PROBABILITY
AND
STATISTICS
FOR
ECONOMISTS

BRUCE E. HANSEN

This Revision: July 20, 2021

Contents

Preface

This textbook is the first in a two-part series covering the core material typically taught in a one-year Ph.D. course in econometrics. The sequence is

1. *Probability and Statistics for Economists* (this volume)
2. *Econometrics* (the next volume)

The textbooks are written as an integrated series, but either can be used as a stand-alone course textbook.

This first volume covers intermediate-level mathematical statistics. It is a gentle yet a rigorous treatment using calculus but not measure theory. The level of detail and rigor is similar to that of Casella and Berger (2002) and Hogg and Craig (1995). The material is explained using examples at the level of Hogg and Tanis (1997), targeted to students of economics. The goal is to be accessible to students with a variety of backgrounds yet attain full mathematical rigor.

Readers who desire a gentler treatment may try Hogg and Tanis (1997). Readers who desire more detail are recommended to read Casella and Berger (2002) or Shao (2003). Readers wanting a measure-theoretic foundation in probability are recommended to read Ash (1972) or Billingsley (1995). For advanced statistical theory see van der Vaart (1998), Lehmann and Casella (1998), and Lehmann and Romano (2005), each of which has a different emphasis. Mathematical statistics textbooks with similar goals as this textbook include Ramanathan (1993), Amemiya (1994), Gallant (1997), and Linton (2017).

Technical material which is not essential for the main concepts are presented in the starred (*) sections. This material is intended for students interested in the mathematical details. Others may skip these sections with no loss of concepts.

Chapters 1-5 cover probability theory. Chapters 6-18 cover statistical theory.

The end-of-chapter exercises are important parts of the text and are central for learning.

This textbook could be used for a one-semester course. It can also be used for a one-quarter course (as done at the University of Wisconsin) if a selection of topics are skipped. For example, the material in Chapter 3 should probably be viewed as reference rather than taught; Chapter 9 is for advanced students; Chapter 11 can be covered in brief; Chapter 12 can be left for reference; and Chapters 15-18 are optional depending on the instructor.

Acknowledgements

This book and its companion *Econometrics* would not have been possible if it were not for the amazing flow of unsolicited advice, corrections, comments, and questions I have received from students, faculty, and other readers over the twenty years I have worked on this project. I have received emails corrections and comments from so many individuals I have completely lost track of the list. So rather than publish an incomplete list, I simply give an honest and thorough *Thank You* to every single one.

Special thanks go to Xiaoxia Shi, who typed up my handwritten notes for Econ 709 a few years ago, creating a preliminary draft for this manuscript.

My most heartfelt thanks goes to my family: Korinna, Zoe, and Nicholas. Without their love and support over these years this project would not have been possible.

100% of the author's royalties will be re-gifted to charitable purposes.

Mathematical Preparation

Students should be familiar with integral, differential, and multivariate calculus, as well as linear matrix algebra. This is the material typically taught in a four-course undergraduate mathematics sequence at a U.S. university. No prior coursework in probability, statistics, or econometrics is assumed, but would be helpful.

It is also highly recommended, but not necessary, to have studied mathematical analysis and/or a "prove-it" mathematics course. The language of probability and statistics is mathematics. To understand the concepts you need to derive the methods from their principles. This is different from introductory statistics which unfortunately often emphasizes memorization. By taking a mathematical approach little memorization is needed. Instead it requires a facility to work with detailed mathematical derivations and proofs. The reason why it is recommended to have studied mathematical analysis is not that we will be using those results. The reason is that the method of thinking and proof structures are similar. We start with the axioms of probability and build the structure of probability theory from these axioms. Once probability theory is built we construct statistical theory on that base. A timeless introduction to mathematical analysis is Rudin (1976). For those wanting more, Rudin (1987) is recommended.

The appendix to the textbook contains a brief summary of important mathematical results and is included as a reference.

Notation

Real numbers (elements of the real line $\mathbb{R}$, also called **scalars**) are written using lower case italics such as x.

Vectors (elements of $\mathbb{R}^k$) are typically written by lower case italics such as x, and sometimes using lower case bold italics such as $\boldsymbol{x}$ (for matrix algebra expressions), For example, we write

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}.$$

Vectors by default are written as column vectors. The **transpose** of x is the row vector

$$x' = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \end{pmatrix}.$$

There is diversity between fields concerning the choice of notation for the transpose. The notation x' is the most common in econometrics. In statistics and mathematics the notation $x^\top$ is typically used, or occassionally x^t.

Matrices are written using upper case bold italics. For example

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Random variables and vectors are written using upper case italics such as X.

We typically use Greek letters such as β, θ, and σ^2 to denote parameters of a probability model. Estimators are typically denoted by putting a hat "^", tilde "~" or bar "-" over the corresponding letter, e.g. $\widehat{\beta}$ and $\widetilde{\beta}$ are estimators of β.

Common Symbols

a	scalar
a or $\boldsymbol{a}$	vector
$\boldsymbol{A}$	matrix
X	random variable or vector
$\mathbb{R}$	real line
$\mathbb{R}_+$	positive real line
$\mathbb{R}^k$	Euclidean k space
$\mathbb{P}[A]$	probability
$\mathbb{P}[A \mid B]$	conditional probability
$F(x)$	cumulative distribution function
$\pi(x)$	probability mass function
$f(x)$	probability density function
$\mathbb{E}[X]$	mathematical expectation
$\mathbb{E}[Y \mid X = x]$, $\mathbb{E}[Y \mid X]$	conditional expectation
$\text{var}[X]$	variance, or covariance matrix
$\text{var}[Y \mid X]$	conditional variance
$\text{cov}(X, Y)$	covariance
$\text{corr}(X, Y)$	correlation
$\overline{X}_n$	sample mean
$\widehat{\sigma}^2$	sample variance
s^2	biased-corrected sample variance
$\widehat{\theta}$	estimator
$s\left(\widehat{\theta}\right)$	standard error of estimator
$\lim_{n\to\infty}$	limit
$\operatorname{plim}_{n\to\infty}$	probability limit
$\to$	convergence
$\xrightarrow[p]{}$	convergence in probability
$\xrightarrow[d]{}$	convergence in distribution
$L_n(\theta)$	likelihood function
$\ell_n(\theta)$	log-likelihood function
$\mathscr{I}_\theta$	information matrix
$\text{N}(0,1)$	standard normal distribution
$\text{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
χ^2_k	chi-square distribution with k degrees of freedom

$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\text{tr}\,\boldsymbol{A}$	trace of matrix $\boldsymbol{A}$
$\boldsymbol{A}'$	vector or matrix transpose
$\boldsymbol{A}^{-1}$	matrix inverse
$\boldsymbol{A} > 0$	positive definite
$\boldsymbol{A} \geq 0$	positive semi-definite
$\|a\|$	Euclidean norm
$\mathbb{1}\{A\}$	indicator function (1 if A is true, else 0)
$\simeq$	approximate equality
$\sim$	is distributed as
$\log(x)$	natural logarithm
$\exp(x)$	exponential function
$\sum_{i=1}^{n}$	summation from $i = 1$ to $i = n$

Greek Alphabet

It is common in economics and econometrics to use Greek characters to augment the Latin alphabet. The following table lists the various Greek characters and their pronunciations in English. The second character, when listed, is upper case (except for ϵ which is an alternative script for ε.)

Greek Character	Name	Latin Keyboard Equivalent
α	alpha	a
β	beta	b
γ, Γ	gamma	g
δ, Δ	delta	d
ε, ϵ	epsilon	e
ζ	zeta	z
η	eta	h
θ, Θ	theta	y
ι	iota	i
κ	kappa	k
λ, Λ	lambda	l
μ	mu	m
ν	nu	n
ξ, Ξ	xi	x
π, Π	pi	p
ρ	rho	r
σ, Σ	sigma	s
τ	tau	t
υ	upsilon	u
ϕ, Φ	phi	f
χ	chi	x
ψ, Ψ	psi	c
ω, Ω	omega	w

Chapter 1

Basic Probability Theory

1.1 Introduction

Probability theory is foundational for economics and econometrics. Probability is the mathematical language used to handle uncertainty, which is central for modern economic theory. Probability theory is also the foundation of mathematical statistics, which is the foundation of econometric theory.

Probability is used to model uncertainty, variability, and randomness. When we say that something is uncertain we mean that the outcome is unknown. For example, how many students will there be in next year's Ph.D. entering class at your university? By variability we mean that the outcome is not the same across all occurrences. For example, the number of Ph.D. students fluctuate from year to year. By randomness we mean that the variability has some sort of pattern. For example, the number of Ph.D. students may fluctuate between 20 and 30, with 25 more likely than either 20 or 30. Probability gives us a mathematical language to describe uncertainty, variability, and randomness.

1.2 Outcomes and Events

Suppose you take a coin, flip it in the air, and let it land on the ground. What will happen? Will the result be "heads" (H) or "tails" (T)? We do not know the result in advance so we describe the outcome as **random**.

Suppose you record the change in the value of a stock index over a period of time. Will the value increase or decrease? Again, we do not know the result in advance so we describe the outcome as random.

Suppose you select an individual at random and survey them about their economic situation. What is their hourly wage? We do not know in advance. The lack of foreknowledge leads us to describe the outcome as random.

We will use the following terms.

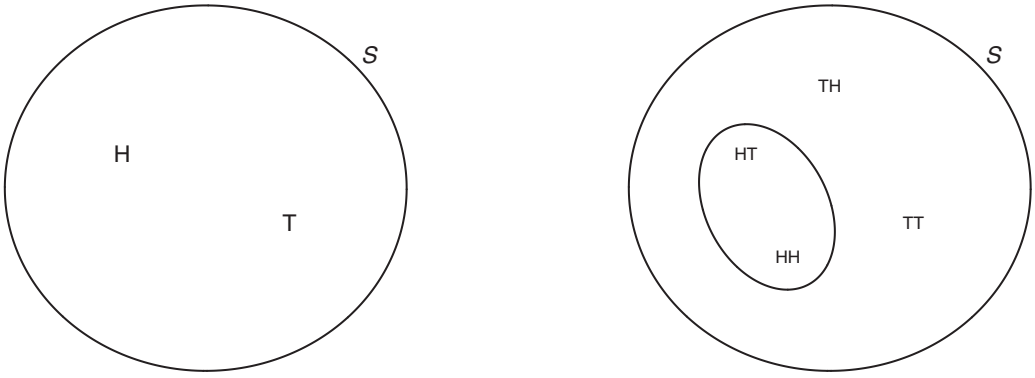
An **outcome** is a specific result. For example, in a coin flip an outcome is either H or T. If two coins are flipped in sequence may write an outcome as HT for a head and then a tails. A roll of a six-sided die has the six outcomes $\{1,2,3,4,5,6\}$.

The **sample space** S is the set of all possible outcomes. In a coin flip the sample space is $S = \{H, T\}$. If two coins are flipped the sample space is $S = \{HH, HT, TH, TT\}$.

An **event** A is a subset of outcomes in S. An example event from the roll of a die is $A = \{1,2\}$.

The one-coin and two-coin sample spaces are illustrated in Figure 1.1. The event $\{HH, HT\}$ is illustrated by the ellipse in Figure 1.1(b).

Set theoretic manipulations are helpful in describing events. We will use the following concepts.



(a) One Coin (b) Two Coins

Figure 1.1: Sample Space

Definition 1.1 For events A and B

1. A is a **subset** of B, written $A \subset B$, if every element of A is an element of B.
2. The event with no outcomes $\varnothing = \{\}$ is called the **null** or **empty set**.
3. The **union** $A \cup B$ is the collection of all outcomes that are in either A **or** B (or both).
4. The **intersection** $A \cap B$ is the collection of elements that are in both A **and** B.
5. The **complement** A^c of A are all outcomes in S which are not in A.
6. The events A and B are **disjoint** if they have no outcomes in common: $A \cap B = \varnothing$.
7. The events $A_1, A_2, ...$ are a **partition** of S if they are mutually disjoint and their union is S.

Events satisfy the rules of set operations, including the commutative, associative, and distributive laws. The following is useful.

Theorem 1.1 Partitioning Theorem. If $\{B_1, B_2, \cdots\}$ is a partition of S, then for any event A,

$$A = \bigcup_{i=1}^{\infty} (A \cap B_i).$$

The sets $(A \cap B_i)$ are mutually disjoint.

A proof is provided in Section 1.15.

1.3 Probability Function

Definition 1.2 A function $\mathbb{P}$ which assigns a numerical value to events[1] is called a **probability function** if it satisfies the following **Axioms of Probability**:

1. $\mathbb{P}[A] \geq 0$.
2. $\mathbb{P}[S] = 1$.
3. If $A_1, A_2, ...$ are disjoint then $\mathbb{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \sum_{j=1}^{\infty} \mathbb{P}\left[A_j\right]$.

In this textbook we use the notation $\mathbb{P}[A]$ for the probability of an event A. Other common notation includes $P(A)$ and $\Pr(A)$.

Let us understand the definition. The phrase "a function $\mathbb{P}$ which assigns a numerical value to events" means that $\mathbb{P}$ is a function from the space of events to the real line. Thus probabilities are numbers. Now consider the Axioms. The first Axiom states that probabilities are non-negative. The second Axiom is essentially a normalization: the probability that "something happens" is one.

The third Axiom imposes considerable structure. It states that probabilities are additive on disjoint events. That is, if A and B are disjoint then

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B].$$

Take, for example, the roll of a six-sided die which has the possible outcomes $\{1,2,3,4,5,6\}$. Since the outcomes are mutually disjoint the third Axiom states that $\mathbb{P}[1 \text{ or } 2] = \mathbb{P}[1] + \mathbb{P}[2]$.

When using the third Axiom it is important to be careful that it is applied only to disjoint events. Take, for example, the roll of a pair of dice. Let A be the event "1 on the first roll" and B the event "1 on the second roll". It is tempting to write $\mathbb{P}[\text{"1 on either roll"}] = \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ but the second equality is incorrect since A and B are not disjoint. The outcome "1 on both rolls" is an element of both A and B.

Any function $\mathbb{P}$ which satisfies the axioms is a valid probability function. Take the coin flip example. One valid probability function sets $\mathbb{P}[H] = 0.5$ and $\mathbb{P}[T] = 0.5$. (This is typically called a **fair coin**.) A second valid probability function sets $\mathbb{P}[H] = 0.6$ and $\mathbb{P}[T] = 0.4$. However a function which sets $\mathbb{P}[H] = -0.6$ is not valid (it violates the first axiom), and a function which sets $\mathbb{P}[H] = 0.6$ and $\mathbb{P}[T] = 0.6$ is not valid (it violates the second axiom).

While the definition states that a probability function must satisfy certain rules it does not describe the *meaning* of probability. The reason is because there are multiple interpretations. One view is that probabilities are the relative frequency of outcomes as in a controlled experiment. The probability that the stock market will increase is the frequency of increases. The probability that an unemployment duration will exceed one month is the frequency of unemployment durations exceeding one month. The probability that a basketball player will make a free throw shot is the frequency with which the player makes free throw shots. The probability that a recession will occur is relative frequency of recessions. In some examples this is conceptually straightforward as the experiment repeats or has multiple occurances. In other cases an situation occurs exactly once and will never be repeated. As I write this paragraph questions of uncertainty of general interest include "Will global warming exceed 2 degrees?" and "When will the COVID-19 epidemic end?" In these cases it is difficult to interpret a probability as a relative frequency as the outcome can only occur once. The interpretation can be salvaged by viewing "relative frequency" abstractly by imagining many alternative universes which start from the same initial conditions but evolve randomly. While this solution works (technically) it is not completely satisfactory.

[1] For events in a sigma field. See Section 1.14.

Another view is that probability is subjective. This view is that probabilities can be interpreted as degrees of belief. If I say "The probability of rain tomorrow is 80%" I mean that this is my personal subjective assessment of the likelihood based on the information available to me. This may seem too broad as it allows for arbitrary beliefs, but the subjective interpretation requires subjective probability to follow the axioms and rules of probability. A major disadvantage associated with this approach is that it is not necessarily appropriate for scientific discourse.

What is common between the two definitions is that the probability function follows the same axioms – otherwise the label "probability" should not be used.

We illustrate the concept with two real-world examples. The first is from finance. Let U be the event that the S&P stock index increases in a given week, and let D be the event that the index decreases. This is similar to a coin flip. The sample space is $\{U, D\}$. We compute[2] that $\mathbb{P}[U] = 0.57$ and $\mathbb{P}[D] = 0.43$. The probability 57% of an increase is somewhat higher than a fair coin. The probability interpretation is that the index will increase in value in 57% of randomly selected weeks.

The second example concerns wage rates in the United States. Take a randomly selected wage earner. Let H be the event that their wage rate exceeds \$25/hour and L be the event that their wage rate is less than \$25/hour. Again the structure is similar to a coin flip. We calculate[3] that $\mathbb{P}[H] = 0.31$ and $\mathbb{P}[L] = 0.69$. To interpret this as a probability we can imagine surveying a random individual. Before the survey we know nothing about the individual. Their wage rate is uncertain and random.

1.4 Properties of the Probability Function

The following properties of probability functions can be derived from the axioms.

Theorem 1.2 For events A and B,

1. $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$.
2. $\mathbb{P}[\varnothing] = 0$.
3. $\mathbb{P}[A] \leq 1$.
4. **Monotone Probability Inequality**: If $A \subset B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$.
5. **Inclusion-Exclusion Principle**: $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.
6. **Boole's Inequality**: $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$.
7. **Bonferroni's Inequality**: $\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1$.

Proofs are provided in Section 1.15.

Part 1 states that the probability that an event does not occur equals one minus the probability that the event occurs.

Part 2 states that "nothing happens" occurs with zero probability. Remember this when asked "What happened today in class?"

Part 3 states that probabilities cannot exceed one.

Part 4 shows that larger sets necessarily have larger probability.

Part 5 is a useful decomposition of the probability of the union of two events.

[2] Calculated from a sample of 3584 weekly prices of the S&P Index between 1950 and 2017.

[3] Calculated from a sample of 50,742 U.S. wage earners in 2009.

Parts 6 & 7 are implications of the inclusion-exclusion principle and are frequently used in probability calculations. Boole's inequality shows that the probability of a union is bounded by the sum of the individual probabilities. Bonferroni's inequality shows that the probability of an intersection is bounded below by an expression involving the individual probabilities. A useful feature of these inequalities is that the right-hand-sides only depend on the individual probabilities.

A further comment related to Part 2 is that any event which occurs with probability zero or one is called **trivial**. Such events are essentially non-random. In the coin flip example we could define the sample space as $S = \{H, T, \text{Edge}, \text{Disappear}\}$ where "Edge" means the coin lands on its edge and "Disappear" means the coin disappears into the air. If $\mathbb{P}\left[\text{Edge}\right] = 0$ and $\mathbb{P}\left[\text{Disappear}\right] = 0$ then these events are trivial.

1.5 Equally-Likely Outcomes

When we build probability calculations from foundations it is often useful to consider settings where symmetry implies that a set of outcomes are equally likely. Standard examples are a coin flip and the toss of a die. We describe a coin as **fair** if the event of a heads is as equally likely as the event of a tails. We describe a die as **fair** if the event of each face is equally likely. Applying the Axioms we deduce the following.

Theorem 1.3 Principle of Equally-Likely Outcomes: If an experiment has N outcomes $a_1, ..., a_N$ which are symmetric in the sense that each outcome is equally likely, then $\mathbb{P}[a_i] = \frac{1}{N}$.

For example, a fair coin satisfies $\mathbb{P}[H] = \mathbb{P}[T] = 1/2$ and a fair die satisfies $\mathbb{P}[1] = \cdots = \mathbb{P}[6] = 1/6$.

In some contexts deciding which outcomes are symmetric and equally likely can be confusing. Take the two coin example. We could define the sample space as {HH,TT,HT} where HT means "one head and one tail". If we guess that all outcomes are equally likely we would set $\mathbb{P}[HH] = 1/3$, etc. However if we define the sample space as {HH,TT,HT,TH} and guess all outcomes are equally likely, we would find $\mathbb{P}[HH] = 1/4$. Both answers (1/3 and 1/4) cannot be correct. The implication is that we should not apply the principle of equally-likely outcomes simply because there is a list of outcomes. Rather there should be a justifiable reason why the outcomes are equally likely. In this two coin example there is no principled reason for symmetry without further analysis so the property should not be applied. We return to this question in Section 1.8.

1.6 Joint Events

Take two events H and C. For concreteness, let H be the event that an individual's wage exceeds \$25/hour and let C be the event that the individual has a college degree. We are interested in the probability of the joint event $H \cap C$. This is "H and C" or in words that the individual's wage exceeds \$25/hour and they have a college degree. Previously we reported that $\mathbb{P}[H] = 0.31$. We can similarly calculate that $\mathbb{P}[C] = 0.36$. What about the joint event $H \cap C$?

From Theorem 1.2 we can deduce that $0 \leq \mathbb{P}[H \cap C] \leq 0.31$. (The upper bound is Bonferroni's inequality.) Thus from the knowledge of $\mathbb{P}[H]$ and $\mathbb{P}[C]$ alone we can bound the joint probability but not determine its value. It turns out that the actual[4] probability is $\mathbb{P}[H \cap C] = 0.19$.

From the three known probabilities and the properties of Theorem 1.2 we can calculate the probabilities of the various intersections. We display the results in the following chart. The four numbers in the central box are the probabilities of the joint events; for example 0.19 is the probability of both a high

[4] Calculated from the same sample of 50,742 U.S. wage earners in 2009.

wage and a college degree. The largest of the four probabilities is 0.52: the joint event of a low wage and no college degree. The four probabilities sum to one as the events are a partition of the sample space. The sum of the probabilities in each column are reported in the bottom row: the probabilities of a college degree and no degree, respectively. The sums by row are reported in the right-most column: the probabilities of a high and low wage, respectively.

Joint Probabilities: Wages and Education

	C	N	Any Education
H	0.19	0.12	0.31
L	0.17	0.52	0.69
Any Wage	0.36	0.64	1.00

As another illustration we examine stock price changes. We reported before that the probability of an increase in the S&P stock index in a given week is 57%. Now consider the change in the stock index over two sequential weeks. What is the joint probability? We display the results in the following chart. U_t means the index increases, D_t means the index decreases, U_{t-1} means the index increases in the previous week, and D_{t-1} means that the index decreases in the previous week.

Joint Probabilities: Stock Returns

	U_{t-1}	D_{t-1}	Any Past Return
U_t	0.322	0.245	0.567
D_t	0.245	0.188	0.433
Any Return	0.567	0.433	1.00

The four numbers in the central box sum to one since they are a partition of the sample space. We can see that the probability that the stock price increases for two weeks in a row is 32% and that it decreases for two weeks in a row is 19%. The probability is 25% for an increase followed by a decrease, and also 25% for a decrease followed by an increase.

1.7 Conditional Probability

Take two events A and B. For example, let A be the event "Receive a grade of A on the econometrics exam" and let B be the event "Study econometrics 12 hours a day". We might be interested in the question: Does B affect the likelihood of A? Alternatively we may be interested in questions such as: Does attending college affect the likelihood of obtaining a high wage? Or: Do tariffs affect the likelihood of price increases? These are questions of **conditional probability**.

Abstractly, consider two events A and B. Suppose that we know that B has occurred. Then the only way for A to occur is if the outcome is in the intersection $A \cap B$. So we are asking: "What is the probability that $A \cap B$ occurs, given that B occurs?" The answer is not simply $\mathbb{P}[A \cap B]$. Instead we can think of the "new" sample space as B. To do so we normalize all probabilities by $\mathbb{P}[B]$. We arrive at the following definition.

Definition 1.3 If $\mathbb{P}[B] > 0$ the **conditional probability** of A given B is

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

The notation "$A \mid B$" means "A given B" or "A assuming that B is true". To add clarity we will sometimes refer to $\mathbb{P}[A]$ as the **unconditional probability** to distinguish it from $\mathbb{P}[A \mid B]$.

For example, take the roll of a fair die. Let $A = \{1,2,3,4\}$ and $B = \{4,5,6\}$. The intersection is $A \cap B = \{4\}$ which has probability $\mathbb{P}[A \cap B] = 1/6$. The probability of B is $\mathbb{P}[B] = 1/2$. Thus $\mathbb{P}[A \mid B] = (1/6)/(1/2) = 1/3$. This can also be calculated by observing that conditional on B, the events $\{4\}$, $\{5\}$, and $\{6\}$ each have probability $1/3$. Event A only occurs given B if $\{4\}$ occurs. Thus $\mathbb{P}[A \mid B] = \mathbb{P}[4 \mid B] = 1/3$.

Consider our example of wages and college education. From the probabilities reported in the previous section we can calculate that

$$\mathbb{P}[H \mid C] = \frac{\mathbb{P}[H \cap C]}{\mathbb{P}[C]} = \frac{0.19}{0.36} = 0.53$$

and

$$\mathbb{P}[H \mid N] = \frac{\mathbb{P}[H \cap N]}{\mathbb{P}[N]} = \frac{0.12}{0.64} = 0.19.$$

There is a considerable difference in the conditional probability of receiving a high wage conditional on a college degree, 53% versus 19%.

As another illustration we examine stock price changes. We calculate that

$$\mathbb{P}[U_t \mid U_{t-1}] = \frac{\mathbb{P}[U_t \cap U_{t-1}]}{\mathbb{P}[U_{t-1}]} = \frac{0.322}{0.567} = 0.568$$

and

$$\mathbb{P}[U_t \mid D_{t-1}] = \frac{\mathbb{P}[U_t \cap D_{t-1}]}{\mathbb{P}[D_{t-1}]} = \frac{0.245}{0.433} = 0.566.$$

In this case the two conditional probabilities are essentially identical. Thus the probability of a price increase in a given week is unaffected by the previous week's result. This is an important special case and is explored further in the next section.

1.8 Independence

We say that events are **independent** if their occurrence is unrelated, or equivalently that the knowledge of one event does not affect the conditional probability of the other event. Take two coin flips. If there is no mechanism connecting the two flips we would typically expect that neither flip is affected by the outcome of the other. Similarly if we take two die throws we typically expect there is no mechanism connecting the dice and thus no reason to expect that one is affected by the outcome of the other. As a third example, consider the crime rate in London and the price of tea in Shanghai. There is no reason to expect one of these two events to affect the other event[5]. In each of these cases we describe the events as independent.

This discussion implies that two unrelated (independent) events A and B will satisfy the properties $\mathbb{P}[A \mid B] = \mathbb{P}[A]$ and $\mathbb{P}[B \mid A] = \mathbb{P}[B]$. In words, the probability that a coin is H is unaffected by the outcome (H or T) of another coin. From the definition of conditional probability this implies $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. We use this as the formal definition.

Definition 1.4 The events A and B are **statistically independent** if $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$.

We typically use the simpler label **independent** for brevity. As an immediate consequence of the derivation we obtain the following equivalence.

[5] Except in a James Bond movie.

Theorem 1.4 If A and B are independent with $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$, then

$$\mathbb{P}[A \mid B] = \mathbb{P}[A]$$
$$\mathbb{P}[B \mid A] = \mathbb{P}[B].$$

Consider the stock index illustration. We found that $\mathbb{P}[U_t \mid U_{t-1}] = 0.57$ and $\mathbb{P}[U_t \mid D_{t-1}] = 0.57$. This means that the probability of an increase is unaffected by the outcome from the previous week. This satisfies the definition of independence. It follows that the events U_t and U_{t-1} are independent.

When events are independent then joint probabilities can be calculated by multiplying individual probabilities. Take two independent coin flips. Write the possible results of the first coin as $\{H_1, T_1\}$ and the possible results of the second coin as $\{H_2, T_2\}$. Let $p = \mathbb{P}[H_1]$ and $q = \mathbb{P}[H_2]$. We obtain the following chart for the joint probabilities

Joint Probabilities: Independent Events

	H_1	T_1	
H_2	pq	$(1-p)q$	q
T_2	$p(1-q)$	$(1-p)(1-q)$	$1-q$
	p	$1-p$	1

The chart shows that the four joint probabilities are determined by p and q, the probabilities of the individual coins. The entries in each column sum to p and $1-p$, and the entries in each row sum to q and $1-q$.

If two events are not independent we say that they are **dependent**. In this case the joint event $A \cap B$ occurs at a different rate than predicted if the events were independent.

For example consider wage rates and college degrees. We have already shown that the conditional probability of a high wage is affected by a college degree, which demonstrates that the two events are dependent. What we now do is see what happens when we calculate the joint probabilities from the individual probabilities under the (false) assumption of independence. The results are shown in the following chart.

Joint Probabilities: Wages and Education

	C	N	Any Education
H	0.11	0.20	0.31
L	0.25	0.44	0.69
Any Wage	0.36	0.64	1.00

The entries in the central box are obtained by multiplication of the individual probabilities, i.e. $\mathbb{P}[H \cap C] = 0.31 \times 0.36 = 0.11$. What we see is that the diagonal entries are much smaller, and the off-diagonal entries are much larger, then the corresponding correct joint probabilities. In this example the joint events $H \cap C$ and $L \cap N$ occur more frequently than that predicted if wages and education were independent.

We can use independence to make probability calculations. Take the two coin example. If two sequential fair coin flips are independent then the probability that both are heads is

$$\mathbb{P}[H_1 \cap H_2] = \mathbb{P}[H_1] \times \mathbb{P}[H_2] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

This answers the question raised in Section 1.5. The probability of HH is 1/4, not 1/3. The key is the assumption of independence not how the outcomes are listed.

As another example, consider throwing a pair of fair die. If the two die are independent then the probability of two 1's is $\mathbb{P}[1] \times \mathbb{P}[1] = 1/36$.

Naïvely, one might think that independence relates to disjoint events, but the converse if true. If A and B are disjoint then they cannot be independent. That is, disjointness means $A \cap B = \varnothing$, and by part 2 of Theorem 1.2

$$\mathbb{P}[A \cap B] = \mathbb{P}[\varnothing] = 0 \neq \mathbb{P}[A]\mathbb{P}[B]$$

and the right-side is non-zero under the definition of independence.

Independence lies at the core of many probability calculations. If you can break an event into the joint occurance of several independent events, then the probability of the event is the product of the individual probabilities.

Take, for example, the two coin example and the event $\{HH, HT\}$. This equals {First coin is H, Second coin is either H or T}. If the two coins are independent this has probability

$$\mathbb{P}[H] \times \mathbb{P}[H \text{ or } T] = \frac{1}{2} \times 1 = \frac{1}{2}.$$

As a bit more complicated example, what is the probability of "rolling a seven" from a pair of die, meaning that the two faces add to seven? We can calculate this as follows. Let (x, y) denote the outcomes from the two (ordered) die. The following outcomes yield a seven: $\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$. The outcomes are disjoint. Thus by the third Axiom the probability of a seven is the sum

$$\mathbb{P}[7] = \mathbb{P}[1,6] + \mathbb{P}[2,5] + \mathbb{P}[3,4] + \mathbb{P}[4,3] + \mathbb{P}[5,2] + \mathbb{P}[6,1].$$

Assume the two die are independent of one another so the probabilities are products. For fair die the above expression equals

$$\begin{aligned}
&\mathbb{P}[1] \times \mathbb{P}[6] + \mathbb{P}[2] \times \mathbb{P}[5] + \mathbb{P}[3] \times \mathbb{P}[4] + \mathbb{P}[4] \times \mathbb{P}[3] + \mathbb{P}[5] \times \mathbb{P}[2] + \mathbb{P}[6] \times \mathbb{P}[1] \\
&= \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} \\
&= 6 \times \frac{1}{6^2} \\
&= \frac{1}{6}.
\end{aligned}$$

Now suppose that the dice are not fair. Suppose they are independent, but each is weighted so that the probability of a "1" is 2/6 and the probability of a "6" is 0. We revise the calculation to find

$$\begin{aligned}
&\mathbb{P}[1] \times \mathbb{P}[6] + \mathbb{P}[2] \times \mathbb{P}[5] + \mathbb{P}[3] \times \mathbb{P}[4] + \mathbb{P}[4] \times \mathbb{P}[3] + \mathbb{P}[5] \times \mathbb{P}[2] + \mathbb{P}[6] \times \mathbb{P}[1] \\
&= \frac{2}{6} \times \frac{0}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{1}{6} \times \frac{1}{6} + \frac{0}{6} \times \frac{2}{6} \\
&= \frac{1}{9}.
\end{aligned}$$

1.9 Law of Total Probability

An important relationship can be derived from the partitioning theorem (Theorem 1.1) which states that if $\{B_i\}$ is a partition of the sample space S then

$$A = \bigcup_{i=1}^{\infty} (A \cap B_i).$$

Since the events $(A \cap B_i)$ are disjoint an application of the third Axiom and the definition of conditional probability implies

$$\mathbb{P}[A] = \sum_{i=1}^{\infty} \mathbb{P}[A \cap B_i] = \sum_{i=1}^{\infty} \mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i].$$

This is called the Law of Total Probability.

Theorem 1.5 Law of Total Probability. If $\{B_1, B_2, \cdots\}$ is a partition of S and $\mathbb{P}[B_i] > 0$ for all i then

$$\mathbb{P}[A] = \sum_{i=1}^{\infty} \mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i].$$

For example, take the roll of a fair die and the events $A = \{1,3,5\}$ and $B_j = \{j\}$. We calculate that

$$\sum_{i=1}^{6} \mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i] = 1 \times \frac{1}{6} + 0 \times \frac{1}{6} + 1 \times \frac{1}{6} + 0 \times \frac{1}{6} + 1 \times \frac{1}{6} + 0 \times \frac{1}{6} = \frac{1}{2}.$$

which equals $\mathbb{P}[A] = 1/2$ as claimed.

1.10 Bayes Rule

A famous result is credited to Reverend Thomas Bayes.

Theorem 1.6 Bayes Rule. If $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$ then

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\,\mathbb{P}[A]}{\mathbb{P}[B \mid A]\,\mathbb{P}[A] + \mathbb{P}[B \mid A^c]\,\mathbb{P}[A^c]}.$$

Proof. The definition of conditional probability (applied twice) implies

$$\mathbb{P}[A \cap B] = \mathbb{P}[A \mid B]\,\mathbb{P}[B] = \mathbb{P}[B \mid A]\,\mathbb{P}[A].$$

Solving we find

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\,\mathbb{P}[A]}{\mathbb{P}[B]}.$$

Applying the law of total probability to $\mathbb{P}[B]$ using the partition $\{A, A_c\}$ we obtain the stated result. ■

Bayes Rule is terrifically useful in many contexts.

As one example suppose you walk by a sports bar where you see a group of people watching a sports match which involves a popular local team. Suppose you suddenly hear a roar of excitement from the bar. Did the local team just score? To investigate this by Bayes Rule, let $A = \{\text{score}\}$ and $B = \{\text{crowd roars}\}$. Assume that $\mathbb{P}[A] = 1/10$, $\mathbb{P}[B \mid A] = 1$ and $\mathbb{P}[B \mid A^c] = 1/10$ (there are other events which can cause a roar). Then

$$\mathbb{P}[A \mid B] = \frac{1 \times \frac{1}{10}}{1 \times \frac{1}{10} + \frac{1}{10} \times \frac{9}{10}} = \frac{10}{19} \simeq 53\%.$$

This is slightly over one-half. Under these assumptions the roar of the crowd is informative though not definitive[6].

[6]Consequently, it is reasonable to enter the sports bar to learn the truth!

