: There may be a MVUE, but its variance may still be larger than this lower bound; i.e., the lower bound may be unobtainable. Once again, the MVUE concept can be extended to parameter vectors by requiring that it apply to each linear combination of parameters. If an observation x has a density $f(x,\theta)$ that is a continuously differentiable function of the parameter θ, define its *Fisher Information* to be

$$\mathbf{J} = \mathbf{E}_{x|\theta}\,[\nabla_\theta \log f(x,\theta)][\nabla_\theta \log f(x,\theta)]'.$$

Because this is the expectation of a square (or in the matrix case, the product of a vector times its transpose), **J** is non-negative (or in the matrix case, positive semi-definite). Except for pathological cases, it will be strictly positive. The following bound establishes a sense in which the Fisher Information provides a lower bound on the precision with which a parameter can be estimated.

Theorem 6.2. (Cramer-Rao Bound) Suppose a simple random sample $\mathbf{x} = (x_1,...,x_N)$ with $f(x,\theta)$ the density of an observation x. Assume that $\log f(x,\theta)$ is twice continuously differentiable in a scalar parameter θ, and that this function and its derivatives are bounded in magnitude by a function that is independent of θ and has a finite integral in x. Let J(x) denote the *Fisher information* in an observation x. Suppose an estimator $T(\mathbf{x})$ has $\mathbf{E}_{x|\theta}T(\mathbf{x}) \equiv \theta + \mu(\theta)$, so that $\mu(\theta)$ is the *bias* of the estimator. Suppose that $\mu(\theta)$ is differentiable. Then, the variance of $T(\mathbf{x})$ satisfies

$$\mathbf{V}_{x|\theta}(T(\mathbf{x})) \geq (\mathbf{I} + \nabla_\theta \mu(\theta))(n\mathbf{J})^{-1}(\mathbf{I} + \nabla_\theta \mu(\theta))'.$$

If the estimator is unbiased, so $\mu(\theta) \equiv 0$, this bound reduces to

$$\mathbf{V}_{x|\theta}(T(\mathbf{X})) \geq (n\mathbf{J})^{-1},$$

so that *the variance of an unbiased estimator is at least as large as the inverse of the Fisher information in the sample*. This result continues to hold when θ is a vector, with $\mathbf{V}_{x|\theta}(T(\mathbf{x}))$ a covariance matrix and "≥" interpreted to mean than the matrix difference is positive semidefinite.

Proof: Assume θ is a scalar. Let $L(\mathbf{x},\theta) = \sum_{i=1}^{n} \log f(x_i,\theta)$, so that the DGP density is $f(\mathbf{x},\theta) = e^{L(x,\theta)}$. By construction,

$$1 \equiv \int_{-\infty}^{+\infty} e^{L(x,\theta)}d\mathbf{x} \quad \text{and} \quad \theta + \mu(\theta) \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})\cdot e^{L(x,\theta)}d\mathbf{x}.$$

The conditions of the Lebesgue dominated convergence theorem are met, allowing differentiation under the integral sign. Then, differentiate each integral with respect to θ to get

$$0 \equiv \int_{-\infty}^{+\infty} \nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(x,\theta)}d\mathbf{x} \quad \text{and} \quad 1 + \mu'(\theta) \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(x,\theta)}d\mathbf{x}\,.$$

Combine these to get an expression for the covariance of T and $\nabla_\theta L$,

$$1 + \mu'(\theta) \equiv \int_{-\infty}^{+\infty} [T(\mathbf{x}) - \theta]\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(x,\theta)}d\mathbf{x}\,.$$

Apply the *Cauchy-Schwartz inequality*; see 3.5.9. In this case, the inequality can be written

$$(1 + \mu'(\theta))^2 = \left(\int_{-\infty}^{+\infty}[T(\boldsymbol{x}) - \theta]\cdot\nabla_\theta L(\boldsymbol{x},\theta)\cdot e^{L(\boldsymbol{x},\theta)}d\boldsymbol{x}\right)^2 \leq [\mathbf{E}_{x|\theta}(T(\mathbf{x}) - \theta)^2]\cdot]\mathbf{E}_{x|\theta}[\nabla_\theta L(\mathbf{x},\theta)]^2]\,.$$

Dividing both sides by the Fisher information in the sample, which is simply the variance of the sample score, $\mathbf{E}_{x|\theta}\,[\nabla_\theta L(\mathbf{x},\theta)]^2$, gives the bound.

When θ is k×1, one again has $\theta + \mu(\theta) = \int_{-\infty}^{+\infty} T(\boldsymbol{x})\, e^{L(\boldsymbol{x},\theta)} d\boldsymbol{x}$. Differentiating with respect to θ gives $\mathbf{I} + \nabla_\theta \mu(\theta) = \int_{-\infty}^{+\infty} T(\boldsymbol{x}) \cdot \nabla_\theta L(\boldsymbol{x},\theta) \cdot e^{L(\boldsymbol{x},\theta)} d\boldsymbol{x} = \int_{-\infty}^{+\infty} (T(\boldsymbol{x}) - \theta - \mu(\theta)) \cdot \nabla_\theta L(\boldsymbol{x},\theta) \cdot e^{L(\boldsymbol{x},\theta)} d\boldsymbol{x}$. The vector $(T(x) - \theta - \mu(\theta))'$, $\nabla_\theta L(\mathbf{x},\theta))$ has a positive semidefinite covariance matrix that can be written in partitioned form as $\begin{bmatrix} \boldsymbol{V}_{x|\theta}(T(x)) & [\boldsymbol{I}+\nabla_\theta \mu(\theta)] \\ [\boldsymbol{I}+\nabla_\theta \mu(\theta)]' & n\boldsymbol{J} \end{bmatrix}$. If one premultiplies this matrix by **W**, and postmultiplies by **W**′, with $\mathbf{W} = \left[\boldsymbol{I} \;\vdots\; -[\boldsymbol{I}+\nabla_\theta \mu(\theta)](n\boldsymbol{J})^{-1}\right]$, the resulting matrix is positive semidefinite, and gives the Cramer-Rao bound for the vector case. □

6.1.10. *Invariance*. In some conditions, one would expect that a change in a problem should not alter an estimate of a parameter, or should alter it in a specific way. Generically, these are called invariance properties of an estimator. For example, when estimating a parameter from data obtained by a simple random sample, the estimate should not depend on the indexing of the observations in the sample; i.e., $T(x_1,...,x_n)$ should be *invariant under permutations of the observations*. To illustrate, suppose one found that $T'(x_1)$ had some reasonable property such as being an unbiased estimator of θ. It is not invariant under permutation, but $T(\mathbf{x}) = (T'(x_1) + ... + T'(x_n))/n$ is, and hence by this invariance criterion would be a preferable estimator.

A second example is *invariance with sample scale*: if $T_n(x_1,...,x_n)$ denotes the estimator for a sample of size n, and the observations all equal a constant c, then the estimator should not change with sample size, or $T_n(c,...,c) = T_1(c)$. A sample mean, for example, has invariance under permutation and invariance with sample scale.

Sometimes a parameter enters a DGP in such a way that there is a simple relationship between shifts in the parameter and the shifts one would expect to observe in the data. For example, suppose the density of an observation is of the form $f(x_i|\theta) \equiv h(x_i-\theta)$; in this case, θ is called a *location parameter*. If the true value of θ shifts up by an amount Δ, one would expect observations on average to shift up by the same amount Δ. If $T_n(x_1,...,x_n)$ is an estimator of θ_o in this problem, a reasonable property to impose on $T_n(\cdot)$ is that $T_n(x_1+\Delta,...,x_n+\Delta) = T_n(x_1,...,x_n) + \Delta$. In this case, $T_n(\cdot)$ is termed *location invariant*. For this parametric family, it is reasonable to restrict attention to estimators with this invariance property.

Another example is *scale invariance*. Suppose the density of an observation has the form $f(x_i|\theta) \equiv \theta^{-1} \cdot h(x_i/\theta)$. Then θ is called a *scale parameter*. If θ is increased by a proportion λ, one would expect observations on average to be scaled up by λ. The corresponding invariance property on an estimator $T_n(\cdot)$ is that $T_n(\lambda \cdot x_1,...,\lambda \cdot x_n) = \lambda T_n(x_1,...,x_n)$.

To illustrate the use of invariance conditions, consider the example of a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from an exponential distribution with an unknown scale parameter λ, with the DGP

density $f(\mathbf{x},\lambda) = \lambda^{-n}\exp(-(x_1 + ... + x_n)/\lambda)$. Then $y = (x_1 + ... + x_n)/n$ is sufficient and we need consider only estimators $T_n(y)$. Invariance with respect to scale implies $T_n(y) = yT_n(1)$. Invariance with sample scale requires that if $x_1 = ... = x_n = 1$, so that $y = 1$, then $T_n(1) = T_1(1)$. Combining these conditions, $T_n(y) = yT_1(1)$, so that an estimator that is a function of the sufficient statistic and has these invariance properties must be proportional to the sample mean, with a proportion independent of sample size.

6.1.11. The next group of properties refer to the limiting behavior of estimators in a sequence of larger and larger samples, and are sometimes called *asymptotic properties*. The rationale for employing these properties is that when one is working with a large sample, then properties that hold in the limit will also hold, approximately, for this sample. The reason for considering such properties at all, rather than concentrating on the sample you actually have, is that one can use these approximate properties to choose among estimators in situations where the exact finite sample property cannot be imposed or is analytically intractable to work out.

Application of asymptotic properties raises several conceptual and technical issues. The first question is what it would mean to increase sample size indefinitely, and whether various methods that might be used to define this limit correspond to approximations that are likely to be relevant to a specific problem. There is no ambiguity when one is drawing simple random samples from an infinite population. However, if one samples from a finite population, a finite sequence of samples of increasing size will terminate in a complete census of the population. While one could imagine sampling with replacement and drawing samples that are larger than the population, it is not obvious why estimators that have some reasonable properties in this limit are necessarily appropriate for the finite population. Put another way, it is not obvious that this limit provides a good approximation to the finite sample.

The issue of the appropriate asymptotic limit is particularly acute for time series. One can imagine extending observations indefinitely through time. This may provide approximations that are appropriate in some situations for some purposes, but not for others. For example, if one is trying to estimate the timing of a particular event, a local feature of the time series, it is questionable that extending the time series indefinitely into the past and future leads to a good approximation to the statistical properties of the estimator of the timing of an event. Other ways of thinking of increasing sample sizes for time series, such as sampling from more and more "parallel" universes, or sampling at shorter and shorter intervals, have their own idiosyncrasies that make them questionable as useful approximations.

A second major issue is how the sequence of estimators associated with various sample sizes is defined. A conceptualization introduced in Chapter 5 defines an estimator to be a functional of the empirical CDF of the data, $T(F_n)$. Then, it is natural to think of $T(F(\cdot,\theta))$ as the limit of this sequence of estimators, and the Glivenko-Cantelli theorem stated in Chapter 5.1 establishes an approximation property that the estimator $T(F_n)$ converges almost surely to $T(F(\cdot,\theta))$ if the latter exists. This suggests that defining estimators as “continuous” functions of the CDF leads to a situation in which the asymptotic limit will have reasonable approximation properties in large samples. However, tt is important to avoid reliance on asymptotic arguments when it is clear that the asymptotic approximation is irrelevant to the behavior of the estimator in the range of sample sizes actually

encountered. Consider an estimation procedure which says "Ignore the data and estimate θ_o to be zero in all samples of size less than 10 billion, and for larger samples employ some computationally complex but statistically sound estimator". This procedure may technically have good asymptotic properties, but this approximation obviously tells you nothing about the behavior of the estimator in economic sample sizes of a few thousand observations.

6.1.12. *Consistency*. A sequence of estimators $T_n(\mathbf{x}) = T_n(x_1,...,x_n)$ for samples of size n are *consistent* for θ_o if the probability that they are more than a distance $\varepsilon > 0$ from θ_o goes to zero as n increases; i.e., $\lim_{n\to\infty} P(|T_n(x_1,...,x_n) - \theta_o| > \varepsilon) = 0$. In the terminology of Chapter 4, this is *weak convergence* or *convergence in probability,* written $T_n(x_1,...,x_n) \to_p \theta_o$. One can also talk about *strong consistency*, which holds when $\lim_{n\to\infty} P(\sup_{m\geq n}|T_m(x_1,...,x_{n'}) - \theta_o| > \varepsilon) = 0$, and corresponds to almost sure convergence, $T_n(x_1,...,x_n) \to_{as} \theta_o$.

6.1.13. *Asymptotic Normality*. A sequence of estimators $T_n(\cdot)$ for samples of size n are *consistent asymptotically normal* (CAN) for θ if there exists a sequence r_n of scaling constants such that $r_n \to +\infty$ and $r_n\cdot(T_n(\mathbf{x}_n) - \theta)$ converges in distribution to a normally distributed random variable with some mean $\mu = \mu(\theta)$ and variance $\sigma^2 = \sigma(\theta)^2$. If $\Psi_n(t)$ is the CDF of $T_n(\mathbf{x}_n)$, then $Q_n = r_n\cdot(T_n(\mathbf{x}_n) - \theta)$ has the CDF $P(Q_n \leq q) = \Psi_n(\theta + q/r_n)$. From Chapter 4, one will have convergence in distribution to a normal, $r_n(T_n(\mathbf{x}_n) - \theta) \to_d Z$ with $Z \sim N(\mu,\sigma^2)$, if and only if for each q, the CDF of Q_n satisfies $\lim_{n\to\infty} |\Psi_n(\theta + q/r_n) - \Phi((q-\mu)/\sigma)| = 0$. This is the conventional definition of convergence in distribution, with the continuity of the normal CDF Φ permitting us to state the condition without excepting jump points in the limit distribution. In this setup, $\Psi_n(t)$ is converging in distribution to $\mathbf{1}(t\geq\theta)$, the CDF of the constant random variable equal to θ. However, r_n is blowing up at just the right rate so that $\Psi_n(\theta + q/r_n)$ has a non-degenerate asymptotic distribution, whose shape is determined by the local shape of Ψ_n in shrinking neighborhoods of θ. Asymptotic normality is used to approximate points of the CDF of $T_n(\mathbf{x}_n)$ in a large but finite sample by using the fact that $\Psi_n(t_n) \approx \Phi((r_n(t_n - \theta) - \mu)/\sigma)$. The mean μ is termed the *asymptotic bias*, and σ^2 is termed the *asymptotic variance*. If $\mu = 0$, the estimator is said to be *asymptotically unbiased*. An unbiased estimator will be asymptotically unbiased, but the reverse is not necessarily true. Often, when a sequence of estimators is said to be asymptotically normal, asymptotic unbiasedness is taken to be part of the definition unless stated explicitly to the contrary. The scaling term r_n can be taken to be $n^{1/2}$ in almost all finite-parameter problems, and unless it is stated otherwise, you can assume that this is the scaling that is being used. When it is important to make this distinction clear, one can speak of *Root-n consistent asymptotically normal* (RCAN) sequences of estimators.