As another example suppose there are two types of workers: hard workers (H) and lazy workers (L). Suppose that we know from previous experience that $\mathbb{P}[H] = 1/4$ and $\mathbb{P}[L] = 3/4$. Suppose we can administer a screening test to determine if an applicant is a hard worker. Let T be the event that an applicant has a high score on the test. Suppose that $\mathbb{P}[T \mid H] = 3/4$ and $\mathbb{P}[T \mid L] = 1/4$. That is, the test has some signal but is not perfect. We are interested in calculating $\mathbb{P}[H \mid T]$, the conditional probability that an applicant is a hard worker, given that they have a high test score. Bayes Rule tells us

$$\mathbb{P}[H \mid T] = \frac{\mathbb{P}[T \mid H]\mathbb{P}[H]}{\mathbb{P}[T \mid H]\mathbb{P}[H] + \mathbb{P}[T \mid L]\mathbb{P}[L]} = \frac{\frac{3}{4} \times \frac{1}{4}}{\frac{3}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{3}{4}} = \frac{1}{2}.$$

The probability the applicant is a hard worker is only 50%! Does this mean the test is useless? Consider the question: What is the probability an applicant is a hard worker, given that they had a poor (P) test score? We find

$$\mathbb{P}[H \mid P] = \frac{\mathbb{P}[P \mid H]\mathbb{P}[H]}{\mathbb{P}[P \mid H]\mathbb{P}[H] + \mathbb{P}[P \mid L]\mathbb{P}[L]} = \frac{\frac{1}{4} \times \frac{1}{4}}{\frac{1}{4} \times \frac{1}{4} + \frac{3}{4} \times \frac{3}{4}} = \frac{1}{10}.$$

This is only 10%. Thus what the test tells us is that if an applicant scores high we are uncertain about that applicant's work habits, but if an applicant scores low it is unlikely that they are a hard worker.

To revisit our real-world example of education and wages, recall that we calculated that the probability of a high wage (H) given a college degree (C) is $\mathbb{P}[H \mid C] = 0.53$. Applying Bayes Rule we can find the probability that an individual has a college degree given that they have a high wage is

$$\mathbb{P}[C \mid H] = \frac{\mathbb{P}[H \mid C]\mathbb{P}[C]}{\mathbb{P}[H]} = \frac{0.53 \times 0.36}{0.31} = 0.62.$$

The probability of a college degree given that they have a low wage (L) is

$$\mathbb{P}[C \mid L] = \frac{\mathbb{P}[L \mid C]\mathbb{P}[C]}{\mathbb{P}[L]} = \frac{0.47 \times 0.36}{0.69} = 0.25.$$

Thus given this one piece of information (if the wage is above or below $25) we have probabilistic information about whether the individual has a college degree.

1.11 Permutations and Combinations

For some calculations it is useful to count the number of individual outcomes. For some of these calculations the concepts of counting rules, permutations, and combinations are useful.

The first definition we explore is the counting rule, which shows how to count options when we combine tasks. For example, suppose you own ten shirts, three pairs of jeans, five pairs of socks, four coats and two hats. How many clothing outfits can you create, assuming you use one of each category? The answer is $10 \times 3 \times 5 \times 4 \times 2 = 1200$ distinct outfits[7].

Theorem 1.7 Counting Rule. If a job consists of K separate tasks, the k^{th} of which can be done in n_k ways, then the entire job can be done in $n_1 n_2 \cdots n_K$ ways.

The counting rule is intuitively simple but is useful in a variety of modeling situations.

The second definition we explore is a **permutation**. A permutation is a re-arrangement of the order. Suppose you take a classroom of 30 students. How many ways can you arrange their order? Each arrangement is called a permutation. To calculate the number of permutations, observe that there are

[7] Remember this when you (or a friend) asserts "I have nothing to wear!"

30 students who can be placed first. Given this choice there are 29 students who can be placed second. Given these two choices there are 28 students for the third position, and so on. The total number of permutations is

$$30 \times 29 \times \cdots \times 1 = 30!$$

Here, the symbol ! denotes the factorial. (See Section A.3.)

The general solution is as follows.

Theorem 1.8 The number of **permutations** of a group of N objects is $N!$

Suppose we are trying to select an ordered five-student team from a 30-student class for a competition. How many ordered groups of five can are there? The calculation is much the same as above, but we stop once the fifth position is filled. Thus the number is

$$30 \times 29 \times 28 \times 27 \times 26 = \frac{30!}{25!}.$$

The general solution is as follows.

Theorem 1.9 The number of **permutations** of a group of N objects taken K at a time is

$$P(N,K) = \frac{N!}{(N-K)!}$$

The third definition we introduce is **combinations**. A combination is an unordered group of objects. For example revisit the idea of sending a five-student team for a competition but now assume that the team is unordered. Then the question is: How many five-member teams can we construct from a class of 30 students? In general, how many groups of K objects can be extracted from a group of N objects? We call this the number of combinations.

The extreme cases are easy. If $K = 1$ then there are N combinations (each individual student). If $K = N$ then there is one combination (the entire class). The general answer can be found by noting that the number of ordered groups is the number of permutations $P(N,K)$. Each group of K can be ordered $K!$ ways (since this is the number of permutations of a group of K). This means that the number of unordered groups is $P(N,K)/K!$. We have found the following.

Theorem 1.10 The number of **combinations** of a group of N objects taken K at a time is

$$\binom{N}{K} = \frac{N!}{K!\,(N-K)!}.$$

The symbol $\binom{N}{K}$, in words "N choose K", is a commonly-used notation for combinations. They are also known as the **binomial coefficients**. The latter name is because they are the coefficients from the binomial expansion.

Theorem 1.11 Binomial Theorem. For any integer $N \geq 0$

$$(a+b)^N = \sum_{K=0}^{N} \binom{N}{K} a^K b^{N-K}.$$

The proof of the binomial theorem is given in Section 1.15.

The permutation and combination rules introduced in this section are useful in certain counting applications but may not be necessary for a general understanding of probability. My view is that the tools should be understood but not memorized. Rather, the tools can be looked up when needed.

1.12 Sampling With and Without Replacement

Consider the problem of sampling from a finite set. For example, consider a $2 Powerball lottery ticket which consists of five integers each between 1 and 69. If all five numbers match the winning numbers, the player wins[8] $1 million!

To calculate the probability of winning the lottery we need to count the number of potential tickets. The answer depends on two factors: (1) Can the numbers repeat? (2) Does the order matter? The number of tickets could be four distinct numbers depending on the two choices just described.

The first question, of whether a number can repeat or not, is called "sampling with replacement" versus "sampling without replacement". In the actual Powerball game 69 ping-pong balls are numbered and put in a rotating air machine with a small exit. As the balls bounce around some find the exit. The first five to exit are the winning numbers. In this setting we have "sampling without replacement" as once a ball exits it is not among the remaining balls. A consequence for the lottery is that a winning ticket cannot have duplicate numbers. However, an alternative way to play the game would be to extract the first ball, replace it into the chamber, and repeat. This would be "sampling with replacement". In this game a winning ticket could have repeated numbers.

The second question, of whether the order matters, is the same as the distinction between permutations and combinations as discussed in the previous section. In the case of the Powerball game the balls emerge in a specific order. However, this order is ignored for the purpose of a winning ticket. This is the case of unordered sets. If the rules of the game were different the order could matter. If so we would use the tools of ordered sets.

We now describe the four sampling problems. We want to find the number of groups of size K which can be taken from N items, e.g. the number of five integers taken from the set $\{1,...,69\}$.

Ordered, with replacement. Consider selecting the items in sequence. The first item can be any of the N, the second can be any of the N, the third can be any of the N, etc. So by the counting rule the total number of possible groups is

$$N \times N \times \cdots \times N = N^K.$$

In the powerball example this is

$$69^5 = 1,564,031,359.$$

This is a very large number of potential tickets!

Ordered, without replacement. This is the number of permutations $P(N,K) = N!/(N-K)!$ In the powerball example this is

$$\frac{69!}{(69-5)!} = \frac{69!}{64!} = 69 \times 68 \times 67 \times 66 \times 65 = 1,348,621,560.$$

This is nearly as large as the case with replacement.

Unordered, without replacement. This is the number of combinations $N!/K!(N-K)!$ In the powerball example this is

$$\frac{69!}{5!\,(69-5)!} = 11,238,513.$$

This is a large number but considerably smaller than the cases of ordered sampling.

Unordered, with replacement. This is a tricky computation. It is not N^K (ordered with replacement) divided by $K!$ because the number of orderings per group depends on whether or not there are repeats.

[8] There are also other prizes for other combinations.

The trick is to recast the question as a different problem. It turns out that the number we are looking for is the same as the number of N-tuples of non-negative integers $\{x_1, ..., x_N\}$ whose sum is K. To see this, a lottery ticket (unordered with replacement) can be represented by the number of "1's" x_1, the number of "2's" x_2, the number of "3's" x_3, etc, and we know that the sum of these numbers $(x_1 + \cdots + x_N)$ must equal K. The solution has a clever name based on the original proof notation.

Theorem 1.12 Stars and Bars Theorem. The number of N-tuples of non-negative integers whose sum is K is equal to $\binom{N+K-1}{K}$.

The proof of the stars and bars theorem is omitted as it is rather tedious. It does give us the answer to the question we started to address, namely the number of unordered sets taken with replacement. In the powerball example this is

$$\binom{69+5-1}{5} = \frac{73!}{5!68!} = 15,020,334.$$

Table 1.1 summarizes the four sampling results.

Table 1.1: Number of possible arrangments of size K from N items

	Without Replacement	With Replacement
Ordered	$\frac{N!}{(N-K)!}$	N^K
Unordered	$\binom{N}{K}$	$\binom{N+K-1}{K}$

The actual Powerball game uses unordered without replacement sampling. Thus there are about 11 million potential tickets. As each ticket has equal chance of occurring (if the random process is fair) this means the probability of winning is about $1/11,000,000$. Since a player wins \$1 million once for every 11 million tickets the expected payout (ignoring the other payouts) is about \$0.09. This is a low payout (considerably below a "fair" bet, given that a ticket costs \$2) but sufficiently high to achieve meaningful interest from players.

1.13 Poker Hands

A fun application of probability theory is to the game of poker. Similar types of calculations can be useful in economic examples involving multiple choices.

A standard game of poker is played with a 52-card deck containing 13 denominations {2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace} in each of four suits {club, diamond, heart, spade}. The deck is shuffled (so the order is random) and a player is dealt[9] five cards called a "hand". Hands are ranked based on whether there are multiple cards (pair, two pair, 3-of-a-kind, full house, or four-of-a-kind), all five cards

[9] A typical game involves additional complications which we ignore.

in sequence (called a "straight"), or all five cards of the same suit (called a "flush"). Players win if they have the best hand.

We are interested in calculating the probability of receiving a winning hand.

The structure is unordered sampling without replacement. The number of possible poker hands is

$$\binom{52}{5} = \frac{52!}{47!5!} = \frac{48\times 49\times 50\times 51\times 52}{2\times 3\times 4\times 5} = 48\times 49\times 5\times 17\times 13 = 2,598,560.$$

Since the draws are symmetric and random all hands have the same probability of receipt, implying that the probability of receiving any specific hand is $1/2,598,560$, an infinitesimally small number.

Another way of calculating this probability is as follows. Imagine picking a specific 5-card hand. The probability of receiving one of the five cards on the first draw is $5/52$, the probability of receiving one of the remaining four on the second draw is $4/51$, the probability of receiving one of the remaining three on the third draw is $3/50$, etc., so the probability of receiving the five card hand is

$$\frac{5\times 4\times 3\times 2\times 1}{52\times 51\times 50\times 49\times 48} = \frac{1}{13\times 17\times 5\times 49\times 48} = \frac{1}{2,598,960}.$$

This is the same.

One way to calculate the probability of a winning hand is to enumerate and count the number of winning hands in each category and then divide by the total number of hands $2,598,560$. We give a few examples.

Four of a Kind: Consider the number of hands with four of a specific denomination (such as Kings). The hand contains all four Kings plus an additional card which can be any of the remaining 48. Thus there are exactly 48 five-card hands with all four Kings. There are thirteen denominations so there are $13\times 48 = 624$ hands with four-of-a-kind. This means that the probability of drawing a four-of-a-kind is

$$\frac{13\times 48}{13\times 17\times 5\times 49\times 48} = \frac{1}{17\times 5\times 49} = \frac{1}{4165} \simeq 0.0\%.$$

Three of a Kind: Consider the number of hands with three of a specific denomination (such as Aces). There are $\binom{4}{3} = 4$ groups of three Aces. There are 48 cards to choose the remaining two. The number of such arrangements is $\binom{48}{2} = \frac{48!}{46!2!} = 47\times 24$. However, this includes pairs. There are twelve denominations each of which has $\binom{4}{2} = 6$ pairs, so there are $12\times 6 = 72$ pairs. This means the number of two-card arrangements excluding pairs is $47\times 24 - 72 = 44\times 24$. Hence the number of hands with three Aces and no pair is $4\times 44\times 24$. As there are 13 possible denominations the number of hands with a three of a kind is $13\times 4\times 44\times 24$. The probability of drawing a three-of-a-kind is

$$\frac{13\times 4\times 44\times 24}{13\times 17\times 5\times 49\times 48} = \frac{88}{17\times 5\times 49} \simeq 2.1\%.$$

One pair: Consider the number of hands with two of a specific denomination (such as a "7"). There are $\binom{4}{2} = 6$ pairs of 7's. From the 48 remaining cards the number of three-card arrangements are $\binom{48}{3} =$

$\frac{48!}{45!3!} = 23 \times 47 \times 16$. However, this includes three-card groups and two-card pairs. There are twelve denominations. Each has $\binom{4}{3} = 4$ three-card groups. Each also has $\binom{4}{2} = 6$ pairs and 44 remaining cards from which to select the third card. Thus there are $12 \times (4 + 6 \times 44)$ three-card arrangements with a either a three-card group or a pair. Subtracting, we find that the number of hands with two 7's and no other pairs is

$$6 \times (23 \times 47 \times 16 - 12 \times (4 + 6 \times 44)).$$

Multiplying by 13, the probability of drawing one pair of any denomination is

$$13 \times \frac{6 \times (23 \times 47 \times 16 - 12 \times (4 + 6 \times 44))}{13 \times 17 \times 5 \times 49 \times 48} = \frac{23 \times 47 \times 2 - 3 \times (2 + 3 \times 44)}{17 \times 5 \times 49} \simeq 42\%.$$

From these simple calculations you can see that if you receive a random hand of five cards you have a good chance of receiving one pair, a small chance of receiving a three-of-a-kind, and a negligible chance of receiving a four-of-a-kind.

1.14 Sigma Fields*

Definition 1.2 is incomplete as stated. When there are an uncountable infinity of events it is necessary to restrict the set of allowable events to exclude pathological cases. This is a technicality which has little impact on practical econometrics. However the terminology is used frequently so it is prudent to be aware of the following definitions. The correct definition of probability is as follows.

Definition 1.5 A **probability function** $\mathbb{P}$ is a function from a sigma field $\mathscr{B}$ to the real line which satisfies the Axioms of Probability.

The difference is that Definition 1.5 restricts the domain to a sigma field $\mathscr{B}$. The latter is a collection of sets which is closed under set operations. The restriction means that there are some events for which probability is not defined.

We now define a sigma field.

Definition 1.6 A collection $\mathscr{B}$ of sets is called a **sigma field** if it satisfies the following three properties:

1. $\varnothing \in \mathscr{B}$.
2. If $A \in \mathscr{B}$ then $A^c \in \mathscr{B}$.
3. If $A_1, A_2, ... \in \mathscr{B}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{B}$.

The infinite union in part 3 includes all elements which are an element of A_i for some i. An example is $\bigcup_{i=1}^{\infty} [0, 1 - 1/i] = [0, 1)$.

An alternative label for a sigma field is "sigma algebra". The following is a leading example of a sigma field.

Definition 1.7 The **Borel sigma field** is the smallest sigma field on $\mathbb{R}$ containing all open intervals (a, b). It contains all open intervals and closed intervals, and their countable unions, intersections, and complements.

A sigma-field can be **generated** from a finite collection of events by taking all unions, intersections, and complements. Take the coin flip example and start with the event $\{H\}$. Its complement is $\{T\}$, their union is $S = \{H, T\}$, and its complement is $\{\varnothing\}$. No further events can be generated. Thus the collection $\{\{\varnothing\}, \{H\}, \{T\}, S\}$ is a sigma field.

For an example on the positive real line take the sets $[0,1]$ and $(1,2]$. Their intersection is $\{\varnothing\}$, union $[0,2]$, and complements $(1,\infty)$, $[0,1] \cup (2,\infty)$, and $(2,\infty)$. A further union is $[0,\infty)$. This collection is a sigma field as no further events can be generated.

When there are an infinite number of events then it may not be possible to generate a sigma field through set operations as there are pathological counter-examples. These counter-examples are difficult to characterize, are non-intuitive, and seem to have no practical implications for econometric practice. Therefore the issue is generally ignored in econometrics.

If the concept of a sigma field seems technical, it is! The concept is not used further in this textbook.

1.15 Technical Proofs*

Proof of Theorem 1.1 Take an outcome ω in A. Since $\{B_1, B_2, \cdots\}$ is a partition of S it follows that $\omega \in B_i$ for some i. Thus $\omega \in A_i \subset \bigcup_{i=1}^{\infty} A_i$. This shows that every element in A is an element of $\bigcup_{i=1}^{\infty} A_i$.

Now take an outcome ω in $\bigcup_{i=1}^{\infty} A_i$. Thus $\omega \in A_i = (A \cap B_i)$ for some i. This implies $\omega \in A$. This shows that every element in $\bigcup_{i=1}^{\infty} A_i$ is an element of A.

Set $A_i = (A \cap B_i)$. For $i \neq j$, $A_i \cap A_j = (A \cap B_i) \cap \left(A \cap B_j\right) = A \cap \left(B_i \cap B_j\right) = \varnothing$ since B_i are mutually disjoint. Thus A_i are mutually disjoint. ■

Proof of Theorem 1.2.1 A and A^c are disjoint and $A \cup A^c = S$. The second and third Axioms imply

$$1 = \mathbb{P}[S] = \mathbb{P}[A] + \mathbb{P}\left[A^c\right]. \tag{1.1}$$

Rearranging we find $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$ as claimed. ■

Proof of Theorem 1.2.2 $\varnothing = S^c$. By Theorem 1.2.1 and the second Axiom, $\mathbb{P}[\varnothing] = 1 - \mathbb{P}[S] = 0$ as claimed. ■

Proof of Theorem 1.2.3 Axiom 1 implies $\mathbb{P}[A^c] \geq 0$. This and equation (1.1) imply

$$\mathbb{P}[A] = 1 - \mathbb{P}\left[A^c\right] \leq 1$$

as claimed. ■

Proof of Theorem 1.2.4 The assumption $A \subset B$ implies $A \cap B = A$. By the partitioning theorem (Theorem 1.1) $B = (B \cap A) \cup (B \cap A^c) = A \cup (B \cap A^c)$ where A and $B \cap A^c$ are disjoint. The third Axiom implies

$$\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}\left[B \cap A^c\right] \geq \mathbb{P}[A]$$

where the inequality is $\mathbb{P}[B \cap A^c] \geq 0$ which holds by the first Axiom. Thus, $\mathbb{P}[B] \geq \mathbb{P}[A]$ as claimed. ■

Proof of Theorem 1.2.5 $\{A \cup B\} = A \cup \{B \cap A^c\}$ where A and $\{B \cap A^c\}$ are disjoint. Also $B = \{B \cap A\} \cup \{B \cap A^c\}$ where $\{B \cap A\}$ and $\{B \cap A^c\}$ are disjoint. These two relationships and the third Axiom imply

$$\begin{aligned} \mathbb{P}[A \cup B] &= \mathbb{P}[A] + \mathbb{P}\left[B \cap A^c\right] \\ \mathbb{P}[B] &= \mathbb{P}[B \cap A] + \mathbb{P}\left[B \cap A^c\right]. \end{aligned}$$

Subtracting,

$$\mathbb{P}[A \cup B] - \mathbb{P}[B] = \mathbb{P}[A] - \mathbb{P}[B \cap A].$$

By rearrangement we obtain the result. ■

Proof of Theorem 1.2.6 From the Inclusion-Exclusion Principle and $\mathbb{P}[A \cap B] \geq 0$ (the first Axiom)

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \leq \mathbb{P}[A] + \mathbb{P}[B]$$

as claimed. ■

Proof of Theorem 1.2.7 Rearranging the Inclusion-Exclusion Principle and using $\mathbb{P}[A \cup B] \leq 1$ (Theorem 1.2.3)

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cup B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1$$

which is the stated result. ■

Proof of Theorem 1.11 Multiplying out, the expression

$$(a+b)^N = (a+b) \times \cdots \times (a+b) \tag{1.2}$$

is a polynomial in a and b with 2^N terms. Each term takes the form of the product of K of the a and $N-K$ of the b, thus of the form $a^K b^{N-K}$. The number of terms of this form is equal to the number of combinations of the a's, which is $\binom{N}{K}$. Consequently expression (1.2) equals $\sum_{K=0}^{N} \binom{N}{K} a^K b^{N-K}$ as stated. ■

1.16 Exercises

Exercise 1.1 Let $A = \{a, b, c, d\}$ and $B = \{a, c, e, f\}$.

(a) Find $A \cap B$.

(b) Find $A \cup B$.

Exercise 1.2 Describe the sample space S for the following experiments.

(a) Flip a coin.

(b) Roll a six-sided die.

(c) Roll two six-sided dice.

(d) Shoot six free throws (in basketball).

Exercise 1.3 From a 52-card deck of playing cards draw five cards to make a hand.

(a) Let A be the event "The hand has two Kings". Describe A^c.

(b) A **straight** is five cards in sequence, e.g. {5,6,7,8,9}. A **flush** is five cards of the same suit. Let A be the event "The hand is a straight" and B be the event "The hand is 3-of-a-kind". Are A and B disjoint or not disjoint?

(c) Let A be the event "The hand is a straight" and B be the event "The hand is flush". Are A and B disjoint or not disjoint?

Exercise 1.4 For events A and B, express "either A or B but not both" as a formula in terms of $\mathbb{P}[A]$, $\mathbb{P}[B]$, and $\mathbb{P}[A \cap B]$.

Exercise 1.5 If $\mathbb{P}[A] = 1/2$ and $\mathbb{P}[B] = 2/3$, can A and B be disjoint? Explain.

Exercise 1.6 Prove that $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

Exercise 1.7 Show that $\mathbb{P}[A \cap B] \leq \mathbb{P}[A] \leq \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$.

Exercise 1.8 Suppose $A \cap B = A$. Can A and B be independent? If so, give the appropriate condition.

Exercise 1.9 Prove that

$$\mathbb{P}[A \cap B \cap C] = \mathbb{P}[A \mid B \cap C]\,\mathbb{P}[B \mid C]\,\mathbb{P}[C].$$

Assume $\mathbb{P}[C] > 0$ and $\mathbb{P}[B \cap C] > 0$.

Exercise 1.10 Is $\mathbb{P}[A \mid B] \leq \mathbb{P}[A]$, $\mathbb{P}[A \mid B] \geq \mathbb{P}[A]$ or is neither necessarily true?

Exercise 1.11 Give an example where $\mathbb{P}[A] > 0$ yet $\mathbb{P}[A \mid B] = 0$.

Exercise 1.12 Calculate the following probabilities concerning a standard 52-card playing deck.

(a) Drawing a King with one card.

(b) Drawing a King on the second card, conditional on a King on the first card.

(c) Drawing two Kings with two cards.

(d) Drawing a King on the second card, conditional on the first card is not a King.

(e) Drawing a King on the second card, when the first card is placed face down (so is unknown).

Exercise 1.13 You are on a game show, and the host shows you five doors marked A, B, C, D, and E. The host says that a prize is behind one of the doors, and you win the prize if you select the correct door. Given the stated information, what probability distribution would you use for modeling the distribution of the correct door?

Exercise 1.14 Calculate the following probabilities, assuming fair coins and dice.

(a) Getting three heads in a row from three coin flips.

(b) Getting a heads given that the previous coin was a tails.

(c) From two coin flips getting two heads given that at least one coin is a heads.

(d) Rolling a six from a pair of dice.

(e) Rolling "snakes eyes" from a pair of dice. (Getting a pair of ones.)

Exercise 1.15 If four random cards are dealt from a deck of playing cards, what is the probability that all four are Aces?

Exercise 1.16 Suppose that the unconditional probability of a disease is 0.0025. A screening test for this disease has a detection rate of 0.9, and has a false positive rate of 0.01. Given that the screening test returns positive, what is the conditional probability of having the disease?

Exercise 1.17 Suppose 1% of athletes use banned steroids. Suppose a drug test has a detection rate of 40% and a false positive rate of 1%. If an athlete tests positive what is the conditional probability that the athlete has taken banned steroids?

Exercise 1.18 Sometimes we use the concept of **conditional independence**. The definition is as follows: let A, B, C be three events with positive probabilities. Then A and B are conditionally independent given C if $\mathbb{P}[A \cap B \mid C] = \mathbb{P}[A \mid C]\,\mathbb{P}[B \mid C]$. Consider the experiment of tossing two dice. Let $A = \{\text{First die is 6}\}$, $B = \{\text{Second die is 6}\}$, and $C = \{\text{Both dice are the same}\}$. Show that A and B are independent (unconditionally), but A and B are dependent given C.

Exercise 1.19 Monte Hall. This is a famous (and surprisingly difficult) problem based on an old U.S. television game show "Let's Make a Deal hosted by Monte Hall". A standard part of the show ran as follows: A contestant was asked to select from one of three identical doors: A, B, and C. Behind one of the three doors there was a prize. If the contestant selected the correct door they would receive the prize. The contestant picked one door (say A) but it is not immediately opened. To increase the drama the host opened one of the two remaining doors (say door B) revealing that that door does not have the prize. The host then made the offer: "You have the option to switch your choice" (e.g. to switch to door C). You can imagine that the contestant may have made one of reasonings (a)-(c) below. Comment on each of these three reasonings. Are they correct?

(a) "When I selected door A the probability that it has the prize was 1/3. No information was revealed. So the probability that Door A has the prize remains 1/3."

(b) "The original probability was 1/3 on each door. Now that door B is eliminated, doors A and C each have each probability of 1/2. It does not matter if I stay with A or switch to C."

(c) "The host inadvertently revealed information. If door C had the prize, he was forced to open door B. If door B had the prize he would have been forced to open door C. Thus it is quite likely that door C has the prize."

(d) Assume a prior probability for each door of 1/3. Calculate the posterior probability that door A and door C have the prize. What choice do you recommend for the contestant?

Exercise 1.20 In the game of blackjack you are dealt two cards from a standard playing deck. Your score is the sum of the value of the two cards, where numbered cards have the value given by their number, face cards (Jack, Queen, King) each receive 10 points, and an Ace either 1 or 11 (player can choose). A **blackjack** is receiving a score of 21 from two cards, thus an Ace and any card worth 10 points.

(a) What is the probability of receiving a blackjack?

(b) The dealer is dealt one of his cards face down and one face up. Suppose the "show" card is an Ace. What is the probability the dealer has a blackjack? (For simplicity assume you have not seen any other cards.)

Exercise 1.21 Consider drawing five cards at random from a standard deck of playing cards. Calculate the following probabilities.

(a) A straight (five cards in sequence, suit not relevant).

(b) A flush (five cards of the same suit, order not relevant).

(c) A full house (3-of-a-kind and a pair, e.g. three Kings and two “3’s”.

Exercise 1.22 In the poker game “Five Card Draw" a player first receives five cards drawn at random. The player decides to discard some of their cards and then receives replacement cards. Assume a player is dealt a hand with one pair and three unrelated cards and decides to discard the three unrelated cards to obtain replacements. Calculate the following conditional probabilities for the resulting hand after the replacements are made.

(a) Obtaining a four-of-a-kind.

(b) Obtaining a three-of-a-kind.

(c) Obtaining two pairs.

(d) Obtaining a straight or a flush.

(e) Ending with one pair.

Chapter 2

Random Variables

2.1 Introduction

In practice it is convenient to represent random outcomes numerically. If the outcome is numerical and one-dimensional we call the outcome a random variable. If the outcome is multi-dimensional we call it a random vector.

Random variables are one of the most important and core concepts in probability theory. It is so central that most of the time we don't think about the foundations.

As an example consider a coin flip which has the possible outcome H or T. We can write the outcome numerically by setting the result as $X = 1$ if the coin is heads and $X = 0$ if the coin is tails. The object X is random since its value depends on the outcome of the coin flip.

2.2 Random Variables

Definition 2.1 A **random variable** is a real-valued outcome; a function from the sample space S to the real line $\mathbb{R}$.

A random variable is typically represented by an uppercase Latin character; common choices are X and Y. In the coin flip example the function is

$$X = \begin{cases} 1 & \text{if H} \\ 0 & \text{if T.} \end{cases}$$

To illustrate, Figure 2.1 illustrates a mapping from the coin flip sample space to the real line, with T mapped to 0 and H mapped to 1. A coin flip may seem overly simple but the structure is identical to any two-outcome application.

Notationally it is useful to distinguish between random variables and their realizations. In probability theory and statistics the convention is to use upper case X to indicate a random variable and use lower case x to indicate a realization or specific value. This may seem a bit abstract. Think of X as the random object whose value is unknown and x as a specific number or outcome.

2.3 Discrete Random Variables

We have a defined a random variable X as a real-valued outcome. In most cases X only takes values on a subset of the real line. Take, for example, a coin flip coded as 1 for heads and 0 for tails. This only takes the values 0 and 1. It is an example of what we call a discrete distribution.

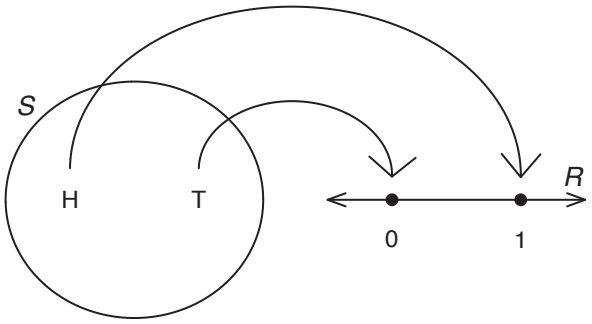


Figure 2.1: A Random Variable is a Function

Definition 2.2 The set $\mathscr{X}$ is **discrete** if it has a finite or countably infinite number of elements.

Most discrete sets in applications are non-negative integers. For example, in a coin flip $\mathscr{X} = \{0,1\}$ and in a roll of a die $\mathscr{X} = \{1,2,3,4,5,6\}$.

Definition 2.3 If there is a discrete set $\mathscr{X}$ such that $\mathbb{P}[X \in \mathscr{X}] = 1$ then X is a **discrete random variable**. The smallest set $\mathscr{X}$ with this property is the **support** of X.

The support is the set of values which receive positive probability of occurance. We sometimes write the support as $\mathscr{X} = \{\tau_1, \tau_2, ..., \tau_n\}$, $\mathscr{X} = \{\tau_1, \tau_2, ...\}$ or $\mathscr{X} = \{\tau_0, \tau_1, \tau_2, ...\}$ when we need an explicit description of the support. We call the values τ_j the **support points**.

The following definition is useful.

Definition 2.4 The **probability mass function** of a random variable is $\pi(x) = \mathbb{P}[X = x]$, the probability that X equals the value x. When evaluated at the support points τ_j we write $\pi_j = \pi(\tau_j)$.

Take, for example, a coin flip with probability p of heads. The support is $\mathscr{X} = \{0,1\} = \{\tau_0, \tau_1\}$. The probability mass function takes the values $\pi_0 = 1 - p$ and $\pi_1 = p$.

Take a fair die. The support is $\mathscr{X} = \{1,2,3,4,5,6\} = \{\tau_j : j = 1, ..., 6\}$ with probability mass function $\pi_j = 1/6$ for $j = 1, ..., 6$.

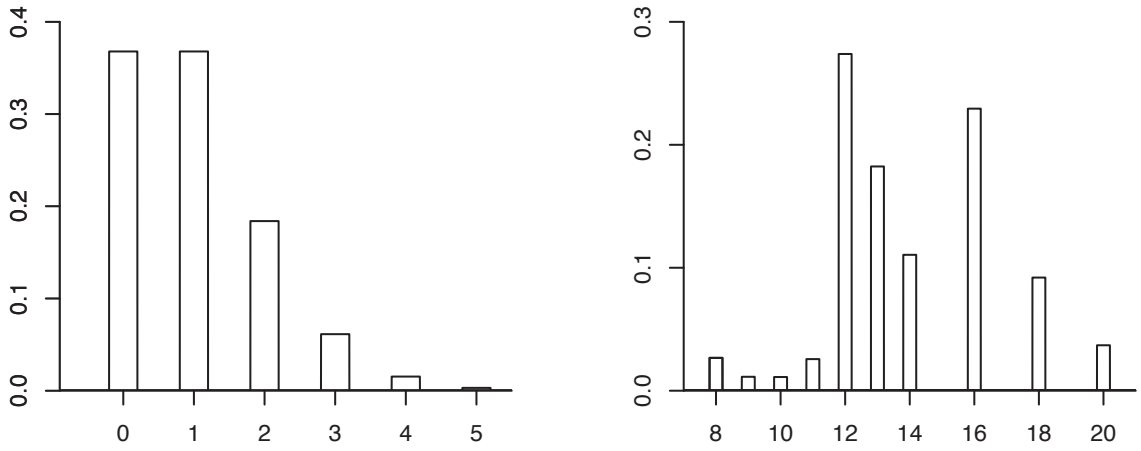
An example of a countably infinite discrete random variable is

$$\mathbb{P}[X = k] = \frac{e^{-1}}{k!}, \qquad k = 0, 1, 2, ... \tag{2.1}$$

This is a valid probability function since $e = \sum_{k=0}^{\infty} 1/k!$. The support is $\mathscr{X} = \{0,1,2,...\}$ with probability mass function $\pi_j = e^{-1}/(j!)$ for $j \geq 0$. (This a special case of the Poisson distribution which will be defined in Section 3.6.)

It can be useful to plot a probability mass function as a bar graph to visualize the relative frequency of occurrence. Figure 2.2(a) displays the probability mass function for the distribution (2.1). The height of each bar is the the probability π_j at the support point. While the distribution is countably infinite the probabilities are negligible for $k \geq 6$ so we have plotted the probability mass function for $k \leq 5$. You can see that the probabilities for $k = 0$ and $k = 1$ are about 0.37, that for $k = 2$ about 0.18, for $k = 3$ is 0.06 and for $k = 4$ the probability is 0.015.

To illustrate using a real-world example, Figure 2.2(b) displays the probability mass function for the years of education[1] among U.S. wage earners in 2009. You can see that the highest probability occurs at 12 years of education (about 27%) and second highest at 16 years (about 23%)



(a) Poisson (b) Education

Figure 2.2: Probability Mass Functions

2.4 Transformations

If X is a random variable and $Y = g(X)$ for some function $g : \mathscr{X} \to \mathscr{Y} \subset \mathbb{R}$ then Y is also a random variable. To see this formally, since X is a mapping from the sample space S to $\mathscr{X}$, and g maps $\mathscr{X}$ to $\mathscr{Y} \subset \mathbb{R}$, then Y is also a mapping from S to $\mathbb{R}$.

We are interested in describing the probability mass function for Y. Write the support of X as $\mathscr{X} = \{\tau_1, \tau_2, ...\}$ and its probability mass function as $\pi_X(\tau_j)$.

If we apply the transformation to each of X's support points we obtain $\mu_j = g(\tau_j)$. If the μ_j are unique (there are no redundancies) then Y has support $\mathscr{Y} = \{\mu_1, \mu_2, ...\}$ with probability mass function

[1]Here, *education* is defined as years of schooling beyond kindergarten. A high school graduate has *education*=12, a college graduate has *education*=16, a Master's degree has *education*=18, and a professional degree (medical, law or PhD) has *education*=20.

$\pi_Y(\mu_j) = \pi_X(\tau_j)$. The impact of the transformation $X \to Y$ is to move the support points from τ_j to μ_j, and the probabilities are maintained.

If the μ_j are not unique then some probabilities are combined. Essentially the transformation reduces the number of support points. As an example suppose that the support of X is $\{-1, 0, 1\}$ and $Y = X^2$. Then the support for Y is $\{0, 1\}$. Since both -1 and 1 are mapped to 1, the probability mass function for Y inherits the sum of these two probabilities. That is, the probability mass function for Y is

$$\begin{aligned}\pi_Y(0) &= \pi_X(0) \\ \pi_Y(1) &= \pi_X(-1) + \pi_X(1).\end{aligned}$$

In general

$$\pi_Y(\mu_i) = \sum_{j:g(\tau_j)=g(\tau_i)} \pi_X\left(\tau_j\right)$$

The sum looks complicated, but it simply states that the probabilites are summed over all indices for which there is equality among the transformed values.

2.5 Expectation

The **expectation** of a random variable X, denoted as $\mathbb{E}[X]$, is a useful measure of the central tendency of the distribution. It is the average value with probability-weighted averaging. The expectation is also called the **expected value**, **average**, or **mean** of the distribution. We prefer the labels "expectation" or "expected value" as they are the least ambiguous. It is typical to write the expectation of X as $\mathbb{E}[X]$, $\mathbb{E}(X)$, or $\mathbb{E}X$. Some authors use the notation $E[X]$ or $\mathrm{E}[X]$.

Definition 2.5 For a discrete random variable X with support $\{\tau_j\}$, the **expectation** of X is

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} \tau_j \pi_j$$

if the series is convergent. (For the definition of convergence see Appendix A.1.)

It is important to understand that while X is random, the expectation $\mathbb{E}[X]$ is non-random. It is a fixed feature of the distribution.

Example: $X = 1$ with probability p and $X = 0$ with probability $1 - p$. Its expected value is

$$\mathbb{E}[X] = 0 \times (1 - p) + 1 \times p = p.$$

Example: Fair die throw

$$\mathbb{E}[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{7}{2}.$$

Example: $\mathbb{P}[X = k] = \dfrac{e^{-1}}{k!}$ for non-negative integer k.

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\frac{e^{-1}}{k!} = 0 + \sum_{k=1}^{\infty} k\frac{e^{-1}}{k!} = \sum_{k=1}^{\infty} \frac{e^{-1}}{(k-1)!} = \sum_{k=0}^{\infty} \frac{e^{-1}}{k!} = 1.$$

Example: Years of Education. For the probability distribution displayed in Figure 2.2(c) the expected value is

$$\mathbb{E}[X] = 8 \times 0.027 + 9 \times 0.011 + 10 \times 0.011 + 11 \times 0.026 + 12 \times 0.274$$
$$+13 \times 0.182 + 14 \times 0.111 + 16 \times 0.229 + 18 \times 0.092 + 20 \times 0.037 = 13.9.$$

Thus the average number of years of education is about 14.

One property of the expectation is that it is the **center of mass** of the distribution. Imagine the probability mass functions of Figure 2.2 as a set of weights on a board placed on top of a fulcrum. For the board to balance the fulcrum needs to be placed at the expectation $\mathbb{E}[X]$. It is instructive to review Figure 2.2 with the knowledge the center of mass for the Poisson distribution is 1, and that for the years of education is 14.

We similarly define the expectation of transformations.

Definition 2.6 For a discrete random variable X with support $\{\tau_j\}$, the **expectation** of $g(X)$ is

$$\mathbb{E}\big[g(X)\big] = \sum_{j=1}^{\infty} g(\tau_j)\pi_j$$

if the series is convergent.

When applied to transformations we may use simplified notation when it leads to less clutter. For example, we may write $\mathbb{E}|X|$ rather than $\mathbb{E}[|X|]$, and $\mathbb{E}|X|^r$ rather than $\mathbb{E}[|X|^r]$.

Expectation is an linear operator.

Theorem 2.1 Linearity of Expectation. For any constants a and b,

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X].$$

Proof: Using the definition of expectation

$$\begin{aligned}\mathbb{E}[a + bX] &= \sum_{j=1}^{\infty} \big(a + b\tau_j\big)\pi_j \\ &= a\sum_{j=1}^{\infty} \pi_j + b\sum_{j=1}^{\infty} \tau_j\pi_j \\ &= a + b\mathbb{E}[X]\end{aligned}$$

since $\sum_{j=1}^{\infty} \pi_j = 1$ and $\sum_{j=1}^{\infty} \pi_j\tau_j = \mathbb{E}[X]$. ■

2.6 Finiteness of Expectations

The definition of expectation includes the phrase "if the series is convergent". This is included since some series are non-convergent. In the latter case the expectation is either infinite or not defined.

As an example, suppose that X has support points 2^k for $k = 1, 2, ...$ and has probability mass function $\pi_k = 2^{-k}$. This is a valid probability function since

$$\sum_{k=1}^{\infty} \pi_k = \sum_{k=1}^{\infty} \frac{1}{2^k} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1.$$

The expected value is

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k \pi_k = \sum_{k=1}^{\infty} 2^k 2^{-k} = \infty.$$

This example is known as the **St. Petersburg Paradox** and corresponds to a bet. Toss a fair coin K times until a heads appears. Set the payout as $X = 2^K$. This means that a player is paid \$2 if $K = 1$, \$4 if $K = 2$, \$8 if $K = 3$, etc. The payout has infinite expectation, as shown above. This game has been called a paradox because few individuals are willing to pay a high price to receive this random payout, despite the fact that the expected value is infinite[2].

The probability mass function for this payout is displayed in Figure 2.3. You can see how the probability decays slowly in the right tail. Recalling the property of the expected value as equal to the center of mass, this means that if we imagine an infinitely-long board with the probability mass function of Figure 2.3 as weights, there would be no position where you could put a fulcrum such that the board balances. Regardless of where the fulcrum is placed, the board will tilt to the right. The small but increasingly distantly-placed probability mass points dominate.

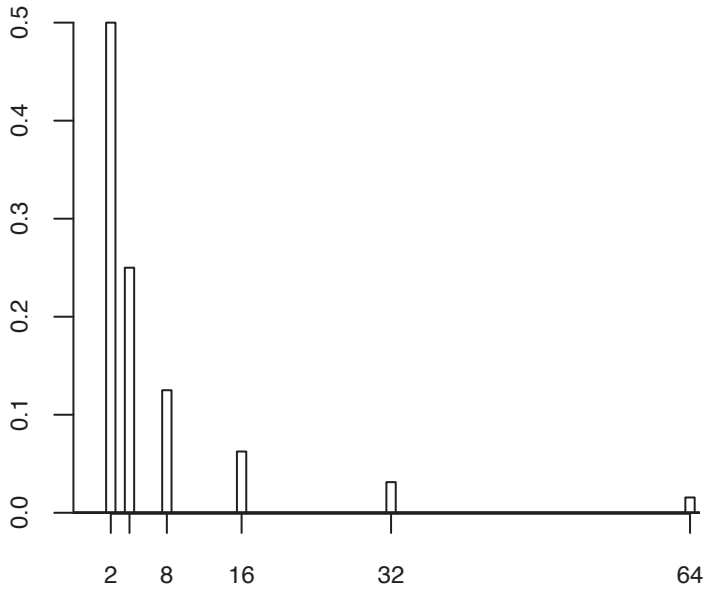


Figure 2.3: St. Petersburg Paradox

The above example is a non-convergent case where the expectation is infinite. In some non-convergent cases the expectation is undefined. Suppose that we modify the payout of the St. Petersburg bet so that

[2]Economists should realize that there is no paradox once you introduce concave utility. If utility is $u(x) = x^{1/2}$ then the expected utility of the bet is $\sum_{k=1}^{\infty} 2^{-k/2} = 2^{-1/2} / \left(1 - 2^{-1/2}\right) \simeq 2.41$. The value of the bet (certainty equivalence) is \$5.83, for this also yields a utility of 2.41.

the support points are 2^k and -2^k each with probability $\pi_k = 2^{-k-1}$. Then the expected value is

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k \pi_k - \sum_{k=1}^{\infty} 2^k \pi_k = \sum_{k=1}^{\infty} \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{2} = \infty - \infty.$$

This series is not convergent but also neither $+\infty$ nor $-\infty$. (It is tempting but incorrect to guess that the two infinite sums cancel.) In this case we say that the expectation is **not defined** or **does not exist**.

The lack of finiteness of the expected value can make economic transactions difficult. Suppose that X is loss due to an unexpected severe event, such as fire, tornado, or earthquake. With high probability $X = 0$. With with low probability X is positive and quite large. In these contexts risk adverse economic agents seek insurance. An ideal insurance contract compensates fully for a random loss X. In a market with no asymmetries or frictions, insurance companies offer insurance contracts with the premium set to equal the expected loss $\mathbb{E}[X]$. However, when the loss is not finite this is impossible so such contracts are infeasible.

2.7 Distribution Function

A random variable can be represented by its distribution function.

Definition 2.7 The **distribution function** is $F(x) = \mathbb{P}[X \le x]$, the probability of the event $\{X \le x\}$.

$F(x)$ is also known as the **cumulative distribution function (CDF)**. A common shorthand is to write "$X \sim F$" to mean "the random variable X has distribution function F" or "the random variable X is distributed as F". We use the symbol "$\sim$" to means that the variable on the left has the distribution indicated on the right.

It is standard notation to use upper case letters to denote a distribution function. The most common choice is F though any symbol can be used. When there is need to be clear about the random variable we add a subscript, writing the distribution function as $F_X(x)$. The subscript X indicates that F_X is the distribution of X. The argument "x" does not signify anything. We could equivalently write the distribution as $F_X(t)$ or $F_X(s)$. When there is only one random variable under discussion we simplify the notation and write the distribution function as $F(x)$.

For a discrete random variable with support points τ_j, the CDF at the support points equals the cumulative sum of the probabilities less than j.

$$F(\tau_j) = \sum_{k=1}^{j} \pi(\tau_k).$$

The CDF is constant between the support points. Therefore the CDF of a discrete random variable is a step function with jumps at each support point of magnitude $\pi(\tau_j)$.

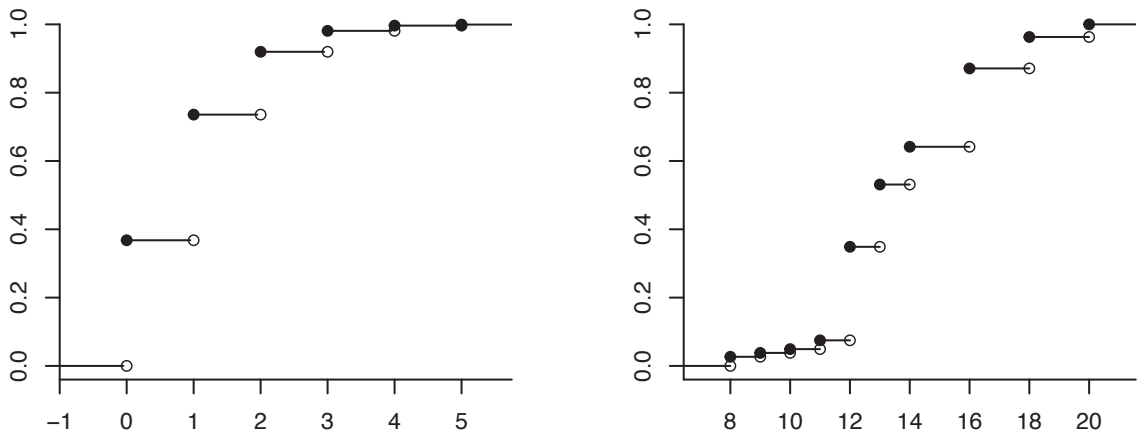
Figure 2.4 displays the distribution functions for the two examples displayed in Figure 2.2.

You can see how each distribution function is a step function with steps at the support points. The size of the jumps are varied since the probabilities of the support points are unequal.

In general (not just for discrete random variables) the CDF has the following properties.

Theorem 2.2 Properties of a CDF. If $F(x)$ is a distribution function then

1. $F(x)$ is non-decreasing.
2. $\lim_{x \to -\infty} F(x) = 0$.



(a) Poisson (b) Education

Figure 2.4: Discrete Distribution Functions

3. $\lim_{x\to\infty} F(x) = 1$.
4. $F(x)$ is right-continuous, meaning $\lim_{x\downarrow x_0} F(x) = F(x_0)$.

Properties 1 and 2 are consequences of Axiom 1 (probabilities are non-negative). Property 3 is Axiom 2. Property 4 states that at points where $F(x)$ has a step, $F(x)$ is discontinuous to the left but continuous to the right. This property is due to the definition of the distribution function as $\mathbb{P}[X \le x]$. If the definition were $\mathbb{P}[X < x]$ then $F(x)$ would be left-continuous.

2.8 Continuous Random Variables

If a random variable X takes a continuum of values it is not discretely distributed. Formally, we define a random variable to be continuous if the distribution function is continuous.

Definition 2.8 If $X \sim F(x)$ and $F(x)$ is continuous then X is a **continuous** random variable.

Example: Uniform Distribution

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1. \end{cases}$$

The function $F(x)$ is globally continuous, limits to zero as $x \to -\infty$ and limits to 1 as $x \to \infty$. Therefore it satisfies the properties of a CDF.

Example: Exponential Distribution.

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp(-x) & x \ge 0. \end{cases}$$

The function $F(x)$ is globally continuous, limits to zero as $x \to -\infty$ and limits to 1 as $x \to \infty$. Therefore it satisfies the properties of a CDF.

Example: Hourly Wages. As a real-world example, Figure 2.5 displays the distribution function for hourly wages in the U.S. in 2009, plotted over the range [\$0, \$60]. The function is continuous and everywhere increasing. This is because wage rates are dispersed. Marked with arrows are the values of the distribution function at \$10 increments from \$10 to \$50. This is read as follows. The distribution function at \$10 is 0.14. Thus 14% of wages are less than or equal to \$10. The distribution function at \$20 is 0.54. Thus 54% of the wages are less than or equal to \$20. Similarly, the distribution at \$30, \$40, and \$50 are 0.78, 0.89, and 0.94.

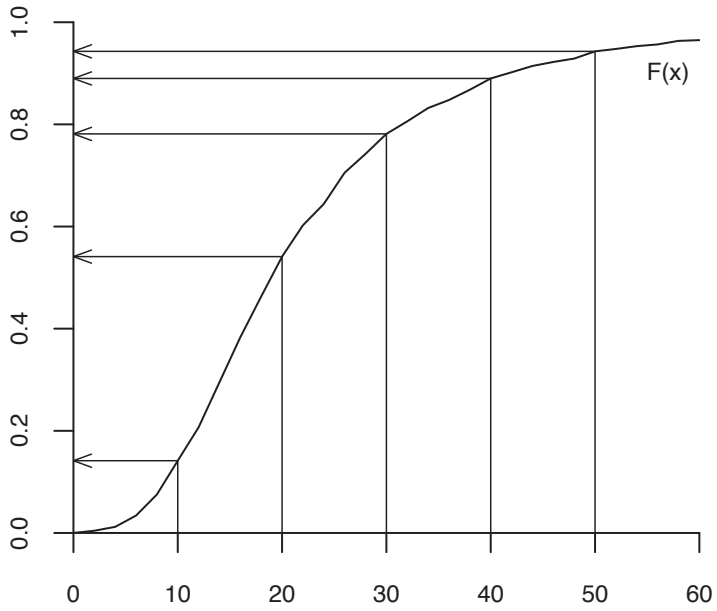


Figure 2.5: Distribution Function – U.S. Wages

One way to think about the distribution function is in terms of differences. Take an interval $(a, b]$. The probability that $X \in (a, b]$ is $\mathbb{P}[a < X \leq b] = F(b) - F(a)$, the difference in the distribution function. Thus the difference between two points of the distribution function is the probability that X lies in the interval. For example, the probability that a random person's wage is between \$10 and \$20 is $0.54 - 0.14 = 0.30$. Similarly, the probability that their wage is in the interval [\$40, \$50] is $94\% - 89\% = 5\%$.

One property of continuous random variables is that the probability that they equal any specific value is 0. To see this take any number x. We can find the probability that X equals x by taking the limit of the sequence of probabilities that X is in the interval $[x, x + \epsilon]$ as ϵ decreases to zero. This is

$$\mathbb{P}[X = x] = \lim_{\epsilon \to 0} \mathbb{P}[x \leq X \leq x + \epsilon] = \lim_{\epsilon \to 0} F(x + \epsilon) - F(x) = 0$$

when $F(x)$ is continuous. This is a bit of a paradox. The probability that X equals any specific value x is zero, but the probability that X equals some value is one. The paradox is due to the magic of the real line and the richness of uncountable infinity.

An implication is that for continuous random variables we have the equalities

$$\mathbb{P}[X < x] = \mathbb{P}[X \leq x] = F(x)$$
$$\mathbb{P}[X \geq x] = \mathbb{P}[X > x] = 1 - F(x).$$

2.9 Quantiles

For a continuous distribution $F(x)$ the **quantiles** $q(\alpha)$ are defined as the solutions to the function

$$\alpha = F(q(\alpha)).$$

Effectively they are the inverse of $F(x)$, thus

$$q(\alpha) = F^{-1}(\alpha).$$

The quantile function $q(\alpha)$ is a function from $[0,1]$ to the range of X.

Expressed as percentages, $100 \times q(\alpha)$ are called the **percentiles** of the distribution. For example, the 95^{th} percentile equals the 0.95 quantile.

Some quantiles have special names. The **median** of the distribution is the 0.5 quantile. The **quartiles** are the 0.25, 0.50, and 0.75 quantiles. The later are called the quartiles as they divide the population into four equal groups. The **quintiles** are the 0.2, 0.4, 0.6, and 0.8 quantiles. The **deciles** are the 0.1, 0.2, ..., 0.9 quantiles.

Quantiles are useful summaries of the spread of the distribution.

Example: Exponential Distribution. $F(x) = 1 - \exp(-x)$ for $x \geq 0$. To find a quantile $q(\alpha)$ set $\alpha = 1 - \exp(-x)$ and solve for x, thus $x = -\log(1-\alpha)$. For example the 0.9 quantile is $-\log(1-0.9) \simeq 2.3$, the 0.5 quantile is $-\log(1-0.5) \simeq 0.7$.

Example: Hourly Wages. Figure 2.6 displays the distribution function for hourly wages. From the points 0.25, 0.50, and 0.75 on the y-axis lines are drawn to the distribution function and then to the x-axis with arrows. These are the quartiles of the wage distribution, and are \$12.82, \$18.88, and \$28.35. The interpretation is that 25% of wages are less than or equal to \$12.82, 50% of wages are less than or equal to \$18.88, and 75% are less than or equal to \$28.35.

2.10 Density Functions

Continuous random variables do not have a probability mass function. An analog is the derivative of the distribution function which is called the density.

Definition 2.9 When $F(x)$ is differentiable, its **density** is $f(x) = \frac{d}{dx}F(x)$.

A density function is also called the **probability density function (PDF)**. It is standard notation to use lower case letters to denote a density function; the most common choice is f. As for distribution functions when we want to be clear about the random variable we write the density function as $f_X(x)$ where the subscript indicates that this is the density function of X. A common shorthand is to write "$X \sim f$" to mean "the random variable X has density function f".

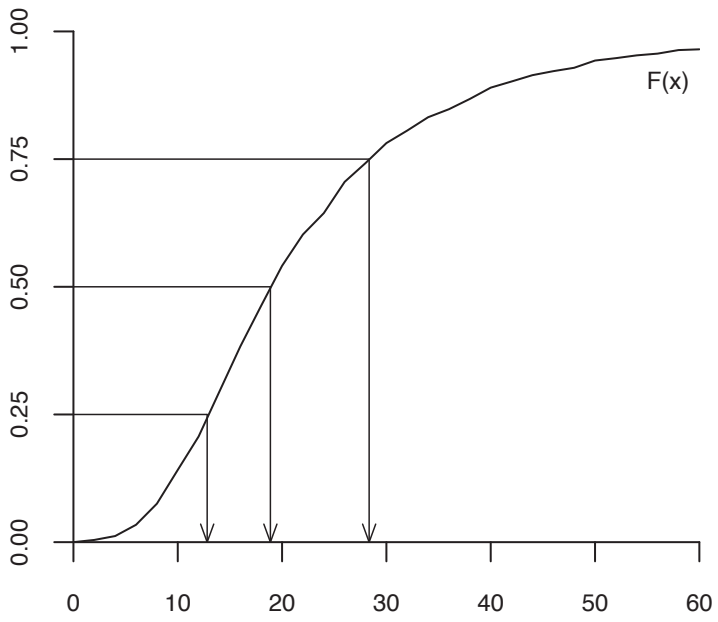


Figure 2.6: Quantile Function – U.S. Wages

Theorem 2.3 Properties of a PDF. A function $f(x)$ is a density function if and only if

1. $f(x) \geq 0$ for all x.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

A density is a non-negative function which is integrable and integrates to one. By the fundamental theorem of calculus we have the relationship

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)dx.$$

This states that the probability that X is in the interval $[a,b]$ is the integral of the density over $[a,b]$. This shows that the area underneath the density function are probabilities.

Example: Uniform Distribution. $F(x) = x$ for $0 \leq x \leq 1$. The density is

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}x = 1$$

on for $0 \leq x \leq 1$, 0 elsewhere. This density function is non-negative and satisfies

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 1dx = 1$$

and so satisfies the properties of a density function.

Example: Exponential Distribution. $F(x) = 1 - \exp(-x)$ for $x \geq 0$. The density is

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}\left(1 - \exp(-x)\right) = \exp(-x)$$

on $x \geq 0$, 0 elsewhere. This density function is non-negative and satisfies

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \exp(-x)dx = 1$$

and so satisfies the properties of a density function. We can use it for probability calculations. For example

$$\mathbb{P}[1 \leq X \leq 2] = \int_{1}^{2} \exp(-x)dx = \exp(-1) - \exp(-2) \simeq 0.23.$$

Example: Hourly Wages. Figure 2.7 plots the density of U.S. hourly wages.

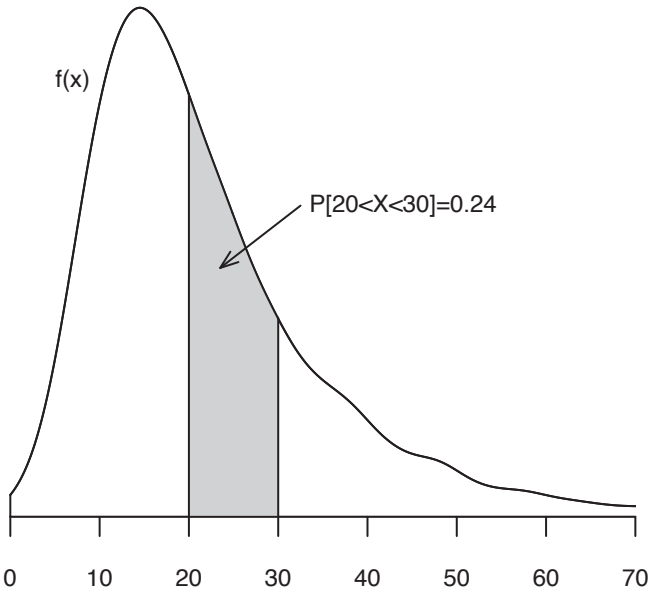


Figure 2.7: Density Function for Wage Distribution

The way to interpret the density function is as follows. The regions where the density $f(x)$ is relatively high are the regions where X has relatively high likelihood of occurrence. The regions where the density is relatively small are the regions where X has relatively low likelihood. The density declines to zero in

the tails – this is a necessary consequence of the property that the density function is integrable. Areas underneath the density are probabilities. For example, in Figure 2.7 the shaded region is for $20 < X < 30$. This region has area of 0.24, corresponding to the fact that the probability that a wage is between $20 and $30 is 0.24. The density has a single peak around $15. This is the **mode** of the distribution.

The wage density has an asymmetric shape. The left tail has a steeper slope than the right tail, which drops off more slowly. This asymmetry is called **skewness**. This is commonly observed in earnings and wealth distributions. It reflects the fact that there is a small but meaningful probability of a very high wage relative to the general population.

For continuous random variables the support is defined as the set of values for which the density is positive.

Definition 2.10 The **support** $\mathscr{X}$ of a continuous random variable is the smallest set containing $\{x : f(x) > 0\}$.

2.11 Transformations of Continuous Random Variables

If X is a random variable with continuous distribution function F then for any function $g(x)$, $Y = g(X)$ is a random variable. What is the distribution of Y?

First consider the support. If $\mathscr{X}$ is the support of X, and $g : \mathscr{X} \to \mathscr{Y}$ then $\mathscr{Y}$ is the support of Y. For example, if X has support $[0,1]$ and $g(x) = 1 + 2x$ the support for $Y = g(X)$ is $[1,3]$. If X has support $\mathbb{R}_+$ and $g(x) = \log x$ then $Y = g(X)$ has support $\mathbb{R}$.

The probability distribution of Y is $F_Y(y) = \mathbb{P}\left[Y \le y\right] = \mathbb{P}\left[g(X) \le y\right]$. Let $B(y)$ be the set of $x \in \mathbb{R}$ such that $g(x) \le y$. The events $\{g(X) \le y\}$ and $\{X \in B(y)\}$ are identical. So the distribution function for Y is

$$F_Y(y) = \mathbb{P}\left[X \in B(y)\right].$$

This shows that the distribution of Y is determined by the probability function of X.

When $g(x)$ is strictly monotonically increasing then $g(x)$ has an inverse function

$$h(y) = g^{-1}(y)$$

which implies $X = h(Y)$ and $B(y) = (-\infty, h(y)]$. The distribution function of Y is

$$F_Y(y) = \mathbb{P}\left[X \le h(y)\right] = F_X(h(y)).$$

The density function is its derivative. By the chain rule we find

$$f_Y(y) = \frac{d}{dy} F_X(h(y)) = f_X(h(y)) \frac{d}{dy} h(y) = f_X(h(y)) \left| \frac{d}{dy} h(y) \right|.$$

The last equality holds since $h(y)$ has a positive derivative.

Now suppose that $g(x)$ is monotonically decreasing with inverse function $h(y)$. Then $B(y) = [h(y), \infty)$ so

$$F_Y(y) = \mathbb{P}\left[X \ge h(y)\right] = 1 - F_X(h(y)).$$

The density function is the derivative

$$f_Y(y) = -\frac{d}{dy} F_X(h(y)) = -f_X(h(y)) \frac{d}{dy} h(y) = f_X(h(y)) \left| \frac{d}{dy} h(y) \right|.$$

The last equality holds since $h(y)$ has a negative derivative.

We have found that when $g(x)$ is strictly monotonic that the density for Y is

$$f_Y(y) = f_X(g^{-1}(y))J(y)$$

where

$$J(y) = \left|\frac{d}{dy}h(y)\right| = \left|\frac{d}{dy}g^{-1}(y)\right|$$

is called the **Jacobian** of the transformation. It should be familiar from calculus.

We have shown the following.

Theorem 2.4 If $X \sim f_X$, $f_X(x)$ is continuous on $\mathscr{X}$, $g(x)$ is strictly monotone, and $g^{-1}(y)$ is continuously differentiable on $\mathscr{Y}$, then for $y \in \mathscr{Y}$

$$f_Y(y) = f_X(g^{-1}(y))J(y)$$

where $J(y) = \left|\frac{d}{dy}g^{-1}(y)\right|$.

Theorem 2.4 gives an explicit expression for the density function of the transformation Y. We illustrate Theorem 2.4 by four examples.

Example: $f_X(x) = \exp(-x)$ for $x \geq 0$. Set $Y = \lambda X$ for some $\lambda > 0$. This means $g(x) = \lambda x$. Y has support $\mathscr{Y} = [0,\infty)$. The function $g(x)$ is monotonically increasing with inverse function $h(y) = y/\lambda$. The Jacobian is the derivative

$$J(y) = \left|\frac{d}{dy}h(y)\right| = \frac{1}{\lambda}.$$

the density of Y is

$$f_Y(y) = f_X(g^{-1}(y))J(y) = \exp\left(-\frac{y}{\lambda}\right)\frac{1}{\lambda}$$

for $y \geq 0$. This is a valid density since

$$\int_0^\infty f_Y(y)\,dy = \int_0^\infty \exp\left(-\frac{y}{\lambda}\right)\frac{1}{\lambda}\,dy = \int_0^\infty \exp(-x)\,dx = 1$$

where the second equality makes the change of variables $x = y/\lambda$.

Example: $f_X(x) = 1$ for $0 \leq x \leq 1$. Set $Y = g(X)$ where $g(x) = -\log(x)$. Since X has support $[0,1]$, Y has support $\mathscr{Y} = (0,\infty)$. The function $g(x)$ is monotonically decreasing with inverse function

$$h(y) = g^{-1}(y) = \exp(-y).$$

Take the derivative to obtain the Jacobian:

$$J(y) = \left|\frac{d}{dy}h(y)\right| = \left|-\exp(-y)\right| = \exp(-y).$$

Notice that $f_X(g^{-1}(y)) = 1$ for $y \geq 0$. We find that the density of Y is

$$f_Y(y) = f_X(g^{-1}(y))J(y) = \exp(-y)$$

for $y \geq 0$. This is the exponential density. We have shown that if X is uniformly distributed, then $Y = -\log(X)$ has an exponential distribution.

Example: Let X have any continuous and invertible (strictly increasing) CDF $F_X(x)$. Define the random variable $Y = F_X(X)$. Y has support $\mathscr{Y} = [0,1]$. The CDF of Y is

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left[Y \le y\right] \\ &= \mathbb{P}\left[F_X(X) \le y\right] \\ &= \mathbb{P}\left[X \le F_X^{-1}(y)\right] \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned}$$

on $[0,1]$. Taking the derivative we find the PDF

$$f_Y(y) = \frac{d}{dy} y = 1.$$

This is the density function of a $U[0,1]$ random variable. Thus $Y \sim U[0,1]$.

The transformation $Y = F_X(X)$ is known as the **Probability Integral Transformation**. The fact that this transformation renders Y to be uniformly distributed regardless of the initial distribution F_X is quite wonderful.

Example: Let $f_X(x)$ be the density function of wages from Figure 2.7. Let $Y = \log(X)$. If X has support on $\mathbb{R}_+$ then Y has support on $\mathbb{R}$. The inverse function is $h(y) = \exp(y)$ and Jacobian is $\exp(y)$. The density of Y is $f_Y(y) = f_X(\exp(y))\exp(y)$. It is displayed in Figure 2.8. This density is more symmetric and less skewed than the density of wages in levels.

2.12 Non-Monotonic Transformations

If $Y = g(X)$ where $g(x)$ is not monotonic, we can (in some cases) derive the distribution of Y by direct manipulations. We illustrate by focusing on the case where $g(x) = x^2$, $\mathscr{X} = \mathbb{R}$, and X has density $f_X(x)$.

Y has support $\mathscr{Y} \subset [0,\infty)$. For $y \ge 0$,

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left[Y \le y\right] \\ &= \mathbb{P}\left[X^2 \le y\right] \\ &= \mathbb{P}\left[|X| \le \sqrt{y}\right] \\ &= \mathbb{P}\left[-\sqrt{y} \le X \le \sqrt{y}\right] \\ &= \mathbb{P}\left[X \le \sqrt{y}\right] - \mathbb{P}\left[X < -\sqrt{y}\right] \\ &= F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right). \end{aligned}$$

We find the density by taking the derivative and applying the chain rule.

$$\begin{aligned} f_Y(y) &= \frac{f_X\left(\sqrt{y}\right)}{2\sqrt{y}} + \frac{f_X\left(-\sqrt{y}\right)}{2\sqrt{y}} \\ &= \frac{f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)}{2\sqrt{y}}. \end{aligned}$$

To further the example, suppose that $f_X(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ (the standard normal density). It follows that the density of $Y = X^2$ is

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}}\exp(-y/2)$$

for $y \ge 0$. This is known as the chi-square density with 1 degree of freedom, written χ^2_1. We have shown that if X is standard normal then $Y = X^2$ is χ^2_1.

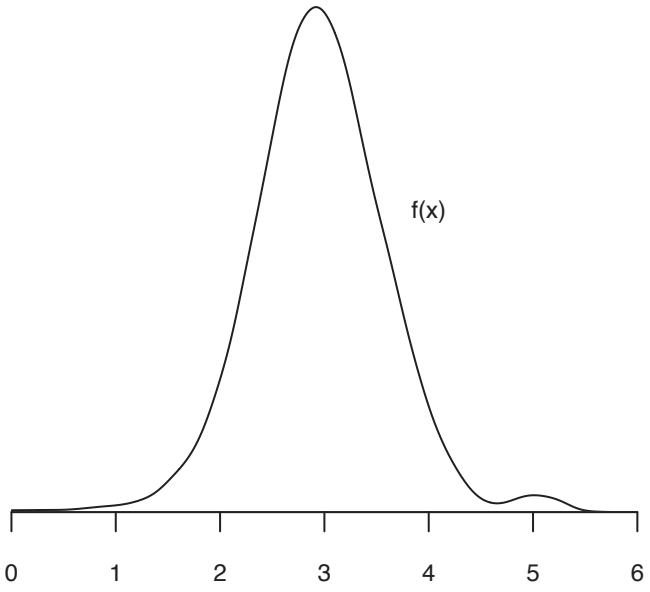


Figure 2.8: Density Function for Log Wage Distribution

2.13 Expectation of Continuous Random Variables

We introduced the expectation for discrete random variables. In this section we consider the continuous case.

Definition 2.11 If X is continuously distributed with density $f(x)$ its **expectation** is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

when the integral is convergent.

The expectation is a weighted average of x using the continuous weight function $f(x)$. Just as in the discrete case the expectation equals the center of mass of the distribution. To visualize, take any density function and imagine placing it on a board on top of a fulcrum. The board will be balanced when the fulcrum is placed at the expected value.

Example: $f(x) = 1$ on $0 \le x \le 1$.

$$\mathbb{E}[X] = \int_0^1 x dx = \frac{1}{2}.$$

Example: $f(x) = \exp(-x)$ on $x \ge 0$. We can show that $\mathbb{E}[X] = 1$. Calculation takes two steps. First,

$$\mathbb{E}[X] = \int_0^{\infty} x \exp(-x)\, dx.$$

Apply integration by parts with $u = x$ and $v = \exp(-x)$. We find

$$\mathbb{E}[X] = \int_0^\infty \exp(-x)\, dx = 1.$$

Thus $\mathbb{E}[X] = 1$ as stated.

Example: Hourly Wage Distribution (Figure 2.5). The expected value is $23.92. Examine the density plot in Figure 2.7. The expected value is approximately in the middle of the grey shaded region. This is the center of mass, balancing the high mode to the left and the thick tail to the right.

Expectations of transformations are similarly defined.

Definition 2.12 If X has density $f(x)$ then the expected value of $g(X)$ is

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Example: $X \sim f(x) = 1$ on $0 \le x \le 1$. Then $\mathbb{E}\left[X^2\right] = \int_0^1 x^2 dx = 1/3$.

Example: Log Wage Distribution (Figure 2.8). The expected value is $\mathbb{E}\left[\log(\text{wage})\right] = 2.95$. Examine the density plot in Figure 2.8. The expected value is the approximate mid-point of the density since it is approximately symmetric.

Just as for discrete random variables, the expectation is a linear operator.