Convergence in distribution to a normal is a condition that holds pointwise for each true parameter θ. One could strengthen the property by requiring that this convergence be uniform in θ; i.e., by requiring for each $\varepsilon > 0$ and q that there be a sample size $n(\varepsilon,q)$ beyond which $\sup_\theta |\Psi(\theta_o + q/r_n) - \Phi((q-\mu(\theta_o))/\sigma(\theta_o))| < \varepsilon$. If this form of convergence holds, and in addition $\mu(\theta)$ and $\sigma(\theta)^2$ are continuous functions of θ, then the estimator is said to be *consistent uniformly asymptotically normal* (CUAN).

6.1.14. *Asymptotic Efficiency*. Consider a family ***T*** of sequences of estimators $T_n(\cdot)$ that are CUAN for a parameter θ and have asymptotic bias $\mu(\theta) \equiv 0$. An estimator $T^*(\cdot)$ is *asymptotically efficient* relative to class ***T*** if its asymptotic variance is no larger than that of any other member of the family. The reason for restricting attention to the CUAN class is that in the absence of uniformity, there exist "super-efficient" estimators, constructed in the following way: Suppose $T_n(\cdot)$ is an asymptotically efficient estimator in the CUAN class. For an arbitrary θ^*, define $T_n^*(\cdot)$ to equal $T_n(\cdot)$ if $n^{1/2}|T_n(\mathbf{x}) - \theta^*| \geq 1$, and equal to θ^* otherwise. This estimator will have the same asymptotic variance as $T_n(\cdot)$ for fixed $\theta \neq \theta^*$, and an asymptotic variance of zero for $\theta = \theta^*$. Thus, it is more efficient. On the other hand, it has a nasty asymptotic bias for parameter vectors that are "local" to θ^*, so that it is not CUAN, and would be an unattractive estimator to use in practice. Once these non-uniform superefficient estimators are excluded by restricting attention to the CUAN class, one has the result that under reasonable regularity conditions, an asymptotic version of the Cramer-Rao bound for unbiased estimators holds for CUAN estimators.

6.1.15. *Asymptotic sufficiency*. In some problems, sufficiency does not provide a useful reduction of dimension in finite samples, but a weaker "asymptotic" form of sufficiency will provide useful restrictions. This could arise if the DGP density can be written $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$ for a low-dimensional statistic $\mathbf{y}$, but both g_1 and g_2 depend on θ so $\mathbf{y}$ is not sufficient. However, $g_2(\mathbf{z}|\mathbf{y},\theta)$ may converge in distribution to a density that does not depend on θ. Then, there is a large sample rationale for concentrating on estimators that depend only on $\mathbf{y}$.

6.2. GENERAL ESTIMATION CRITERIA

6.2.1. It is useful to have some general methods of generating estimators that as a consequence of their construction will have some desirable statistical properties. Such estimators may prove adequate in themselves, or may form a starting point for refinements that improve statistical properties. We introduce several such methods:

6.2.2. *Analogy Estimators*. Suppose one is interested in a feature of a target population that can be described as a functional of its CDF $F(\cdot)$, such as its mean, median, or variance, and write this feature as $\theta = \mu(F)$. An analogy estimator exploits the similarity of a population and of a simple random sample drawn from this population, and forms the estimator $T(\mathbf{x}) = \mu(F_n)$, where μ is the functional that produces the target population feature and F_n is the empirical distribution function. For example, a sample mean will be an analogy estimator for a population mean.

6.2.3. *Moment Estimators*. Population moments will depend on the parameter index in the underlying DGP. This is true for ordinary moments such as means, variances, and covariances, as well as more complicated moments involving data transformations, such as quantiles. Let $m(x)$ denote a function of an observation and $\mathbf{E}_{x|\theta}m(x) = \gamma(\theta)$ denote the population moment formed by taking the expectation of $m(x)$. In a sample $\mathbf{x} = (x_1,...,x_n)$, the idea of a moments estimator is to form

a sample moment $n^{-1}\sum_{i=1}^{n} m(x_i) \equiv \mathbf{E}_n m(x)$, and then to use the analogy of the population and sample moments to form the approximation $\mathbf{E}_n m(x) \approx \mathbf{E}_{x|\theta} = \gamma(\theta)$. The sample average of a function m(x) of an observation can also be interpreted as its expectation with respect to the empirical distribution of the sample; we use the notation $\mathbf{E}_n m(x)$ to denote this empirical expectation. The moment estimator T(**x**) solves $\mathbf{E}_n m(x) = \gamma(T(\mathbf{x}))$. When the number of moment conditions equals the number of parameters, an exact solution is normally obtainable, and T(**x**) is termed a *classical method of moments estimator*. When the number of moment conditions exceeds the number of parameters, it is not possible in general to find T(**x**) that sets them all to zero at once. In this case, one may form a number of linear combinations of the moments equal to the number of parameters to be estimated, and find T(**x**) that sets these linear combinations to zero. The linear combinations in turn may be derived starting from some metric that provides a measure of the distance of the moments from zero, with T(**x**) interpreted as a minimand of this metric. This is called *generalized method of moments* estimation.

6.2.4. *Maximum likelihood estimators*. Consider the DGP density f(**x**,θ) for a given sample as a function of θ. The maximum likelihood estimator of the unknown true value θ is the statistic T(**x**) that maximizes f(**x**,θ). The intuition behind this estimator is that if we guess a value for θ that is far away from the true θ_o, then the probability law for this θ would be very unlikely to produce the data that are actually observed, whereas if we guess a value for θ that is near the true θ_o, then the probability law for this θ would be likely to produce the observed data. Then, the T(**x**) which maximized this likelihood, as measured by the probability law itself, should be close to the true θ. The maximum likelihood estimator plays a central role in classical statistics, and can be motivated solely in terms of its desirable classical statistical properties in large samples.

When the data are a sample of n independent observations, each with density $f(x,\theta)$, then the likelihood of the sample is $f(\mathbf{x},\theta) = \prod_{i=1}^{n} f(x_i,\theta)$. It is often convenient to work with the logarithm of the density, $l(x,\theta) \equiv \text{Log } f(x,\theta)$. Then, the *Log Likelihood* of the sample is $L(\mathbf{x},\theta) \equiv \text{Log } f(\mathbf{x},\theta) = \sum_{i=1}^{n} l(x_i,\theta)$. The *maximum likelihood estimator* is the function t = T(**x**) of the data that when substituted for θ maximizes f(**x**,θ), or equivalently L(**x**,θ).

The gradient of the log likelihood of an observation with respect to θ is denoted $s(\mathbf{x},\theta) \equiv \nabla_\theta l(x,\theta)$, and termed the *score*. The maximum likelihood estimator is a zero of the sample expectation of the score, $\mathbf{E}_n s(\mathbf{x},T(\mathbf{x}))$. Then, the maximum likelihood estimator is a special case of a moments estimator.

Maximum likelihood estimators will under quite general regularity conditions be consistent and asymptotically normal. Under uniformity conditions that rule out some odd non-uniform "super-efficient" alternatives, they are also asymptotically efficient. They often have good finite-sample properties, or can be easily modified so that they do. However, their finite-sample properties have to be determined on a case-by-case basis. In multiple parameter problems, particularly when there

are primary parameters α and nuisance parameters β, the maximum likelihood principle can sometimes be used to handle the nuisance parameters. Specifically, maximum likelihood estimation for all parameters will find the parameter values that solve $\max_{\alpha,\beta} L(\mathbf{x},\alpha,\beta)$. But one could get the same solution by first maximizing in the nuisance parameters β, obtaining a solution $\beta = r(x,\alpha)$, and substituting this back into the likelihood function to obtain $L(x,\alpha,r(x,\alpha))$. This is called the concentrated likelihood function, and it can now be maximized in α alone. The reason this can be an advantage is that one may be able to obtain $r(x,\alpha)$ "formally" without having to compute it.

6.3. ESTIMATION IN NORMALLY DISTRIBUTED POPULATIONS

6.3.1. Consider a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from a population in which observations are normally distributed with mean μ and variance σ^2. Let $\varphi(v) = (2\pi)^{-1/2}\exp(-v^2/2)$ denote the standard normal density. Then the density of observation x_i is $\varphi((x_i - \mu)/\sigma)/\sigma$. The log likelihood of the sample is

$$L(\mathbf{x},\mu,\sigma^2) = -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log}\,\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2.$$

We will find estimates μ_e and σ_e^2 for the parameters μ and σ^2 using the maximum likelihood method, and establish some of the statistical properties of these estimators.

6.3.2. The first-order-conditions for maximizing $L(\mathbf{x},\mu,\sigma^2)$ in μ and σ^2 are

$$0 = \sum_{i=1}^{n}(x_i-\mu)/\sigma^2 \Longrightarrow \mu_e = \bar{x} \equiv n^{-1}\sum_{i=1}^{n} x_i,$$

$$0 = -n/2\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^4 \Longrightarrow \sigma_e^2 = n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

The maximum likelihood estimator of μ is then the sample mean, and the maximum likelihood estimator of σ^2 is the sample variance. Define $s^2 = \sigma_e^2\cdot n/(n-1) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$, the sample variance with a sample size correction. The following result summarizes the properties of these estimators:

Theorem 6.3. If $\mathbf{x} = (x_1,...,x_n)$ is a simple random sample from a population in which observations are normally distributed with mean μ and variance σ^2, then

(1) $(\bar{x},s^2)$ are joint minimal sufficient statistics for (μ,σ^2)**.**
(2) $\bar{x}$ is an unbiased estimator for μ, and s^2 an unbiased estimator for σ^2**.**
(3) $\bar{x}$ is a Minimum Variance Unbiased Estimator (MVUE) for μ**;** s^2 is MVUE for σ^2**.**
(4) $\bar{x}$ is Normally distributed with mean μ and variance σ^2/n.
(5) $(n-1)s^2/\sigma^2$ has a Chi-square distribution with n-1 degrees of freedom.

(6) $\bar{x}$ and s^2 are statistically independent.
(7) $n^{1/2}(\bar{x} - \mu)/s$ has a Student's-T distribution with n-1 degrees of freedom.
(8) $(\bar{x} - \mu)^2/s^2$ has an F-distribution with 1 and n-1 degrees of freedom.

Proof: (1) Factor the log likelihood function as

$$L(\mathbf{x},\mu,\sigma^2) = -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log}\,\sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2/\sigma^2$$

$$= -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log}\,\sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n}(\bar{x}-\mu)^2/\sigma^2$$

$$= -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log}\,\sigma^2 - \frac{1}{2}\cdot\frac{(n-1)s^2}{\sigma^2} - \frac{n}{2}(\bar{x}-\mu)^2/\sigma^2 .$$

This implies that $\bar{x}$ and s^2 are jointly sufficient for μ and σ^2. Because the dimension of $(\bar{x},s^2)$ is the same as the dimension of (μ,σ^2), they are obviously minimal sufficient statistics.

(2) The expectation of $\bar{x}$ is $\mathbf{E}\,\bar{x} = n^{-1}\sum_{i=1}^{n}\mathbf{E}x_i = \mu$, since the expectation of each observation is μ. Hence $\bar{x}$ is unbiased. To establish the expectation of s^2, first form the n×n matrix $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$, where $\mathbf{I}_n$ is the n×n identity matrix and $\mathbf{1}_n$ is a n×1 vector of ones. The matrix $\mathbf{M}$ is idempotent (check) and its trace satisfies $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{1}_n\mathbf{1}_n'/n) = n - \text{tr}(\mathbf{1}_n'\mathbf{1}_n/n) = n - 1$. The result then follows from Theorem 3.11 (viii). For a direct demonstration, let $Z' = (x_1-\mu,...,x_n-\mu)$ denote the vector of deviations of observations from the population mean. This vector contains independent identically distributed normal random variables with mean zero and variance σ^2, so that $\mathbf{E}ZZ' = \sigma^2\mathbf{I}_n$. Further, $Z'\mathbf{M} = (x_1 - \bar{x},...,x_n - \bar{x})$ and $s^2 = Z'\mathbf{M}\cdot\mathbf{M}Z/(n-1) = Z'\mathbf{M}Z/(n-1)$. Therefore, $\mathbf{E}s^2 = \mathbf{E}(Z'\mathbf{M}Z)/(n-1) = \mathbf{E}\,\text{tr}(Z'\mathbf{M}Z)/(n-1) = \mathbf{E}\,\text{tr}(\mathbf{M}ZZ')/(n-1) = \text{tr}(\mathbf{M}\cdot\mathbf{E}(ZZ'))/(n-1) = \sigma^2\cdot\text{tr}(\mathbf{M})/(n-1) = \sigma^2$. Hence, s^2 is unbiased.

(3) The MVUE property of $\bar{x}$ and s^2 is most easily proved by application of the Blackwell theorem. We already know that these estimators are unbiased. Any other unbiased estimator of μ then has the property that the difference of this estimator and $\bar{x}$, which we will denote by $h(\mathbf{x})$, must satisfy $\mathbf{E}h(\mathbf{x}) \equiv 0$. Alternately, $h(\mathbf{x})$ could be the difference of s^2 and any other unbiased estimator of σ^2. We list a series of conditions, and then give the arguments that link these conditions.

(a) $$0 \equiv \mathbf{E}h(\mathbf{x}) \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\exp(L(\mathbf{x},\mu,\sigma^2))d\mathbf{x} \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}.$$

(b) $$0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\left[\sum_{i=1}^{n}(x_i-\mu)\right]\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

$$\equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\bar{x}\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

$$(c)\ 0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\bar{x}\ \left[\sum_{i=1}^{n}(x_i-\mu)\right]\ \exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

$$\equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\bar{x}^2\ \exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

$$\equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot[\bar{x}-\mu]^2\ \exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

$$(d)\ 0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\sum_{i=1}^{n}(x_i-\mu)^2\cdot\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

$$\equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\cdot\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\cdot\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right)d\mathbf{x}$$

Condition (a) is the statement that h(x) has expectation zero, and the second form is obtained by striking terms that can be taken outside the integral. Differentiate the last form of (a) with respect to μ and strike out terms to get condition (b). The second form of (b) is obtained by using (a) to show that the second part of the term in brackets makes a contribution that is zero. Differentiate the last form of (b) with respect to μ and strike out terms to get (c). The second form of (c) is obtained by using (a) to simplify the term in brackets. The last form of (c) follows by expanding the squared term and applying (a) and (b). Condition (d) is obtained by differentiating the last form of (a) with respect to σ^2 and striking out terms. Write $x_i - \mu = x_i - \bar{x} + \bar{x} - \mu$, expand the square, and apply (c) to obtain the last form of (d). Condition (b) implies $\mathbf{E}h(\mathbf{x})\cdot\bar{x} \equiv 0$, and condition (d) implies $\mathbf{E}h(\mathbf{x})\cdot s^2 = 0$. Then, the estimators $\bar{x}$ and s^2 are uncorrelated with any unbiased estimator of zero. The Blackwell theorem then establishes that they are the unique minimum variance estimators among all unbiased estimators.