Theorem 2.5 Linearity of Expectation. For any constants a and b,

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X].$$

Proof: Suppose X is continuously distributed. Then

$$\begin{aligned}
\mathbb{E}[a + bX] &= \int_{-\infty}^{\infty} (a + bx) f(x) dx \\
&= a \int_{-\infty}^{\infty} f(x) dx + b \int_{-\infty}^{\infty} x f(x) dx \\
&= a + b\mathbb{E}[X]
\end{aligned}$$

since $\int_{-\infty}^{\infty} f(x) dx = 1$ and $\int_{-\infty}^{\infty} x f(x) dx = \mathbb{E}[X]$. ■

Example: $f(x) = \exp(-x)$ on $x \ge 0$. Make the transformation $Y = \lambda X$. By the linearity of expectations and our previous calculation $\mathbb{E}[X] = 1$

$$\mathbb{E}[Y] = \mathbb{E}[\lambda X] = \lambda \mathbb{E}[X] = \lambda.$$

Alternatively, by transformation of variables we have previous shown that Y has density $\exp(-y/\lambda)/\lambda$. By direct calculation we can show that that

$$\mathbb{E}[Y] = \int_0^\infty y \exp\left(-\frac{y}{\lambda}\right) \frac{1}{\lambda} dy = \lambda.$$

Either calculation shows that the expected value of Y is λ.

2.14 Finiteness of Expectations

In our discussion of the St. Petersburg Paradox we found that there are discrete distributions which do not have convergent expectations. The same issue applies in the continuous case. It is possible for expectations to be infinite or to be not defined.

Example: $f(x) = x^{-2}$ for $x > 1$. This is a valid density since $\int_1^\infty f(x)dx = \int_1^\infty x^{-2}dx = -x^{-1}\big|_1^\infty = 1$. However the expectation is

$$\mathbb{E}[X] = \int_1^\infty x f(x)dx = \int_1^\infty x^{-1}dx = \log(x)\big|_1^\infty = \infty.$$

Thus the expectation is infinite. The reason is because the integral $\int_1^\infty x^{-1}dx$ is not convergent. The density $f(x) = x^{-2}$ is a special case of the Pareto distribution, which is used to model heavy-tailed distributions.

Example: $f(x) = \dfrac{1}{\pi\left(1+x^2\right)}$ for $x \in \mathbb{R}$. The expected value is

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)dx \\
&= \int_0^\infty \frac{x}{\pi\left(1+x^2\right)}dx + \int_{-\infty}^0 \frac{x}{\pi\left(1+x^2\right)}dx \\
&= \left.\frac{\log\left(1+x^2\right)}{2\pi}\right|_0^\infty - \left.\frac{\log\left(1+x^2\right)}{2\pi}\right|_0^\infty \\
&= \log(\infty) - \log(\infty)
\end{aligned}$$

which is undefined. This is called the Cauchy distribution.

2.15 Unifying Notation

An annoying feature of intermediate probability theory is that expectations (and other objects) are defined separately for discrete and continuous random variables. This means that all proofs have to done twice, yet the exact same steps are used in each. In advanced probability theory it is typical to instead define expectation using the Riemann-Stieltijes integral (see Appendix A.8) which combines these cases. It is useful to be familiar with the notation even if you are not familiar with the mathematical details.

Definition 2.13 For any random variable X with distribution $F(x)$ the **expectation** is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x dF(x)$$

if the integral is convergent.

For the remainder of this chapter we will not make a distinction between discrete and continuous random variables. For simplicity we will typically use the notation for continuous random variables (using densities and integration) but the arguments apply to the general case by using Riemann-Stieltijes integration.

2.16 Mean and Variance

Two of the most important features of a distribution are its mean and variance, typically denoted by the Greek letters μ and σ^2.

Definition 2.14 The **mean** of X is $\mu = \mathbb{E}[X]$.

The mean is either finite, infinite, or undefined.

Definition 2.15 The **variance** of X is $\sigma^2 = \operatorname{var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

The variance is necessarily non-negative $\sigma^2 \geq 0$. It is either finite or infinite. It is zero only for degenerate random variables.

Definition 2.16 A random variable X is **degenerate** if for some c, $\mathbb{P}[X = c] = 1$.

A degenerate random variable is essentially non-random, and has a variance of zero.

The variance is measured in square units. To put the variance in the same units as X we take its square root and give it an entirely new name.

Definition 2.17 The **standard deviation** of X is the positive square root of the variance, $\sigma = \sqrt{\sigma^2}$.

It is typical to use the mean and standard deviation to summarize the center and spread of a distribution.

The following are two useful calculations.

Theorem 2.6 $\operatorname{var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$.

To see this expand the quadratic

$$(X - \mathbb{E}[X])^2 = X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2$$

and then take expectations to find

$$\begin{aligned}
\operatorname{var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}\left[(\mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.
\end{aligned}$$

The third equality uses the fact that $\mathbb{E}[X]$ is a constant. The fourth combines terms.

Theorem 2.7 $\operatorname{var}[a + bX] = b^2 \operatorname{var}[X]$.

To see this notice that by linearity

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X].$$

Thus

$$\begin{aligned}
(a + bX) - \mathbb{E}[a + bX] &= a + bX - (a + b\mathbb{E}[X]) \\
&= b(X - \mathbb{E}[X]).
\end{aligned}$$

Hence

$$\begin{aligned}\operatorname{var}[a+bX] &= \mathbb{E}\left[((a+bX)-\mathbb{E}[a+bX])^2\right] \\ &= \mathbb{E}\left[(b(X-\mathbb{E}[X]))^2\right] \\ &= b^2\mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \\ &= b^2\operatorname{var}[X].\end{aligned}$$

An important implication of Theorem 2.7 is that the variance is invariant to additive shifts: X and $a+X$ have the same variance.

Example: A Bernoulli random variable takes the value 1 with probability p and 0 with probability $1-p$.

$$X = \begin{cases} 1 & \text{with probability} \quad p \\ 0 & \text{with probability} \quad 1-p. \end{cases}$$

This has mean and variance

$$\begin{aligned}\mu &= p \\ \sigma^2 &= p(1-p).\end{aligned}$$

See Exercise 2.4.

Example: An exponential random variable has density

$$f(x) = \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$$

for $x \geq 0$. This has mean and variance

$$\begin{aligned}\mu &= \lambda \\ \sigma^2 &= \lambda^2.\end{aligned}$$

See Exercise 2.5.

Example: Hourly Wages. (Figure 2.5). The mean, variance, and standard deviation are

$$\begin{aligned}\mu &= 24 \\ \sigma^2 &= 429 \\ \sigma &= 20.7.\end{aligned}$$

Example: Years of Education. (Figure 2.2(c)). The mean, variance, and standard deviation are

$$\begin{aligned}\mu &= 13.9 \\ \sigma^2 &= 7.5 \\ \sigma &= 2.7.\end{aligned}$$

2.17 Moments

The moments of a distribution are the expected values of the powers of X. We define both uncentered and central moments

Definition 2.18 The m^{th} **moment** of X is $\mu'_m = \mathbb{E}[X^m]$.

Definition 2.19 For $m > 1$ the m^{th} **central moment** of X is $\mu_m = \mathbb{E}[(X - \mathbb{E}[X])^m]$.

The moments are the expected values of the variables X^m. The central moments are those of $(X - \mathbb{E}[X])^m$. Odd moments may be finite, infinite, or undefined. Even moments are either finite or infinite. For non-negative X, m can be real-valued.

For ease of reference we define the first central moment as $\mu_1 = \mathbb{E}[X]$.

Theorem 2.8 For $m > 1$ the central moments are invariant to additive shifts, that is $\mu_m(a + X) = \mu_m(X)$.

2.18 Jensen's Inequality

Expectation is a linear operator $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$. It is tempting to apply the same reasoning to nonlinear functions but this is not valid. We can say more for convex and concave functions.

Definition 2.20 The function $g(x)$ is **convex** if for any $\lambda \in [0,1]$ and all x and y

$$g\left(\lambda x + (1-\lambda)y\right) \leq \lambda g(x) + (1-\lambda)g\left(y\right).$$

The function is **concave** if

$$\lambda g(x) + (1-\lambda)g\left(y\right) \leq g\left(\lambda x + (1-\lambda)y\right).$$

Examples of convex functions are the exponential $g(x) = \exp(x)$ and quadratic $g(x) = x^2$. Examples of concave functions are the logarithm $g(x) = \log(x)$ and square root $g(x) = x^{1/2}$ for $x \geq 0$.

Concavity is illustrated in Figure 2.9. A concave function $f(x)$ is displayed, and a chord between the points a and b is drawn. The chord lies below the function. The point c lies on the chord, and is less than the point d which lies on the function.

Theorem 2.9 Jensen's Inequality. For any random variable X, if $g(x)$ is a convex function then

$$g(\mathbb{E}[X]) \leq \mathbb{E}\left[g(X)\right].$$

If $g(x)$ is a concave function then

$$\mathbb{E}\left[g(X)\right] \leq g(\mathbb{E}[X]).$$

Proof. We focus on the convex case. Let $a + bx$ be the tangent line to $g(x)$ at $x = \mathbb{E}[X]$. Since $g(x)$ is convex, $g(x) \geq a + bx$. Evaluated at $x = X$ and taking expectations we find

$$\mathbb{E}\left[g(X)\right] \geq a + b\mathbb{E}[X] = g(\mathbb{E}[X])$$

as claimed. ■

Jensen's equality states that a convex function of an expectation is less than the expectation of the transformation. Conversely, the expectation of a concave transformation is less than the function of the expectation.

Examples of Jensen's inequality are:

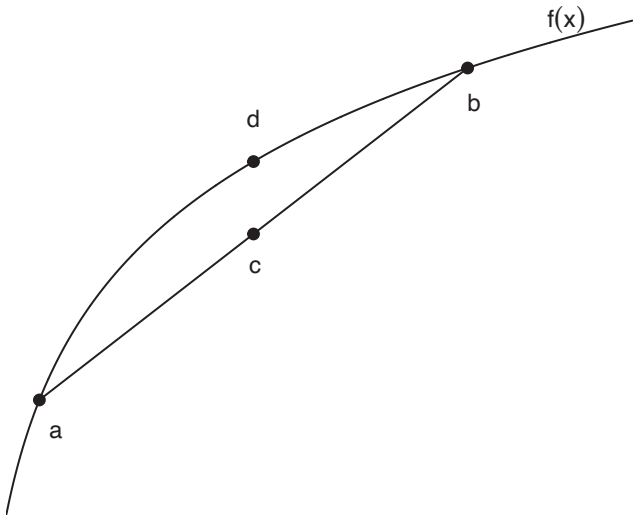


Figure 2.9: Concavity

1. $\exp(\mathbb{E}[X]) \leq \mathbb{E}\left[\exp(X)\right]$
2. $(\mathbb{E}[X])^2 \leq \mathbb{E}\left[X^2\right]$
3. $\mathbb{E}\left[\log(X)\right] \leq \log(\mathbb{E}[X])$
4. $\mathbb{E}\left[X^{1/2}\right] \leq (\mathbb{E}[X])^{1/2}$.

2.19 Applications of Jensen's Inequality*

Jensen's inequality can be used to establish other useful results.

Theorem 2.10 Expectation Inequality. For any random variable X

$$|\mathbb{E}[X]| \leq \mathbb{E}|X|.$$

Proof: The function $g(x) = |x|$ is convex. An application of Jensen's inequality with $g(x)$ yields the result. ■

Theorem 2.11 Lyapunov's Inequality. For any random variable X and any $0 < r \leq p$

$$\left(\mathbb{E}|X|^r\right)^{1/r} \leq \left(\mathbb{E}|X|^p\right)^{1/p}.$$

Proof: The function $g(x) = x^{p/r}$ is convex for $x > 0$ since $p \geq r$. Let $Y = |X|^r$. By Jensen's inequality

$$g\left(\mathbb{E}\left[Y\right]\right) \leq \mathbb{E}\left[g\left(Y\right)\right]$$

or

$$\left(\mathbb{E}\left|X\right|^r\right)^{p/r} \leq \mathbb{E}\left|X\right|^p.$$

Raising both sides to the power $1/p$ completes the proof. ■

Theorem 2.12 Discrete Jensen's Inequality. If $g(x) : \mathbb{R} \to \mathbb{R}$ is convex, then for any non-negative weights a_j such that $\sum_{j=1}^m a_j = 1$ and any real numbers x_j

$$g\left(\sum_{j=1}^m a_j x_j\right) \leq \sum_{j=1}^m a_j g\left(x_j\right). \tag{2.2}$$

Proof: Let X be a discrete random variable with distribution $\mathbb{P}\left[X = x_j\right] = a_j$. Jensen's inequality implies (2.2). ■

Theorem 2.13 Geometric Mean Inequality. For any non-negative real weights a_j such that $\sum_{j=1}^m a_j = 1$, and any non-negative real numbers x_j

$$x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m} \leq \sum_{j=1}^m a_j x_j. \tag{2.3}$$

Proof: Since the logarithm is strictly concave, by the discrete Jensen inequality

$$\log\left(x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m}\right) = \sum_{j=1}^m a_j \log x_j \leq \log\left(\sum_{j=1}^m a_j x_j\right).$$

Applying the exponential yields (2.3). ■

Theorem 2.14 Loève's c_r Inequality. For any real numbers x_j, if $0 < r \leq 1$

$$\left|\sum_{j=1}^m x_j\right|^r \leq \sum_{j=1}^m \left|x_j\right|^r \tag{2.4}$$

and if $r \geq 1$

$$\left|\sum_{j=1}^m x_j\right|^r \leq m^{r-1} \sum_{j=1}^m \left|x_j\right|^r. \tag{2.5}$$

For the important special case $m = 2$ we can combine these two inequalities as

$$|a+b|^r \leq C_r\left(|a|^r + |b|^r\right) \tag{2.6}$$

where $C_r = \max\left[1, 2^{r-1}\right]$.

Proof: For $r \geq 1$ this is a rewriting of Jensen's inequality (2.2) with $g(u) = u^r$ and $a_j = 1/m$. For $r < 1$, define $b_j = \left|x_j\right| / \left(\sum_{j=1}^m \left|x_j\right|\right)$. The facts that $0 \leq b_j \leq 1$ and $r < 1$ imply $b_j \leq b_j^r$ and thus

$$1 = \sum_{j=1}^m b_j \leq \sum_{j=1}^m b_j^r.$$

This implies

$$\left(\sum_{j=1}^m x_j\right)^r \leq \left(\sum_{j=1}^m \left|x_j\right|\right)^r \leq \sum_{j=1}^m \left|x_j\right|^r.$$

■

Theorem 2.15 Norm Monotonicity. If $0 < t \leq s$, for any real numbers x_j

$$\left|\sum_{j=1}^m \left|x_j\right|^s\right|^{1/s} \leq \left|\sum_{j=1}^m \left|x_j\right|^t\right|^{1/t}. \tag{2.7}$$

Proof: Set $y_j = \left|x_j\right|^s$ and $r = t/s \leq 1$. The c_r inequality (2.4) implies $\left|\sum_{j=1}^m y_j\right|^r \leq \sum_{j=1}^m \left|y_j\right|^r$ or

$$\left|\sum_{j=1}^m \left|x_j\right|^s\right|^{t/s} \leq \sum_{j=1}^m \left|x_j\right|^t.$$

Raising both sides to the power $1/t$ yields (2.7). ■

2.20 Symmetric Distributions

We say that a distribution of a random variable is **symmetric** about 0 if the distribution function satisfies

$$F(x) = 1 - F(-x).$$

If X has a density $f(x)$, X is symmetric about 0 if

$$f(x) = f(-x).$$

For example, the standard normal density $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ is symmetric about zero.

If a distribution is symmetric about zero then all finite odd moments are zero (if the moment is finite). To see this let m be odd. Then

$$\mathbb{E}\left[X^m\right] = \int_{-\infty}^{\infty} x^m f(x) dx = \int_0^{\infty} x^m f(x) dx + \int_{-\infty}^0 x^m f(x) dx.$$

For the integral from $-\infty$ to 0 make the change-of-variables $x = -t$. Then the right hand side equals

$$\int_0^{\infty} x^m f(x) dx + \int_0^{\infty} (-t)^m f(-t) dt = \int_0^{\infty} x^m f(x) dx - \int_0^{\infty} t^m f(t) dt = 0$$

where the first equality uses the symmetry property. The second equality holds if the moment is finite. This last step is subtle and easy to miss. If $\mathbb{E}|X|^m = \infty$ then the last step is

$$\mathbb{E}\left[X^m\right] = \infty - \infty \neq 0.$$

In this case $\mathbb{E}[X^m]$ is undefined.

More generally, if a distribution is symmetric about zero then the expectation of any odd function (if finite) is zero. An odd function satisfies $g(-x) = -g(x)$. For example, $g(x) = x^3$ and $g(x) = \sin(x)$ are odd functions. To see this let $g(x)$ be an odd function and without loss of generality assume $\mathbb{E}[X] = 0$. Then making a change-of-variables $x = -t$

$$\int_{-\infty}^{0} g(x)f(x)dx = \int_{0}^{\infty} g(-t)f(-t)dt = -\int_{0}^{\infty} g(t)f(t)dt$$

where the second equality uses the assumptions that $g(x)$ is odd and $f(x)$ is symmetric about zero. Then

$$\mathbb{E}\left[g(X)\right] = \int_{0}^{\infty} g(x)f(x)dx + \int_{-\infty}^{0} g(x)f(x)dx = \int_{0}^{\infty} g(x)f(x)dx - \int_{0}^{\infty} g(t)f(t)dt = 0.$$

Theorem 2.16 If $f(x)$ is symmetric about zero, $g(x)$ is odd, and $\int_0^\infty \left|g(x)\right| f(x)dx < \infty$, then $\mathbb{E}\left[g(X)\right] = 0$.

2.21 Truncated Distributions

Sometimes we only observe part of a distribution. For example, consider a sealed bid auction for a work of art with a minimum bid requirement. Assuming that participants make a bid based on their personal valuation, they will only make bids if their personal valuation is at least as high as the required minimum. Thus we do not observe "bids" from these participants. This is an example of truncation from below. Truncation is a specific transformation of a random variable.

If X is a random variable with distribution $F(x)$ and X is truncated to satisfy $X \le c$ (truncated from above) then the truncated distribution function is

$$F^*(x) = \mathbb{P}[X \le x \mid X \le c] = \begin{cases} \dfrac{F(x)}{F(c)}, & x < c \\ 1, & x \ge c. \end{cases}$$

If $F(x)$ is continuous then the density of the truncated distribution is

$$f^*(x) = f(x \mid X \le c) = \frac{f(x)}{F(c)}$$

for $x \le c$. The mean of the truncated distribution is

$$\mathbb{E}[X \mid X \le c] = \frac{\int_{-\infty}^{c} xf(x)dx}{F(c)}.$$

If X is truncated to satisfy $X \ge c$ (truncated from below) then the truncated distribution and density functions are

$$F^*(x) = \mathbb{P}[X \le x \mid X \ge c] = \begin{cases} 0, & x < c \\ \dfrac{F(x) - F(c)}{1 - F(c)}, & x \ge c \end{cases}$$

and

$$f^*(x) = f(x \mid X \ge c) = \frac{f(x)}{1 - F(c)}$$

for $x \ge c$. The mean of the truncated distribution is

$$\mathbb{E}[X \mid X \ge c] = \frac{\int_{c}^{\infty} xf(x)dx}{1 - F(c)}.$$

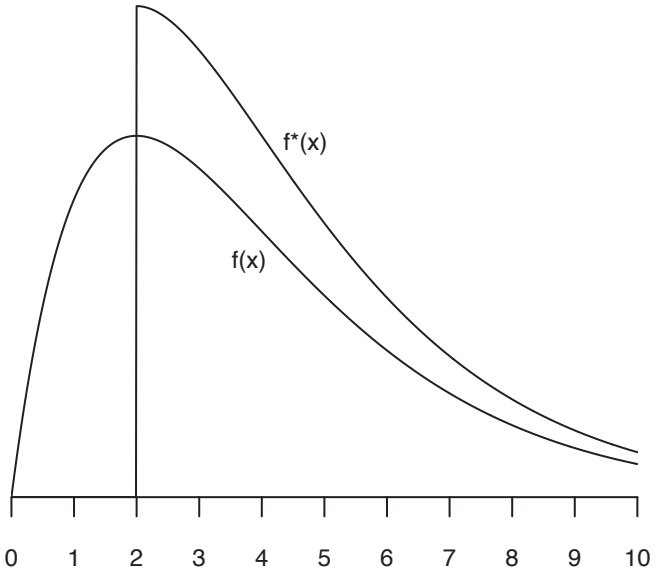


Figure 2.10: Truncation

Trucation from below is illustrated in Figure 2.10. The untruncated density function is marked as $f(x)$. The truncated density is marked as $f^*(x)$. The portion of $f(x)$ below 2 is eliminated and the density above 2 is shifted up to compensate.

An interesting example is the exponential $F(x) = 1 - e^{-x/\lambda}$ with density $f(x) = \lambda^{-1} e^{-x/\lambda}$ and mean λ. The truncated density is

$$f(x \mid X \geq c) = \frac{\lambda^{-1} e^{-x/\lambda}}{e^{-c/\lambda}} = \lambda^{-1} e^{-(x-c)/\lambda}.$$

This is also the expontial distribution, but shifted by c. The mean of this distribution is $c + \lambda$, which is the same as the original exponential distribution shifted by c. This is a "memoryless" property of the exponential distribution.

2.22 Censored Distributions

Sometimes a boundary constraint is forced on a random variable. For example, let X be a desired level of consumption, but X^* is constrained to satisfy $X^* \leq c$ (censored from above). In this case if $X > c$, the constrained consumption will satisfy $X^* = c$. We can write this as

$$X^* = \begin{cases} X, & X \leq c \\ c, & X > c. \end{cases}$$

Similarly if X^* is constrained to satisfy $X^* \geq c$ then when $X < c$ the constrained version will satisfy $X^* = c$. We can write this as

$$X^* = \begin{cases} X, & X \geq c \\ c, & X < c. \end{cases}$$

Censoring is related to truncation but is different. Under truncation the random variables exceeding the boundary are excluded. Under censoring they are transformed to satisfy the constraint.

When the original random variable X is continuously distributed then the censored random variable X^* will have a mixed distribution, with a continuous component over the unconstrained set and a discrete mass at the constrained boundary.

Censoring is common in economic applications. A standard example is consumer purchases of individual items. In this case one constraint is $X^* \geq 0$ and consequently we typically observe a discrete mass at zero. Another standard example is "top-coding" where a continuous variable such as income is recorded either in categories or is continuous up to a top category "income above $Y". All incomes above this threshold are recorded at the threshold $Y.

The expected value of a censored random variable is

$$X^* \leq c: \qquad \mathbb{E}\left[X^*\right] = \int_{-\infty}^{c} x f(x)\,dx + c\,(1 - F(c))$$
$$X^* \geq c: \qquad \mathbb{E}\left[X^*\right] = \int_{c}^{\infty} x f(x)\,dx + cF(c).$$

2.23 Moment Generating Function

The following is a technical tool used to facilitate some proofs. It is not particularly intuitive.

Definition 2.21 The **moment generating function (MGF)** of X is $M(t) = \mathbb{E}\left[\exp(tX)\right]$.

Since the exponential is non-negative the MGF is either finite or infinite. For the MGF to be finite the density of X must have thin tails. When we use the MGF we are implicitly assuming that it is finite.

Example: $U[0,1]$. The density is $f(x) = 1$ for $0 \leq x \leq 1$. The MGF is

$$M(t) = \int_{-\infty}^{\infty} \exp(tx) f(x)\,dx = \int_0^1 \exp(tx)\,dx = \frac{\exp(t) - 1}{t}.$$

Example: Exponential Distribution. The density is $f(x) = \lambda^{-1}\exp(x/\lambda)$ for $x \geq 0$. The MGF is

$$M(t) = \int_{-\infty}^{\infty} \exp(tx) f(x)\,dx = \frac{1}{\lambda}\int_0^{\infty} \exp(tx)\exp\left(-\frac{x}{\lambda}\right)dx = \frac{1}{\lambda}\int_0^{\infty} \exp\left(\left(t - \frac{1}{\lambda}\right)x\right)dx.$$

This integral is only convergent if $t < 1/\lambda$. Assuming this holds make the change of variables $y = \left(t - \frac{1}{\lambda}\right)x$ and we find the above integral equals

$$M(t) = -\frac{1}{\lambda\left(t - \frac{1}{\lambda}\right)} = \frac{1}{1 - \lambda t}.$$

In this example the MGF is finite only on the region $t < 1/\lambda$.

Example: $f(x) = x^{-2}$ for $x > 1$. The MGF is

$$M(t) = \int_1^{\infty} \exp(tx)\, x^{-2}\,dx = \infty$$

for $t > 0$. This non-convergence means that in this example the MGF cannot be succesfully used for calculations.

The moment generating function has the important property that it completely characterizes the distribution of X. It also has the following properties.

Theorem 2.17 Moments and the MGF. If $M(t)$ is finite for t in a neighborhood of 0 then $M(0) = 1$,

$$\frac{d}{dt}M(t)\bigg|_{t=0} = \mathbb{E}[X],$$
$$\frac{d^2}{dt^2}M(t)\bigg|_{t=0} = \mathbb{E}\left[X^2\right],$$

and

$$\frac{d^m}{dt^m}M(t)\bigg|_{t=0} = \mathbb{E}\left[X^m\right],$$

for any moment which is finite.

This is why it is called the "moment generating" function. The curvature of $M(t)$ at $t = 0$ encodes all moments of the distribution of X.

Example: $U[0,1]$. The MGF is $M(t) = t^{-1}\left(\exp(t) - 1\right)$. Using L'Hôpital's rule (Theorem A.12)

$$M(0) = \exp(0) = 1.$$

Using the derivative rule of differentiation and L'Hôpital's rule

$$\mathbb{E}[X] = \frac{d}{dt}\frac{\exp(t)-1}{t}\bigg|_{t=0} = \frac{t\exp(t) - \left(\exp(t)-1\right)}{t^2}\bigg|_{t=0} = \frac{\exp(0)}{2} = \frac{1}{2}.$$

Example: Exponential Distribution. The MGF is $M(t) = (1-\lambda t)^{-1}$. The first moment is

$$\mathbb{E}[X] = \frac{d}{dt}\frac{1}{1-\lambda t}\bigg|_{t=0} = \frac{\lambda}{(1-\lambda t)^2}\bigg|_{t=0} = \lambda.$$

The second moment is

$$\mathbb{E}\left[X^2\right] = \frac{d^2}{dt^2}\frac{1}{1-\lambda t}\bigg|_{t=0} = \frac{2\lambda^2}{(1-\lambda t)^3}\bigg|_{t=0} = 2\lambda^2.$$

Proof of Theorem 2.17 We use the assumption that X is continuously distributed. Then

$$M(t) = \int_{-\infty}^{\infty} \exp(tx) f(x) dx$$

so

$$M(0) = \int_{-\infty}^{\infty} \exp(0x) f(x) dx = \int_{-\infty}^{\infty} f(x) dx = 1.$$

The first derivative is

$$\begin{aligned}\frac{d}{dt}M(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} \exp(tx) f(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d}{dt}\exp(tx) f(x) dx \\ &= \int_{-\infty}^{\infty} \exp(tx)\, x f(x) dx.\end{aligned}$$

Evaluated at $t = 0$ we find

$$\frac{d}{dt}M(t)\bigg|_{t=0} = \int_{-\infty}^{\infty} \exp(0x)\, x f(x)\, dx = \int_{-\infty}^{\infty} x f(x)\, dx = \mathbb{E}[X]$$

as claimed. Similarly

$$\begin{aligned}\frac{d^m}{dt^m}M(t) &= \int_{-\infty}^{\infty} \frac{d^m}{dt^m}\exp(tx)\, f(x)\, dx \\ &= \int_{-\infty}^{\infty} \exp(tx)\, x^m f(x)\, dx\end{aligned}$$

so

$$\frac{d^m}{dt^m}M(t)\bigg|_{t=0} = \int_{-\infty}^{\infty} \exp(0x)\, x^m f(x)\, dx = \int_{-\infty}^{\infty} x^m f(x)\, dx = \mathbb{E}\left[X^m\right]$$

as claimed. ■

2.24 Cumulants

The **cumulant generating function** is the natural log of the moment generating function

$$K(t) = \log M(t).$$

Since $M(0) = 1$ we see $K(0) = 0$. Expanding as a power series we obtain

$$K(t) = \sum_{r=1}^{\infty} \kappa_r \frac{t^r}{r!}$$

where

$$\kappa_r = K^{(r)}(0)$$

is the r^{th} derivative of $K(t)$, evaluated at $t = 0$. The constants κ_r are known as the **cumulants** of the distribution. Note that $\kappa_0 = K(0) = 0$ since $M(0) = 1$.

The cumulants are related to the central moments. We can calculate that

$$\begin{aligned}K^{(1)}(t) &= \frac{M^{(1)}(t)}{M(t)} \\ K^{(2)}(t) &= \frac{M^{(2)}(t)}{M(t)} - \left(\frac{M^{(1)}(t)}{M(t)}\right)^2\end{aligned}$$

so $\kappa_1 = \mu_1$ and $\kappa_2 = \mu_2' - \mu_1^2 = \mu_2$. The first six cumulants are as follows.

$$\begin{aligned}\kappa_1 &= \mu_1 \\ \kappa_2 &= \mu_2 \\ \kappa_3 &= \mu_3 \\ \kappa_4 &= \mu_4 - 3\mu_2^2 \\ \kappa_5 &= \mu_5 - 10\mu_3\mu_2 \\ \kappa_6 &= \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3.\end{aligned}$$

We see that the first three cumulants correspond to the central moments, but higher cumulants are polynomial functions of the central moments.

Inverting, we can also express the central moments in terms of the cumulants, for example, the 4^{th} through 6^{th} are as follows.

$$\begin{aligned}\mu_4 &= \kappa_4 + 3\kappa_2^2\\ \mu_5 &= \kappa_5 + 10\kappa_3\kappa_2\\ \mu_6 &= \kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3.\end{aligned}$$

Example: Exponential Distribution. The MGF is $M(t) = (1-\lambda t)^{-1}$. The cumulant generating function is $K(t) = -\log(1-\lambda t)$. The first four derivatives are

$$\begin{aligned}K^{(1)}(t) &= \frac{\lambda}{1-\lambda t}\\ K^{(2)}(t) &= \frac{\lambda^2}{(1-\lambda t)^2}\\ K^{(3)}(t) &= \frac{2\lambda^3}{(1-\lambda t)^3}\\ K^{(4)}(t) &= \frac{6\lambda^4}{(1-\lambda t)^4}.\end{aligned}$$

Thus the first four cumulants of the distribution are λ, λ^2, $2\lambda^3$ and $6\lambda^4$.

2.25 Characteristic Function

Because the moment generating function is not necessarily finite the characteristic function is used for advanced formal proofs.

Definition 2.22 The **characteristic function (CF)** of X is $C(t) = \mathbb{E}\left[\exp(itX)\right]$ where $i = \sqrt{-1}$.

Since $\exp(iu) = \cos(u) + i\sin(u)$ is bounded, the characteristic function exists for all random variables.

If X is symmetrically distributed about zero then since the sine function is odd and bounded, $\mathbb{E}[\sin(X)] = 0$. It follows that for symmetrically distributed random variables the characteristic function can be written in terms of the cosine function only.

Theorem 2.18 If X is symmetrically distributed about 0 its characteristic function equals $C(t) = \mathbb{E}[\cos(tX)]$.

The characteristic function has similar properties as the MGF, but is a bit more tricky to deal with since it involves complex numbers.

Example: Exponential Distribution. The density is $f(x) = \lambda^{-1}\exp(-x/\lambda)$ for $x \geq 0$. The CF is

$$C(t) = \int_0^\infty \exp(itx)\frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)dx = \frac{1}{\lambda}\int_0^\infty \exp\left(\left(it - \frac{1}{\lambda}\right)x\right)dx.$$

Making the change of variables $y = \left(it - \frac{1}{\lambda}\right)x$ we find the above integral equals

$$M(t) = -\frac{1}{\lambda\left(it - \frac{1}{\lambda}\right)} = \frac{1}{1-\lambda it}.$$

This is finite for all t.

2.26 Expectation: Mathematical Details*

In this section we give a rigorous definition of expectation. Define the Riemann-Stieltijes integrals

$$I_1 = \int_0^\infty x dF(x) \tag{2.8}$$

$$I_2 = \int_{-\infty}^0 x dF(x). \tag{2.9}$$

The integral I_1 is the integral over the positive real line and the integral I_2 is the integral over the negative real line. The number I_1 can be 0, positive, or positive infinity. The number I_2 can be 0, negative, or negative infinity.

Definition 2.23 The **expectation** $\mathbb{E}[X]$ of a random variable X is

$$\mathbb{E}[X] = \begin{cases} I_1 + I_2 & \text{if both } I_1 < \infty \text{ and } I_2 > -\infty \\ \infty & \text{if } I_1 = \infty \text{ and } I_2 > -\infty \\ -\infty & \text{if } I_1 < \infty \text{ and } I_2 = -\infty \\ \text{undefined} & \text{if both } I_1 = \infty \text{ and } I_2 = -\infty. \end{cases}$$

This definition allows for an expectation to be finite, infinite, or undefined. The expectation $\mathbb{E}[X]$ is finite if and only if

$$\mathbb{E}|X| = \int_{-\infty}^\infty |x| \, dF(x) < \infty.$$

In this case it is common to say that $\mathbb{E}[X]$ is **well-defined**.

More generally, X has a finite r^{th} moment if

$$\mathbb{E}|X|^r < \infty. \tag{2.10}$$

By Lyapunov's Inequality (Theorem 2.11), (2.10) implies $\mathbb{E}|X|^s < \infty$ for all $0 \le s \le r$. Thus, for example, if the fourth moment is finite then the first, second and third moments are also finite, and so is the 3.9^{th} absolute moment.

2.27 Exercises

Exercise 2.1 Let $X \sim U[0,1]$. Find the PDF of $Y = X^2$.

Exercise 2.2 Let $X \sim U[0,1]$. Find the distribution function of $Y = \log\left(\dfrac{X}{1-X}\right)$.

Exercise 2.3 Define $F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \exp(-x) & \text{if } x \ge 0. \end{cases}$

(a) Show that $F(x)$ is a CDF.

(b) Find the PDF $f(x)$.

(c) Find $\mathbb{E}[X]$.

(d) Find the PDF of $Y = X^{1/2}$.

Exercise 2.4 A Bernoulli random variable takes the value 1 with probability p and 0 with probability $1-p$

$$X = \begin{cases} 1 & \text{with probability} \quad p \\ 0 & \text{with probability} \quad 1-p. \end{cases}$$

Find the mean and variance of X.

Exercise 2.5 Find the mean and variance of X with density $f(x) = \dfrac{1}{\lambda}\exp\left(-\dfrac{x}{\lambda}\right)$.

Exercise 2.6 Compute $\mathbb{E}[X]$ and $\operatorname{var}[X]$ for the following distributions.

(a) $f(x) = ax^{-a-1}$, $0 < x < 1$, $a > 0$.

(b) $f(x) = \frac{1}{n}$, $x = 1, 2, ..., n$.

(c) $f(x) = \frac{3}{2}(x-1)^2$, $0 < x < 2$.

Exercise 2.7 Let X have density

$$f_X(x) = \frac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} x^{r/2-1}\exp\left(-\frac{x}{2}\right)$$

for $x \geq 0$. This is known as the **chi-square** distribution. Let $Y = 1/X$. Show that the density of Y is

$$f_Y(y) = \frac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} y^{-r/2-1}\exp\left(-\frac{1}{2y}\right)$$

for $y \geq 0$. This is known as the **inverse chi-square** distribution.

Exercise 2.8 Show that if the density satisfies $f(x) = f(-x)$ for all $x \in \mathbb{R}$ then the distribution function satisfies $F(-x) = 1 - F(x)$.

Exercise 2.9 Suppose X has density $f(x) = e^{-x}$ on $x > 0$. Set $Y = \lambda X$ for $\lambda > 0$. Find the density of Y.

Exercise 2.10 Suppose X has density $f(x) = \lambda^{-1}e^{-x/\lambda}$ on $x > 0$ for some $\lambda > 0$. Set $Y = X^{1/\alpha}$ for $\alpha > 0$. Find the density of Y.

Exercise 2.11 Suppose X has density $f(x) = e^{-x}$ on $x > 0$. Set $Y = -\log X$. Find the density of Y.

Exercise 2.12 Find the median of the density $f(x) = \frac{1}{2}\exp(-|x|)$, $x \in \mathbb{R}$.

Exercise 2.13 Find a which minimizes $\mathbb{E}\left[(X-a)^2\right]$. Your answer should be a moment of X.

Exercise 2.14 Show that if X is a continuous random variable, then

$$\min_a \mathbb{E}|X-a| = \mathbb{E}|X-m|,$$

where m is the median of X.
Hint: Work out the integral expression of $\mathbb{E}|X-a|$ and notice that it is differentiable.

Exercise 2.15 The **skewness** of a distribution is

$$\text{skew} = \frac{\mu_3}{\sigma^3}$$

where μ_3 is the 3^{rd} central moment.

(a) Show that if the density function is symmetric about some point a, then skew = 0.

(b) Calculate skew for $f(x) = \exp(-x),\ x \geq 0$.

Exercise 2.16 Let X be a random variable with $\mathbb{E}[X] = 1$. Show that $\mathbb{E}\left[X^2\right] > 1$ if X is not degenerate.
Hint: Use Jensen's inequality.

Exercise 2.17 Let X be a random variable with mean μ and variance σ^2. Show that $\mathbb{E}\left[\left(X-\mu\right)^4\right] \geq \sigma^4$.

Exercise 2.18 Suppose the random variable X is a duration (a period of time). Examples include: a spell of unemployment, length of time on a job, length of a labor strike, length of a recession, length of an economic expansion. The **hazard function** $h(x)$ associated with X is

$$h(x) = \lim_{\delta \to 0} \frac{\mathbb{P}[x \leq X \leq x+\delta \mid X \geq x]}{\delta}.$$

This can be interpreted as the rate of change of the probability of continued survival. If $h(x)$ increases (decreases) with x this is described as **increasing (decreasing) hazard**.

(a) Show that if X has distribution F and density f then $h(x) = f(x)/(1-F(x))$.

(b) Suppose $f(x) = \lambda^{-1}\exp(-x/\lambda)$. Find the hazard function $h(x)$. Is this increasing or decreasing hazard or neither?

(c) Find the hazard function for the Weibull distribution from Section 3.23. Is this increasing or decreasing hazard or neither?

Exercise 2.19 Let $X \sim U[0,1]$ be uniformly distributed on $[0,1]$. (X has density $f(x) = 1$ on $[0,1]$, zero elsewhere.) Suppose X is truncated to satisfy $X \leq c$ for some $0 < c < 1$.

(a) Find the density function of the truncated variable X.

(b) Find $\mathbb{E}[X \mid X \leq c]$.

Exercise 2.20 Let X have density $f(x) = e^{-x}$ for $x \geq 0$. Suppose X is censored to satisfy $X^* \geq c > 0$. Find the mean of the censored distribution.

Exercise 2.21 Surveys routinely ask discrete questions when the underlying variable is continuous. For example, *wage* may be continuous but the survey questions are categorical. Take the following example.

wage	Frequency
\$0 ≤ wage ≤ \$10	0.1
\$10 < wage ≤ \$20	0.4
\$20 < wage ≤ \$30	0.3
\$30 < wage ≤ \$40	0.2

Assume that \$40 is the maximal wage.

(a) Plot the discrete distribution function putting the probability mass at the right-most point of each interval. Repeat putting the probability mass at the left-most point of each interval. Compare. What can you say about the true distribution function?

(b) Calculate the expected wage using the two discrete distributions from part (a). Compare.

(c) Make the assumption that the distribution is uniform on each interval. Plot this density function, distribution function, and expected wage. Compare with the above results.

Exercise 2.22 First-Order Stochastic Dominance. A distribution $F(x)$ is said to first-order dominate distribution $G(x)$ if $F(x) \leq G(x)$ for all x and $F(x) < G(x)$ for at least one x. Show the following proposition: F stochastically dominates G if and only if every utility maximizer with increasing utility in X prefers outcome $X \sim F$ over outcome $X \sim G$.

Chapter 3

Parametric Distributions

3.1 Introduction

A **parametric distribution** $F(x \mid \theta)$ is a distribution indexed by a **parameter** $\theta \in \Theta$. For each θ, the function $F(x \mid \theta)$ is a valid distribution function. As θ varies, the distribution function changes. The set Θ is called the **parameter space**. We sometimes call $F(x \mid \theta)$ a **family** of distributions. Parametric distributions are typically simple in shape and functional form, and are selected for their ease of manipulation.

Parametric distributions are often used by economists for economic modeling. A specific distribution may be selected based on appropriateness, convenience, and tractability.

Econometricians use parametric distributions for statistical modeling. A set of observations may be modeled using a specific distribution. The parameters of the distribution are unspecified, and their values chosen (estimated) to match features of the data. It is therefore desirable to understand how variation in the parameters leads to variation in the shape of the distribution.

In this chapter we list common parametric distributions used by economists and features of these distribution such as their mean and variance. The list is not exhaustive. It is not necessary to memorize this list nor the details. Rather, this information is presented for reference.

3.2 Bernoulli Distribution

A **Bernoulli** random variable is a two-point distribution. It is typically parameterized as

$$\begin{aligned}\mathbb{P}[X = 0] &= 1 - p \\ \mathbb{P}[X = 1] &= p.\end{aligned}$$

The probability mass function is

$$\begin{aligned}\pi\left(x \mid p\right) &= p^x \left(1 - p\right)^{1-x} \\ 0 < p &< 1.\end{aligned}$$

A Bernoulli distribution is appropriate for any variable which only has two outcomes, such as a coin flip. The parameter p indexes the likelihood of the two events.

$$\begin{aligned}\mathbb{E}[X] &= p \\ \operatorname{var}[X] &= p(1 - p).\end{aligned}$$

3.3 Rademacher Distribution

A **Rademacher** random variable is a two-point distribution. It is parameterized as

$$\begin{aligned}\mathbb{P}[X=-1] &= 1/2 \\ \mathbb{P}[X=1] &= 1/2.\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X] &= 0 \\ \text{var}[X] &= 1.\end{aligned}$$

3.4 Binomial Distribution

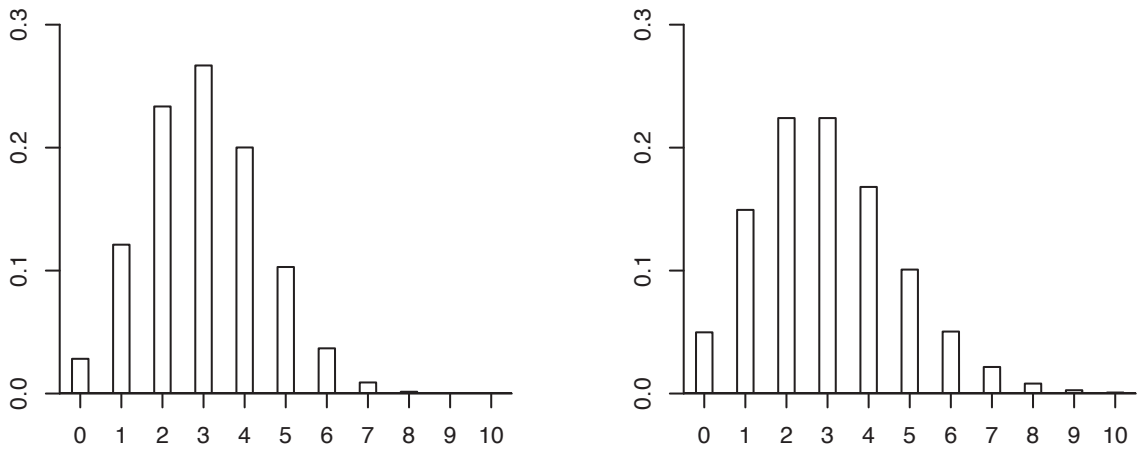
A **Binomial** random variable has support $\{0,1,...,n\}$ and probability mass function

$$\begin{aligned}\pi\left(x \mid n,p\right) &= \binom{n}{x} p^x \left(1-p\right)^{n-x}, \qquad x=0,1,...,n \\ 0 &< p < 1.\end{aligned}$$

The binomial random variable equals the outcome of n independent Bernoulli trials. If you flip a coin n times, the number of heads has a binomial distribution.

$$\begin{aligned}\mathbb{E}[X] &= np \\ \text{var}[X] &= np(1-p).\end{aligned}$$

The binomial probability mass function is displayed in Figure 3.1(a) for the case $p = 0.3$ and $n = 10$.



(a) Binomial (b) Poisson

Figure 3.1: Discrete Distributions

3.5 Multinomial Distribution

The term **multinomial** has two uses in econometrics.

(1) A single **multinomial** random variable or **multinomial trial** is a K-point distribution with support $\{x_1, ..., x_K\}$. The probability mass function is

$$\pi\left(x_j \mid p_1, ..., p_K\right) = p_j$$
$$\sum_{j=1}^{K} p_j = 1.$$

This can be written as

$$\pi\left(x_j \mid p_1, ..., p_K\right) = p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K}.$$

A multinomial can be used to model categorical outcomes (e.g. car, bicycle, bus, or walk), ordered numerical outcomes (a roll of a K-sided die), or numerical outcomes on any set of support points. This is the most common usage of the term "multinomial" in econometrics.

$$\mathbb{E}[X] = \sum_{j=1}^{K} p_j x_j$$
$$\operatorname{var}[X] = \sum_{j=1}^{K} p_j x_j^2 - \left(\sum_{j=1}^{K} p_j x_j\right)^2.$$

(2) A **multinomial** is the set of outcomes from n independent single multinomial trials. It is the sum of outcomes for each category, and is thus a set $(X_1, ..., X_K)$ of random variables satisfying $\sum_{j=1}^{K} X_k = n$. The probability mass function is

$$\mathbb{P}\left[X_1 = x_1, ..., X_K = x_k \mid n, p_1, ..., p_K\right] = \frac{n!}{x_1! \cdots x_K!} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K}$$
$$\sum_{j=1}^{K} x_k = n$$
$$\sum_{j=1}^{K} p_j = 1.$$

3.6 Poisson Distribution

A **Poisson** random variable has support on the non-negative integers.

$$\pi(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, \qquad x = 0, 1, 2...,$$
$$\lambda > 0.$$

The parameter λ indexes the mean and spread. In economics the Poisson distribution is often used for arrival times. Econometricians use it for count (integer-valued) data.

$$\mathbb{E}[X] = \lambda$$
$$\operatorname{var}[X] = \lambda.$$

The Poisson probability mass function is displayed in Figure 3.1(b) for the case $\lambda = 3$.

3.7 Negative Binomial Distribution

A limitation of the Poisson model for count data is that the single parameter λ controls both the mean and variance. An alternative is the **Negative Binomial**.

$$\pi\left(x \mid r, p\right) = \binom{x+r-1}{x} p^x \left(1-p\right)^r, \qquad x = 0, 1, 2...,$$
$$0 < p < 1$$
$$r > 0.$$

The distribution has two parameters, so the mean and variance are freely varying.

$$\mathbb{E}[X] = \frac{pr}{1-p}$$
$$\operatorname{var}[X] = \frac{pr}{\left(1-p\right)^2}.$$

3.8 Uniform Distribution

A **Uniform** random variable, typically written $U[a,b]$, has density

$$f\left(x \mid a, b\right) = \frac{1}{b-a}, \qquad a \leq x \leq b$$

$$\mathbb{E}[X] = \frac{b+a}{2}$$
$$\operatorname{var}[X] = \frac{(b-a)^2}{12}.$$

3.9 Exponential Distribution

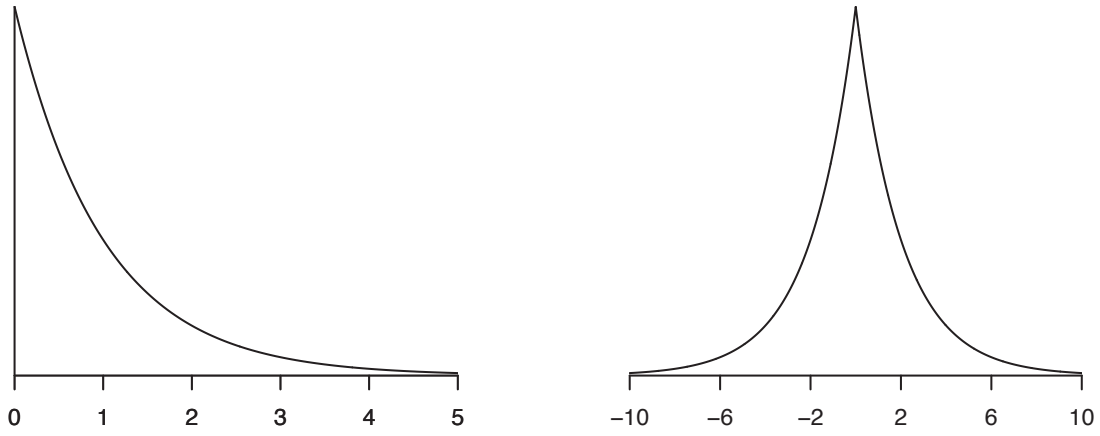
An **Exponential** random variable has density

$$f\left(x \mid \lambda\right) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right), \qquad x \geq 0$$
$$\lambda > 0.$$

The exponential is frequently used by economists in theoretical models due to its simplicity. It is not commonly used in econometrics.

$$\mathbb{E}[X] = \lambda$$
$$\operatorname{var}[X] = \lambda^2.$$

If $U \sim U[0,1]$ then $X = -\log U \sim$exponential(1) The exponential density function is displayed in Figure 3.2(b) for the case $\lambda = 1$.



(a) Exponential (b) Double Exponential

Figure 3.2: Exponential and Double Exponential Densities

3.10 Double Exponential Distribution

A **Double Exponential** or **Laplace** random variable has density

$$f(x \mid \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|x|}{\lambda}\right), \qquad x \in \mathbb{R}$$
$$\lambda > 0.$$

The double exponential is used in robust analysis.

$$\mathbb{E}[X] = 0$$
$$\operatorname{var}[X] = 2\lambda^2.$$

The double exponential density function is displayed in Figure 3.2(c) for the case $\lambda = 2$.

3.11 Generalized Exponential Distribution

A **Generalized Exponential (GED)** random variable has density

$$f(x \mid \lambda, r) = \frac{1}{2\Gamma(1/r)\lambda} \exp\left(-\left|\frac{x}{\lambda}\right|^r\right), \qquad x \in \mathbb{R}$$
$$\lambda > 0$$
$$r > 0$$

where $\Gamma(\alpha)$ is the gamma function (Definition A.20). The generalized exponential nests the double exponential and normal distributions.

3.12 Normal Distribution

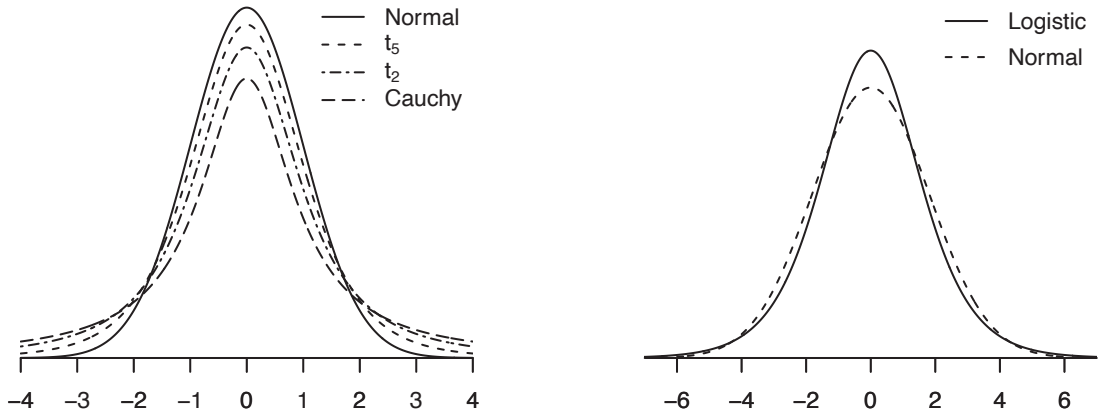
A **Normal** random variable, typically written $X \sim \mathrm{N}(\mu, \sigma^2)$, has density

$$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}$$
$$\mu \in \mathbb{R}$$
$$\sigma^2 > 0.$$

The normal distribution is the most commonly-used distribution in econometrics. When $\mu = 0$ and $\sigma^2 = 1$ it is called the **standard normal**. The standard normal density function is typically written as $\phi(x)$, and the distribution function as $\Phi(x)$. The normal density function can be written as $\phi_\sigma(x-\mu)$ where $\phi_\sigma(u) = \sigma^{-1}\phi(u/\sigma)$. μ is a location parameter and σ^2 is a scale parameter.

$$\mathbb{E}[X] = \mu$$
$$\mathrm{var}[X] = \sigma^2.$$

The standard normal density function is displayed in both panels of Figure 3.3.



(a) Cauchy, Student t, Normal

(b) Logistic and Normal

Figure 3.3: Normal, Cauchy, Student t, and Logistic Densities

3.13 Cauchy Distribution

A **Cauchy** random variable has density and distribution function

$$f(x) = \frac{1}{\pi\left(1+x^2\right)}, \qquad x \in \mathbb{R}$$
$$F(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}$$

The density is bell-shaped but with thicker tails than the normal. An interesting feature is that it has no finite integer moments.

A Cauchy density function in Figure 3.3(a).

3.14 Student t Distribution

A **Student t** random variable, typically written $X \sim t_r$ or $t(r)$, has density

$$f(x \mid r) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty, \tag{3.1}$$

where $\Gamma(\alpha)$ is the gamma function (Definition A.20). The parameter r is called the "degrees of freedom". The student t is used for critical values in the normal sampling model.

$$\begin{aligned} \mathbb{E}[X] &= 0 \qquad \text{if } r > 1 \\ \operatorname{var}[X] &= \frac{r}{r-2} \qquad \text{if } r > 2. \end{aligned}$$

The t distribution has the property that moments below r are finite. Moments greater than or equal to r are undefined.

The student t specializes to the Cauchy distribution when $r = 1$. As a limiting case as $r \to \infty$ it specializes to the normal (as shown in the next result). Thus the student t includes both the Cauchy and normal as limiting special cases.

Theorem 3.1 As $r \to \infty$, $f(x \mid r) \to \phi(x)$.

The proof is presented in Section 3.26.

A **scaled student t random variable** has density

$$f(x \mid r, \upsilon) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi\upsilon}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r\upsilon}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty,$$

where υ is a scale parameter. The variance of the scaled student t is $\upsilon r/(r-2)$.

Plots of the student t density function are displayed in Figure 3.3(a) for $r = 1$ (Cauchy), 2, 5, and ∞ (Normal). The density function of the student t is bell-shaped like the normal, but the t has thicker tails.

3.15 Logistic Distribution

A **Logistic** random variable has density and distribution function

$$\begin{aligned} F(x) &= \frac{1}{1+e^{-x}}, \qquad x \in \mathbb{R} \\ f(x) &= F(x)\left(1-F(x)\right). \end{aligned}$$

The density is bell-shaped with a strong resemblance to the normal. It is used frequently in econometrics as a substitute for the normal because the CDF is available in closed form.

$$\begin{aligned} \mathbb{E}[X] &= 0 \\ \operatorname{var}[X] &= \pi^2/3. \end{aligned}$$

If U_1 and U_1 are independent exponential(1) then $X = \log U_1 - \log U_2$ is Logistic. If $U \sim U[0,1]$ then $X = \log(U/(1-U))$ is Logistic.

A Logistic density function scaled to have unit variance is displayed in Figure 3.3(b). The standard normal density is plotted for contrast.

3.16 Chi-Square Distribution

A **Chi-Square** random variable, written $Q \sim \chi^2_r$ or $\chi^2(r)$, has density

$$f(x \mid r) = \frac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} \exp(-x/2), \qquad x \geq 0 \tag{3.2}$$
$$r > 0,$$

where $\Gamma(\alpha)$ is the gamma function (Definition A.20).

$$\mathbb{E}[X] = r$$
$$\operatorname{var}[X] = 2r.$$

The chi-square specializes to the exponential with $\lambda = 2$ when $r = 2$. The chi-square is commonly used for critical values for asymptotic tests.

It will be useful to derive the MGF of the chi-square distribution.

Theorem 3.2 The moment generating function of $Q \sim \chi^2_r$ is $M(t) = (1-2t)^{-r/2}$.

The proof is presented in Section 3.26.

An interesting calculation (see Exercise 3.8) reveals the inverse moment.

Theorem 3.3 If $Q \sim \chi^2_r$ with $r \geq 2$ then $\mathbb{E}\left[\frac{1}{Q}\right] = \frac{1}{r-2}$.

The chi-square density function is displayed in Figure 3.4(a) for the cases $r = 2$, 3, 4, and 6.

3.17 Gamma Distribution

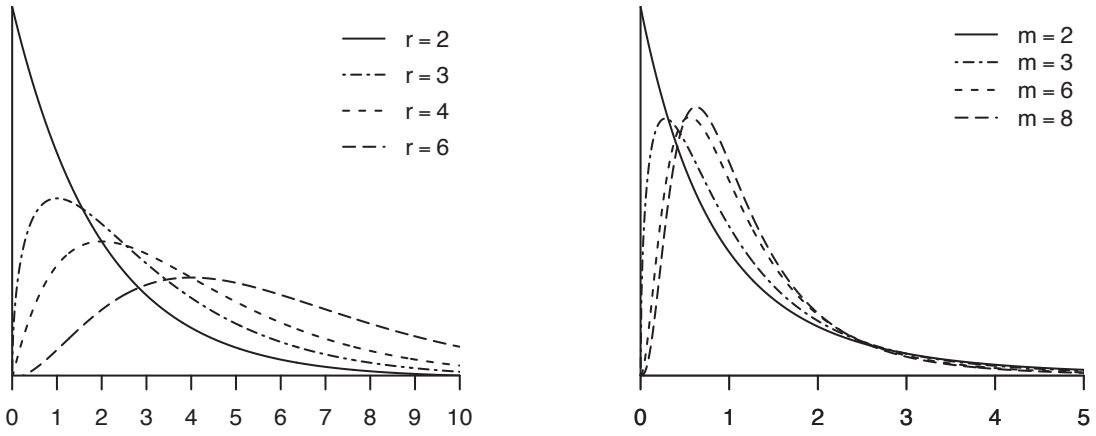
A **Gamma** random variable has density

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x\beta), \qquad x \geq 0$$
$$\alpha > 0$$
$$\beta > 0,$$

where $\Gamma(\alpha)$ is the gamma function (Definition A.20). The gamma distribution includes the chi-square as a special case when $\beta = 1/2$ and $\alpha = r/2$. That is, $\chi^2_r \sim \text{gamma}(r/2, 1/2)$ and $\text{gamma}(\alpha, \beta) \sim \chi^2_{2\alpha}/2\beta$. The gamma distribution when $\alpha = 1$ is exponential with $\lambda = 1/\beta$.

The gamma distribution is sometimes motivated as a flexible parametric family on the positive real line. α is a shape parameter while β is a scale parameter. It is also used in Bayesian analysis.

$$\mathbb{E}[X] = \frac{\alpha}{\beta}$$
$$\operatorname{var}[X] = \frac{\alpha}{\beta^2}.$$



(a) Chi-Square (b) F

Figure 3.4: Chi-Square and F Densities

3.18 F Distribution

An F random variable, typically written $X \sim F_{m,r}$ or $F(m,r)$, has density

$$f(x \mid m,r) = \frac{\left(\frac{m}{r}\right)^{m/2} x^{m/2-1} \Gamma\left(\frac{m+r}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)\left(1+\frac{m}{r}x\right)^{(m+r)/2}}, \qquad x>0, \tag{3.3}$$

where $\Gamma(\alpha)$ is the gamma function (Definition A.20). The F is used for critical values in the normal sampling model.

$$\mathbb{E}[X] = \frac{r}{r-2} \qquad \text{if } r>2.$$

As a limiting case, as $r \to \infty$ the F distribution simplifies to Q_m/m, a normalized χ^2_m. Thus the F distribution is a generalization of the χ^2_m distribution.

Theorem 3.4 Let $X \sim F_{m,r}$. As $r \to \infty$, the density of mX approaches that of χ^2_m.

The proof is presented in Section 3.26.

The F distribution was tabulated by a 1934 paper by Snedecor. He introduced the notation F as the distribution is related to Sir Ronald Fisher's work on the analysis of variance.

Plots of the $F_{m,r}$ density for $m = 2, 3, 6, 8$, and $r = 10$ are displayed in Figure 3.4(b).

3.19 Non-Central Chi-Square

A **Non-Central Chi-Square** random variable, typically written $X \sim \chi^2_r(\lambda)$ or $\chi^2(r,\lambda)$, has density

$$f(x) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!}\left(\frac{\lambda}{2}\right)^i f_{r+2i}(x), \qquad x>0 \tag{3.4}$$

where $f_r(x)$ is the χ^2_r density function (3.2). This is a weighted average of chi-square densities with Poisson weights. The non-central chi-square is used for theoretical analysis in the multivariate normal model and in asymptotic statistics. The parameter λ is called a **non-centrality parameter**.

The non-central chi-square includes the chi-square as a special case when $\lambda = 0$.

$$\begin{aligned}\mathbb{E}[X] &= r + \lambda \\ \operatorname{var}[X] &= 2(r + 2\lambda).\end{aligned}$$

Plots of the $\chi^2_r(\lambda)$ density for $r = 3$ and $\lambda = 0$, 2, 4, and 6 are displayed in Figure 3.4(c).

3.20 Beta Distribution

A **Beta** random variable has density

$$\begin{aligned}f(x \mid \alpha, \beta) &= \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \le x \le 1 \\ \alpha &> 0 \\ \beta &> 0\end{aligned}$$

where

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is the beta function, and $\Gamma(\alpha)$ is the gamma function (Definition A.20). The beta distribution is used as a flexible parametric family on $[0,1]$.

$$\begin{aligned}\mathbb{E}[X] &= \frac{\alpha}{\alpha+\beta} \\ \operatorname{var}[X] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.\end{aligned}$$

The beta density function is displayed in Figure 3.5(a) for the cases $(\alpha,\beta) = (2,2), (2,5)$ and $(5,1)$.

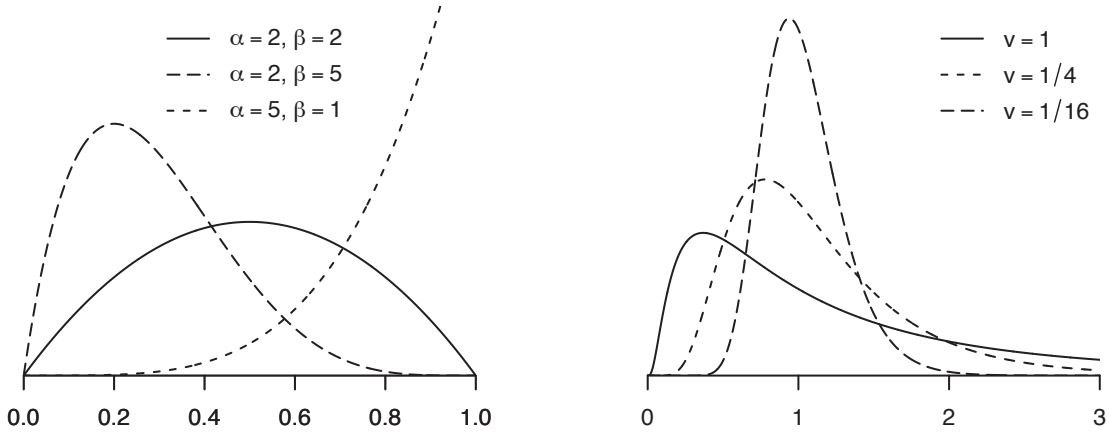
3.21 Pareto Distribution

A **Pareto** random variable has density

$$\begin{aligned}f(x \mid \alpha, \beta) &= \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \qquad x \ge \beta \\ \alpha &> 0, \\ \beta &> 0.\end{aligned}$$

It is used to model thick-tailed distributions. The parameter α controls the rate at which the tail of the density declines to zero.

$$\begin{aligned}\mathbb{E}[X] &= \frac{\alpha\beta}{\alpha-1} \qquad \text{if } \alpha > 1 \\ \operatorname{var}[X] &= \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} \qquad \text{if } \alpha > 2.\end{aligned}$$



(a) Beta (b) Lognormal

Figure 3.5: Beta and Lognormal Densities

3.22 Lognormal Distribution

A **Lognormal** random variable has density

$$f(x \mid \theta, \nu) = \frac{1}{\sqrt{2\pi\nu}} x^{-1} \exp\left(-\frac{(\log x - \theta)^2}{2\nu}\right), \qquad x > 0$$
$$\theta \in \mathbb{R}$$
$$\nu > 0.$$

The name comes from the fact that $\log(X) \sim \mathrm{N}(\theta, \nu)$. It is very common in applied econometrics to apply a normal model to variables after taking logarithms, which implicitly is applying a lognormal model to the levels. The lognormal distribution is highly skewed with a thick right tail.

$$\mathbb{E}[X] = \exp(\theta + \nu/2)$$
$$\operatorname{var}[X] = \exp(2\theta + 2\nu) - \exp(2\theta + \nu).$$

The lognormal density function with $\theta = 0$ and $\nu = 1$, $1/4$, and $1/16$ is displayed in Figure 3.5(b).

3.23 Weibull Distribution

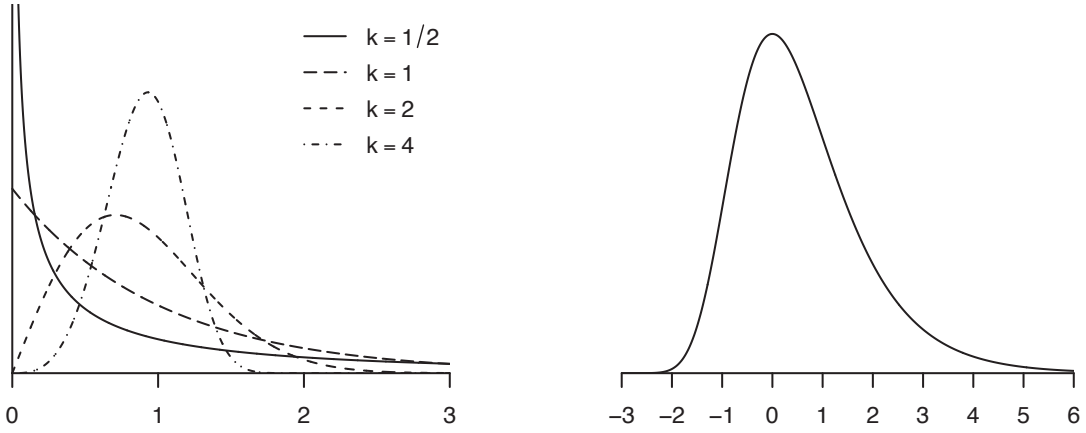
A **Weibull** random variable has density and distribution function

$$f(x \mid \alpha, \lambda) = \frac{\alpha}{\lambda}\left(\frac{x}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right), \qquad x \geq 0$$
$$F(x \mid \alpha, \lambda) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right)$$
$$\alpha > 0$$
$$\lambda > 0.$$

The Weibull distribution is used in survival analysis. The parameter α controls the shape and the parameter λ controls the scale.

$$\mathbb{E}[X] = \lambda\Gamma(1+1/\alpha)$$
$$\operatorname{var}[X] = \lambda^2\left(\Gamma(1+2/\alpha) - (\Gamma(1+1/\alpha))^2\right)$$

where $\Gamma(\alpha)$ is the gamma function (Definition A.20).



(a) Weibull

(b) Extreme Value

Figure 3.6: Weibull and Type I Extreme Value Densities

If $Y \sim$exponential(λ) then $X = Y^{1/\alpha} \sim$ Weibull($\alpha, \lambda^{1/\alpha}$).

The Weibull density function with $\lambda = 1$ and $k = 1/2$, 1, 2, and 4 is displayed in Figure 3.6(a).

3.24 Extreme Value Distribution

The **Type I Extreme Value** distribution (also known as the **Gumbel**) takes two forms. The density and distribution functions for the most common case are

$$f(x) = \exp(-x)\exp\left(-\exp(-x)\right), \qquad x \in \mathbb{R}$$
$$F(x) = \exp\left(-\exp(-x)\right).$$

The density and distriubtion functions for the alternative (minimum) case are

$$f(x) = \exp(x)\exp\left(-\exp(x)\right), \qquad x \in \mathbb{R}$$
$$F(x) = 1 - \exp\left(-\exp(x)\right).$$

The type I extreme value is used in discrete choice modeling.

If $Y \sim$exponential(1) then $X = -\log Y \sim$ Type I extreme value. If X_1 and X_2 are independent Type I extreme value, then $Y = X_1 - X_2 \sim$ Logistic.

The Type I extreme value density function is displayed in Figure 3.6(b).

3.25 Mixtures of Normals

A **Mixture of Normals** density function is

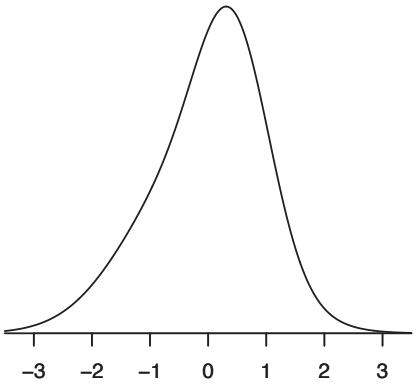
$$f\left(x \mid p_1, \mu_1, \sigma_1^2, ..., p_M, \mu_M, \sigma_M^2\right) = \sum_{m=1}^{M} p_m \phi_{\sigma_m}\left(x - \mu_m\right)$$

$$\sum_{m=1}^{M} p_m = 1.$$

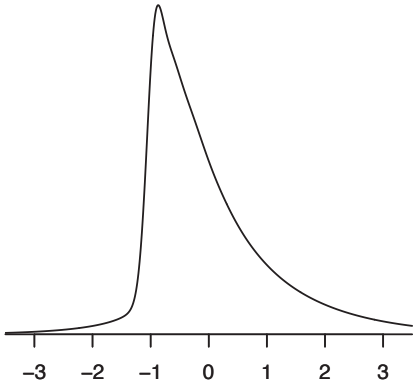
M is the number of mixture components. Mixtures of normals can be motivated by the idea of latent types. The latter means there are M latent types, each with a distinct mean and variance. Mixtures are frequently used in economics to model heterogeneity. Mixtures can also be used to flexibly approximate unknown density shapes. Mixtures of normals can also be convenient for certain theoretical calculations due to their simple structure.

To illustrate the flexibility which can be obtained by mixtures of normals, Figure 3.7 plots six examples of mixture normal density functions[1]. All are normalized to have mean zero and variance one. The labels are descriptive, not formal names.

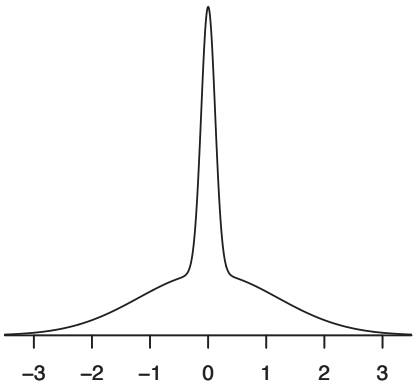
[1] These are constructed based on examples presented in Marron and Wand (1992), Figure 1 and Table 1.