(4) Next consider the distribution of $\bar{x}$. We use the fact that linear transformations of multivariate normal random vectors are again multivariate normal: If $Z \sim N(\boldsymbol{\mu},\boldsymbol{\Omega})$ and $W = \mathbf{C}Z$, then $W \sim N(\mathbf{C}\boldsymbol{\mu},\mathbf{C}\boldsymbol{\Omega}\mathbf{C}')$. This result holds even if Z and W are of different dimensions, or **C** is of less than full rank. (If the rank of $\mathbf{C}\boldsymbol{\Omega}\mathbf{C}'$ is less than full, then the random variable has all its density concentrated on a subspace.) Now $\bar{x} = \mathbf{C}\mathbf{x}$ when $\mathbf{C} = (1/n,...,1/n)$. We have **x** multivariate normal with mean $\mathbf{1}_n\mu$ and covariance matrix $\sigma^2\mathbf{I}_n$, where $\mathbf{1}_n$ is a n×1 vector of ones and $\mathbf{I}_n$ is the n×n identity matrix. Therefore, $\bar{x} \sim N(\mu\mathbf{C}\mathbf{1}_n,\sigma^2\mathbf{C}\mathbf{C}') = N(\mu,\sigma^2/n)$.

(5) Next consider the distribution of s^2. Consider the quadratic form $(\mathbf{x}/\sigma)'\mathbf{M}(\mathbf{x}/\sigma)$, where **M** is the idempotent matrix $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ from (2). The vector $(\mathbf{x}/\sigma)$ is independent standard normal, so that Theorem 3.11(iii) gives the result.

(6) The matrices $\mathbf{C} = (1/n,...,1/n) = \mathbf{1}_n'$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ have $\mathbf{C}\mathbf{M} = \mathbf{0}$. Then Theorem 3.11(vii) gives the result that $\mathbf{C}(\mathbf{x}/\sigma) = \bar{x}/\sigma$ and $(\mathbf{x}/\sigma)'\mathbf{M}(\mathbf{x}/\sigma) = (n-1)s^2/\sigma^2$ are independent.

For (7), Use Theorem 3.9(ii), and for (8), use Theorem 3.9(iii). □

4. LARGE SAMPLE PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES

This section provides a brief and informal introduction to the statistical properties of maximum likelihood estimators and similar estimation methods in large samples. Consider a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from a population in which the density of an observation is $f(x,\theta_o)$. The DGP density or likelihood of the sample is then $f(\mathbf{x},\theta) = f(x_1,\theta)\cdot...\cdot f(x_n,\theta)$, with θ_o the true value of θ. The log likelihood of an observation is $l(x,\theta) = \log f(x,\theta_o)$, and the log likelihood of the sample is $L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} l(x_n,\theta)$. The maximum likelihood estimator $T_n(\mathbf{x})$ is a value of θ which maximizes $L_n(\mathbf{x},\theta)$. The first-order condition for this maximum is that the *sample score*,

$$\nabla_\theta L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} \nabla_\theta l(x_i,\theta),$$

equal zero at $\theta = T_n(\mathbf{x})$. The second order condition is that the *sample hessian* $\nabla_{\theta\theta} L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} \nabla_{\theta\theta} l(x_i,\theta)$, be negative at $\theta = T(\mathbf{x})$. When the parameter θ is more than one-dimensional, the second-order condition is that the sample hessian is a negative definite matrix.

Under very mild regularity conditions, the expectation of the score of an observation is zero at the true parameter vector. Start from the identity $\int_{-\infty}^{+\infty} \exp(l(x,\theta))\cdot dx \equiv 1$ and differentiate with respect to θ under the integral sign to obtain the condition $\int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta)\cdot\exp(l(x,\theta))\cdot dx \equiv 0$. (Regularity conditions are needed to assure that one can indeed differentiate under the integral; this will be supplied by assuming a dominance condition so that the Lebesgue dominated convergence theorem can be applied; see Theorem 3.1 and the discussion following.) Then, at the true parameter θ,one has $\mathbf{E}_{x|\theta}\nabla_\theta l(x,\theta) = 0$, the condition that the *population score* is zero when $\theta = \theta_o$. Another regularity condition requires that $E_{\boldsymbol{x}|\theta_o} \nabla_\theta l(x,\theta) = 0$ <u>only</u> if $\theta = \theta_o$; this has the interpretation of an *identification condition*. The maximum likelihood estimator can be interpreted as an analogy estimator that chooses $T_n(\mathbf{x})$ to satisfy a sample condition (that the sample score be zero) that is analogous to the population score condition. One could sharpen the statement of this analogy by writing the population score as an explicit function of the population DGP, $\mu(\theta,F(\cdot,\theta_o)) \equiv E_{\boldsymbol{x}|\theta_o} \nabla_\theta l(x,\theta)$, and writing the sample score as $\mu(\theta,F_n) \equiv \mathbf{E}_n\nabla_\theta l(x,\theta)$, where "$\mathbf{E}_n$" stands for empirical expectation, or sample average. The mapping $\mu(\theta,\cdot)$ is linear in its second argument, and this is enough to assure that it is continuous (in an appropriate sense) in this argument. Then one has almost sure convergence of $\mu(\theta,F_n)$ to $\mu(\theta,F(\cdot,\theta_o))$ for each θ, from the Glivenko-Cantelli theorem. A few additional regularity conditions are enough to ensure that this convergence is uniform in θ, and that a solution $T_n(\mathbf{x})$ that sets the sample score to zero converges almost surely to the value θ_o that sets the population score to zero.

The basic large sample properties of maximum likelihood estimators are that, subject to suitable regularity conditions, T_n converges in probability to the true parameter vector θ_o, and $n^{1/2}(T_n - \theta_o)$ converges in distribution to a normal random variable with mean zero and a variance which achieves the Cramer- Rao bound for an unbiased estimator. These results imply that in large samples, T_n will become a more and more precise estimate of the true parameter. Further, the convergence in distribution to a Normal permits one to use the properties of a Normal population to construct approximate hypothesis tests and confidence bounds, and get approximations for significance levels and power whose accuracy increases with sample size. The achievement of the Cramer-Rao lower bound on variance indicates that in large samples there are no alternative estimators which are uniformly more precise, so MLE is the "best" one can do.

We next list a series of regularity conditions under which the results stated above can be shown to hold. Only the single parameter case will be presented. However, the conditions and results have direct generalizations to the multiple parameter case. This list is chosen so the conditions are easy to interpret and to check in applications. Note that these are conditions on the population DGP, not on a specific sample. Hence, "checking" means verifying that your model of the DGP and your assumptions on distributions of random variables are <u>logically</u> consistent with the regularity conditions. They cannot be verified empirically by looking at the data, but it is often possible to set up and carry out empirical tests that may allow you to conclude that some of the regularity conditions fail. There are alternative forms for the regularity conditions, as well as weaker conditions, which give the same or similar limiting results. The regularity conditions are quite generic, and will be satisfied in many economic applications. However, it is a serious mistake to assume without checking that the DGP you assume for your problem is consistent with these conditions. While in most cases the mantra "I assume the appropriate regularity conditions" will work out, you can be acutely embarrassed if your DGP happens to be one of the exceptions that is logically inconsistent with the regularity conditions, particularly if it results in estimators that fail to have desirable statistical properties. Here are the conditions:

A.1. There is a single parameter θ which is permitted to vary in a closed bounded subset $\boldsymbol{\Theta}$. The true value θ_o is in the interior of $\boldsymbol{\Theta}$.

A.2. The sample observations are realizations of independently identically distributed random variables $x_1,...,x_n$, with a common density $f(x,\theta_o)$.

A.3. The density $f(x,\theta)$ is continuous in θ, and three times continuously differentiable in θ, for each x, and is "well behaved" (e.g., measurable or piecewise continuous or continuous) in x for each θ.

A.4. There exists a bound $\beta(x)$ on the density and its derivatives which is uniform in θ and satisfies $|l(x,\theta)| \leq \beta(x)$, $(\nabla_\theta l(x,\theta))^2 \leq \beta(x)$, $|\nabla_{\theta\theta} l(x,\theta)| \leq \beta(x)$, $|\nabla_{\theta\theta\theta} l(x,\theta)| \leq \beta(x)$, and

$$\int_{-\infty}^{+\infty} \beta(x)^2 f(x|\theta_o)dx < +\infty.$$ (Then, $\beta(x)$ is a dominating, square-integrable function.)

A.5 The function $\lambda(\theta) = \mathbf{E}_{x|\theta}\, l(x,\theta)$ has $\lambda(\theta) < \lambda(\theta_o)$ and $\nabla_\theta \lambda(\theta) \neq 0$ for $\theta \neq \theta_o$ and $J = -\nabla_{\theta\theta}\lambda(\theta_o) > 0$.

The expression J in A.5 is termed the *Fisher information* in an observation. The first two assumptions mostly set the problem. The restriction of the parameter to a closed bounded set guarantees that a MLE exists, and can be relaxed by adding conditions elsewhere. Requiring θ_o interior to Θ guarantees that the first-order condition $\mathbf{E}_n\nabla_\theta l(x,T_n(\cdot)) = 0$ for a maximum holds for large n, rather than an inequality condition for a maximum at a boundary. This really matters because MLE at boundaries can have different asymptotic distributions and rates of convergence than the standard $n^{1/2}$ rate of convergence to the normal. The continuity conditions A.3 are satisfied for most economic problems, and in some weak form are critical to the asymptotic distribution results. Condition A.4 gives bounds that permit exchange of the order of differentiation and integration in forming expectations with respect to the population density. Condition A.5 is an identification requirement which implies there cannot be a parameter vector other than θ_o that on average always explains the data as well as θ_o.

The next result establishes that under these regularity conditions, a MLE is consistent and asymptotically normal (CAN):

Theorem 6.4. If A.1-A.5 hold, then a maximum likelihood estimator T_n satisfies
(1) T_n is consistent for θ_o.
(2) T_n is asymptotically normal: $n^{1/2}(T_n(\mathbf{x}) - \theta_o) \rightarrow_d Z_o \sim N(0,J^{-1})$, with J equal to the Fisher information in an observation, $J = E_{x|\theta_o} \nabla_\theta l(x,\theta_o)^2$.

(3) $\mathbf{E}_n[\nabla_\theta l(x,T_n)]^2 \rightarrow_p J$ and $-\mathbf{E}_n\nabla_{\theta\theta} l(x,T_n) \rightarrow_p J$.
(4) Suppose T_n' is any sequence of estimators that solve equations of the form $\mathbf{E}_n g(x,\theta) = 0$, where g is twice continually differentiable and satisfies $E_{x|\theta_o} g(x,\theta) = 0$ if and only if $\theta = \theta_o$;

uniform bounds $|g(x,\theta)| \leq \beta(x)$, $|\nabla_\theta g(y,\theta)^2| \leq \beta(x)$, $|\nabla_{\theta\theta} g(x,\theta)| \leq \beta(x)$, where $\mathbf{E}\beta(x)^2 < +\infty$; and $R = -\mathbf{E}\nabla_\theta g(y,\theta_o) \neq 0$. Let $S = \mathbf{E}g(x,\theta_o)^2$. Then $T_n' \rightarrow_p \theta_o$ and $n^{1/2}(T_n' - \theta^*) \rightarrow_d Z_1 \sim N(0,V)$, where $V = R^{-1}SR'^{-1}$. Further, $V \geq J^{-1}$, so that the MLE T_n is efficient relative to T_n'. Further, Z_0 and Z_1 have the covariance property $\text{cov}(Z_0, Z_1 - Z_0) = 0$.

Result (2) in this theorem implies that to a good approximation in large samples, the estimator T_n is normal with mean θ_o and variance $(nJ)^{-1}$, where J is the Fisher information in an observation. Since this variance is the Cramer-Rao bound for an unbiased estimator, this also suggests that one is not going to be able to find other estimators that are also unbiased in this approximation sense and which have lower variance. Result 3 gives two ways of estimating the asymptotic variance J^{-1} consistently, where we use the fact that J^{-1} is a continuous function of J for $J \neq 0$, so that it can be estimated consistently by the inverse of a consistent estimator of J.. Result (4) establishes that MLE is efficient relative to a broad class of estimators called *M-estimators*.

Proof: An intuitive demonstration of the Theorem will be given rather than formal proofs. Consider first the consistency result. The reasoning is as follows. Consider the expected likelihood of an observation,

$$\lambda(\theta) \equiv E_{x|\theta_o}\, l(x,\theta) = \int_{-\infty}^{+\infty} l(x,\theta) f(x,\theta_o) dx.$$

We will argue that $\lambda(\theta)$ has a unique maximum at θ_o. Then we will argue that any function which is uniformly very close to $\lambda(\theta)$ must have its maximum near θ_o. Finally, we argue by applying a uniform law of large numbers that the likelihood function is with probability approaching one uniformly very close to λ for n sufficiently large. Together, these results will imply that with probability approaching one, T_n is close to θ_o for n large.

Assumption A.4 ensures that $\lambda(\theta)$ is continuous, and that one can reverse the order of differentiation and integration to obtain continuous derivatives

$$\nabla_\theta \lambda(\theta) \equiv \int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta) f(x,\theta_o) dx \equiv E_{x|\theta_o}\, \nabla_\theta l(x,\theta)$$

$$\nabla_{\theta\theta} \lambda(\theta) \equiv \int_{-\infty}^{+\infty} \nabla_{\theta\theta} l(x,\theta) f(x,\theta_o) dx \equiv E_{x|\theta_o}\, \nabla_{\theta\theta} l(x,\theta)$$

Starting from the identity

$$1 \equiv \int_{-\infty}^{+\infty} f(x,\theta) dx \equiv \int_{-\infty}^{+\infty} e^{l(x,\theta)} dx,$$

one obtains by differentiation

$$0 \equiv \int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta) e^{l(x,\theta)} dx$$

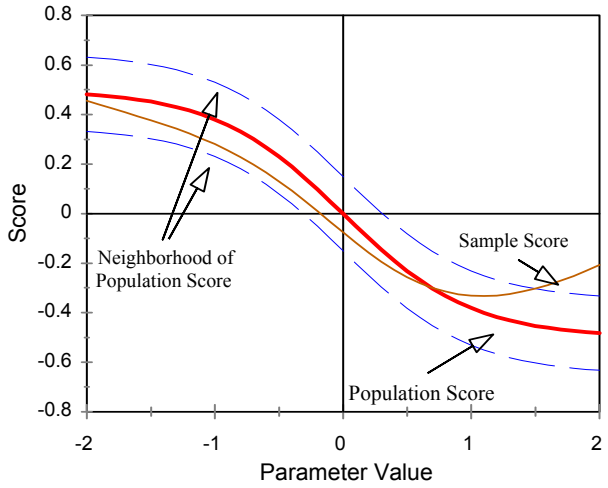
$$0 \equiv \int_{-\infty}^{+\infty} [\nabla_{\theta\theta} l(x,\theta) + \nabla_\theta l(x,\theta)^2] e^{l(x,\theta)} dx$$

Evaluated at θ_o, these imply $0 = \nabla_\theta \lambda(\theta_o)$ and $-\nabla_{\theta\theta}\lambda(\theta_o) = E_{x|\theta_o}\, \nabla_\theta l(x,\theta)^2 = J$.