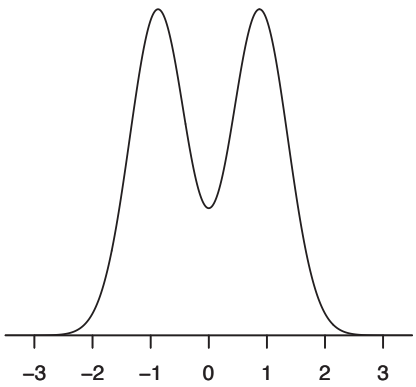
(a) Skewed



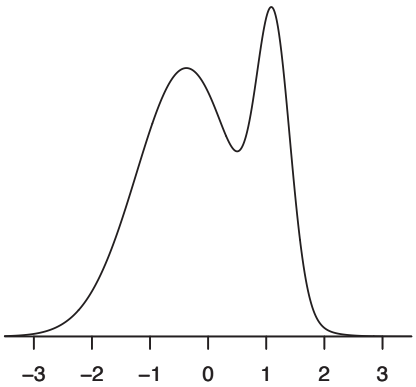
(b) Strongly Skewed



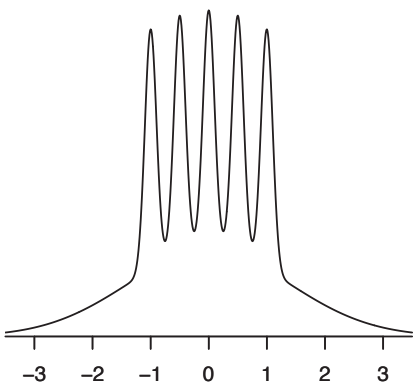
(c) Kurtotic



(d) Bimodal



(e) Skewed Bimodal



(f) Claw

Figure 3.7: Mixture of Normals Densities

3.26 Technical Proofs*

Proof of Theorem 3.1 Theorem A.28.6 states

$$\lim_{n\to\infty}\frac{\Gamma(n+x)}{\Gamma(n)\,n^x}=1.$$

Setting $n=r/2$ and $x=1/2$ we find

$$\lim_{r\to\infty}\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}=\frac{1}{\sqrt{2\pi}}.$$

Using the definition of the exponential function (see Appendix A.4)

$$\lim_{r\to\infty}\left(1+\frac{x^2}{r}\right)^r=\exp\left(x^2\right).$$

Taking the square root we obtain

$$\lim_{r\to\infty}\left(1+\frac{x^2}{r}\right)^{r/2}=\exp\left(\frac{x^2}{2}\right). \tag{3.5}$$

Furthermore,

$$\lim_{r\to\infty}\left(1+\frac{x^2}{r}\right)^{\frac{1}{2}}=1.$$

Together

$$\lim_{r\to\infty}\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)=\phi(x).$$

■

Proof of Theorem 3.2 The MGF of the density (3.2) is

$$\begin{aligned}\int_0^\infty \exp(tq)\,f(q)\,dq &= \int_0^\infty \exp(tq)\,\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp(-q/2)\,dq\\ &= \int_0^\infty \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp\left(-q\left(1/2-t\right)\right)dq\\ &= \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}\left(1/2-t\right)^{-r/2}\Gamma\left(\frac{r}{2}\right)\\ &= (1-2t)^{-r/2},\end{aligned} \tag{3.6}$$

the third equality using Theorem A.28.3. ■

Proof of Theorem 3.4 Applying change-of-variables to the density in Theorem 3.3, the density of mF is

$$\frac{x^{m/2-1}\Gamma\left(\frac{m+r}{2}\right)}{r^{m/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)\left(1+\frac{x}{r}\right)^{(m+r)/2}}. \tag{3.7}$$

Using Theorem A.28.6 with $n=r/2$ and $x=m/2$ we have

$$\lim_{r\to\infty}\frac{\Gamma\left(\frac{m+r}{2}\right)}{r^{m/2}\Gamma\left(\frac{r}{2}\right)}=2^{-m/2}$$

and similarly to (3.5) we have

$$\lim_{r\to\infty}\left(1+\frac{x}{r}\right)^{\left(\frac{m+r}{2}\right)}=\exp\left(\frac{x}{2}\right).$$

Together, (3.7) tends to

$$\frac{x^{m/2-1}\exp\left(-\frac{x}{2}\right)}{2^{m/2}\Gamma\left(\frac{m}{2}\right)}$$

which is the χ^2_m density. ■

3.27 Exercises

Exercise 3.1 For the Bernoulli distribution show

(a) $\sum_{x=0}^{1}\pi\left(x\mid p\right)=1.$

(b) $\mathbb{E}[X]=p.$

(c) $\operatorname{var}[X]=p(1-p).$

Exercise 3.2 For the Binomial distribution show

(a) $\sum_{x=0}^{n}\pi\left(x\mid n,p\right)=1.$
Hint: Use the Binomial Theorem.

(b) $\mathbb{E}[X]=np.$

(c) $\operatorname{var}[X]=np(1-p).$

Exercise 3.3 For the Poisson distribution show

(a) $\sum_{x=0}^{\infty}\pi\left(x\mid\lambda\right)=1.$

(b) $\mathbb{E}[X]=\lambda.$

(c) $\operatorname{var}[X]=\lambda.$

Exercise 3.4 For the $U[a,b]$ distribution show

(a) $\int_a^b f\left(x\mid a,b\right)dx=1.$

(b) $\mathbb{E}[X]=(b-a)/2$

(c) $\operatorname{var}[X]=(b-a)/12.$

Exercise 3.5 For the exponential distribution show

(a) $\int_0^\infty f\left(x\mid\lambda\right)dx=1.$

(b) $\mathbb{E}[X]=\lambda.$

(c) $\text{var}[X] = \lambda^2$.

Exercise 3.6 For the double exponential distribution show

(a) $\int_{-\infty}^{\infty} f(x \mid \lambda)\, dx = 1$.

(b) $\mathbb{E}[X] = 0$.

(c) $\text{var}[X] = 2\lambda^2$.

Exercise 3.7 For the chi-square density $f(x \mid r)$ show

(a) $\int_0^{\infty} f(x \mid r)\, dx = 1$.

(b) $\mathbb{E}[X] = r$.

(c) $\text{var}[X] = 2r$.

Exercise 3.8 Show Theorem 3.3. Hint: Show that $x^{-1} f(x \mid r) = \frac{1}{r-2} f(x \mid r-2)$.

Exercise 3.9 For the gamma distribution show

(a) $\int_0^{\infty} f\left(x \mid \alpha, \beta\right) dx = 1$.

(b) $\mathbb{E}[X] = \dfrac{\alpha}{\beta}$.

(c) $\text{var}[X] = \dfrac{\alpha}{\beta^2}$.

Exercise 3.10 Suppose $X \sim \text{gamma}(\alpha, \beta)$. Set $Y = \lambda X$. Find the density of Y. Which distribution is this?

Exercise 3.11 For the Pareto distribution show

(a) $\int_{\beta}^{\infty} f\left(x \mid \alpha, \beta\right) dx = 1$.

(b) $F\left(x \mid \alpha, \beta\right) = 1 - \dfrac{\beta^{\alpha}}{x^{\alpha}}$, $x \geq \beta$.

(c) $\mathbb{E}[X] = \dfrac{\alpha\beta}{\alpha - 1}$.

(d) $\text{var}[X] = \dfrac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$.

Exercise 3.12 For the logistic distribution show

(a) $F(x)$ is a valid distribution function.

(b) The density function is $f(x) = \exp(-x) / \left(1 + \exp(-x)\right)^2 = F(x)\left(1 - F(x)\right)$.

(c) The density $f(x)$ is symmetric about zero.

Exercise 3.13 For the lognormal distribution show

(a) The density is obtained by the transformation $X = \exp(Y)$ with $Y \sim \mathrm{N}(\theta, \nu)$.

(b) $\mathbb{E}[X] = \exp(\theta + \nu/2)$.

Exercise 3.14 For the mixture of normals distribution show

(a) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

(b) $F(x) = \sum_{m=1}^{M} p_m \Phi\left(\frac{x - \mu_m}{\sigma_m}\right)$.

(c) $\mathbb{E}[X] = \sum_{m=1}^{M} p_m \mu_m$.

(d) $\mathbb{E}\left[X^2\right] = \sum_{m=1}^{M} p_m \left(\sigma_m^2 + \mu_m^2\right)$.

Chapter 4

Multivariate Distributions

4.1 Introduction

In Chapter 2 we introduced the concept of random variables. We now generalize this concept to multiple random variables known as **random vectors**. To make the distinction clear we will refer to one-dimensional random variables as **univariate**, two-dimensional random pairs as **bivariate**, and vectors of arbitrary dimension as **multivariate**.

We start the chapter with bivariate random variables. Later sections generalize to multivariate random vectors.

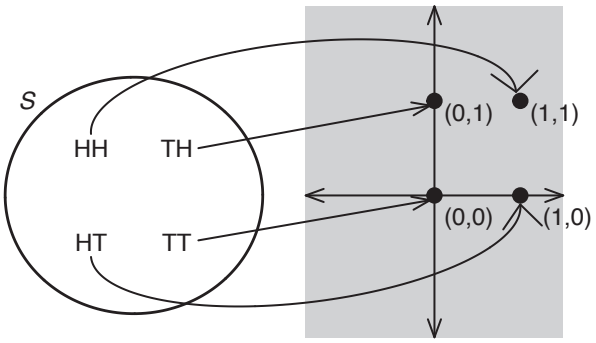


Figure 4.1: Two Coin Flip Sample Space

4.2 Bivariate Random Variables

A pair of bivariate random variables are two random variables with a joint distribution. They are typically represented by a pair of uppercase Latin characters such as (X, Y) or (X_1, X_2). Specific values will be written by a pair of lower case characters, e.g. (x, y) or (x_1, x_2).

Definition 4.1 A pair of **bivariate random variables** is a pair of numerical outcomes; a function from the sample space to $\mathbb{R}^2$.

To illustrate, Figure 4.1 illustrates a mapping from the two coin flip sample space to $\mathbb{R}^2$, with TT mapped to (0,0), TH mapped to (0,1), HT mapped to (1,0) and HH mapped to (1,1).

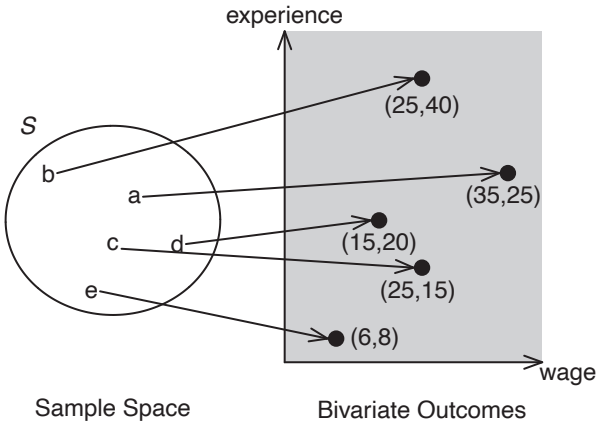


Figure 4.2: Bivariate Random Variables

For a real-world example consider the bivariate pair (wage, work experience). We are interested in how wages vary with experience, and therefore in their joint distribution. The mapping is illustrated by Figure 4.2. The ellipse is the sample space with random outcomes a, b, c, d, e. (You can think of outcomes as individual wage-earners at a point in time.) The graph is the positive orthant in $\mathbb{R}^2$ representing the bivariate pairs (wage, work experience). The arrows depict the mapping. Each outcome is a point in the sample space. Each outcome is mapped to a point in $\mathbb{R}^2$. The latter are a pair of random variables (wage and experience). Their values are marked on the plot, with wage measured in dollars per hour and experience in years.

4.3 Bivariate Distribution Functions

We now define the distribution function for bivariate random variables.

Definition 4.2 The **joint distribution function** of (X, Y) is $F(x, y) = \mathbb{P}\left[X \leq x, Y \leq y\right] = \mathbb{P}\left[\{X \leq x\} \cap \{Y \leq y\}\right]$.

We use "joint" to specifically indicate that this is the distribution of multiple random variables. For simplicity we omit the term "joint" when the meaning is clear from the context. When we want to be clear that the distribution refers to the pair (X, Y) we add subscripts, e.g. $F_{X,Y}(x, y)$. When the variables are clear from the context we omit the subscripts.

An example of a joint distribution function is $F(x, y) = (1 - e^{-x})(1 - e^{-y})$ for $x, y \geq 0$.

The properties of the joint distribution function are similar to the univariate case. The distribution function is weakly increasing in each argument and satisfies $0 \leq F(x, y) \leq 1$.

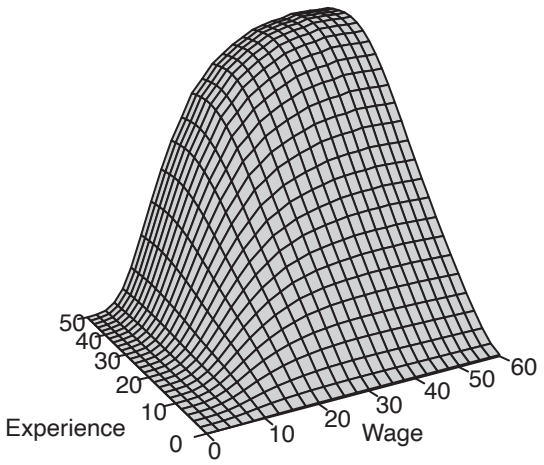


Figure 4.3: Bivariate Distribution of Experience and Wages

To illustrate with a real-world example, Figure 4.3 displays the bivariate joint distribution[1] of hourly wages and work experience. Wages are plotted from \$0 to \$60, and experience from 0 to 50 years. The joint distribution function increases from 0 at the origin to near one in the upper-right corner. The function is increasing in each argument. To interpret the plot, fix the value of one variable and trace out the curve with respect to the other variable. For example, fix experience at 30 and then trace out the plot with respect to wages. You see that the function steeply slopes up between \$14 and \$24 and then flattens

[1] Among wage earners in the United States in 2009.

out. Alternatively fix hourly wages at $30 and trace the function with respect to experience. In this case the function has a steady slope up to about 40 years and then flattens.

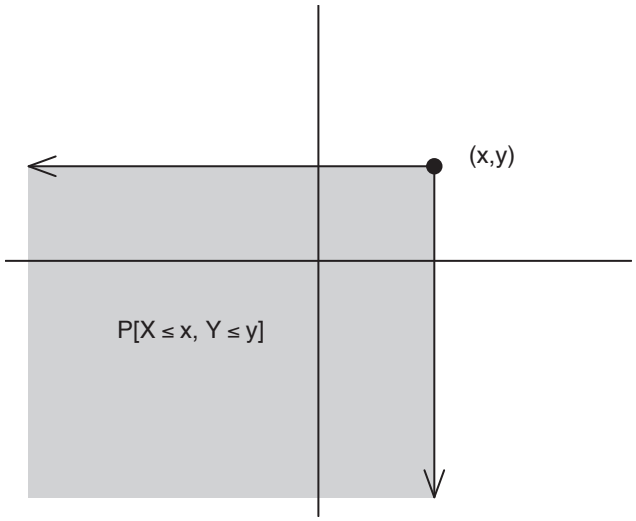


Figure 4.4: Bivariate Joint Distribution Calculation

Figure 4.4 illustrates how the joint distribution function is calculated for a given (x, y). The event $\{X \leq x, Y \leq y\}$ occurs if the pair (X, Y) lies in the shaded region (the region to the lower-left of the point (x, y)). The distribution function is the probability of this event. In our empirical example, if $(x, y) = (30, 30)$ then the calculation is the joint probability that wages are less than or equal to $30 and experience is less than or equal to 30 years. It is difficult to read this number from the plot in Figure 4.3 but it equals 0.58. This means that 58% of wage earners satisfy these conditions.

The distribution function satisfies the following relationship

$$\mathbb{P}[a < X \leq b, c < Y \leq d] = F(b,d) - F(b,c) - F(a,d) + F(a,c).$$

See Exercise 4.5. This is illustrated in Figure 4.5.

The shaded region is the set $\{a < x \leq b, c < y \leq d\}$. The probability that (X, Y) is in the set is the joint probability that X is in $(a, b]$ and Y is in $(c, d]$, and can be calculated from the distribution function evaluated at the four corners. For example

$$\begin{aligned}\mathbb{P}\left[10 < \text{wage} \leq 20, 10 < \text{experience} \leq 20\right] &= F(20,20) - F(20,10) - F(10,20) + F(10,10) \\ &= 0.265 - 0.131 - 0.073 + 0.042 \\ &= 0.103.\end{aligned}$$

Thus about 10% of wage earners satisfy these conditions.

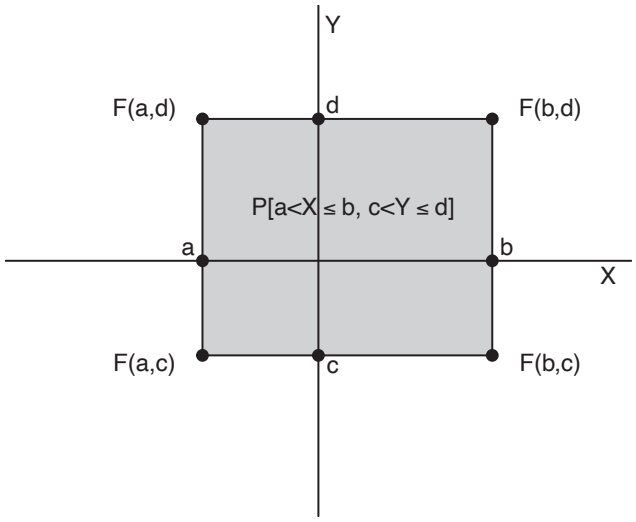


Figure 4.5: Probability and Distribution Functions

4.4 Probability Mass Function

As for univariate random variables it is useful to consider separately the case of discrete and continuous bivariate random variables.

A pair of random variables is **discrete** if there is a discrete set $\mathscr{S} \subset \mathbb{R}^2$ such that $\mathbb{P}[(X,Y) \in \mathscr{S}] = 1$. The set $\mathscr{S}$ is the support of (X,Y) and consists of a set of points in $\mathbb{R}^2$. In many cases the support takes the product form, meaning that the support can be written as $\mathscr{S} = \mathscr{X} \times \mathscr{Y}$ where $\mathscr{X} \subset \mathbb{R}$ and $\mathscr{Y} \subset \mathbb{R}$ are the supports for X and Y. We can write these support points as $\{\tau_1^x, \tau_2^x, ...\}$ and $\{\tau_1^y, \tau_2^y, ...\}$. The **joint probability mass function** is $\pi(x,y) = \mathbb{P}\left[X = x, Y = y\right]$. At the support points we set $\pi_{ij} = \pi(\tau_i^x, \tau_j^y)$. There is no loss in generality in assuming the support takes a product form if we allow $\pi_{ij} = 0$ for some pairs.

Example 1: A pizza restaurant caters to students. Each customer purchases either one or two slices of pizza and either one or two drinks during their meal. Let X be the number of pizza slices purchased, and Y be the number of drinks. The joint probability mass function is

$$\pi_{11} = \mathbb{P}[X = 1, Y = 1] = 0.4$$
$$\pi_{12} = \mathbb{P}[X = 1, Y = 2] = 0.1$$
$$\pi_{21} = \mathbb{P}[X = 2, Y = 1] = 0.2$$
$$\pi_{22} = \mathbb{P}[X = 2, Y = 2] = 0.3.$$

This is a valid probability function since all probabilities are non-negative and the four probabilities sum to one.

4.5 Probability Density Function

The pair (X, Y) has a **continuous** distribution if the joint distribution function $F(x, y)$ is continuous in (x, y).

In the univariate case the probability density function is the derivative of the distribution function. In the bivariate case it is a double partial derivative.

Definition 4.3 When $F(x, y)$ is continuous and differentiable its **joint density** $f(x, y)$ equals

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

When we want to be clear that the density refers to the pair (X, Y) we add subscripts, e.g. $f_{X,Y}(x, y)$. Joint densities have similar properties to the univariate case. They are non-negative functions and integrate to one over $\mathbb{R}^2$.

Example 2: (X, Y) are continuously distributed on $\mathbb{R}^2_+$ with joint density

$$f(x, y) = \frac{1}{4}\left(x + y\right) x y \exp\left(-x - y\right).$$

The joint density is displayed in Figure 4.6. We now verify that this is a valid density by checking that it integrates to one. The integral is

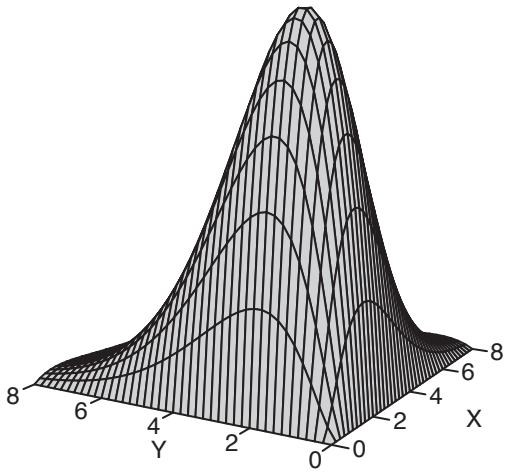


Figure 4.6: Joint Density

$$\begin{aligned}\int_0^\infty \int_0^\infty f(x,y)\,dx\,dy &= \int_0^\infty \int_0^\infty \frac{1}{4}(x+y)\exp(-x-y)\,dx\,dy \\ &= \frac{1}{4}\left(\int_0^\infty \int_0^\infty x^2 y \exp(-x-y)\,dx\,dy + \int_0^\infty \int_0^\infty x y^2 \exp(-x-y)\,dx\,dy\right) \\ &= \frac{1}{4}\left(\int_0^\infty y\exp(-y)\,dy \int_0^\infty x^2\exp(-x)\,dx + \int_0^\infty y^2\exp(-y)\,dy \int_0^\infty x\exp(-x)\,dx\right) \\ &= 1.\end{aligned}$$

Thus $f(x,y)$ is a valid density.

The probability interpretation of a bivariate density is that the probability that the random pair (X,Y) lies in a region in $\mathbb{R}^2$ equals the area under the density over this region. To see this, by the Fundamental Theorem of Calculus (Theorem A.20)

$$\int_c^d \int_a^b f(x,y)\,dx\,dy = \int_c^d \int_a^b \frac{\partial^2}{\partial x \partial y} F(x,y)\,dx\,dy = \mathbb{P}\left[a \le X \le b, c \le Y \le d\right].$$

This is similar to the property of univariate density functions, but in the bivariate case this requires two-dimensional integration. This means that for any $A \subset \mathbb{R}^2$

$$\mathbb{P}\left[(X,Y) \in A\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{1}\left\{(x,y) \in A\right\} f(x,y)\,dx\,dy.$$

In particular, this implies

$$\mathbb{P}\left[a \le X \le b, c \le Y \le d\right] = \int_c^d \int_a^b f(x,y)\,dx\,dy.$$

This is the joint probability that X and Y jointly lie in the intervals $[a,b]$ and $[c,d]$. Take a look at Figure 4.5 which shows this region in the (x,y) plane. The above expression shows that the probability that (X,Y) lie in this region is the integral of the joint density $f(x,y)$ over this region.

As an example, take the density $f(x,y) = 1$ for $0 \le x, y \le 1$. We calculate the probability that $X \le 1/2$ and $Y \le 1/2$. It is

$$\begin{aligned}\mathbb{P}\left[X \le 1/2, Y \le 1/2\right] &= \int_0^{1/2}\int_0^{1/2} f(x,y)\,dx\,dy \\ &= \int_0^{1/2} dx \int_0^{1/2} dy \\ &= \frac{1}{4}.\end{aligned}$$

Example 3: (Wages & Experience). Figure 4.7 displays the bivariate joint probability density of hourly wages and work experience corresponding to the joint distribution from Figure 4.4. Wages are plotted from \$0 to \$60 and experience from 0 to 50 years. Reading bivariate density plots takes some practice. To start, pick some experience level, say 10 years, and trace out the shape of the density function in wages. You see that the density is bell-shaped with its peak around \$15. The shape is similar to the univariate plot for wages from Figure 2.7.

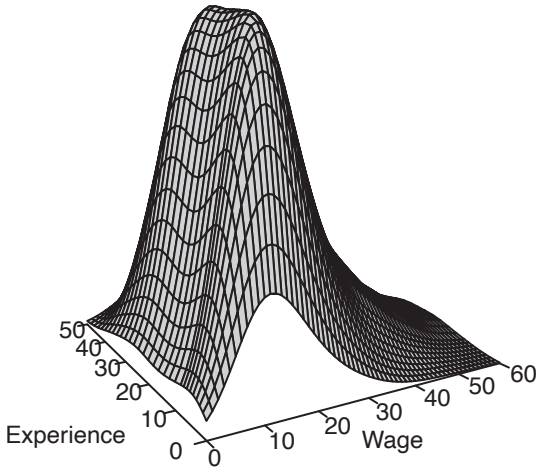


Figure 4.7: Bivariate Density of Experience and Log Wages

4.6 Marginal Distribution

The joint distribution of the random vector (X, Y) fully describes the distribution of each component of the random vector.

Definition 4.4 The **marginal distribution** of X is

$$F_X(x) = \mathbb{P}[X \le x] = \mathbb{P}[X \le x, Y \le \infty] = \lim_{y \to \infty} F(x, y).$$

In the continuous case we can write this as

$$F_X(x) = \lim_{y \to \infty} \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f(u, v)\, du\, dv.$$

The marginal density of X is the derivative of the marginal distribution, and equals

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{-\infty}^{\infty} \int_{-\infty}^{x} f(u, v)\, du\, dv = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

Similarly, the marginal PDF of Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

These marginal PDFs are obtained by "integrating out" the other variable.

Marginal CDFs (PDFs) are simply CDFs (PDFs), but are referred to as "marginal" to distinguish from the joint CDFs (PDFs) of the random vector. In practice we treat a marginal PDF the same as a PDF.

Definition 4.5 The **marginal densities** of X and Y given a joint density $f(x,y)$ are

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Example 1: (continued). The marginal probabilities are

$$\begin{aligned}\mathbb{P}[X=1] &= \mathbb{P}[X=1,Y=1] + \mathbb{P}[X=1,Y=2] = 0.4+0.1 = 0.5\\ \mathbb{P}[X=2] &= \mathbb{P}[X=2,Y=1] + \mathbb{P}[X=2,Y=2] = 0.2+0.3 = 0.5\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}[Y=1] &= \mathbb{P}[X=1,Y=1] + \mathbb{P}[X=2,Y=1] = 0.4+0.2 = 0.6\\ \mathbb{P}[Y=2] &= \mathbb{P}[X=1,Y=2] + \mathbb{P}[X=2,Y=2] = 0.1+0.3 = 0.4.\end{aligned}$$

Thus 50% of the customers order one slice of pizza and 50% order two slices. 60% also order one drink, while 40% order two drinks.

Example 2: (continued). The marginal density of X is

$$\begin{aligned}f_X(x) &= \int_0^{\infty} \frac{1}{4}(x+y)\,xy\exp(-x-y)\,dy\\ &= \left(x^2\int_0^{\infty} y\exp(-y)\,dy + x\int_0^{\infty} y^2\exp(-y)\,dy\right)\frac{1}{4}\exp(-x)\\ &= \frac{x^2+2x}{4}\exp(-x).\end{aligned}$$

for $x \geq 0$.

Example 3: (continued). The marginal density of wages was displayed in Figure 2.8, and can be found from Figure 4.7 by integrating over experience. The marginal density for experience can be similarly found by integrating over wages.

4.7 Bivariate Expectation

Definition 4.6 The **expectation** of real-valued $g(X,Y)$ is

$$\mathbb{E}\left[g(X,Y)\right] = \sum_{(x,y)\in\mathbb{R}^2:\pi(x,y)>0} g(x,y)\pi(x,y)$$

for the discrete case, and

$$\mathbb{E}\left[g(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$

for the continuous case.

If $g(X)$ only depends on one of the variables the expectation can be written in terms of the marginal density

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)f(x,y)dxdy = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

In particular the expected value of a variable is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dxdy = \int_{-\infty}^{\infty} xf_X(x)dx.$$

Example 1: (continued). We calculate that

$$\begin{aligned}\mathbb{E}[X] &= 1\times 0.5 + 2\times 0.5 = 1.5\\ \mathbb{E}[Y] &= 1\times 0.6 + 2\times 0.4 = 1.4.\end{aligned}$$

This is the average number of pizza slices and drinks purchased per customer. The second moments are

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= 1^2\times 0.5 + 2^2\times 0.5 = 2.5\\ \mathbb{E}\left[Y^2\right] &= 1^2\times 0.6 + 2^2\times 0.4 = 2.2.\end{aligned}$$

The variances are

$$\begin{aligned}\text{var}[X] &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = 2.5 - 1.5^2 = 0.25\\ \text{var}[Y] &= \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 = 2.2 - 1.4^2 = 0.24.\end{aligned}$$

Example 2: (continued). The expected value of X is

$$\begin{aligned}\mathbb{E}[X] &= \int_0^{\infty} xf_X(x)dx\\ &= \int_0^{\infty} x\left(\frac{x^2+2x}{4}\right)\exp(-x)\,dx\\ &= \int_0^{\infty}\frac{x^3}{4}\exp(-x)dx + \int_0^{\infty}\frac{x^2}{2}\exp(-x)dx\\ &= \frac{5}{2}.\end{aligned}$$

The second moment is

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \int_0^{\infty} x^2f_X(x)dx\\ &= \int_0^{\infty} x^2\left(\frac{x^2+2x}{4}\right)\exp(-x)\,dx\\ &= \int_0^{\infty}\frac{x^4}{4}\exp(-x)dx + \int_0^{\infty}\frac{x^3}{2}\exp(-x)dx\\ &= 9.\end{aligned}$$

Its variance is

$$\text{var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = 9 - \left(\frac{5}{2}\right)^2 = \frac{11}{4}.$$

Example 3 & 4: (continued). The means and variances of wages, experience, and experience are presented in the following chart.

Means and Variances

	Mean	Variance
Hourly Wage	23.90	20.7^2
Experience (Years)	22.21	11.7^2
Education (Years)	13.98	2.58^2

4.8 Conditional Distribution for Discrete X

In this and the following section we define the conditional distribution and density of a random variable Y conditional that another random variable X takes a specific value x. In this section we consider the case where X has a discrete distribution and in the next section take the case where X has a continuous distribution.

Definition 4.7 If X has a discrete distribution the **conditional distribution function** of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \mathbb{P}\left[Y \leq y \mid X = x\right]$$

for any x such that $\mathbb{P}[X = x] > 0$.

This is a valid distribution as a function of y. That is, it is weakly increasing in y and asymptotes to 0 and 1. You can think of $F_{Y|X}(y \mid x)$ as the distribution function for the sub-population where $X = x$. Take the case where Y is hourly wages and X is a worker's gender. Then $F_{Y|X}(y \mid x)$ specifies the distribution of wages separately for the sub-populations of men and women. If X denotes years of education (measured discretely) then $F_{Y|X}(y \mid x)$ specifies the distribution of wages separately for each education level.

Example 4: Wages & Education. In Figure 2.2(b) we displayed the distribution of education for the population of U.S. wage earners using ten categories. Each category is a sub-population of the entirety of wage earners, and for each of these sub-populations there is a distribution of wages. These conditional distribution functions $F(y \mid x)$ are displayed in Figure 4.8 for four groups, education (x) equalling 12, 16, 18, and 20, which correspond to high school degree, college degree, master's degree, and professional/PhD degree. The difference between the distribution functions is large and striking. The distributions shift uniformly to the right with each increase in education level. The largest shifts are between the distributions of those with high school and college degrees, and between those with master's and professional degrees.

If Y is continuously distributed we define the conditional density as the derivative of the conditional distribution function.

Definition 4.8 If $F_{Y|X}(y \mid x)$ is differentiable with respect to y and $\mathbb{P}[X = x] > 0$ then the **conditional density function** of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{\partial}{\partial y} F_{Y|X}(y \mid x).$$

The conditional density $f_{Y|X}(y \mid x)$ is a valid density function since it is the derivative of a distribution function. You can think of it as the density of Y for the sub-population with a given value of X.

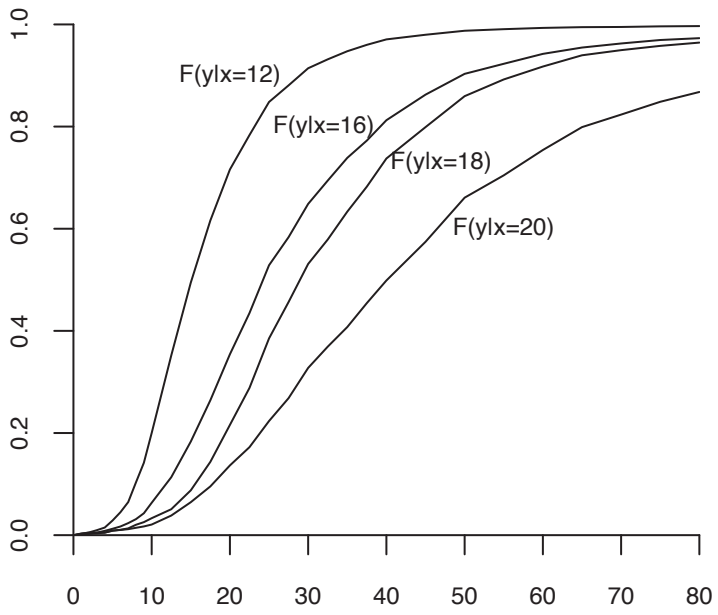


Figure 4.8: Conditional Distribution of Hourly Wages Given Education

Example 4: (continued). The conditional density functions $f(y \mid x)$ corresponding to the conditional distributions displayed in Figure 4.8 are displayed in Figure 4.9. By examining the density function it is easier to see where the probability mass is distributed. Compare the conditional densities for those with a high school ($x = 12$) and college ($x = 16$) degree. The latter density is shifted to the right and is more spread out. Thus college graduates have higher average wages but they are also more dispersed. While the conditional density for college graduates is substantially shifted to the right there is a considerable area of overlap between the density functions. Next compare the conditional densities of the college graduates and those with master's degrees ($x = 18$). The latter density is shifted to the right, but rather modestly. Thus these two densities are more similar than dissimilar. Now compare these conditional densities with the final density, that for the highest education level ($x = 20$). This conditional density function is substantially shifted to the right and substantially more dispersed.

4.9 Conditional Distribution for Continuous X

The conditional density for continuous random variables is defined as follows.

Definition 4.9 For continuous X and Y the **conditional density** of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

for any x such that $f_X(x) > 0$.

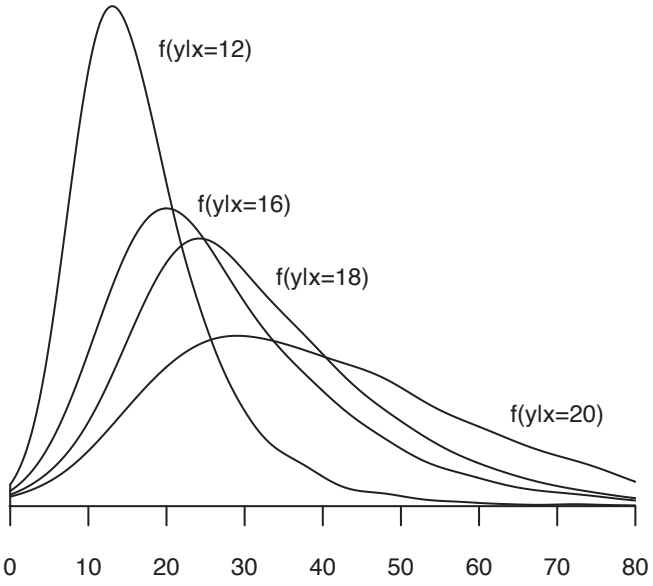


Figure 4.9: Conditional Density of Hourly Wages Given Education

If you are satisfied with this definition you can skip the remainder of this section. However, if you would like a justification, read on.