Assumption A.5 requires further that $J \neq 0$, and that θ_o is the only root of $\nabla_\theta \lambda(\theta)$. Hence, $\lambda(\theta)$ has a unique maximum at θ_o, and at no other θ satisfies a first-order condition or boundary condition for a local maximum.

We argue next that any function which is close enough to $\nabla_\theta \lambda(\theta)$ will have at least one root near θ_o and no roots far away from θ_o. The figure below graphs $\nabla_\theta \lambda(\theta)$, along with a "sleeve" which is a vertical distance δ from $\nabla_\theta \lambda$. Any function trapped in the sleeve must have at least one root between $\theta_o - \varepsilon_1$ and $\theta_o + \varepsilon_2$, where $[\theta_o - \varepsilon_1, \theta_o + \varepsilon_2]$ is the interval where the sleeve intersects the axis, and must have no roots outside this interval. Furthermore, the uniqueness of the root θ_o of $\nabla_\theta \lambda(\theta)$ plus the condition $\nabla_{\theta\theta}\lambda(\theta_o) < 0$ imply that as δ shrinks toward zero, so do ε_1 and ε_2. In the graph, the sample score intersects the axis within the sleeve, but for parameter values near two is outside the sleeve. The last step in the consistency argument is to show that with probability approaching one the sample score will be entirely contained within the sleeve; i.e., that $L_n(\mathbf{x},\theta)$ is with probability approaching one contained in a δ-sleeve around $\lambda(\theta)$. For fixed θ, $L_n(\mathbf{x},\theta) = l(x_i,\theta)$ is a sample average of i.i.d. random variables $l(x,\theta)$ with mean $\lambda(\theta)$. Then Kolmogorov's SLLN implies $L_n(\mathbf{x},\theta)$

$\rightarrow_{as} \lambda(\theta)$. This is not quite enough, because there is a question of whether $L_n(\mathbf{x},\theta)$ could converge non-uniformly to $\lambda(\theta)$, so that for any n there are some values of θ where $L_n(\mathbf{x},\theta)$ is outside the sleeve. However, assumptions A.1, A.3, and A.4 imply $\max_{\theta\in\Theta} |L_n(\mathbf{x},\theta) - \lambda(\theta)| \rightarrow_{as} 0$. This follows in particular because the differentiability of $f(x,\theta)$ in θ from A.3 and the bound on $\nabla_\theta l(x,\theta)$ from A.4 imply that $l(\cdot,\theta)$ is almost surely continuous on the compact set Θ, so that the uniform SLLN in Chapter 4.5 applies. This establishes that $T_n \rightarrow_{as} \theta$.



We next demonstrate the asymptotic normality of T_n. A Taylor's expansion about θ of the first-order condition for maximization of the log likelihood function gives

$$(1) \quad 0 = \nabla_\theta L_n(T_n) = \nabla_\theta L_n(\theta) + \nabla_{\theta\theta} L_n(\theta)\cdot(T_n-\theta) + \nabla_{\theta\theta\theta} L_n(T_{an})\cdot(T_n-\theta)^2/2\,,$$

where T_{an} is some point between T_n and θ. Define the quantities

$$B_n = n^{-1}\sum_{i=1}^n \nabla_\theta l(y_i,\theta),\ C_n = n^{-1}\sum_{i=1}^n \nabla_{\theta\theta} l(y_i,\theta),\ D_n = n^{-1}\sum_{i=1}^n \nabla_{\theta\theta\theta} l(y_i,T_{an})$$

Multiply equation (1) by $n^{1/2}/(1+n^{1/2}|T_n-\theta|)$ and let $Z_n = n^{1/2}(T_n-\theta)/(1 + n^{1/2}|T_n-\theta|)$. Then, one gets

$$0 = n^{1/2}B_n/(1+n^{1/2}|T_n-\theta|) + C_n Z_n + D_n Z_n(T_n-\theta)/2.$$

We make a limiting argument on each of the terms. First, the $\nabla_\theta l(y_i,\theta_o)$ are i.i.d. random variables with $\mathbf{E}\nabla_\theta l(y_i,\theta_o) = \nabla_\theta\lambda(\theta_o) = 0$ and $\mathbf{E}[\nabla_\theta l(y_i,\theta_o)]^2 = -\mathbf{E}\nabla_{\theta\theta}\lambda(\theta_o) = J$. Hence the Lindeberg-Levy CLT implies $B_n \rightarrow_d W_o \sim N(0,J)$. Second, $\nabla_{\theta\theta} l(Y_i,\theta_o)$ are i.i.d. random variables with $\mathbf{E}\nabla_{\theta\theta} l(Y_i,\theta_o) = -J$. Hence the Khinchine WLLN implies $C_n \rightarrow_p -J$. Third, $|D_n| \le n^{-1}\sum_{i=1}^n |\nabla_{\theta\theta\theta} l(y_i,T_{an})| \le$

$n^{-1}\sum_{i=1}^{n}\ \beta(y_i) \rightarrow_p \mathbf{E}\beta(Y) < +\infty$, by A.4 and Khinchine's WLLN, so that $|D_n|$ is stochastically bounded. Furthermore, $|Z_n| \le 1$, implying $Z_n = O_p(1)$. Since T_n is consistent, $(T_n - \theta_o) = o_p(1)$. Therefore, by rule 6 in Figure 4.3, $D_nZ_n(T_n - \theta_o)/2 = o_p(1)$.

Given $J/2 > \varepsilon > 0$, these arguments establish we can find n_o such that for $n > n_o$ with probability at least $1-\varepsilon$, we have $|D_nZ_n(T_n - \theta_o)/2| < \varepsilon$, $|C_n+J| < \varepsilon$ and $|B_n| < M$ for a large constant M (since $B_n \rightarrow_d W_o \Rightarrow B_n$ implies $O_p(1)$). In this event, $|C_n| > J-\varepsilon$, $|B_n + C_n n^{1/2}(T_n-\theta_o)| < \varepsilon(1 + n^{1/2}\cdot|T_n-\theta_o|)$, and $|B_n| \le M$ imply $|C_n|n^{1/2}|T_n-\theta_o| - |B_n| \le |B_n+C_nn^{1/2}|T_n-\theta_o)| < \varepsilon\cdot(1 - n^{1/2}\cdot|T_n-\theta_o|)$. This implies the inequality $(J - 2\varepsilon)n^{1/2}\cdot|T_n-\theta_o| < M + \varepsilon$. Therefore $n^{1/2}(T_n-\theta_o) = O_p(1)$; i.e., it is stochastically bounded. Therefore, by rule 6 in Figure 3.3, multiplying (2) by $1 + n^{1/2}\cdot|T_n-\theta_o|$ yields $0 = B_n + C_nn^{1/2}\cdot|T_n-\theta_o| + o_p(1)$. But $C_n \rightarrow_p -J < 0$ implies $C_n^{-1} \rightarrow_p -J^{-1}$. By rule 6, $(C_n+J^{-1})B_n = o_p(1)$ and $n^{1/2}(T_n-\theta_o) = J^{-1}B_n + o_p(1)$. The limit rules in Figure 3.1 then imply $J^{-1}B_n \rightarrow_d Z_o \sim N(0,J^{-1})$, $n^{1/2}\cdot|T_n-\theta_o| - J^{-1}B_n \rightarrow_p 0$, and hence $n^{1/2}\cdot|T_n-\theta_o| \rightarrow_d Z_o$.

The third result in the theorem is that J is estimated consistently by

$$(3)\quad J_n = \ n^{-1}\sum_{i=1}^{n}\ \nabla_\theta l(y_i,T_n)^2.$$

To show this, make a Taylor's expansion of this expression around θ_o,

$$(4)\quad J_n = \ n^{-1}\sum_{i=1}^{n}\ l_\theta(y_i,\theta_o)^2 + 2\ \ n^{-1}\sum_{i=1}^{n}\ \nabla_\theta l(y_i,T_{an})\cdot\nabla_{\theta\theta} l(y_i,T_{an})(T_n-\theta_o).$$

We have already shown that the first term in (4) converges in probability to J. The second term is the product of $(T_n - \theta_o) \rightarrow_p 0$ and an expression which is bounded by $n^{-1}\sum_{i=1}^{n}\ 2\beta(y_i)^2 \rightarrow_p 2\mathbf{E}_Y\beta(Y)^2 < +\infty$, by Khinchine's WLLN. Hence the second term is $o_p(1)$ and $J_n \rightarrow_p J$.

The final result in the theorem establishes that the MLE is efficient relative to any M-estimator T_n' satisfying $n^{-1}\sum_{i=1}^{n}\ g(y_i,T_n') = 0$, where g meets a series of regularity conditions. The first conclusion in this result is that T_n' is consistent and $n^{1/2}(T_n'-\theta_o)$ is asymptotically normal. This is actually of considerable independent interest, since many of the alternatives to MLE that are used in econometrics for reasons of computational convenience or robustness are M-estimators. Ordinary least squares is a leading example of an estimator in this class. The argument for the properties of T_n' are exactly the same as for the MLE case above, with g replacing $\nabla_\theta l$. The only difference is that R and S are not necessarily equal, whereas for $g = \nabla_\theta l$ in the MLE case, we had R = S = J. To make the efficiency argument, consider together the Taylor's expansions used to get the asymptotic distributions of T_n and T_n',

$$0 = \nabla_\theta l(y_i,T_n) = \quad n^{-1}\sum_{i=1}^{n} \quad \nabla_\theta l(y_i,\theta_o) + \quad n^{-1}\sum_{i=1}^{n} \quad \nabla_{\theta\theta} l(y_i,\theta_o)\, n^{1/2}(T_n-\theta_o) + o_p(1)$$

$$0 = g(y_i,T_n') = \quad n^{-1}\sum_{i=1}^{n} \quad g(y_i,\theta_o) + \quad n^{-1}\sum_{i=1}^{n} \quad g_\theta(Y_i,\theta_o) n^{1/2}(T_n'-\theta_o) + o_p(1)$$

Solving these two equations gives

$$n^{1/2}(T_n-\theta_o) = J^{-1}W_n + o_p(1)$$

$$n^{1/2}(T_n'-\theta_o) = R^{-1}U_n + o_p(1)$$

with $W_n = \quad n^{-1/2}\sum_{i=1}^{n} \quad \nabla_\theta l(y_i,\theta_o)$ and $U_n = \quad n^{-1/2}\sum_{i=1}^{n} \quad g(y_i,\theta_o)$. Consider any weighted average of these equations,

$$n^{1/2}((1-\gamma)T_n + \gamma T_n' - \theta_o) = J^{-1}(1-\gamma)W_n + R^{-1}\gamma U_n + o_p(1).$$

The Lindeberg-Levy CLT implies that this expression is asymptotically normal with mean zero and variance

$$\Omega = J^{-2}(1-\gamma)^2 \mathbf{E}\nabla_\theta l(Y|\theta_o)^2 + R^{-2}\gamma^2 \mathbf{E}g(Y,\theta_o)^2 + 2J^{-1}R^{-1}(1-\gamma)\gamma \mathbf{E}l_\theta(Y|\theta_o)g(Y,\theta_o)$$

.

The condition $0 \equiv \int g(y,\theta)f(y|\theta)dy \equiv \int g(y,\theta)e^{l(y|\theta)}dy$, implies, differentiating under the integral sign,

$$0 \equiv \int \nabla_\theta g(y,\theta)e^{l(y,\theta)}dy + \int \nabla_\theta l(y,\theta)g(y,\theta)e^{l(y,\theta)}dy.$$

Evaluated at θ_o, this implies $0 \equiv -R + \mathbf{E}\nabla_\theta l(Y|\theta_o)g(Y,\theta_o)$. Hence,

$$\Omega = J^{-1}(1-\gamma)^2 + R^{-2}S\,\gamma^2 + 2(1-\gamma)\gamma\, J^{-1}R^{-1}R = J^{-1} + [R^{-2}S - J^{-1}]\gamma^2.$$

Since $\Omega \geq 0$ for any γ, this requires $V = R^{-2}S \geq J^{-1}$, and hence $\Omega \geq J^{-1}$. Further, note that

$$\Omega = \mathrm{var}(Z_o+\gamma(Z_1-Z_o)) = \mathrm{var}(Z_o) + \gamma^2\,\mathrm{var}(Z_1-Z_o) + 2\gamma\,\mathrm{cov}(Z_o,Z_1-Z_o),$$

and $\mathrm{var}(Z_o) = J^{-1}$, implying

$$2\gamma\,\mathrm{cov}(Z_o, Z_1 - Z_o) \geq -\gamma^2\,\mathrm{var}(Z_1 - Z_o).$$

Taking γ small positive or negative implies $\mathrm{cov}(Z_o, Z_1 - Z_o) = 0$. □

6.5. EXERCISES

1. You have a random sample i = 1,...,n of observations x_i drawn from a normal distribution with unknown mean μ and known variance 1. Your prior density p(μ) for μ is normal with mean zero and variance 1/k, where k is a number you know. You must choose an estimate T of μ. You have a quadratic loss function $C(T,\mu) = (T - \mu)^2$. (a) What is the density of the observations, or likelihood, f(**x**,μ)? (b) What is the posterior density p(μ|**x**)? (c) What is the Bayes risk R(T(**x**)|**x**)? (d) What is the optimal estimator T^*(**x**) that minimizes Bayes risk?

2. A simple random sample with n observations is drawn from an exponential distribution with density $\lambda\cdot\exp(-\lambda x)$. (a) What is the likelihood function f(**x**,λ)? (b) What is the maximum likelihood estimator for λ? (c) If you have a prior density $\alpha\cdot\exp(-\alpha\lambda)$ for λ, where α is a constant you know, what is the posterior density of λ? What is the optimal estimator that minimizes Bayes risk if you have a quadratic loss function. (d) Using characteristic functions, show that the exact distribution of $W = 2n\lambda\bar{x}$, where $\bar{x}$ is the sample mean, is chi-square with 2n degrees of freedom. Use this to find the exact sampling distribution of the maximum likelihood estimator.

3. If h(t) is a convex function and t = T(**x**) is a statistic, then Jensen's inequality says that **E**h(T) ≥ h(**E**T), with the inequality strict when h is not linear over the support of T. When h is a concave function, **E**h(T) ≤ h(**E**T). If T is an unbiased estimator of a parameter σ^2, what can you say about $T^{1/2}$ as an estimator of σ and exp(T) as an estimator of $\exp(\sigma^2)$?

4. A simple random sample i = 1,...,n is drawn from a binomial distribution b(K,1,p); i.e., $K = k_1 + ... + k_n$ is the count of the number of times an event occurs in n independent trials, where $k_i = 1$ with (unknown) probability p and $k_i = 0$ with probability 1-p for I = 1,...,n. Which of the following statistics are <u>sufficient</u> for the parameter p: a. $(k_1,...,k_n)$; b. $(k_1^2,[k_2+...+k_n]^2)$; c. $f \equiv K/n$; d. $(f,[k_1^2+...+k_n^2])$; e. $[k_1^2+...+k_n^2]$?

5. You want to estimate mean consumption from a random sample of households i = 1,...,n. You have two alternative income measures, C_{1i} which includes the value of in-kind transfers and C_{2i} which excludes these transfers. You believe that the sample mean m_1 of C_{1i} will overstate economic consumption because in-kind transfers are not fully fungible, but the sample mean m_2 of C_{2i} will understate economic consumption because these transfers do have value. After some investigation, you conclude that $0.7\cdot m_1 + 0.3\cdot m_2$ is an unbiased estimator of mean economic consumption; i.e., an in-kind transfer that costs a dollar has a value of 70 cents to the consumer because it is not fully fungable. Your friend Dufus proposes instead the following estimator: Draw a random number between 0 and 1, report the estimate m_2 if this random number is less than 0.3, and report the estimate m_1 otherwise. Is the Dufus estimator unbiased? Is it as satisfactory as your estimator? (Hint: Does it pass the test of <u>ancillarity</u>?)

6. Suppose T(**x**) is an unbiased estimator of a parameter θ, and that T has a finite variance. Show that T is *inadmissible* by demonstrating that $(1-\lambda)\cdot T(\mathbf{x}) + \lambda\cdot 17$ for λ some small positive constant has a smaller mean square error. (This is called a *Stein shrinkage estimator*. The constant 17 is obviously immaterial, zero is often used.)

7. In Problem Set 2, you investigated some of the features of the data set nyse.txt, located in the class data area, which contains 7806 observations from January 2, 1968 through December 31, 1998 on stock prices. The file contains columns for the date in yymmdd format (DAT), the daily return on the New York Stock Exchange, including distributions (RNYSE), the Standard & Poor Stock Price Index (SP500), and the daily return on U.S. Treasury 90-day bills from the secondary market (RTB90). Define an additional variable GOOD which is one on days when RNYSE exceeds RTB90, and zero otherwise. The variable GOOD identifies the days on which an overnight buyer of the NYSE portfolio makes money. For the purposes of this exercise, make the maintained hypothesis that these observations are independent, identically distributed draws from an underlying population; i.e., suspend concerns about dependence in the observations across successive days or secular trends in their distribution.

a. Estimate **E**(GOOD). Describe the finite sample distribution of your estimator, and estimate its sample variance. Use a normal approximation to the finite sample distribution (i.e., match the mean and variance of the exact distribution) to estimate a 90 percent confidence bound.

b. Estimate the population expectation μ of RNYSE employing the sample mean, and alternately the sample median. To obtain estimates of the distribution of these estimators, employ the following procedure, called the *bootstrap*. From the given sample, draw a *resample* of the same size *with replacement*. (To do this, draw 7806 random integers k = floor(1+7806*u), where the u are uniform (0,1) random numbers. Then take observation k of RNYSE for each random integer draw; some observations will be repeated and others will be omitted. Record the resample mean and median. Repeat this process 100 times, and then estimate the mean and variance of the 100 bootstrap resample means and medians. Compare the bootstrap estimate of the precision of the sample mean estimator with what you would expect if RNYSE were normally distributed. Do confidence statements based on an assumption of normality appear to be justified? Compare the bootstrap estimates of the precision of the mean and median estimators of μ. Does choice of the sample mean rather than the sample median to estimate μ appear to be justified?

CHAPTER 7. HYPOTHESIS TESTING

7.1. THE GENERAL PROBLEM

It is often necessary to make a decision, on the basis of available data from an experiment (carried out by yourself or by Nature), on whether a particular proposition H_o (theory, model, hypothesis) is true, or the converse H_1 is true. This decision problem is often encountered in scientific investigation. Economic examples of hypotheses are

(a) The commodities market is efficient (i.e., opportunities for arbitrage are absent).
(b) There is no discrimination on the basis of gender in the market for academic economists.
(c) Household energy consumption is a necessity, with an income elasticity not exceeding one.
(d) The survival curve for Japanese cars is less convex than that for Detroit cars.

Notice that none of these economically interesting hypotheses are framed directly as precise statements about a probability law (e.g., a statement that the parameter in a family of probability densities for the observations from an experiment takes on a specific value). A challenging part of statistical analysis is to set out maintained hypotheses that will be accepted by the scientific community as true, and which in combination with the proposition under test give a probability law. Deciding the truth or falsity of a null hypothesis H_o presents several general issues: the cost of mistakes, the selection and/or design of the experiment, and the choice of the test.

7.2. THE COST OF MISTAKES

Consider a two-by-two table that compares the truth with the result of the statistical decision. For now, think of each of the alternatives H_o and H_1 as determining a unique probability law for the observations; these are called simple alternatives. Later, we consider compound hypotheses or alternatives that are consistent with families of probability laws.

		Truth	
		H_0	H_1
Decision	H_0 Accepted	Cost = 0 Probability = $1 - \alpha$	Cost = C_{II} Porbability = β
	H_0 Rejected	Cost = C_I Probability = α	Cost = 0 Probability = $\pi = 1 - \beta$

There are two possible mistakes, *Type I* in which a true hypothesis is rejected, and *Type II* in which a false hypothesis is accepted. There are costs associated with these mistakes -- let C_I denote the cost

associated with a Type I mistake, and C_{II} denote the cost associated with a Type II mistake. If the hypothesis is true, then there is a probability α that a particular decision procedure will result in rejection; this is also called the *Type I error probability* or the *significance level*. If the hypothesis is false, there is a probability β that it will be accepted; this is called the *Type II error probability*. The probability $\pi \equiv 1-\beta$ is the probability that the hypothesis will be rejected when it is false, and is called the *power* of the decision procedure.

This table is in principle completely symmetric between the states H_o and H_1: You can call your favorite theory H_o and hope the evidence leads to it being accepted, or call it H_1 and hope the evidence leads to H_o being rejected. However, classical statistical analysis is oriented so that α is chosen by design, and β requires a sometimes complex calculation. Then, the Type I error is easier to control. Thus, in classical statistics, it is usually better to assign your theory between H_o and H_1 so that the more critical mistake becomes the Type I mistake. For example, suppose you set out to test your favorite theory. Your study will be convincing only if your theory passes a test which it would have a high (and known) probability of failing if it is in fact false. You can get such a stringent test by making your theory H_1 and selecting a null and a decision procedure for which α is known and small; then your theory will be rejected in favor of H_o with large known probability $1-\alpha$ if in fact H_o rather than H_1 is true. (This will not work if you pick a "straw horse" for the null that no one thinks is plausible.) Conversely, if you set out to do a convincing demolition of a theory that you think is false, then make it the null, so that there is a small known probability α of rejecting the hypothesis if it is in fact true.

A common case for hypothesis testing is that the null hypothesis H_o is simple, but the alternative hypothesis H_1 is compound, containing a family of possible probability laws. Then, the probability of a Type II error depends on which member of this family is true. Thus, the power of a test is a function of the specific probability law in a compound alternative. When both the null hypothesis and alternative are compound, the probability of a Type I error is a function of which member of the family of probability laws consistent with H_o is true. In classical statistics, the significance level is always defined to be the "worst case": the largest α for any probability law consistent with the null.

Given the experimental data available and the statistical procedure adopted, there will be a trade off between the probabilities of Type I and Type II errors. When the cost C_I is much larger than the cost C_{II}, a good decision procedure will make α small relative to β. Conversely, when C_I is much smaller than C_{II}, the procedure will make α large relative to β. For example, suppose the null hypothesis is that a drug is sufficiently safe and effective to be released to the market. If the drug is critical for treatment of an otherwise fatal disease, then C_I is much larger than C_{II}, and the decision procedure should make α small. Conversely, a drug to reduce non-life-threatening wrinkles should be tested by a procedure that makes β small.

7.3. DESIGN OF THE EXPERIMENT

One way to reduce the probability of Type I and Type II errors is to collect more observations by increasing sample size. One may also by clever design be able to get more information from a given sample size, or more relevant information from a given data collection budget. One has the

widest scope for action when the data is being collected in a laboratory experiment that you can specify. For example, the Negative Income Experiments in the 1960's and 1970's were able to specify experimental treatments that presented subjects with different trade offs between wage and transfer income, so that labor supply responses could be observed. However, even in investigations where only natural experiments are available, important choices must be made on what events to study and what data to collect. For example, if a survey of 1000 households is to be made to determine the income elasticity of the demand for energy, one can get more precision by oversampling high income and low income households to get a greater spread of incomes.

There is an art to designing experiments or identifying natural experiments that allow tests of a null hypothesis without confounding by extraneous factors. For example, suppose one wishes to test the null hypothesis that Japanese cars have the same durability as Detroit cars. One might consider the following possible experiments:

(a) Determine the average age, by origin, of registered vehicles.
(b) Sample the age/make of scrapped cars as they arrive at junk yards.
(c) Draw a sample of individual new cars, and follow them longitudinally until they are scrapped.
(d) Draw a sample of individual new cars, and operate them on a test track under controlled conditions until they fail.

Experiment (a) is confounded by potential differences in historical purchase patterns; some of this could be removed by econometric methods that condition on the number of original purchases in earlier years. Experiments (a)-(c) are confounded by possible variations in usage patterns (urban/rural, young/old, winter roads/not). For example, if rural drivers who stress their cars less tend to buy Detroit cars, this factor rather than the intrinsic durability of the cars might make Detroit cars appear to last longer. One way to reduce this factor would be to assign drivers to car models randomly, as might be done for example for cars rented by Avis in the "compact" category. The ideal way to do this is a "double blind" experiment in which neither the subject nor the data recorder knows which "treatment" is being received, so there is no possibility that bias in selection or response could creep in. Most economic experimental treatments are obvious to aware subjects, so that "double blind" designs are impossible. This puts an additional burden on the researcher to carefully randomize assignment of treatments and to structure the treatments so that their form does not introduce factors that confound the experimental results.

Economists are often confronted with problems and data where a designed experiment is infeasible and Nature has not provided a clean "natural experiment", and in addition sample frames and protocols are not ideal. It may nevertheless be possible to model the data generation process to take account of sampling problems, and to use multivariate statistical methods to estimate and test hypotheses about the separate effects of different factors. This exercise can provide useful insights, but must be used cautiously and carefully to avoid misattribution and misinterpretation. Econometricians should follow the rule "Do No Harm". When a natural experiment or data are not adequate to resolve an economic hypothesis, econometric analysis should stop, and not be used to dress up propositions that a righteous analysis cannot support. Every econometric study should

consider very carefully all the possible processes that could generate the observed data, candidly discuss alternative explanations of observations, and avoid unsupportable claims..

7.4. CHOICE OF THE DECISION PROCEDURE

Suppose one thinks of hypothesis testing as a statistical decision problem, like the problem faced by Cab Franc in Chapter 1, with a prior p_o that H_o is true and $p_1 = 1 - p_o$ that H_1 is true. Let $f(\mathbf{x}|H_o)$ denote the likelihood of $\mathbf{x}$ if H_o is true, and $f(\mathbf{x}|H_1)$ denote the likelihood if H_1 is true. Then, the posterior likelihood of H_o given $\mathbf{x}$ is, by application of Bayes Law, $q(H_o|\mathbf{x}) = f(\mathbf{x}|H_o)p_o/[f(\mathbf{x}|H_o)p_o + f(\mathbf{x}|H_1)p_1)]$. The expected cost of rejecting H_o given $\mathbf{x}$ is then $C_I q(H_o|\mathbf{x})$, and the expected cost of accepting H_o given $\mathbf{x}$ is $C_{II}q(H_1|\mathbf{x})$. The optimal decision rule is then to *reject* H_o for $\mathbf{x}$ in the *critical region* $\mathbf{C}$ where $C_I q(H_o|\mathbf{x}) < C_{II}q(H_1|\mathbf{x})$. This inequality simplifies to $C_I f(\mathbf{x}|H_o)p_o < C_{II}f(\mathbf{x}|H_1)p_1$, implying

$$\mathbf{x} \in \mathbf{C} \text{ (i.e., } \textit{reject } H_o\text{) if and only if } f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o) > k \equiv C_i p_o/C_{II}p_1.$$

The expression $f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o)$ is termed the *likelihood ratio*. The optimal criterion is then to reject H_o if and only if the likelihood ratio exceeds a threshold k. The larger C_I or p_o, the larger this threshold.

A classical statistical treatment of this problem will also pick a *critical region* $\mathbf{C}$ of $\mathbf{x}$ for which H_o will be rejected, and will do so by maximizing power $\pi = \int_C f(\mathbf{x}|H_1)d\mathbf{x}$ subject to the constraint $\alpha = \int_C f(\mathbf{x}|H_o)d\mathbf{x}$. But this is accomplished by picking $\mathbf{C} = \{\mathbf{x}|f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o) > k\}$, where k is a constant chosen so the constraint is satisfied. To see why observe that if C contains a little rectangle $[\mathbf{x},\mathbf{x}+\delta\mathbf{1}]$, where δ is a tiny positive constant, then this rectangle contributes $f(\mathbf{x}|H_o)\delta^n$ to meeting the constraint and $f(\mathbf{x}|H_1)\delta^n$ to power. The ratio $f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o)$ then gives the rate at which power is produced per unit of type I error probability used up. The optimal critical region will start where this rate is the highest, and keep adding to C by decreasing the rate threshold k until the type I error probability constraint is met.

The optimal decision rule for various prior probabilities and costs and the classical statistical test procedure trace out the same families of procedures, and will coincide when the critical likelihood ratio k in the two approaches is the same. In more general classical hypothesis testing situations where the alternative is compound, there is no longer an exact coincidence of the classical and statistical decision theory approaches to decisions. However, the likelihood ratio often remains a useful basis for constructing good test procedures. In many cases, a "best" test by some classical statistical criterion and a test utilizing the likelihood ratio criterion will be the same or nearly the same.