Recall that the definition of the conditional distribution function for the case of discrete X is

$$F_{Y|X}(y \mid x) = \mathbb{P}\left[Y \leq y \mid X = x\right].$$

This does not apply for the case of continuous X because $\mathbb{P}[X = x] = 0$. Instead we can define the conditional distribution function as a limit. Thus we propose the following definition.

Definition 4.10 For continuous X and Y the **conditional distribution** of Y given $X = x$ is

$$F_{Y|X}(y \mid x) = \lim_{\epsilon \downarrow 0} \mathbb{P}\left[Y \leq y \mid x - \epsilon \leq X \leq x + \epsilon\right].$$

This is the probability that Y is smaller than y, conditional on X being in an arbitrarily small neighborhood of x. This is essentially the same concept as the definition for the discrete case. Fortunately the expression can be simplified.

Theorem 4.1 If $F(x, y)$ is differentiable with respect to x and $f_X(x) > 0$ then $F_{Y|X}(y \mid x) = \dfrac{\frac{\partial}{\partial x} F(x, y)}{f_X(x)}$.

This result shows that the conditional distribution function is the ratio of a partial derivative of the joint distribution to the marginal density of X.

To prove this theorem we use the definition of conditional probability and L'Hôpital's rule (Theorem A.12), which states that the ratio of two limits which each tend to zero equals the ratio of the two derivatives. We find that

$$\begin{aligned}
F_{Y|X}\left(y \mid x\right) &= \lim_{\epsilon \downarrow 0} \mathbb{P}\left[Y \leq y \mid x-\epsilon \leq X \leq x+\epsilon\right] \\
&= \lim_{\epsilon \downarrow 0} \frac{\mathbb{P}\left[Y \leq y, x-\epsilon \leq X \leq x+\epsilon\right]}{\mathbb{P}\left[x-\epsilon \leq X \leq x+\epsilon\right]} \\
&= \lim_{\epsilon \downarrow 0} \frac{F(x+\epsilon, y)-F(x-\epsilon, y)}{F_X(x+\epsilon)-F_X(x-\epsilon)} \\
&= \lim_{\epsilon \downarrow 0} \frac{\frac{\partial}{\partial x}F(x+\epsilon, y)+\frac{\partial}{\partial x}F(x-\epsilon, y)}{\frac{\partial}{\partial x}F_X(x+\epsilon)+\frac{\partial}{\partial x}F_X(x-\epsilon)} \\
&= \frac{2\frac{\partial}{\partial x}F(x, y)}{2\frac{\partial}{\partial x}F_X(x)} \\
&= \frac{\frac{\partial}{\partial x}F(x, y)}{f_X(x)}.
\end{aligned}$$

This proves the result.

To find the conditional density we take the partial derivative of the conditional distribution function:

$$\begin{aligned}
f_{Y|X}\left(y \mid x\right) &= \frac{\partial}{\partial y}F_{Y|X}(y \mid x) \\
&= \frac{\partial}{\partial y}\frac{\frac{\partial}{\partial x}F(x, y)}{f_X(x)} \\
&= \frac{\frac{\partial^2}{\partial y \partial x}F(x, y)}{f_X(x)} \\
&= \frac{f(x, y)}{f_X(x)}.
\end{aligned}$$

This is identical to the definition given at the beginning of this section.

What we have shown is that this definition (the ratio of the joint density to the marginal density) is the natural generalization of the case of discrete X.

4.10 Visualizing Conditional Densities

To visualize the conditional density $f_{Y|X}\left(y \mid x\right)$, start with the joint density $f(x, y)$, which is a 2-D surface in 3-D space. Fix x. Slice through the joint density along y. This creates an unnormalized density in one dimension. It is unnormalized because it does not integrate to one. To normalize, divide by $f_X(x)$. To verify that the conditional density integrates to one and is thus a valid density, observe that

$$\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\,dy = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_X(x)}\,dy = \frac{f_X(x)}{f_X(x)} = 1.$$

Example 2: (continued). The joint density was displayed in Figure 4.6. To visualize the conditional density select a value of x. Trace the shape of the joint density as a function of y. After re-normalization this is the conditional density function. As you vary x you obtain different conditional density functions.

To explicitly calculate the conditional density we devide the joint density by the marginal density. We find

$$\begin{aligned} f_{Y|X}(y \mid x) &= \frac{f(x,y)}{f_X(x)} \\ &= \frac{\frac{1}{4}\left(x+y\right)xy\exp\left(-x-y\right)}{\frac{1}{4}\left(x^2+2x\right)\exp\left(-x\right)} \\ &= \frac{\left(xy+y^2\right)\exp\left(-y\right)}{x+2}. \end{aligned}$$

Example 3: (continued). The conditional density of wages given experience is calculated by taking the joint density in Figure 4.7 and dividing by the marginal density of experience. We do so at three levels of experience (0 years, 8 years, and 30 years). The resulting densities are displayed in Figure 4.10. The shapes of the three conditional densities are similar, but they shift to the right and spread out as experience increases. This means that the distribution of wages shifts upwards and widens as work experience increases.

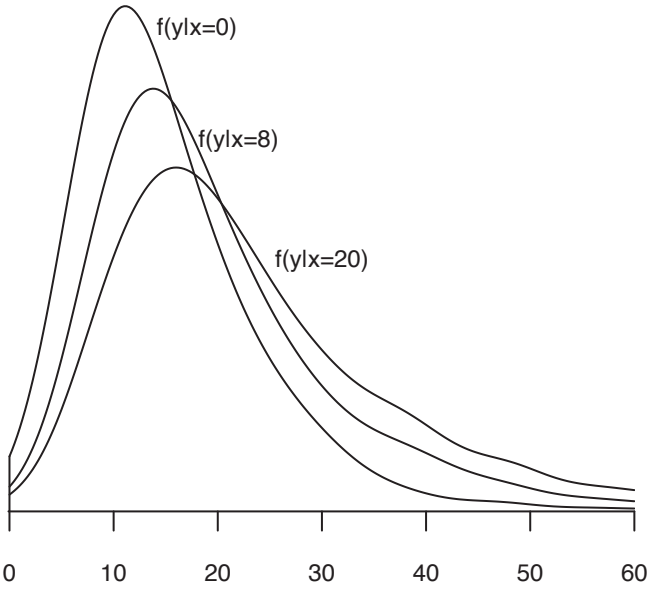


Figure 4.10: Conditional Density of Wages Given Experience

If we compare the conditional densities in Figure 4.10 with those in Figure 4.9 we can see that the effect of education on wages is considerably stronger than the effect of experience.

4.11 Independence

In this section we define independence between random variables.

Recall that two events A and B are independent if the probability that they both occur equals the product of their probabilities, thus $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$. Consider the events $A = \{X \le x\}$ and $B = \{Y \le y\}$. The probability that they both occur is

$$\mathbb{P}[A \cap B] = \mathbb{P}\left[X \le x, Y \le y\right] = F(x, y).$$

The product of their probabilities is

$$\mathbb{P}[A]\mathbb{P}[B] = \mathbb{P}[X \le x]\mathbb{P}\left[Y \le y\right] = F_X(x)F_Y(y).$$

These two expressions equal if $F(x, y) = F_X(x)F_Y(y)$. This means that if the events A and B are independent then the joint distribution function factors as $F(x, y) = F_X(x)F_Y(y)$ for all (x, y). This can be used as a definition of independence between random variables.

Definition 4.11 The random variables X and Y are **statistically independent** if for all x & y

$$F(x, y) = F_X(x)F_Y(y).$$

This is often written as $X \perp\!\!\!\perp Y$.

An implication of statistical independence is that all events of the form $A = \{X \in C_1\}$ and $B = \{Y \in C_2\}$ are independent.

If X and Y fail to satisfy the property of independence we say that they are **statistically dependent**.

It is more convenient to work with mass functions and densities rather than distributions. In the case of continuous random variables, by differentiating the above expression with respect to x and y, we find $f(x, y) = f_X(x)f_Y(y)$. Thus the definition is equivalent to stating that the joint density function factors into the product of the marginal densities.

In the case of discrete random variables a similar argument leads to $\pi(x, y) = \pi_X(x)\pi_Y(y)$, which means that the joint probability mass function factors into the product of the marginal mass functions.

Theorem 4.2 The discrete random variables X and Y are statistically independent if for all x & y

$$\pi(x, y) = \pi_X(x)\pi_Y(y).$$

If X and Y have a differentiable distribution function they are statistically independent if for all x & y

$$f(x, y) = f_X(x)f_Y(y).$$

An interesting connection arises between the conditional density function and independence.

Theorem 4.3 If X and Y are independent and continuously distributed,

$$f_{Y|X}(y \mid x) = f_Y(y)$$
$$f_{X|Y}(x \mid y) = f_X(x).$$

Thus the conditional density equals the marginal (unconditional) density. This means that $X = x$ does not affect the shape of the density of Y. This seems reasonable given that the two random variables are independent. To see this

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

We now present a sequence of related results.

First, by rewriting the definition of the conditional density we obtain a density version of Bayes Theorem.

Theorem 4.4 Bayes Theorem for Densities.

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy}.$$

Our next result shows that the expectation of the product of independent random variables is the product of the expectations.

Theorem 4.5 If X and Y are independent then for any functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}\left|g(X)\right| < \infty$ and $\mathbb{E}\left|h(Y)\right| < \infty$

$$\mathbb{E}\left[g(X)h(Y)\right] = \mathbb{E}\left[g(X)\right]\mathbb{E}\left[h(Y)\right].$$

Proof: We give the proof for continuous random variables.

$$\begin{aligned}
\mathbb{E}\left[g(X)h(Y)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y) f(x,y) dx dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y) f_X(x) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} g(x) f_X(x) dx \int_{-\infty}^{\infty} h(y) f_Y(y) dy \\
&= \mathbb{E}\left[g(X)\right]\mathbb{E}\left[h(Y)\right].
\end{aligned}$$

■

Take $g(x) = \mathbb{1}\{x \le a\}$ and $h(y) = \mathbb{1}\{y \le b\}$ for arbitrary constants a and b. Theorem 4.5 implies that independence implies $\mathbb{P}[X \le a, Y \le b] = \mathbb{P}[X \le a]\mathbb{P}[Y \le b]$, or

$$F(x,y) = F_X(x) F_X(y)$$

for all $(x,y) \in \mathbb{R}^2$. Recall that the latter is the definition of independence. Consequently the definition is "if and only if". That is, X and Y are independent if and only if this equality holds.

A useful application of Theorem 4.5 is to the moment generating function.

Theorem 4.6 If X and Y are independent with MGFs $M_X(t)$ and $M_Y(t)$, then the MGF of $Z = X + Y$ is $M_Z(t) = M_X(t) M_Y(t)$.

Proof: By the properties of the exponential function $\exp(t(X+Y)) = \exp(tX)\exp(tY)$. By Theorem 4.5, since X and Y are independent

$$\begin{aligned}
M_Z(t) &= \mathbb{E}\left[\exp(t(X+Y))\right] \\
&= \mathbb{E}\left[\exp(tX)\exp(tY)\right] \\
&= \mathbb{E}\left[\exp(tX)\right]\mathbb{E}\left[\exp(tY)\right] \\
&= M_X(t) M_Y(t).
\end{aligned}$$

■

Furthermore, transformations of independent variables are also independent.

Theorem 4.7 If X and Y are independent then for any functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$, $U = g(X)$ and $V = h(Y)$ are independent.

Proof: For any $u \in \mathbb{R}$ and $v \in \mathbb{R}$ define the sets $A(u) = \{x : g(x) \le u\}$ and $B(v) = \{y : h(y) \le v\}$. The joint distribution of (U, V) is

$$\begin{aligned} F_{U,V}(u, v) &= \mathbb{P}[U \le u, V \le v] \\ &= \mathbb{P}\left[g(X) \le u, h(Y) \le v\right] \\ &= \mathbb{P}[X \in A(u), Y \in B(v)] \\ &= \mathbb{P}[X \in A(u)]\,\mathbb{P}[Y \in B(v)] \\ &= \mathbb{P}\left[g(X) \le u\right]\mathbb{P}[h(Y) \le v] \\ &= \mathbb{P}[U \le u]\,\mathbb{P}[V \le v] \\ &= F_U(u) F_V(v). \end{aligned}$$

Thus the distribution function factors, satisfying the definition of independence. The key is the fourth equality, which uses the fact that the events $\{X \in A(u)\}$ and $\{Y \in B(v)\}$ are independent, which can be shown by an extension of Theorem 4.5. ■

Example 1: (continued). Are the number of pizza slices and drinks independent or dependent? To answer this we can check if the joint probabilities equal the product of the individual probabilities. Recall that

$$\begin{aligned} \mathbb{P}[X = 1] &= 0.5 \\ \mathbb{P}[Y = 1] &= 0.6 \\ \mathbb{P}[X = 1, Y = 1] &= 0.4. \end{aligned}$$

Since

$$0.5 \times 0.6 = 0.3 \neq 0.4$$

we conclude that X and Y are dependent.

Example 2: (continued). $f_{Y|X}(y \mid x) = \frac{(xy+y^2)\exp(-y)}{x+2}$ and $f_Y(y) = \frac{1}{4}\left(y^2+2y\right)\exp\left(-y\right)$. Since these are not equal we deduce that X and Y are not independent. Another way of seeing this is that $f(x, y) = \frac{1}{4}\left(x+y\right)xy\exp\left(-x-y\right)$ cannot be factored.

Example 3: (continued). Figure 4.10 displayed the conditional density of wages given experience. Since the conditional density changes with experience, wages and experience are not independent.

4.12 Covariance and Correlation

A feature of the joint distribution of (X, Y) is the covariance.

Definition 4.12 If X and Y have finite variances, the **covariance** between X and Y is

$$\begin{aligned} \operatorname{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]. \end{aligned}$$

Definition 4.13 If X and Y have finite variances, the **correlation** between X and Y is

$$\operatorname{corr}(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}.$$

If $\operatorname{cov}(X,Y)=0$ then $\operatorname{corr}(X,Y)=0$ and it is typical to say that X and Y are **uncorrelated**.

Theorem 4.8 If X and Y are independent with finite variances, then X and Y are uncorrelated.

The reverse is not true. For example, suppose that $X\sim U[-1,1]$. Since it is symmetrically distributed about 0 we see that $\mathbb{E}[X]=0$ and $\mathbb{E}\left[X^3\right]=0$. Set $Y=X^2$. Then

$$\operatorname{cov}(X,Y)=\mathbb{E}\left[X^3\right]-\mathbb{E}[X]\mathbb{E}\left[X^2\right]=0.$$

Thus X and Y are uncorrelated yet are fully dependent! This shows that uncorrelated random variables may be dependent.

The following results are quite useful.

Theorem 4.9 If X and Y have finite variances, $\operatorname{var}[X+Y]=\operatorname{var}[X]+\operatorname{var}[Y]+2\operatorname{cov}(X,Y)$.

To see this it is sufficient to assume that X and Y are zero mean. Completing the square and using the linear property of expectations

$$\begin{aligned}\operatorname{var}[X+Y]&=\mathbb{E}\left[(X+Y)^2\right]\\&=\mathbb{E}\left[X^2+Y^2+2XY\right]\\&=\mathbb{E}\left[X^2\right]+\mathbb{E}\left[Y^2\right]+2\mathbb{E}[XY]\\&=\operatorname{var}[X]+\operatorname{var}[Y]+2\operatorname{cov}(X,Y).\end{aligned}$$

Theorem 4.10 If X and Y are uncorrelated, then $\operatorname{var}[X+Y]=\operatorname{var}[X]+\operatorname{var}[Y]$.

This follows from the previous theorem, since uncorrelatedness means that $\operatorname{cov}(X,Y)=0$.

Example 1: (continued). The cross moment is

$$\mathbb{E}[XY]=1\times1\times0.4+1\times2\times0.1+2\times1\times0.2+2\times2\times0.3=2.2.$$

The covariance is

$$\operatorname{cov}(X,Y)=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]=2.2-1.5\times1.4=0.1$$

and correlation

$$\operatorname{corr}(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}[X]\operatorname{var}[Y]}}=\frac{0.1}{\sqrt{0.25\times0.24}}=0.41.$$

This is a high correlation. As might be expected, the number of pizza slices and drinks purchased are positively and meaningfully correlated.

Example 2: (continued). The cross moment is

$$\mathbb{E}[XY]=\int_0^\infty\int_0^\infty xy\frac{1}{4}(x+y)\,xy\exp(-x-y)\,dx\,dy=6.$$

The covariance is

$$\operatorname{cov}(X,Y)=6-\left(\frac{5}{2}\right)^2=-\frac{1}{4}$$

and correlation

$$\operatorname{corr}(X,Y)=\frac{-\left(\frac{1}{4}\right)}{\sqrt{\frac{11}{4}\times\frac{11}{4}}}=-\frac{1}{11}.$$

This is a negative correlation, meaning that the two variables co-vary in opposite directions. The magnitude of the correlation, however, is small, indicating that the co-movement is mild.

Example 3 & 4: (continued). Correlations of wages, experience, and experience are presented in the following chart known as a correlation matrix.

Correlation Matrix

	Wage	Experience	Education
Wage	1	0.06	0.40
Experience	0.06	1	−0.17
Education	0.40	−0.17	1

This shows that education and wages are highly correlated, wages and experiences mildly correlated, and education and experience negatively correlated. The last feature is likely due to the fact that the variation in experience at a point in time is mostly due to differences across cohorts (people of different ages). This negative correlation is because education levels are different across cohorts – later generations have higher average education levels.

4.13 Cauchy-Schwarz

This following inequality is used frequently.

Theorem 4.11 Cauchy-Schwarz Inequality. For any random variables X and Y

$$\mathbb{E}|XY|\leq\sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}.$$

Proof: By the Geometric Mean Inequality (Theorem 2.13)

$$|ab|=\sqrt{a^2}\sqrt{b^2}\leq\frac{a^2+b^2}{2}. \tag{4.1}$$

Set $U=|X|/\sqrt{\mathbb{E}\left[X^2\right]}$ and $V=|Y|/\sqrt{\mathbb{E}\left[Y^2\right]}$. Using (4.1) we obtain

$$|UV|\leq\frac{U^2+V^2}{2}.$$

Taking expectations we find

$$\frac{\mathbb{E}|XY|}{\sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}}=\mathbb{E}|UV|\leq\mathbb{E}\left[\frac{U^2+V^2}{2}\right]=1$$

the final equality since $\mathbb{E}\left[U^2\right]=1$ and $\mathbb{E}\left[V^2\right]=1$. This is the theorem. ■

The Cauchy-Schwarz inequality implies bounds on covariances and correlations.

Theorem 4.12 Covariance Inequality. For any random variables X and Y with finite variances

$$|\mathrm{cov}(X,Y)| \leq (\mathrm{var}[X]\,\mathrm{var}[Y]])^{1/2}$$
$$|\mathrm{corr}(X,Y)| \leq 1.$$

Proof. We apply the Expectation (Theorem 2.10) and Cauchy-Schwarz (Theorem 4.11) inequalities to find

$$|\mathrm{cov}(X,Y)| \leq \mathbb{E}\,|(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])| \leq (\mathrm{var}[X]\,\mathrm{var}[Y])^{1/2}$$

as stated. ■

The fact that the correlation is bounded below one helps us understand how to interpret a correlation. Correlations close to zero are small. Correlations close to 1 and -1 are large.

4.14 Conditional Expectation

An important concept in econometrics is conditional expectation. Just as the expectation is the central tendency of a distribution, the conditional expectation is the central tendency of a conditional distribution.

Definition 4.14 The **conditional expectation** of Y given $X = x$ is the expected value of the conditional distribution $F_{Y|X}(y \mid x)$ and is written as $m(x) = \mathbb{E}[Y \mid X = x]$. For discrete random variables this is

$$\mathbb{E}[Y \mid X = x] = \frac{\sum_{j=1}^{\infty} \tau_j \pi(x, \tau_j)}{\pi_X(x)}.$$

For continuous Y this is

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy.$$

We also call $\mathbb{E}[Y \mid X = x]$ the **conditional mean**.

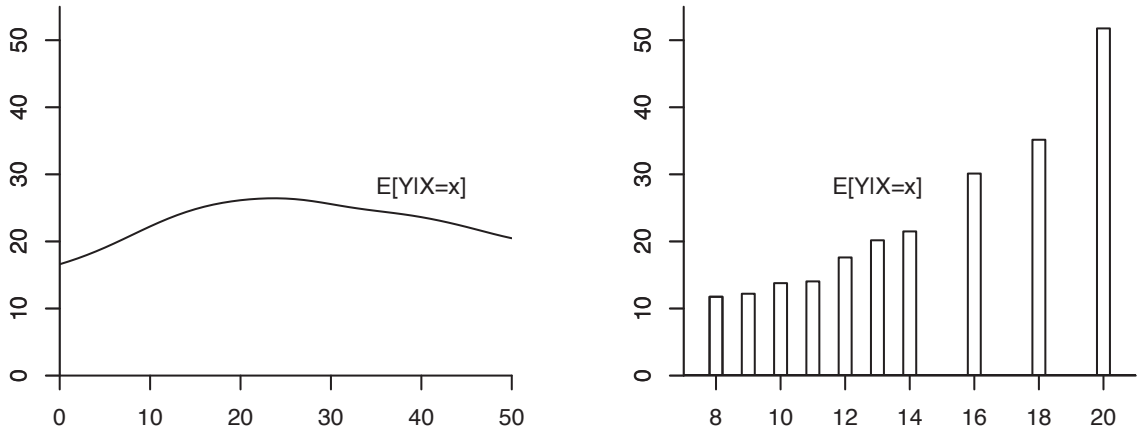
In the continuous case, using the definition of the conditional PDF we can write $\mathbb{E}[Y \mid X = x]$ as

$$\mathbb{E}[Y \mid X = x] = \frac{\int_{-\infty}^{\infty} y f(y,x) dy}{\int_{-\infty}^{\infty} f(x,y) dy}.$$

The conditional expectation tells us the average value of Y given that X equals the specific value x. When X is discrete the conditional expectation is the expected value of Y within the sub-population for which $X = x$. For example, if X is gender then $\mathbb{E}[Y \mid X = x]$ is the expected value for men and women, separately. If X is education then $\mathbb{E}[Y \mid X = x]$ is the expected value for each education level. When X is continuous the conditional expectation is the expected value of Y within the infinitesimally small population for which $X \simeq x$.

Example 1: (continued): The conditional expectation for the number of drinks per customer is

$$\mathbb{E}[Y \mid X = 1] = \frac{1 \times 0.4 + 2 \times 0.1}{0.5} = 1.2$$
$$\mathbb{E}[Y \mid X = 2] = \frac{1 \times 0.2 + 2 \times 0.3}{0.5} = 1.6.$$



(a) Wages Given Experience

(b) Wages Given Education

Figure 4.11: Conditional Expectation Functions

Thus the restaurant can expect to serve (on average) more drinks to customers who purchase two slices of pizza.

Example 2: (continued): The conditional expectation of Y given $X = x$ is

$$\begin{aligned}\mathbb{E}[Y \mid X = x] &= \int_0^\infty y f_{Y|X}(y \mid x)\,dy \\ &= \int_0^\infty y \frac{\left(xy + y^2\right)\exp\left(-y\right)}{x+2}\,dy \\ &= \frac{2x+6}{x+2}.\end{aligned}$$

This conditional expectation function is downward sloping for $x \geq 0$. Thus as x increases the expected value of Y declines.

Example 3: (continued). The conditional expectation of wages given experience is displayed in Figure 4.11(a). The x-axis is years of experience. The y-axis is wages. You can see that the expected wage is about \$16.50 for 0 years of experience, and increases near linearily to about \$26 by 20 years of experience. Above 20 years of experience the expected wage falls, reaching about \$21 by 50 years of experience. Overall the shape of the wage-experience profile is an inverted U-shape, increasing for the early years of experience, and decreasing in the more advanced years.

Example 4: (continued). The conditional expectation of wages given education is displayed in Figure 4.11(b). The x-axis is years of education. Since education is discrete the conditional mean is a discrete function as well. Examining Figure 4.11(b) we see that the conditional expectation is monotonically increasing in years of education. The mean is \$11.75 for an individual with 8 years of education, \$17.61 for an individual with 12 years of education, \$21.49 for an individual with 14 years, \$30.12 for 16 years, \$35.16 for 18 years, and \$51.76 for 20 years.

4.15 Law of Iterated Expectations

The function $m(x) = \mathbb{E}[Y \mid X = x]$ is not random. Rather, it is a feature of the joint distribution. Sometimes, however, it is useful to treat the conditional expectation as a random variable. To do so, we evaluate the function $m(x)$ at the random variable X. This is $m(X) = \mathbb{E}[Y \mid X]$. This is a random variable, a transformation of X.

What does this mean? Take our example 1. We found that $\mathbb{E}[Y \mid X = 1] = 1.2$ and $\mathbb{E}[Y \mid X = 2] = 1.6$. We also know that $\mathbb{P}[X = 1] = \mathbb{P}[X = 2] = 0.5$. Thus $m(X)$ is a random variable with a two-point distribution, equalling 1.2 and 1.6 each with probability one-half.

This may seem abstract and a bit confusing. Another way of expressing the distinction is that $m(x)$ is the value of the conditional expectation at $X = x$ while $m(X)$ is a function (a transformation) of the random variable X.

By treating $\mathbb{E}[Y \mid X]$ as a random variable we can do some interesting manipulations. For example, what is the expectation of $\mathbb{E}[Y \mid X]$? Take the continuous case. We find

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \int_{-\infty}^{\infty} \mathbb{E}[Y \mid X = x] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) f_X(x) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(y, x) dy dx \\
&= \mathbb{E}[Y].
\end{aligned}$$

In words, the average across group averages is the grand average.

Now take the discrete case.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_{i=1}^{\infty} \mathbb{E}[Y \mid X = \tau_i] \pi_X(\tau_i) \\
&= \sum_{i=1}^{\infty} \frac{\sum_{j=1}^{\infty} \tau_j \pi(\tau_i, \tau_j)}{\pi_X(\tau_i)} \pi_X(\tau_i) \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \tau_j \pi(\tau_i, \tau_j) \\
&= \mathbb{E}[Y].
\end{aligned}$$

This is a very critical result.

Theorem 4.13 Law of Iterated Expectations. If $\mathbb{E}|Y| < \infty$ then $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$.

Example 1: (continued): We calculated earlier that $\mathbb{E}[Y] = 1.4$. Using the law of iterated expectations

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[Y \mid X = 1] \mathbb{P}[X = 1] + \mathbb{E}[Y \mid X = 2] \mathbb{P}[X = 2] \\
&= 1.2 \times 0.5 + 1.6 \times 0.5 \\
&= 1.4
\end{aligned}$$

which is the same.

Example 2: (continued): We calculated that $\mathbb{E}[Y] = 5/2$. Using the law of iterated expectations

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^\infty \mathbb{E}[Y \mid X = x] f_X(x) dx \\ &= \int_0^\infty \left(\frac{2x+6}{x+2}\right)\left(\frac{x^2+2x}{4}\exp(-x)\right) dx \\ &= \int_0^\infty \left(\frac{x^2+3x}{2}\right)\exp(-x)\, dx \\ &= \frac{5}{2}\end{aligned}$$

which is the same.

Example 4: (continued). The conditional expectation of wages given education is displayed in Figure 4.11(b). The marginal probabilities of education were displayed in Figure 2.2(b). By the law of iterated expectations the unconditional expectation of wages is the sum of the products of these two displays. This is

$$\begin{aligned}\mathbb{E}[\text{wage}] &= 11.75 \times 0.027 + 12.20 \times 0.011 + 13.78 \times 0.011 + 13.78 \times 0.011 + 14.04 \times 0.026 + 17.61 \times 0.274 \\ &\quad + 20.17 \times 0.182 + 21.49 \times 0.111 + 30.12 \times 0.229 + 35.16 \times 0.092 + 51.76 \times 0.037 \\ &= \$23.90.\end{aligned}$$

This equals the average wage.

4.16 Conditional Variance

Another feature of the conditional distribution is the conditional variance.

Definition 4.15 The **conditional variance** of Y given $X = x$ is the variance of the conditional distribution $F_{Y|X}(y \mid x)$ and is written as $\text{var}[Y \mid X = x]$ or $\sigma^2(x)$. It equals

$$\text{var}[Y \mid X = x] = \mathbb{E}\left[(Y - m(x))^2 \mid X = x\right].$$

The conditional variance $\text{var}[Y \mid X = x]$ is a function of x and can take any non-negative shape. When X is discrete $\text{var}[Y \mid X = x]$ is the variance of Y within the sub-population with $X = x$. When X is continuous $\text{var}[Y \mid X = x]$ is the variance of Y within the infinitesimally small sub-population with $X \simeq x$.

By expanding the quadratic we can re-express the conditional variance as

$$\text{var}[Y \mid X = x] = \mathbb{E}\left[Y^2 \mid X = x\right] - (\mathbb{E}[Y \mid X = x])^2. \tag{4.2}$$

We can also define $\text{var}[Y \mid X] = \sigma^2(X)$, the conditional variance treated as a random variable. We have the following relationship.

Theorem 4.14 $\text{var}[Y] = \mathbb{E}[\text{var}[Y \mid X]] + \text{var}[\mathbb{E}[Y \mid X]]$.

The first term on the right-hand side is often called the **within group variance**, while the second term is called the **across group variance**.

We prove the theorem for the continuous case. Using (4.2)

$$\begin{aligned}\mathbb{E}[\operatorname{var}[Y \mid X]] &= \int \operatorname{var}[Y \mid X = x] f_X(x) dx \\ &= \int \mathbb{E}\left[Y^2 \mid X = x\right] f_X(x) dx - \int m(x)^2 f_X(x) dx \\ &= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[m(X)^2\right] \\ &= \operatorname{var}[Y] - \operatorname{var}[m(X)].\end{aligned}$$

The third equality uses the law of iterated expecations for Y^2. The fourth uses $\operatorname{var}[Y] = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2$ and $\operatorname{var}[m(X)] = \mathbb{E}\left[m(X)^2\right] - (\mathbb{E}[m(X)])^2 = \mathbb{E}\left[m(X)^2\right] - (\mathbb{E}[Y])^2$.

Example 1: (continued): The conditional variance for the number of drinks per customer is

$$\mathbb{E}\left[Y^2 \mid X = 1\right] - (\mathbb{E}[Y \mid X = 1])^2 = \frac{1^2 \times 0.4 + 2^2 \times 0.1}{0.5} - 1.2^2 = 0.16$$
$$\mathbb{E}\left[Y^2 \mid X = 2\right] - (\mathbb{E}[Y \mid X = 2])^2 = \frac{1^2 \times 0.2 + 2^2 \times 0.3}{0.5} - 1.6^2 = 0.24.$$

The variability of the number of drinks purchased is greater for customers who purchase two slices of pizza.

Example 2: (continued): The conditional second moment is

$$\begin{aligned}\mathbb{E}\left[Y^2 \mid X = x\right] &= \int_0^\infty y^2 f_{Y|X}(y \mid x) dy \\ &= \int_0^\infty y^2 \frac{\left(xy + y^2\right)\exp\left(-y\right)}{x+2} dy \\ &= \frac{6x+24}{x+2}.\end{aligned}$$

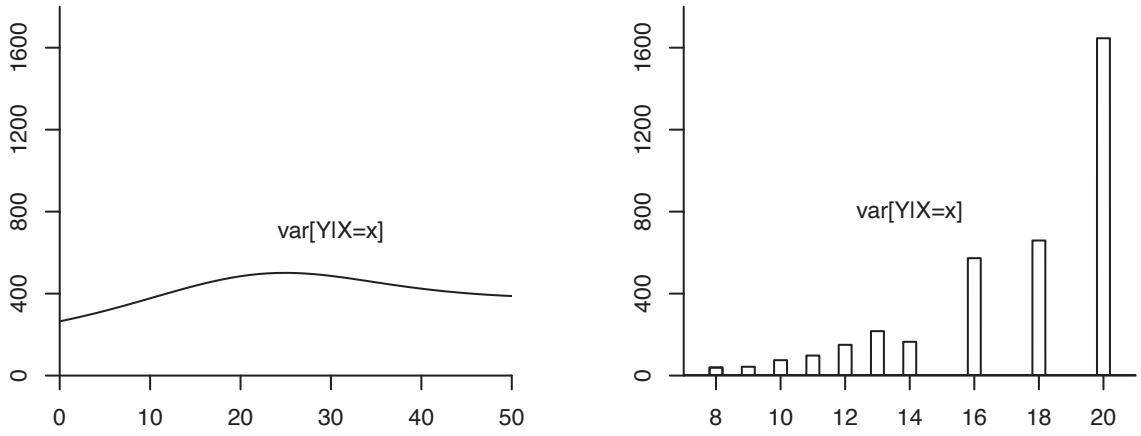
The conditional variance is

$$\begin{aligned}\operatorname{var}[Y \mid X = x] &= \mathbb{E}\left[Y^2 \mid X = x\right] - (\mathbb{E}[Y \mid X = x])^2 \\ &= \frac{6x+24}{x+2} - \left(\frac{2x+6}{x+2}\right)^2 \\ &= \frac{2x^2+12x+12}{(x+2)^2}.\end{aligned}$$

This varies with x.

Example 3: (continued). The conditional variance of wages given experience is displayed in Figure 4.12(a). We see that the variance is a hump-shaped function of experience. The variance substantially increases between 0 and 25 years of experience, and then falls somewhat between 25 and 50 years of experience.

Example 4: (continued). The conditional variance of wages given education is displayed in Figure 4.12(b). The conditional variance is strongly varying as a function of education. The variance for high school graduates is 150, that for college graduates 573, and those with professional degrees 1646. These are large and meaningful changes. It means that while the average level of wages increases significantly with education level, so does the spread of the wage distribution. The effect of education on the conditional variance is much stronger than the effect of experience.



(a) Wages Given Experience

(b) Wages Given Education

Figure 4.12: Conditional Variance Functions

4.17 Hölder's and Minkowski's Inequalities*

The following inequalities are useful generalizations of the Cauchy-Schwarz inequality.

Theorem 4.15 Hölder's Inequality. For any random variables X and Y and any $p \geq 1$ and $q \geq 1$ satisfying $1/p + 1/q = 1$,

$$\mathbb{E}|XY| \leq \left(\mathbb{E}|X|^p\right)^{1/p} \left(\mathbb{E}|Y|^q\right)^{1/q}.$$

Proof: By the Geometric Mean Inequality (Theorem 2.13) for non-negative a and b

$$ab = \left(a^p\right)^{1/p} \left(b^q\right)^{1/q} \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{4.3}$$

Without loss of generality assume $\mathbb{E}|X|^p = 1$ and $\mathbb{E}|Y|^q = 1$. Applying (4.3)

$$\mathbb{E}|XY| \leq \frac{\mathbb{E}|X|^p}{p} + \frac{\mathbb{E}|Y|^q}{q} = \frac{1}{p} + \frac{1}{q} = 1$$

as needed. ■

Theorem 4.16 Minkowski's Inequality. For any random variables X and Y and any $p \geq 1$

$$\left(\mathbb{E}|X+Y|^p\right)^{1/p} \leq \left(\mathbb{E}|X|^p\right)^{1/p} + \left(\mathbb{E}|Y|^p\right)^{1/p}.$$

Proof. Using the triangle inequality and then applying Hölder's inequality (Theorem 4.15) to the two

expectations

$$\begin{aligned}
\mathbb{E}|X+Y|^p &= \mathbb{E}\left[|X+Y||X+Y|^{p-1}\right] \\
&\leq \mathbb{E}\left[|X||X+Y|^{p-1}\right] + \mathbb{E}\left(|Y||X+Y|^{p-1}\right) \\
&\leq \left(\mathbb{E}|X|^p\right)^{1/p}\left(\mathbb{E}|X+Y|^{(p-1)q}\right)^{1/q} \\
&+ \left(\mathbb{E}|Y|^p\right)^{1/p}\left(\mathbb{E}|X+Y|^{(p-1)q}\right)^{1/q} \\
&= \left(\left(\mathbb{E}|X|^p\right)^{1/p} + \left(\mathbb{E}|Y|^p\right)^{1/p}\right)\left(\mathbb{E}|X+Y|^p\right)^{(p-1)/p}
\end{aligned}$$

where the second inequality picks q to satisfy $1/p + 1/q = 1$, and the final equality uses this fact to make the substitution $q = p/(p-1)$ and then collect terms. Dividing both sides by $\left(\mathbb{E}|X+Y|^p\right)^{(p-1)/p}$, we complete the proof. ■

4.18 Vector Notation

Write an m-vector as

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}.$$

This is an element of $\mathbb{R}^m$, m-dimensional Euclidean space. In some cases (primarily for matrix algebra) we may use boldface $\boldsymbol{x}$ to indicate a vector.