In general, we will consider DGP which we maintain are members of a family $f(\mathbf{x},\theta)$ indexed by a parameter θ. The null hypothesis is that the true value θ_o of θ is contained in a set $\mathbf{N}$, and the

alternative is that it is contained in a set **A**, with **A** and **N** partitioning the universe **Θ** of possible values of θ_o. The value θ_e of θ that maximizes $f(\mathbf{x},\theta)$ over $\theta \in \mathbf{\Theta}$ is the *maximum likelihood estimator*. The theory of the maximum likelihood estimator given in Chapter 6 shows that it will have good statistical properties in large samples under mild regularity conditions. The value θ_{oe} that maximizes $f(\mathbf{x},\theta)$ over $\theta \in \mathbf{N}$ is called the *constrained maximum likelihood estimator* subject to the null hypothesis. When the null hypothesis is true, the constrained maximum likelihood estimator will also have good statistical properties. Intuitively, the reason is that when the null hypothesis is true, the true parameter satisfies the hypothesis, and hence the maximum value of the constrained likelihood will be at least as high as the value of the likelihood at the true parameter. If an identification condition is met, the likelihood at the true parameter converges in probability to a larger number than the likelihood at any other parameter value. Then, the constrained maximum likelihood must converge in probability to the true parameter. A rigorous proof of the properties of constrained estimators is given in Chapter 22.

A likelihood ratio critical region for the general testing problem is usually defined as a set of the form

$$\mathbf{C} = \{\mathbf{x} \mid \sup_{\theta\in\mathbf{A}} f(\mathbf{x},\theta)/\sup_{\theta\in\mathbf{N}} f(\mathbf{x},\theta) > k\}.$$

The likelihood ratio in this criterion is less than or equal to one when the maximum likelihood estimator of θ_o falls in **N**, and otherwise is greater than one. Then a critical region defined for some $k > 1$ will include the observed vectors **x** that are the least likely to have been generated by a DGP with a parameter in **N**. The significance level of the test is set by adjusting k.

Since $\sup_{\theta\in\mathbf{\Theta}} f(\mathbf{x},\theta)/\sup_{\theta\in\mathbf{N}} f(\mathbf{x},\theta) = \max\{1, \sup_{\theta\in\mathbf{A}} f(\mathbf{x},\theta)/\sup_{\theta\in\mathbf{N}} f(\mathbf{x},\theta)\}$, an equivalent expression for the critical region when $k > 1$ is

$$\mathbf{C} = \{\mathbf{x} \mid \sup_{\theta\in\mathbf{\Theta}} f(\mathbf{x},\theta)/\sup_{\theta\in\mathbf{N}} f(\mathbf{x},\theta) > k\}.$$

This can also be expressed in terms of the log likelihood function,

$$\mathbf{C} = \{\mathbf{x} \mid \sup_{\theta\in\mathbf{\Theta}} \log f(\mathbf{x},\theta) - \sup_{\theta\in\mathbf{N}} \log f(\mathbf{x},\theta) > \kappa = \log k\}.$$

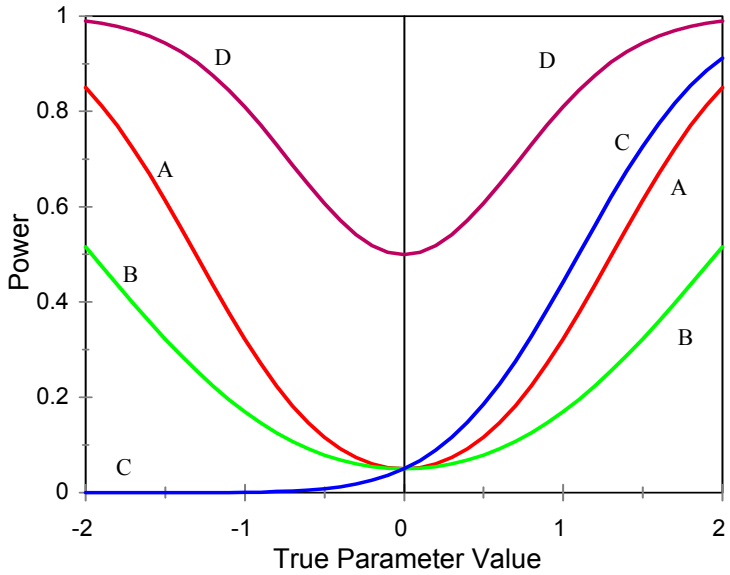
Clearly, the log ratio in this expression equals the difference in the log likelihood evaluated at the maximum likelihood estimator and the log likelihood evaluated at the constrained maximum likelihood estimator. This difference is zero if the maximum likelihood estimator is in **N**, and is otherwise positive.

The analyst will often have available alternative testing procedures in a classical testing situation. For example, one procedure to test a hypothesis about a location parameter might be based on the sample mean, a second might be based on the sample median, and a third might be based on the likelihood ratio. Some of these procedures may be better than others in the sense of giving higher power for the same significance level. The ideal, as in the simple case, is to maximize the power given the significance level. When there is a compound alternative, so that power is a function of the alternative, one may be able to tailor the test to have high power against alternatives of particular

importance. In a few cases, there will be a single procedure that will have uniformly best power against a whole range of alternatives. If so, this will be called a *uniformly most powerful* test.

The figure below shows power functions for some alternative test procedures. The null hypothesis is that a parameter θ is zero. Power curves A and B equal 0.05 when H_o: θ = 0 is true. Then, the significance level of these three procedures is α = 0.05. The significance level of D is much higher, 0.5. Compare the curves A and B. Since A lies everywhere above B and has the same significance level, A is clearly the superior procedure. A comparison like A and B most commonly arises when A uses more data than B; that is, A corresponds to a larger sample. However, it is also possible to get a picture like this when A and B are using the same sample, but B makes poor use of the information in the sample.

Compare curves A and C. Curve C has significance level α = 0.05, and has lower power than A against alternatives less than θ = 0, but better power against alternatives greater than θ = 0. Thus, A is a better test it we want to test against all alternatives, while C is a better test if we are mainly interested in alternatives to the right of θ = 0 (i.e., we want to test H_o: θ ≤ 0). Compare curves A and D. Curve D has high power, but at the cost of a high probability of a Type I error. Thus, A and D represent a trade off between Type I and Type II errors.



Finally, suppose we are most interested in the alternative H_1: θ = 1.5. The procedure giving curve A has power 0.61 against this alternative, and hence has a reasonable chance of discriminating between H_o and H_1. On the other hand, the procedure B has power 0.32, and much less chance of discriminating. We would conclude that the procedure A is a moderately satisfactory statistical test procedure, while B is of limited use.

7.5. HYPOTHESIS TESTING IN NORMAL POPULATIONS

This section provides a summary of hypothesis test calculations for standard setups involving data drawn from a normal population, including power calculations. Assume that we start from a simple random sample of size n, giving i.i.d. observations $x_1,...,x_n$. Recall from Chapter 6.3 that the log likelihood of a normal random sample is

$$L(\mathbf{x},\mu,\sigma^2) = -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log}\,\sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$$

$$= -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log}\,\sigma^2 - \frac{1}{2}\cdot\frac{(n-1)s^2}{\sigma^2} - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2 \equiv L(\bar{x},s^2,\mu,\sigma^2)\,.$$

where the sample mean $\bar{x} = \sum_{i=1}^{n} x_i$ and the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ are unbiased estimators of μ and σ^2, respectively. If **N** denotes the set of parameter values (μ,σ^2) consistent with a null hypothesis, then a likelihood ratio critical set for this hypothesis will take the form

$$\mathbf{C} = \{(\bar{x},s^2)|\ \sup_{\theta\in\mathbf{\Theta}} L(\bar{x},s^2,\mu,\sigma^2) - \sup_{\theta\in\mathbf{N}} L(\bar{x},s^2,\mu,\sigma^2) > \kappa\}.$$

We consider a sequence of hypotheses and conditions. See Chapter 3.7 for the densities and other properties of the distributions used in this section, and Chapter 6.3 for the relation of these distributions to data from a normal population. The following table gives the statistical functions that are available in many econometrics software packages; the specific notation is that used in the Statistical Software Tools (SST) package. Tables at the end of most statistics texts can also be used to obtain values of the central versions of these distributions. The direct functions give the CDF probability for a specified argument, while the inverse functions give the argument that yields a specified probability.

Distribution	CDF	Inverse CDF
Normal	cumnorm(x)	invnorm(p)
Chi-Square	cumchi(x.df)	invchi(p,df)
F-Distribution	cumf(xdf1,df2)	invf(p,df1,df2)
T-Distribution	cumt(x,df)	invt(p,df)
Non-central Chi-Square	cumchi(x,df,δ)	NA
Non-central F-Distribution	cumf(x,df1,df2,δ)	NA
Non-Central T-Distribution	cumt(x,df,λ)	NA

In this table, df denotes degrees of freedom, and λ and δ are non-centrality parameters. Inverse CDF's are not available for non-central distributions in most packages, and are not needed. In most statistical packages, values of these functions can either be printed out or saved for further calculations. For example, in SST, the command "calc cumnorm(1.7)" will print out the probability that a standard normal random variable is less than 1.7, the command "calc p = cumnorm(1.7)" will store the result of this calculation in the variable p for further use, and a subsequent command "calc p" will also print out its value.

Problem 1: Testing the mean of a normal population that has known variance

Suppose a random sample of size n from a normal population with an unknown mean μ and a known variance σ^2. The null hypothesis is H_o: $\mu = \mu_o$, and the alternative is H_1: $\mu \neq \mu_o$. Verify that the likelihood ratio, $\text{Max}_\mu n(\bar{x},\mu,1)/n(\bar{x},\mu_o,1)$, is an increasing function of $(\bar{x} - \mu_o)^2$. Hence, a test equivalent to a likelihood ratio test can be based on $(\bar{x} - \mu_o)^2$. From Chapter 6.3(8), one has the result that under the null hypothesis, the statistic $n(\bar{x} - \mu_o)^2/\sigma^2$ is distributed χ_1^2. Alternately, from Chapter 6.3(5), the square root of this expression, $n^{1/2}(\bar{x} - \mu_o)/\sigma$, has a standard normal distribution.

Using the Chi-Square form of the statistic, the critical region will be values exceeding a critical level z_c, where z_c is chosen so that the selected significance level α satisfies $\chi_1^2(z_c) = 1 - \alpha$. For example, taking $\alpha = 0.05$ yields $z_c = 3.84146$. This comes from a statistical table, or from the SST command "calc invchi(1- α,k)", where α is the significance level and k is degrees of freedom. The test procedure <u>rejects</u> H_o whenever

(1) $n(\bar{x} - \mu_o)^2/\sigma^2 > z_c = 3.84146.$

Consider the power of the Chi-square test against an alternative such as $\mu = \mu_1 \neq \mu_o$. The non-centrality parameter is

(2) $\delta = n(\mu_1-\mu_o)^2/\sigma^2.$

For example, if $\mu_1-\mu_o = 1.2$, $\sigma^2 = 25$, and $n = 100$, then $\delta = 1.44 \cdot 100/25 = 5.76$. The power is calculated from the non-central Chi-square distribution (with 1 degree of freedom), and equals the probability that a random draw from this distribution exceeds z_c. This probability π is readily calculated using the SST command "calc 1 - cumchi(z_c,k,δ)". In the example, π = calc 1 - cumchi(3.84146,1,5.76) = 0.67006. Then, a test with a five percent significance level has power of 67 percent against the alternative that the true mean is 1.2 units larger than hypothesized.

An equivalent test can be carried out using the standard normal distributed form $n^{1/2}(\bar{x} - \mu_o)/\sigma$. The critical region will be values of this expression that in magnitude exceed a critical level w_c, where w_c is chosen for a specified significance level α so that a draw from a standard normal density has probability $\alpha/2$ of being below $-w_c$, and symmetrically a probability $\alpha/2$ of being above $+w_c$. One can find w_c from statistical tables, or by using a SST command "calc invnorm(1-α/2)". For example, if $\alpha = 0.05$, then w_c = calc invnorm(0.975) = 1.95996. The test <u>rejects</u> H_o whenever

(3) $n^{1/2}(\bar{x} - \mu_o)/\sigma < -w_c$ <u>or</u> $n^{1/2}(\bar{x} - \mu_o)/\sigma > w_c.$

For example, if n = 100, $\sigma = 5$, and $\mu_o = 0$, the critical region for a test with significance level $\alpha = 0.05$ is $10\,\bar{x}/5 < -1.95996$ or $10\bar{x}/5 > +1.95996$. Note that $w_c^2 = z_c$, so this test rejects exactly when the Chi-square test rejects. The power of the test above against the alternative $\mu = \mu_1 \neq \mu_o$ is the probability that the random variable $n^{1/2}(\bar{x} - \mu_o)/\sigma$ lies in the critical region when $\bar{x} \sim N(\mu_1,\sigma^2)$. In this case, $n^{1/2}(\bar{x} - \mu_1)/\sigma \equiv Y$ is standard normal, and therefore $n^{1/2}(\bar{x} - \mu_o)/\sigma \equiv Y + \lambda$, where

(4) $\lambda = n^{1/2}(\mu_1 - \mu_o)/\sigma.$

Note that $\lambda^2 = \delta$, where δ is given in (2). The probability of rejection in the left tail is $\Pr(n^{1/2}(\bar{x} - \mu_o)/\sigma < -w_c \,|\, \mu = \mu_1) = \Pr(Y < -w_c - \lambda)$. For the right tail, $\Pr(n^{1/2}(\bar{x} - \mu_o)/\sigma > w_c \,|\, \mu = \mu_1) = \Pr(Y > w_c - \lambda)$. Using the fact that the standard normal is symmetric, we then have

(5) $\pi = \Phi(-w_c - \lambda) + 1 - \Phi(w_c - \lambda) \equiv \Phi(-w_c - \lambda) + \Phi(-w_c + \lambda).$

This can be calculated using the SST command

π = calc cumnorm($-w_c-\lambda$) + cumnorm($-w_c+\lambda$).

For example, $\sigma = 5$, N = 100, $\mu_1 - \mu_o = 1.2$, $w_c = 1.95996$ give $\delta = 2.4$ and power π = calc cumnorm($-w_c$ - 2.4) + cumnorm($-w_c$ + 2.4) = 0.670. Note this is the same as the power of the Chi-square version of the test.

Suppose that instead of testing the null hypothesis H_o: $\mu = \mu_o$ against the alternative H_1: $\mu \neq \mu_o$, you want to test the one-sided hypothesis H_o: $\mu \leq \mu_o$ against the alternative H_1: $\mu > \mu_o$. The likelihood ratio in this case is $Sup_{\mu>\mu_o}\ n(\bar{x},\mu,\sigma^2) / Sup_{\mu\leq\mu_o}\ n(\bar{x},\mu,\sigma^2)$, which is constant for $\bar{x} \leq \mu_o$ and is monotone increasing in $(\bar{x} - \mu_o)$ for $\bar{x} > \mu_o$. Hence, a test that rejects H_o for $\bar{x} - \mu_o$ large appears desirable. This suggests using a test based on the statistic $n^{1/2}(\bar{x}-\mu_o)/\sigma$, which is normal with variance one, and has a non-positive mean under the null. Pick a critical level $w_c > 0$ such that

$$Sup_{\mu\leq\mu_o}\ \Pr(n^{1/2}(\bar{x}-\mu)/\sigma > w_c) = \alpha.$$

Note that the sup is taken over all the possible true μ consistent with H_o, and that α is the selected significance level. The maximum probability of Type I error is achieved when $\mu = \mu_o$. (To see this, note that $\Prob(n^{1/2}(\bar{x}-\mu_o)/\sigma > w_c) \equiv \Pr(Y \equiv n^{1/2}(\bar{x}-\mu)/\sigma > w_c + n^{1/2}(\mu_o - \mu)/\sigma$, where μ is the true value. Since Y is standard normal, this probability is largest over $\mu \leq \mu_o$ at $\mu = \mu_o$.) Then, w_c is determined to give probability α that a draw from a standard normal exceeds w_c. For example, if n = 100, α = 0.05, $\sigma = 5$, and H_o is that $\mu \leq 0$, then w_c = calc invnorm(0.95) = 1.64485. The power of the test of $\mu \leq \mu_o = 0$ against the alternative $\mu = \mu_1 = 1.2$ is given by

(6) $\pi = \Pr(n^{1/2}(\bar{x}-\mu_o)/\sigma > w_c \,|\, \mu = \mu_1) \equiv \Pr(Y \equiv n^{1/2}(\bar{x}-\mu_1)/\sigma > w_c - \lambda)$
$\equiv 1 - \Phi(w_c - \lambda) \equiv \Phi(-w_c + \lambda) \equiv$ calc cumnorm($-w_c + \lambda$),

where λ is given in (4). In the example, π = calc cumnorm(- 1.64485 + 2.4) = 0.775. Hence, a test which has a probability of at most $\alpha = 0.05$ of rejecting the null hypothesis when it is true has power 0.775 against the specific alternative $\mu_1 = 1.2$.

Problem 2. Testing the Mean of a Normal Population with Unknown Variance

This problem is identical to Problem 1, except that σ^2 must now be estimated. Use the estimator s^2 for σ^2 in the Problem 1 test statistics. From Chapter 6.3(8), the Chi-square test statistic with σ replaced by s, $F = n(\bar{x} - \mu_o)^2/s^2$, has an F-distribution with degrees of freedom 1 and N-1. Hence, to test H_o: $\mu = \mu_o$ against the alternative H_1: $\mu \neq \mu_o$, find a critical level z_c such that a specified significance level α equals the probability that a draw from $F_{1,n-1}$ exceeds z_c. The SST function calc z_c = invf(1-α,1,n-1) gives this critical level; it can also be found in standard tables. For n = 100 and $\alpha = 0.05$, the critical level is $z_c = 3.93694$.

The power of the test against an alternative μ_1 is the probability that the statistic F exceeds z_c. Under this alternative, F has a non-central F-distribution (from Chapter 3.9) with the non-centrality parameter $\delta = n(\mu_1 - \mu_o)^2/\sigma^2$ given in (2). Then, the power is given by

(7) π = calc 1 - cumf(z_c,1,n-1,δ).

In the example with $\mu_1 - \mu_o = 1.2$ and $\sigma^2 = 25$, one has $\delta = 144/25$, and the power is

(8) π = calc 1 - cumf(3.93694,1,99,144/25) = 0.662.

The non-centrality parameter is defined using the <u>true</u> σ^2 rather than the estimate s^2. Calculating power at an <u>estimated</u> non-centrality parameter $\delta_e = n(\mu_1 - \mu_o)^2/s^2$ introduces some error -- you will evaluate the power curve at a point somewhat different than you would like. For most practical purposes, you do not need an exact calculation of power; you are more interested in whether it is 0.1 or 0.9. Then, the error introduced by this approximation can be ignored. In particular, for large sample sizes where the power against economically interesting alternatives is near one, this error is usually negligible. Note that $\delta/\delta_e = s^2/\sigma^2$, so $(n-1)\delta/\delta_e$ is distributed $\chi^2(n-1)$. For the rare application where you really need to know how precise your power calculation is, you can form a <u>confidence interval</u> as follows: Given a "significance level" α, compute z_1 = calc invchi(α/2,n-1) and z_2 = calc invchi(1-α/2,n-1). Then, with probability α, $\delta_1 \equiv z_1\delta_e/(n-1) < \delta < z_1\delta_e/(n-1) \equiv \delta_2$. The power π_1 calculated at δ_1 and the power π_2 calculated at δ_2 give a α-level confidence bound on the exact power. For example, $\alpha = 0.5$, n = 100, $\mu_1 - \mu_o = 1.2$, and $s^2 = 25$ imply $\delta_e = 144/25$, z_1 = calc invchi(.25,99) = 89.18, $\delta_1 = 5.189$, and π_1 = calc 1-cumf(3.93694,1,99,5.189) = 0.616. Also, z_2 = calc invchi(.75,99) = 108.093, $\delta_2 = 6.289$, and π_2 = calc 1-cumf(3.93694,1,99,6.289) = 0.700. Then, with probability 0.5, the exact power for the alternative $\mu_1 - \mu_o = 2$ is in the interval [0.616,0.700].

The test of H_o: $\mu = \mu_o$ can be carried out equivalently using

(9) $T = n^{1/2}(\bar{x} - \mu_o)/s$,

which by Chapter 6.3(7) has a t-distribution with n-1 degrees of freedom under H_o: $\mu = \mu_o$. For a significance level α, choose a critical level w_c, and reject the null hypothesis when $|T| > w_c$. The value of w_c satisfies $\alpha/2 = t_{n-1}(-w_c)$, and is given in standard tables, or in SST by w_c = invt(1-α/2,n-1). For the example $\alpha = 0.05$ and $n = 100$, this value is w_c = calc invt(.975,99) = 1.9842.

The power of the test is calculated as in Problem 1, replacing the normal distribution by the non-central t-distribution: $\pi = t_{n-1,\lambda}(-w_c) + 1 - t_{n-1,\lambda}(w_c)$, where $\lambda = n^{1/2}(\mu_1 - \mu_o)/\sigma$ as in equation (4). Points of the non-central t are not in standard tables, but are provided by a SST function, π = cumt($-w_c$,n-1,λ) + 1 - cumt(w_c,n-1,λ). For the example $\alpha = 0.05$, $N = 100$, $\sigma = 5$, and $\mu_1 - \mu_o = 1.2$ imply $\lambda = 2.4$, and this formula gives π =

The T-statistic (9) can be used to test the one-sided hypothesis H_o: $\mu \le \mu_o$. The hypothesis will be rejected if $T > w_c$, where w_c satisfies $\alpha = t_{n-1}(-w_c)$, and is given in standard tables, or in SST by w_c = invt(1-α,n-1). The power of the test is calculated in the same way as the one-sided test in Problem 1, with the non-central t-distribution replacing the normal: π = 1 - cumt(w_c,n-1,λ).

Problem 3. Testing the Variance of a Normal Population with Unknown Mean

Suppose H_o: $\sigma^2 = \sigma_o^2$ versus the alternative H_1 that this equality does not hold.. Under the null, the statistic $X \equiv (n-1)s^2/\sigma_o^2$ is distributed $\chi^2(n-1)$. Then, a test with significance level α can be made by rejecting H_o if $X < z_{c1}$ or $X > z_{c2}$, where z_{c1} and z_{c2} are chosen so the probability is $\alpha/2$ that a draw from $\chi^2(n-1)$ is less than z_{c1}, and $\alpha/2$ that it is greater than z_{c2}. These can be calculated using z_{c1} = calc invchi(α/2,n-1) and z_{c2} = calc invchi(1-α/2,n-1). To calculate the power of the test against the alternative H_1: $\sigma^2 = \sigma_1^2$, note that in this case $(n-1)s^2/\sigma_1^2 = X\sigma_o^2/\sigma_1^2 \equiv Y$ is $\chi^2(n-1)$. Then,

$$\pi = 1 - \Pr(z_{c1} \le X \le z_{c2} | \sigma^2=\sigma_1^2) = 1 - \Pr(z_{c1}\sigma_o^2/\sigma_1^2 \le Y \le z_{c2}\sigma_o^2/\sigma_1^2)$$
$$= \text{calc cumchi}(z_{c1}\sigma_o^2/\sigma_1^2,n-1) + 1 - \text{cumchi}(z_{c2}\sigma_o^2/\sigma_1^2,n-1).$$

Problem 4. Testing the Equality of Unknown Variances in Two Populations

Suppose independent random samples of sizes n_i are drawn from normal populations with means μ_i and variances σ_i^2, respectively, for i = 1,2. The null hypothesis is H_o: $\sigma_1^2 = \sigma_2^2$, and the alternative is $\sigma_1^2 \ne \sigma_2^2$. For each population, we know from 3.6 that $(n_i-1)s_i^2/\sigma_i^2$ has a Chi-square distribution with n_i-1 degrees of freedom. Further, we know that the ratio of two independent Chi-square distributed random variables, each divided by its degrees of freedom, has an F-distribution with these respective degrees of freedom. Then, $R = s_1^2/s_2^2$ is distributed $F(n_1-1,n_2-1)$ under H_o. One can form a critical region $\mathbf{C} = \{R|R<c_L \text{ or } R > c_U\}$ that has significance level α by choosing the lower and upper tails c_L and c_U of the F-distribution so that each has probability $\alpha/2$.

Under alternatives to the null, the ratio s_1^2/s_2^2, multiplied by the ratio σ_2^2/σ_1^2, has a central $F(n_1-1,n_2-1)$-distribution, and the power of the test is

$$\pi = 1 - \text{Prob}(c_L \le R \le c_U) = 1 - \text{Prob}(c_L\,\sigma_2^2/\sigma_1^2 \le R\,\sigma_2^2/\sigma_1^2 \le c_U\,\sigma_2^2/\sigma_1^2)$$
$$= F(c_L\cdot\sigma_2^2/\sigma_1^2,n_1-1,n_2-1) + 1 - F(c_U\cdot\sigma_2^2/\sigma_1^2,n_1-1,n_2-1).$$

Problem 5. Testing the Equality of Unknown Means in Two Populations with a Common Unknown Variance

Suppose independent random samples of sizes n_i are drawn from normal populations with means μ_i for i = 1,2 and a common variance σ^2. The null hypothesis is H_o: $\mu_1 = \mu_2$, and the alternative is $\mu_1 \neq \mu_2$. Then $\bar{x}_1 - \bar{x}_2$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma^2(n_1^{-1} + n_2^{-1})$. Further, $(n_1-1)s_1^2/\sigma^2$ is chi-square with n_1-1 degrees of freedom, $(n_2-1)s_2^2/\sigma^2$ is chi-square with n_2-1 degrees of freedom, and all three random variables are independent. Then $((n_1-1)s_1^2 + (n_2-1)s_2^2)/\sigma^2$ is chi-square with $n_1 + n_2 - 2$ degrees of freedom. It follows that

$$s^2 = (n_1^{-1} + n_2^{-1})\cdot((n_1-1)s_1^2 + (n_2-1)s_2^2)/(n_1 + n_2 - 2)$$

is an unbiased estimator of $\sigma^2(n_1^{-1} + n_2^{-1})$, with $(n_1 + n_2 - 2)s^2/\sigma^2(n_1^{-1} + n_2^{-1})$ distributed Chi-square with $n_1 + n_2 - 2$ degrees of freedom. Therefore, the statistic

$$(\bar{x}_1 - \bar{x}_2)/s = (\bar{x}_1 - \bar{x}_2)/[(n_1^{-1} + n_2^{-1})\cdot((n_1-1)s_1^2 + (n_2-1)s_2^2)/(n_1 + n_2 - 2)]^{1/2}$$

is distributed under the null hypothesis with a T-distribution with $n_1 + n_2 - 2$ degrees of freedom. The power against an alternative $\mu_1 \neq \mu_2$ is calculated exactly as in Problem 2, following (9), except the degrees of freedom is now $n_1 + n_2 - 2$ and the non-centrality parameter is

$$\lambda = (\mu_1 - \mu_2)/\sigma(n_1^{-1} + n_2^{-1})^{1/2}.$$

7.6. HYPOTHESIS TESTING IN LARGE SAMPLES

Consider data $\mathbf{x} = (x_1,...,x_n)$ obtained by simple random sampling from a population with density $f(x,\theta_o)$, where θ_o is a k×1 vector of unknown parameters contained in the interior of a set $\boldsymbol{\Theta}$. The sample DGP is $f(\mathbf{x},\theta_o) = \prod_{i=1}^{n} f(x_i,\theta_o)$ and log likelihood is $L_n(\mathbf{x},\theta) = \sum_{i=1}^{n} l(x_n,\theta)$, where $l(x,\theta)$ = log $f(x,\theta)$ is the log likelihood of an observation. Consider the maximum likelihood estimator $T_n(\mathbf{x})$, given by the value of θ that maximizes $L_n(\mathbf{x},\theta)$. Under general regularity conditions like those given in Chapter 6.4, the maximum likelihood estimator is consistent and asymptotically normal. This implies specifically that $n^{1/2}(T_n(\mathbf{x})-\theta_o) \rightarrow_d Z_o$ with $Z_o \sim N(O,J^{-1})$ and J the Fisher information in an observation, $J = \mathbf{E}\,[\nabla_\theta l(x,\theta_o)][\nabla_\theta l(x,\theta_o)]'$. The Chapter 3.1.18 rule for limits of continuous transformations implies $n^{1/2}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o) \rightarrow_d N(0,I)$, and hence that the quadratic form $W(\mathbf{x},\theta_o) \equiv n\cdot(T_n(\mathbf{x})-\theta_o)'J(T_n(\mathbf{x})-\theta_o) \equiv (T_n(\mathbf{x})-\theta_o)'\mathbf{V}(T_n(\mathbf{x}))^{-1}(T_n(\mathbf{x})-\theta_o) \rightarrow_d \chi^2(k)$, the Chi-square distribution with k degrees of freedom. When k = 1, this quadratic form equals the square of the difference between $T_n(\mathbf{x})$ and θ_o, divided by the variance $\mathbf{V}(T_n(\mathbf{x}))$ of $T_n(\mathbf{x})$. The square root of this expression, $(T_n(\mathbf{x})-\theta_o)/(\mathbf{V}(T_n(\mathbf{x})))^{1/2}$, converges in distribution to a standard normal.

Consider the null hypothesis H_o: $\theta = \theta_o$. When this null hypothesis is true, the quadratic form $W(\mathbf{x},\theta_o)$ has a limiting Chi-square distribution with k degrees of freedom. Then, a test of the hypothesis with a significance level α can be carried out by choosing a critical level *c* from the upper

tail of the $\chi^2(k)$ distribution so that the tail has probability α, and rejecting H_o when $W(\mathbf{x},\theta_o) > c$. We term $W(\mathbf{x},\theta_o)$ the *Wald statistic*.

Suppose an alternative H_1: $\theta = \theta_1$ to the null hypothesis is true. The power of the Wald test is the probability that the null hypothesis will be rejected when H_1 holds. But in this case, $n^{1/2}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o) = n^{1/2}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_1) + n^{1/2}\cdot J^{1/2}(\theta_1-\theta_o)$, with the first term converging in distribution to N(0,I). For fixed $\theta_1 \neq \theta_o$, the second term blows up. This implies that the probability that $n^{1/2}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o)$ is small enough to accept the null hypothesis goes to zero, and the power of the test goes to one. A test with this property is called *consistent*, and consistency is usually taken to be a minimum requirement for a hypothesis testing procedure to be statistically satisfactory. A closer look at the power of a test in large samples is usually done by considering what is called *local power*. Suppose one takes a sequence of alternatives to the null hypothesis that get closer and closer to the null as sample size grows. Specifically, consider H_1: $\theta = \theta_o + \lambda/n^{1/2}$. For this sequence of alternatives, the term $n^{1/2}\cdot J^{1/2}(\theta_1-\theta_o) = J^{1/2}\delta$ is a constant, and we have the result that $n^{1/2}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o) \rightarrow_d N(J^{1/2}\ \lambda,I)$. This implies that $(T_n(\mathbf{x})-\theta_o)'(nJ)\cdot(T_n(\mathbf{x})-\theta_o)$, the Wald statistic, converges in distribution to a noncentral Chi-square distribution with k degrees of freedom and a noncentrality parameter $\lambda' J\lambda$. The local power of the test is the probability in the upper tail of this distribution above the critical level *c* for the Wald statistic. The local power will be a number between zero and one which provides useful information on the ability of the test to distinguish the null from nearby alternatives. In finite sample applications, the local power approximation can be used for a specific alternative θ_1 of interest by taking $\lambda = n^{1/2}\cdot(\theta_1-\theta_o)$ and using the noncentral Chi-square distribution as described above.

In practice, we do not know the Fisher Information J exactly, but must estimate it from the sample by

$$(10) \quad J_{en} = \mathbf{E}_n[\nabla_\theta l(x,T_n)][\nabla_\theta l(x_i,T_n)]' \equiv n^{-1}\sum_{i=1}^{n} [\nabla_\theta l(x_i,T_n)][\nabla_\theta l(x_i,T_n)]'.$$

The expression in (10) is termed the *outer product of the score* $\nabla_\theta l(x_i,T_n)$ of an observation. When there is a single parameter, this reduces to the square of $\nabla_\theta l(x_i,T_n)$; otherwise, it is a k×k array of squares and cross-products of the components of $\nabla_\theta l(x_i,T_n)$. From the theorem in Chapter 6.4, $J_{en} \rightarrow_p J$, and the rule 1.17 in Chapter 4 implies that replacing J by J_{en} in the Wald test statistic does not change its asymptotic distribution.

In the discussion of maximum likelihood estimation in Chapter 6.4 and the proof of its asymptotic normality, we established that when θ_o is the true parameter,

$$(11) \quad n^{1/2}\cdot(T_n(\mathbf{x})-\theta_o) = J^{-1}\cdot\nabla_\theta L_n(\mathbf{x},\theta_o)/n^{1/2} + o_p(1);$$

that is, the difference of the maximum likelihood estimator from the true parameter, normalized by $n^{1/2}$, equals the normalized score of the likelihood at θ_o, transformed by J^{-1}, plus asymptotically negligible terms. If we substitute (11) into the Wald statistic, we obtain $LM(\mathbf{x},\theta_o) = W(\mathbf{x},\theta_o) + o_p(1)$, where

$$(12) \quad LM(\mathbf{x},\theta_o) = [\nabla_\theta L(\mathbf{x},\theta_o)]'(nJ)^{-1}[\nabla_\theta L(\mathbf{x},\theta_o)].$$

The statistic (12) is called the *Lagrange Multiplier* (LM) *statistic*, or the *score statistic*. The name Lagrange Multiplier comes from the fact that if we maximize $L_n(\mathbf{x},\theta)$ subject to the constraint $\theta_o - \theta = 0$ by setting up the Lagrangian $L_n(\mathbf{x},\theta) + \lambda(\theta_o-\theta)$, we obtain the first order condition $\lambda = \nabla_\theta L_n(\mathbf{x},\theta)$ and hence $LM(\mathbf{x},\theta_o) = \lambda'(nJ)^{-1}\lambda$. Because $LM(\mathbf{x},\theta_o)$ is *asymptotically equivalent* to the Wald statistic, it will have the same asymptotic distribution, so that the same rules apply for determining critical levels and calculating power. The Wald and LM statistics will have different numerical values in finite samples, and sometimes one will accept a null hypothesis when the other rejects. However, when sample sizes are large, their asymptotic equivalence implies that most of the time they will either both accept or both reject, and that they have the same power. In applications, J in (11) must be replaced by an estimate, either J_{en} from (10), or $J_{oen} = \mathbf{E}_n[\nabla_\theta l(x,\theta_o)][\nabla_\theta l(x_i,\theta_o)]'$ in which the score is evaluated at the hypothesized θ_o. Both converge in probability to J, and substitution of either in (12) leaves the asymptotic distribution of the LM statistic unchanged. A major advantage of the LM form of the asymptotic test statistic is that it does not require that one compute the estimate $T_n(\mathbf{x})$. Computation of maximum likelihood estimates can sometimes be difficult. In these cases, the LM statistic avoids the difficulty.

The *generalized likelihood ratio criterion* was suggested in a number of simple tests of hypotheses as a good general procedure for obtaining test statistics. This method rejects H_o if

(13) $\kappa < \max_\theta L_n(\mathbf{x},\theta) - L_n(\mathbf{x},\theta_o)$,

where κ is a constant that is adjusted to give the desired significance level for the test. A Taylor's expansion of $L_n(\mathbf{x},\theta_o)$ about $T_n(\mathbf{x})$ yields

(14) $L_n(\mathbf{x},T_n(\mathbf{x})) - L_n(\mathbf{x},\theta_o) = \nabla_\theta L_n(\mathbf{x},T_n(\mathbf{x}))\cdot(T_n(\mathbf{x})-\theta_o) - (T_n(\mathbf{x})-\theta_o)'\nabla_{\theta\theta}L_n(\mathbf{x},\theta_{en})(T_n(\mathbf{x})-\theta_o)$,

where θ_{en} is between θ_o and θ_n. But $\nabla_\theta L_n(\mathbf{x},T_n(\mathbf{x})) = 0$. Under the regularity conditions in Chapter 6.4, $\nabla_{\theta\theta}L_n(\mathbf{x},\theta_{en})/n \to_p J$. (To make the last statement rigorous, one needs to either establish that the convergence in probability of $\nabla_{\theta\theta}L_n(\mathbf{x},\theta)/n$ to $J(\theta)$ is uniform in θ, or expand $\nabla_{\theta\theta}L_n(\mathbf{x},\theta_{en})/n$ to first order about θ_o and argue that the first term goes in probability to -J and the second term goes in probability to zero.) Then, $LR(\mathbf{x},\theta_o) = 2\cdot[L_n(\mathbf{x},T_n(\mathbf{x})) - L_n(\mathbf{x},\theta_o)]$, termed the *likelihood ratio statistic*, satisfies

(15) $LR(\mathbf{x},\theta_o) = (T_n(\mathbf{x})-\theta_o)'(nJ)(T_n(\mathbf{x})-\theta_o) + o_p(1) \equiv W(\mathbf{x},\theta_o) + o_p(1)$,

and the LR statistic is asymptotically equivalent to the Wald statistic. Therefore, the LR statistic will be asymptotically distributed Chi-square with k degrees of freedom, where k is the dimension of θ_o, and its local power is the same as that of the Wald statistic, and calculated in the same way.

The major advantage of the LR statistic is that its computation requires only the values of the log likelihood unrestricted and with the null imposed; it is unnecessary to obtain an estimate of J or perform any matrix calculations. We conclude that the trinity consisting of the Wald, LM, and LR statistics are all asymptotically equivalent, and provide completely substitutable ways of testing a hypothesis using a large sample approximation.

7.6. EXERCISES

1. Use the data set nyse.txt in the class data area, and the variable RNYSE giving the daily rate of return on the New York Stock Exchange. For the purpose of this exercise, make the maintained hypothesis that the observations are independent and identically normally distributed. Let μ denote the population mean and σ^2 denote the population variance of RNYSE.

a. Test H_o: $\mu = 0.0003$ versus H_1: $\mu \neq 0.0003$ at significance level $\alpha = 0.04$. At $\alpha = 0.01$. What is the power of each of these tests against the alternative $\mu = 0.0005$?

b. Test H_o: $\mu \geq 0.0003$ versus H_1: $\mu < 0.0003$ at significance level $\alpha = 0.05$. What is the power of this test against the alternative $\mu = 0.0005$?

c. Test H_o: $\sigma^2 = 0.0001$ versus H_1: $\sigma^2 \neq 0.0001$ at significance level $\alpha = 0.01$. What is the power of this test against the alternative $\sigma^2 = 0.000095$?

d. Some analysts claim that opening of international capital markets in the 1980's improved the productivity of capital in large multinational corporations, and this has in turn led to higher mean returns to equity. Make the maintained hypothesis that the variance of returns is constant over the full observation period in nyse.txt. Test the hypothesis that mean return after January 1, 1985 was the same as the mean return prior to that date, versus the alternative that it was not. Use $\alpha = 0.01$.

e. Some analysts claim that the introduction of dynamic hedging strategies and electronic trading, beginning around January 1, 1985, has made the stock market more volatile than it was previously. Test the hypothesis that the variance in RNYSE after that date was higher than the variance before that date, versus the alternative that it was smaller. Use $\alpha = 0.02$. Do not maintain the hypothesis of a common mean return in the two periods.

2. The table gives the investment rate X_i in 8 developed countries. Assume that the X_i are i.i.d. draws from a normally distributed population.

Country	Ratio of Gross Fixed Capital Formation to GDP (Pct.,1993)
Japan	30.1
Germany	22.7
Netherlands	19.7
France	18.9
Canada	18.2
Italy	17.1
U.S.A.	16.1
U.K.	14.9

(a) Test the hypothesis that the population mean investment rate is no greater than 17.0 using a significance level of 95 percent. Be specific about the test statistic, its distribution under the null, and the critical level you would use.

(b) Compute the power of this test against the alternative that the mean is 20.0. Be specific about the distribution that would be used for the power calculation. Give numerical values for the parameters of this distribution, substituting s for σ if necessary. Give a numerical value for the power.

3. Suppose a random sample of size 4 is drawn from a uniform distribution on $[0,\theta]$. You want to test H_o: $\theta \leq 2$ versus H_1: $\theta > 2$ by rejecting the null if $\text{Max}(X_n) > K$. Find the value of K that gives significance level $\alpha = 0.05$. Construct the power curve for this test.

4. A random sample $X_1,...,X_N$ is drawn from a normal density. The variance is known to be 25. You want to test the hypothesis H_o: $\mu \geq 2$ versus the alternative H_1: $\mu < 2$ at significance level $\alpha = 0.01$, and you would like to have power $\pi = 0.99$ against the alternative $\mu = 1$. What sample size do you need?

5. Let $X_1,...,X_N$ be a random sample from a density whose mean is μ and variance is σ^2. Consider estimators of μ of the form $m = \sum_{n=1}^{N} a_{Nn}X_n$, where the a_{Nn} are non-random weights. Under what conditions on the weights is m unbiased? Among unbiased estimators of this form, what weights give minimum variance?

6. A husband and wife are both laid off when the local auto assembly plant closes, and begin searching for new jobs on the same day. The number of weeks Y the wife takes to find a job has a geometric density, $f_Y(y) = p^{y-1}(1-p)$, for y = 1,2,..., where p is a parameter. The number of weeks X it takes the husband to find a job is independent of Y, and also has a geometric density, $f_X(x) = q^{x-1}(1-q)$, for x = 1,2,..., where q is a parameter. The parameters have the values p = 0.5 and q = 0.75. Useful facts about the geometric density $f(z) = r^{z-1}(1-r)$ for z = 1,2,...are (I) $\mathbf{E}_Z(Z-1)\cdots(Z-n) \equiv \sum_{z=1}^{\infty} (z-1)\cdots(z-n)r^{z-1}(1-r) \equiv (r/(1-r))^n\, n!$ for n = 1,2,.. and (ii) $\Pr(Z > t) \equiv \sum_{z=1}^{\infty} r^{z-1}(1-r) \equiv r^t$.

a. What is the expected value of the difference between the lengths of the unemployment spells of the husband and the wife?

b. If the wife is unemployed for at least 6 weeks, what is the expectation of the total number of person-weeks of unemployment insurance the couple will receive, assuming benefits continue for a person as long as he or she is unemployed?

c. What is the probability that the unemployment spell for the husband is greater than that for the wife?

d. What is the expected time until at least one member of the couple is employed?

e. What is the expected time until <u>both</u> husband and wife are employed?

7. Let $X_1,...,X_N$ be a random sample from a uniform distribution on $[0,\theta]$, where θ is an unknown parameter. Show that $T = [(N+1)/N]\cdot\mathrm{Max}(X_n)$ is an unbiased estimator for θ. What is its variance? What is the asymptotic distribution of $N(T-\theta)$?

8. You decide to become a bookmaker for the next Nobel prize in Economics. Three events will determine the outcomes of wagers that are placed:

U. The prize will go to a permanent resident of the U.S.

R. The prize will go to an economist politically more conservative than Milton Friedman

A. The prize will go to someone over 70 years of age

(a) You are given the following probabilities: $P(A|U) = 5/6$, $P(A|R) = 4/5$, $P(A|R\&U) = 6/7$, $P(U|A\&R) = 3/4$, $P(R|A) = 4/7$, $P(R|A\&U) = 3/5$. Find $P(R|U)$, $P(A\&R|U)$, $P(A\cup R|U)$.

(b) If in addition, you are given $P(U) = 3/5$, find $P(A)$, $P(R)$, $P(A\&R)$, $P(A\&R\&U)$.

© You want to sell a ticket that will pay \$2 if one of the events U,R,A occurs, \$4 if two occur, and \$8 if all three occur. What is the minimum price for the ticket such that you will not have an expected loss?

9. If X is standard normal, what is the density of $|X|$? Of exp(X)? Of 1/X?

10. You wish to enter a sealed bid auction for a computer that has a value to you of \$3K if you win the auction. You believe that each competitor will make a bid that is uniformly distributed on the interval [\$2K,\$3K].

(a) If you know that the number N of other bidders is 3, what is the probability that all competitor's bids are less than \$2.9K? What should you bid to maximize your expected profit?

(b) Suppose the number N of other bidders is unknown, but you believe N is Poisson distributed with an expected value $\mathbf{E}N = 5$. What is the probability that the maximum bid from your competitors is less than x? What should you bid to maximize expected profit?