The **transpose** of a column vector x is the row vector

$$x' = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \end{pmatrix}.$$

There is diversity between fields concerning the choice of notation for the transpose. The above notation is the most common in econometrics. In statistics and mathematics the notation $x^\top$ is typically used.

The **Euclidean norm** is the Euclidean length of the vector x, defined as

$$\|x\| = \left(\sum_{i=1}^m x_i^2\right)^{1/2} = \left(x'x\right)^{1/2}.$$

Multivariate random vectors are written as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}.$$

Some authors use the notation $\vec{X}$ or $\mathbf{X}$ to denote a random vector.

The equality $\{X = x\}$ and inequality $\{X \leq x\}$ hold if and only if they hold for all components. Thus $\{X = x\}$ means $\{X_1 = x_1, X_2 = x_2, ..., X_m = x_m\}$, and similarly $\{X \leq x\}$ means $\{X_1 \leq x_1, X_2 \leq x_2, ..., X_m \leq x_m\}$. The probability notation $\mathbb{P}[X \leq x]$ means $\mathbb{P}[X_1 \leq x_1, X_2 \leq x_2, ..., X_m \leq x_m]$.

When integrating over $\mathbb{R}^m$ it is convenient to use the following notation. If $f(x) = f(x_1, ..., x_m)$, write

$$\int f(x)dx = \int \cdots \int f(x_1, ..., x_m)dx_1 \cdots dx_m.$$

Thus an integral with respect to a vector argument dx is short-hand for an m-fold integral. The notation on the left is more compact and easier to read. We use the notation on the right when we want to be specific about the arguments.

4.19 Triangle Inequalities*

Theorem 4.17 For any real numbers x_j

$$\left|\sum_{j=1}^{m} x_j\right| \leq \sum_{j=1}^{m} \left|x_j\right|. \tag{4.4}$$

Proof: Take the case $m = 2$. Observe that

$$-|x_1| \leq x_1 \leq |x_1|$$
$$-|x_2| \leq x_2 \leq |x_2|.$$

Adding, we find

$$-|x_1| - |x_2| \leq x_1 + x_2 \leq |x_1| + |x_2|$$

which is (4.4) for $m = 2$. For $m > 2$, we apply (4.4) $m - 1$ times. ■

Theorem 4.18 For any vector $x = (x_1, ..., x_m)'$

$$\|x\| \leq \sum_{i=1}^{m} |x_i|. \tag{4.5}$$

Proof: Without loss of generality assume $\sum_{i=1}^{m} |x_i| = 1$. This implies $|x_i| \leq 1$ and thus $x_i^2 \leq |x_i|$. Hence

$$\|x\|^2 = \sum_{i=1}^{m} x_i^2 \leq \sum_{i=1}^{m} |x_i| = 1.$$

Taking the square root of the two sides completes the proof. ■

Theorem 4.19 Schwarz Inequality. For any m-vectors x and y

$$\left|x'y\right| \leq \|x\| \left\|y\right\|. \tag{4.6}$$

Proof: Without loss of generality assume $\|x\| = 1$ and $\left\|y\right\| = 1$ so our goal is to show that $\left|x'y\right| \leq 1$. By Theorem 4.17 and then applying (4.1) to $\left|x_i y_i\right| = |x_i| \left|y_i\right|$ we find

$$\left|x'y\right| = \left|\sum_{i=1}^{m} x_i y_i\right| \leq \sum_{i=1}^{m} \left|x_i y_i\right| \leq \frac{1}{2}\sum_{i=1}^{m} x_i^2 + \frac{1}{2}\sum_{i=1}^{m} y_i^2 = 1.$$

the final equality since $\|x\| = 1$ and $\left\|y\right\| = 1$. This is (4.6). ■

Theorem 4.20 For any m-vectors x and y

$$\left\|x + y\right\| \leq \|x\| + \left\|y\right\|. \tag{4.7}$$

Proof: We apply (4.6)

$$\begin{aligned}\left\|x + y\right\|^2 &= x'x + 2x'y + y'y \\ &\leq \|x\|^2 + 2\|x\| \left\|y\right\| + \left\|y\right\|^2 \\ &= \left(\|x\| + \left\|y\right\|\right)^2.\end{aligned}$$

Taking the square root of the two sides completes the proof. ■

4.20 Multivariate Random Vectors

We now consider the case of a random vector $X \in \mathbb{R}^m$.

Definition 4.16 A **multivariate random vector** is a function from the sample space to $\mathbb{R}^m$, written as $X = (X_1, X_2, ..., X_m)'$.

We now define the distribution, mass, and density functions for multivariate random vectors.

Definition 4.17 The **joint distribution function** is $F(x) = \mathbb{P}[X \le x] = \mathbb{P}[X_1 \le x_1, ..., X_m \le x_m]$.

Definition 4.18 For discrete random vectors, the **joint probability mass function** is $\pi(x) = \mathbb{P}[X = x]$.

Definition 4.19 When $F(x)$ is continuous and differentiable its **joint density** $f(x)$ equals

$$f(x) = \frac{\partial^m}{\partial x_1 \cdots \partial x_m} F(x).$$

Definition 4.20 The **expectation** of $X \in \mathbb{R}^m$ is the vector of expectations of its elements:

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_m] \end{pmatrix}.$$

Definition 4.21 The $m \times m$ **covariance matrix** of $X \in \mathbb{R}^m$ is

$$\operatorname{var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])'\right].$$

and is commonly written as $\operatorname{var}[X] = \Sigma$.

When applied to a random vector, $\operatorname{var}[X]$ is a matrix. It has elements

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix}$$

where $\sigma_j^2 = \operatorname{var}\left[X_j\right]$ and $\sigma_{ij} = \operatorname{cov}(X_i, X_j)$ for $i, j = 1, 2, \ldots, m$ and $i \ne j$.

Theorem 4.21 Properties of the covariance matrix. Any $m \times m$ covariance matrix Σ satisfies:

1. Symmetric: $\Sigma = \Sigma'$.
2. Positive semi-definite: For any $m \times 1$ $a \ne 0$, $a'\Sigma a \ge 0$.

Proof: Symmetry holds because $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$. For positive semi-definiteness,

$$\begin{aligned} a'\Sigma a &= a'\mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])'\right] a \\ &= \mathbb{E}\left[a'(X - \mathbb{E}[X])(X - \mathbb{E}[X])'a\right] \\ &= \mathbb{E}\left[\left(a'(X - \mathbb{E}[X])\right)^2\right] \\ &= \mathbb{E}\left[Z^2\right] \end{aligned}$$

where $Z = a'(X - \mathbb{E}[X])$. Since $Z^2 \ge 0$, $\mathbb{E}\left[Z^2\right] \ge 0$ and $a'\Sigma a \ge 0$. ■

Theorem 4.22 If $X \in \mathbb{R}^m$ with $m \times 1$ expectation μ and $m \times m$ covariance matrix Σ, and $\boldsymbol{A}$ is $q \times m$, then $\boldsymbol{A}X$ is a random vector with mean $\boldsymbol{A}\mu$ and covariance matrix $\boldsymbol{A}\Sigma\boldsymbol{A}'$.

Theorem 4.23 For $X \in \mathbb{R}^m$, $\mathbb{E}\|X\| < \infty$ if and only if $\mathbb{E}\left|X_j\right| < \infty$ for $j = 1, ..., m$.

Proof*: Assume $\mathbb{E}\left|X_j\right| \le C < \infty$ for $j = 1, ..., m$. Applying the triangle inequality (Theorem 4.18)

$$\mathbb{E}\|X\| \le \sum_{j=1}^{m} \mathbb{E}\left|X_j\right| \le mC < \infty.$$

For the reverse inequality, the Euclidean norm of a vector is larger than the length of any individual component, so for any j, $\left|X_j\right| \le \|X\|$. Thus, if $\mathbb{E}\|X\| < \infty$, then $\mathbb{E}\left|X_j\right| < \infty$ for $j = 1, ..., m$. ■

4.21 Pairs of Multivariate Vectors

Most concepts for pairs of random variables apply to multivariate vectors. Let (X, Y) be a pair of multivariate vectors of dimension m_X and m_Y respectively. For ease of presentation we focus on the case of continuous random vectors.

Definition 4.22 The **joint distribution** function of $(X, Y) \in \mathbb{R}^{m_X} \times \mathbb{R}^{m_Y}$ is

$$F(x, y) = \mathbb{P}\left[X \le x, Y \le y\right].$$

The **joint density** function of $(X, Y) \in \mathbb{R}^{m_X} \times \mathbb{R}^{m_Y}$ is

$$f(x, y) = \frac{\partial^{m_x + m_y}}{\partial x_1 \cdots \partial y_{m_y}} F(x, y).$$

The **marginal density** functions of X and Y are

$$f_X(x) = \int f(x, y) dy$$
$$f_Y(y) = \int f(x, y) dx.$$

The **conditional densities** of Y given $X = x$ and X given $Y = y$ are

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$
$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}.$$

The **conditional expectation** of Y given $X = x$ is

$$\mathbb{E}[Y \mid X = x] = \int y f_{Y|X}(y \mid x) dy.$$

The random vectors Y and X are **independent** if their joint density factors as

$$f(x, y) = f_X(x) f_Y(y).$$

The continuous random variables $(X_1, ..., X_m)$ are **mutually independent** if their joint density factors into the products of marginal densities, thus

$$f(x_1, ..., x_m) = f_1(x_1) \cdots f_m(x_m).$$

4.22 Multivariate Transformations

When $X \in \mathbb{R}^m$ has a density and $Y = g(X) \in \mathbb{R}^q$ where $g(x) : \mathbb{R}^m \to \mathbb{R}^q$ is one-to-one then there is a well-known formula for the joint density of Y.

Theorem 4.24 Suppose X has PDF $f_X(x)$, $g(x)$ is one-to-one, and $h(y) = g^{-1}(y)$ is differentiable. Then $Y = g(X)$ has density

$$f_Y(y) = f_X(h(y))J(y)$$

where

$$J(y) = \left|\det\left(\frac{\partial}{\partial y'}h(y)\right)\right|$$

is the Jacobian of the transformation.

Writing out the derivative matrix in detail, let $h(y) = \left(h_1(y), h_2(y), ..., h_m(y)\right)'$ and

$$\frac{\partial}{\partial y'}h(y) = \begin{pmatrix} \partial h_1(y)/\partial y_1 & \partial h_1(y)/\partial y_2 & \cdots & \partial h_1(y)/\partial y_m \\ \partial h_2(y)/\partial y_1 & \partial h_2(y)/\partial y_2 & \cdots & \partial h_2(y)/\partial y_m \\ \vdots & \vdots & \ddots & \vdots \\ \partial h_m(y)/\partial y_1 & \partial h_m(y)/\partial y_2 & \cdots & \partial h_m(y)/\partial y_m \end{pmatrix}.$$

Example: Let X_1 and X_2 be independent with densities e^{-x_1} and e^{-x_2}. Take the transformation $Y_1 = X_1$ and $Y_2 = X_1 + X_2$. The inverse transformation is $X_1 = Y_1$ and $X_2 = Y_2 - Y_1$ with derivative matrix

$$\frac{\partial}{\partial y'}h(y) = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

Thus $J = 1$. The support for Y is $\{0 \le Y_1 \le Y_2 < \infty\}$. The joint density is therefore

$$f_Y(y) = e^{-y_2}\mathbb{1}\left\{y_1 \le y_2\right\}.$$

We can calculate the marginal density of Y_2 by integrating over Y_1. This is

$$f_2(y_2) = \int_0^\infty e^{-y_2}\mathbb{1}\left\{y_1 \le y_2\right\}dy_1 = \int_0^{y_2} e^{-y_2}dy_1 = y_2e^{-y_2}$$

on $y_2 \in \mathbb{R}$. This is a gamma density with parameters $\alpha = 2$, $\beta = 1$.

4.23 Convolutions

A useful method to calculate the distribution of the sum of random variables is by convolution.

Theorem 4.25 Convolution Theorem. If X and Y are independent random variables with densities $f_X(x)$ and $f_Y(y)$ then the density of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(s)f_Y(z-s)ds = \int_{-\infty}^{\infty} f_X(z-s)f_Y(s)ds.$$

Proof: Define the transformation (X, Y) to (W, Z) where $Z = X + Y$ and $W = X$. The Jacobian is 1. The joint density of (W, Z) is $f_X(w) f_Y(z - w)$. The marginal density of Z is obtained by integrating out W, which is the first stated result. The second can be obtained by transformation of variables. ■

The representation $\int_{-\infty}^{\infty} f_X(s) f_Y(z - s) ds$ is known as the **convolution** of f_X and f_Y.

Example: Suppose $X \sim U[0,1]$ and $Y \sim U[0,1]$ are independent and $Z = X + Y$. Z has support $[0,2]$. By the convolution theorem Z has density

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(s) f_Y(z - s) ds &= \int_{-\infty}^{\infty} \mathbb{1}\{0 \le s \le 1\} \mathbb{1}\{0 \le z - s \le 1\} ds \\
&= \int_0^1 \mathbb{1}\{0 \le z - s \le 1\} ds \\
&= \begin{cases} \int_0^z ds & z \le 1 \\ \int_{z-1}^1 ds & 1 \le z \le 2 \end{cases} \\
&= \begin{cases} z & z \le 1 \\ 2 - z & 1 \le z \le 2. \end{cases}
\end{aligned}$$

Thus the density of Z has a triangle or "tent" shape on $[0,1]$.

Example: Suppose X and Y are independent each with density $\lambda^{-1} \exp(-x/\lambda)$ on $x \ge 0$, and $Z = X + Y$. Z has support $[0, \infty)$. By the convolution theorem

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_X(t) f_Y(z - t) dt \\
&= \int_{-\infty}^{\infty} \mathbb{1}\{t \ge 0\} \mathbb{1}\{z - t \ge 0\} \frac{1}{\lambda} e^{-t/\lambda} \frac{1}{\lambda} e^{-(z-t)/\lambda} dt \\
&= \int_0^z \frac{1}{\lambda^2} e^{-z/\lambda} dt \\
&= \frac{z}{\lambda^2} e^{-z/\lambda}.
\end{aligned}$$

This is a gamma density with parameters $\alpha = 2$ and $\beta = 1/\lambda$.

4.24 Hierarchical Distributions

Often a useful way to build a probability structure for an economic model is to use a hierarchy. Each stage of the hierarchy is a random variable with a distribution whose parameters are treated as random variables. This can result in a compelling economic structure. The resulting probability distribution can in certain cases equal a known distribution, or can lead to a new distribution.

For example suppose we want a model for the number of sales at a retail store. A baseline model is that the store has N customers who each make a binary decision to buy or not with some probability p. This is a binomial model for sales, $X \sim$binomial(N, p). If the number of customers N is unobserved we can also model it as a random variable. A simple model is $N \sim$Poisson(λ). We examine this model below.

In general a two-layer hierarchical model takes the form

$$\begin{aligned}
X \mid Y &\sim f(x \mid y) \\
Y &\sim g(y).
\end{aligned}$$

The joint density of X and Y equals $f(x,y) = f(x \mid y)g(y)$. The marginal density of X is

$$f(x) = \int f(x,y)dy = \int f(x \mid y)g(y)dy.$$

More complicated structures can be built. A three-layer hierarchical model takes the form

$$\begin{aligned} X \mid Y,Z &\sim f(x \mid y,z) \\ Y \mid Z &\sim g(y \mid z) \\ Z &\sim h(z). \end{aligned}$$

The marginal density of X is

$$f(x) = \int\int f(x \mid y,z)g(y \mid z)h(z)dzdy.$$

Binomial-Poisson. This is the retail sales model described above. The distribution of sales X given N customers is binomial and the distribution of the number of customers is Poisson.

$$\begin{aligned} X \mid N &\sim \text{binomial}(N,p) \\ N &\sim \text{Poisson}(\lambda). \end{aligned}$$

The marginal distribution of X equals

$$\begin{aligned} \mathbb{P}[X=x] &= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{e^{-\lambda}\lambda^n}{n!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{n=x}^{\infty} \frac{\left((1-p)\lambda\right)^{n-x}}{(n-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{\left((1-p)\lambda\right)^{t}}{t!} \\ &= \frac{(\lambda p)^x e^{-\lambda p}}{x!}. \end{aligned}$$

The first equality is the sum of the binomial multiplied by the Poisson. The second line writes out the factorials and combines terms. The third line makes the change-of-variables $t = n - x$. The last line recognizes that the sum is over a Poisson density with parameter $(1-p)\lambda$. The result in the final line is the probability mass function for the Poisson(λp) distribution. Thus

$$X \sim \text{Poisson}(\lambda p).$$

Hence the Binomial-Poisson model implies a Poisson distribution for sales!

This shows the (perhaps surprising) result that if customers arrive with a Poisson distribution and each makes a Bernoulli decision then total sales are distributed as Poisson.

Beta-Binomial. Returning to the retail sales model, again assume that sales given customers is binomial, but now consider the case that the probability p of a sale is heterogeneous. This can be modeled by treating p as random. A simple model appropriate for a probability is $p \sim$beta(α,β). The hierarchical model is

$$\begin{aligned} X \mid N &\sim \text{binomial}(N,p) \\ p &\sim \text{beta}(\alpha,\beta). \end{aligned}$$

The marginal distribution of X is

$$\begin{aligned}\mathbb{P}[X=x] &= \int_0^1 \binom{N}{x} p^x (1-p)^{N-x} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} dp \\ &= \frac{B(x+\alpha, N-x+\beta)}{B(\alpha,\beta)} \binom{N}{x}\end{aligned}$$

for $x = 0,...,N$. This is different from the binomial, and is known as the **beta-binomial** distribution. It is more dispersed than the binomial distribution. The beta-binomial is used occassionally in economics.

Variance Mixtures. The normal distributions has "thin tails" , meaning that the density decays rapidly to zero. We can create a random variable with thicker tails by a normal variance mixture. Consider the hierarchical model

$$\begin{aligned}X \mid Q &\sim \mathrm{N}(0, Q) \\ Q &\sim F\end{aligned}$$

for some distribution F such that $\mathbb{E}[Q] = 1$ and $\mathbb{E}\left[Q^2\right] = \kappa$. The first four moments of X are

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \mid Q]] = 0 \\ \mathbb{E}\left[X^2\right] &= \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Q\right]\right] = \mathbb{E}[Q] = 1 \\ \mathbb{E}\left[X^3\right] &= \mathbb{E}\left[\mathbb{E}\left[X^3 \mid Q\right]\right] = 0 \\ \mathbb{E}\left[X^4\right] &= \mathbb{E}\left[\mathbb{E}\left[X^4 \mid Q\right]\right] = \mathbb{E}\left[3Q^2\right] = 3\kappa.\end{aligned}$$

These calculations use the moment properties of the normal which will be introduced in the next chapter. The first three moments of X match those of the standard normal. The fourth moment is 3κ (see Exercise 2.16) while that of the standard normal is 3. Since $\kappa \geq 1$ this means that X can have thicker tails than the standard normal.

Normal Mixtures. The model is

$$\begin{aligned}X \mid T &\sim \begin{cases} \mathrm{N}(\mu_1, \sigma_1^2) & \text{if} \quad T = 1 \\ \mathrm{N}(\mu_2, \sigma_2^2) & \text{if} \quad T = 2 \end{cases} \\ \mathbb{P}[T=1] &= p \\ \mathbb{P}[T=2] &= 1-p.\end{aligned}$$

Normal mixtures are commonly used in economics to model contexts with multiple "latent types". The random variable T determines the "type" from which the random variable X is drawn. The marginal density of X equals the mixture of normals density

$$f\left(x \mid p_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\right) = p\phi_{\sigma_1}\left(x-\mu_1\right) + (1-p)\phi_{\sigma_2}\left(x-\mu_2\right).$$

4.25 Existence and Uniqueness of the Conditional Expectation*

In Section 4.14 we defined the conditional expectation separately for discrete and continuous random variables. We have explored these cases because these are the situations where the conditional expectation is easiest to describe and understand. However, the conditional expectation exists quite generally without appealing to the properties of either discrete or continuous random variables.

To justify this claim we present a deep result from probability theory. What it states is that the conditional expectation exists for all joint distributions (Y, X) for which Y has a finite expectation.

Theorem 4.26 Existence of the Conditional Expectation. Let (Y, X) have a joint distribution. If $\mathbb{E}|Y| < \infty$ then there exists a function $m(x)$ such that for all sets A for which $\mathbb{P}[X \in A]$ is defined,

$$\mathbb{E}[\mathbb{1}\{X \in A\} Y] = \mathbb{E}[\mathbb{1}\{X \in A\} m(X)]. \tag{4.8}$$

The function $m(x)$ is almost everywhere unique, in the sense that if $h(X)$ satisfies (4.8), then there is a set S such that $\mathbb{P}[S] = 1$, and $m(x) = h(x)$ for $x \in S$. The functions $m(x) = \mathbb{E}[Y \mid X = x]$ and $m(X) = \mathbb{E}[Y \mid X]$ are called the **conditional expectation**.

For a proof see Ash (1972) Theorem 6.3.3.

The conditional expectation $m(x)$ defined by (4.8) specializes to our previous definitions when (Y, X) are discrete or have a joint density. The usefulness of definition (4.8) is that Theorem 4.26 shows that the conditional expectation $m(x)$ exists for all finite-mean distributions. This definition allows Y to be discrete or continuous, for X to be scalar or vector-valued, and for the components of X to be discrete or continuously distributed.

4.26 Identification

A critical and important issue in structural econometric modeling is identification, meaning that a parameter is uniquely determined by the distribution of the observed variables. It is relatively straightforward in the context of the unconditional and conditional expectation, but it is worthwhile to introduce and explore the concept at this point for clarity.

Let F denote a probability distribution, for example the distribution of the pair (Y, X). Let $\mathscr{F}$ be a collection of distributions. Let θ be a parameter of interest (for example, the mean $\mathbb{E}[Y]$).

Definition 4.23 A parameter $\theta \in \mathbb{R}^k$ is **identified** on $\mathscr{F}$ if for all $F \in \mathscr{F}$ there is a unique value of θ.

Equivalently, θ is identified if we can write it as a mapping $\theta = g(F)$ on the set $\mathscr{F}$. The restriction to the set $\mathscr{F}$ is important. Most parameters are identified only on a strict subset of the space of distributions.

Take, for example, the mean $\mu = \mathbb{E}[Y]$. It is uniquely determined if $\mathbb{E}|Y| < \infty$, so it is clear that μ is identified for the set $\mathscr{F} = \left\{F : \int_{-\infty}^{\infty} |y| \, dF(y) < \infty\right\}$. However, μ is also defined when it is either positive or negative infinity. Hence, defining I_1 and I_2 as in (2.8) and (2.9), we can deduce that μ is identified on the set $\mathscr{F} = \{F : \{I_1 < \infty\} \cup \{I_2 > -\infty\}\}$.

Next, consider the conditional expectation. Theorem 4.26 demonstrates that $\mathbb{E}|Y| < \infty$ is a sufficient condition for identification.

Theorem 4.27 Identification of the Conditional Expectation. Let (Y, X) have a joint distribution. If $\mathbb{E}|Y| < \infty$ the conditional expectation $m(X) = \mathbb{E}[Y \mid X]$ is identified almost everywhere.

It might seem as if identification is a general property for parameters so long as we exclude degenerate cases. This is true for moments but not necessarily for more complicated parameters. As a case in point, consider censoring. Let Y be a random variable with distribution F. Let Y be censored from above, and Y^* defined by the censoring rule

$$Y^* = \begin{cases} Y & \text{if } Y \le \tau \\ \tau & \text{if } Y > \tau. \end{cases}$$

That is, Y^* is capped at the value τ. The variable Y^* has distribution

$$F^*(u) = \begin{cases} F(u) & \text{for } u \le \tau \\ 1 & \text{for } u \ge \tau. \end{cases}$$

Let $\mu = \mathbb{E}[Y]$ be the parameter of interest. The difficulty is that we cannot calculate μ from F^* except in the trivial case where there is no censoring $\mathbb{P}[Y \geq \tau] = 0$. Thus the mean μ is not identified from the censored distribution.

A typical solution to the identification problem is to assume a parametric distribution. For example, let $\mathscr{F}$ be the set of normal distributions $Y \sim \mathrm{N}(\mu, \sigma^2)$. It is possible to show that the parameters (μ, σ^2) are identified for all $F \in \mathscr{F}$. That is, if we know that the uncensored distribution is normal, we can uniquely determine the parameters from the censored distribution. This is called **parametric identification** as identification is restricted to a parametric class of distributions. In modern econometrics this is viewed as a second-best solution as identification has been achieved only through the use of an arbitrary and unverifiable parametric assumption.

A pessimistic conclusion might be that it is impossible to identify parameters of interest from censored data without parametric assumptions. Interestingly, this pessimism is unwarranted. It turns out that we can identify the quantiles $q(\alpha)$ of F for $\alpha \leq \mathbb{P}[Y \leq \tau]$. For example, if 20% of the distribution is censored we can identify all quantiles for $\alpha \in (0, 0.8)$. This is called **nonparametric identification** as the parameters are identified without restriction to a parametric class.

What we have learned from this little exercise is that in the context of censored data moments can only be parametrically identified while non-censored quantiles are nonparametrically identified. Part of the message is that a study of identification can help focus attention on what can be learned from the data distributions available.

4.27 Exercises

Exercise 4.1 Let $f(x, y) = 1/4$ for $-1 \leq x \leq 1$ and $-1 \leq y \leq 1$ (and zero elsewhere).

(a) Verify that $f(x, y)$ is a valid density function.

(b) Find the marginal density of X.

(c) Find the conditional density of Y given $X = x$.

(d) Find $\mathbb{E}[Y \mid X = x]$.

(e) Determine $\mathbb{P}\left[X^2 + Y^2 < 1\right]$.

(f) Determine $\mathbb{P}[|X + Y| < 2]$.

Exercise 4.2 Let $f(x, y) = x + y$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$ (and zero elsewhere).

(a) Verify that $f(x, y)$ is a valid density function.

(b) Find the marginal density of X.

(c) Find $\mathbb{E}[Y]$, $\mathrm{var}[X]$, $\mathbb{E}[XY]$ and $\mathrm{corr}(X, Y)$.

(d) Find the conditional density of Y given $X = x$.

(e) Find $\mathbb{E}[Y \mid X = x]$.

Exercise 4.3 Let

$$f(x, y) = \frac{2}{\left(1 + x + y\right)^3}$$

for $0 \leq x$ and $0 \leq y$.

(a) Verify that $f(x,y)$ is a valid density function.

(b) Find the marginal density of X.

(c) Find $\mathbb{E}[Y]$, $\text{var}[Y]$, $\mathbb{E}[XY]$ and $\text{corr}(X,Y)$.

(d) Find the conditional density of Y given $X = x$.

(e) Find $\mathbb{E}[Y \mid X = x]$.

Exercise 4.4 Let the joint PDF of X and Y be given by $f(x,y) = g(x)h(y)$ for some functions $g(x)$ and $h(y)$. Let $a = \int_{-\infty}^{\infty} g(x)dx$ and $b = \int_{-\infty}^{\infty} h(x)dx$.

(a) What conditions a and b should satisfy in order for $f(x,y)$ to be a bivariate PDF?

(b) Find the marginal PDF of X and Y.

(c) Show that X and Y are independent.

Exercise 4.5 Let $F(x,y)$ be the distribution function of (X,Y). Show that

$$\mathbb{P}[a < X \leq b, c < Y \leq d] = F(b,d) - F(b,c) - F(a,d) + F(a,c).$$

Exercise 4.6 Let the joint PDf of X and Y be given by

$$f(x,y) = \begin{cases} cxy & \text{if } x,y \in [0,1], x+y \leq 1 \\ 0 & \text{otherwise}. \end{cases}$$

(a) Find the value of c such that $f(x,y)$ is a joint PDF.

(b) Find the marginal distributions of X and Y.

(c) Are X and Y independent? Compare your answer to Problem 2 and discuss.

Exercise 4.7 Let X and Y have density $f(x,y) = \exp(-x-y)$ for $x > 0$ and $y > 0$. Find the marginal density of X and Y. Are X and Y independent or dependent?

Exercise 4.8 Let X and Y have density $f(x,y) = 1$ on $0 < x < 1$ and $0 < y < 1$. Find the density function of $Z = XY$.

Exercise 4.9 Let X and Y have density $f(x,y) = 12xy(1-y)$ for $0 < x < 1$ and $0 < y < 1$. Are X and Y independent or dependent?

Exercise 4.10 Show that any random variable is uncorrelated with a constant.

Exercise 4.11 Let X and Y be independent random variables with means μ_X, μ_Y, and variances σ_X^2, σ_Y^2. Find an expression for the correlation of XY and Y in terms of these means and variances.
Hint: "XY" is not a typo.

Exercise 4.12 Prove the following: If $(X_1, X_2, \ldots, X_m)$ are pairwise uncorrelated

$$\text{var}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \text{var}[X_i].$$

Exercise 4.13 Suppose that X and Y are jointly normal, i.e. they have the joint PDF:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2} - 2\frac{\rho xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right)\right).$$

(a) Derive the marginal distribution of X and Y and observe that both are normal distributions.

(b) Derive the conditional distribution of Y given $X = x$. Observe that it is also a normal distribution.

(c) Derive the joint distribution of (X, Z) where $Z = (Y/\sigma_Y) - (\rho X/\sigma_X)$, and then show that X and Z are independent.

Exercise 4.14 Let $X_1 \sim \text{gamma}(r,1)$ and $X_2 \sim \text{gamma}(s,1)$ be independent. Find the distribution of $Y = X_1 + X_2$.

Exercise 4.15 Suppose that the distribution of Y conditional on $X = x$ is $\text{N}(x, x^2)$ and the marginal distribution of X is $U[0,1]$.

(a) Find $\mathbb{E}[Y]$.

(b) Find $\text{var}[Y]$.

Exercise 4.16 Prove that for any random variables X and Y with finite variances:

(a) $\text{cov}(X,Y) = \text{cov}(X, \mathbb{E}[Y \mid X])$.

(b) X and $Y - \mathbb{E}[Y \mid X]$ are uncorrelated.

Exercise 4.17 Suppose that Y conditional on X is $\text{N}(X,X)$, $\mathbb{E}[X] = \mu$ and $\text{var}[X] = \sigma^2$. Find $\mathbb{E}[Y]$ and $\text{var}[Y]$.

Exercise 4.18 Consider the hierarchical distribution

$$X \mid Y \sim \text{N}(Y, \sigma^2)$$
$$Y \sim \text{gamma}(\alpha, \beta).$$

Find

(a) $\mathbb{E}[X]$. Hint: Use the law of iterated expectations (Theorem 4.13).

(b) $\text{var}[X]$. Hint: Use Theorem 4.14.

Exercise 4.19 Consider the hierarchical distribution

$$X \mid N \sim \chi^2_{2N}$$
$$N \sim \text{Poisson}(\lambda).$$

Find

(a) $\mathbb{E}[X]$.

(b) $\text{var}[X]$.

Exercise 4.20 Find the covariance and correlation between $a + bX$ and $c + dY$.

Exercise 4.21 If two random variables are independent are they necessarily uncorrelated? Find an example of random variables which are independent yet not uncorrelated.

Hint: Take a careful look at Theorem 4.8.

Exercise 4.22 Let X be a random variable with finite variance. Find the correlation between

(a) X and X.

(b) X and $-X$.

Exercise 4.23 Use Hölder's inequality (Theorem 4.15 to show the following.

(a) $\mathbb{E}\left|X^3 Y\right| \leq \mathbb{E}\left(|X|^4\right)^{3/4} \mathbb{E}\left(|Y|^4\right)^{1/4}$.

(b) $\mathbb{E}\left|X^a Y^b\right| \leq \mathbb{E}\left(|X|^{a+b}\right)^{a/(a+b)} \mathbb{E}\left(|Y|^{a+b}\right)^{b/(a+b)}$.

Exercise 4.24 Extend Minkowski's inequality (Theorem 4.16) to to show that if $p \geq 1$

$$\left(\mathbb{E}\left|\sum_{i=1}^{\infty} X_i\right|^p\right)^{1/p} \leq \sum_{i=1}^{\infty} \left(\mathbb{E}\,|X_i|^p\right)^{1/p}.$$

Appendix A

Mathematics Reference

A.1 Limits

Definition A.1 A sequence a_n has the **limit** a, written $a_n \to a$ as $n \to \infty$, or alternatively as $\lim_{n\to\infty} a_n = a$, if for all $\delta > 0$ there is some $n_\delta < \infty$ such that for all $n \geq n_\delta$, $|a_n - a| \leq \delta$.

Definition A.2 When a_n has a finite limit a we say that a_n **converges** or is **convergent**.

Definition A.3 $\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf_{m\geq n} a_m$.

Definition A.4 $\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup_{m\geq n} a_m$.

Theorem A.1 If a_n has a limit a then $\liminf_{n\to\infty} a_n = \lim_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$.

Theorem A.2 Cauchy Criterion. The sequence a_n converges if for all $\epsilon > 0$

$$\inf_m \sup_{j>m} \left| a_j - a_m \right| \leq \epsilon.$$

A.2 Series

Definition A.5 Summation Notation. The sum of $a_1, ..., a_n$ is

$$S_n = a_1 + \cdots + a_n = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} a_i = \sum_{1}^{n} a_i = \sum a_i.$$

The sequence of sums S_n is called a **series**.

Definition A.6 The series S_n is **convergent** if it has a finite limit as $n \to \infty$, thus $S_n \to S < \infty$.

Definition A.7 The series S_n is **absolutely convergent** if $\sum_{i=1}^{n} |a_i|$ is convergent.

Theorem A.3 Tests for Convergence. The series S_n is absolutely convergent if any of the following hold:

1. **Comparison Test**. If $0 \leq a_i \leq b_i$ and $\sum_{i=1}^{n} b_i$ converges.

2. **Ratio Test**. If $a_i \geq 0$ and $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} < 1$.

3. **Integral Test**. If $a_i = f(i) > 0$ where $f(x)$ is monotonically decreasing and $\int_1^\infty f(x)dx < \infty$.

Theorem A.4 Theorem of Cesaro Means. If $a_i \to a$ as $i \to \infty$ then $n^{-1}\sum_{i=1}^{n} a_i \to a$ as $n \to \infty$.

Theorem A.5 Toeplitz Lemma. Suppose w_{ni} satisfies $w_{ni} \to 0$ as $n \to \infty$ for all i, $\sum_{i=1}^{n} w_{ni} = 1$, and $\sum_{i=1}^{n} |w_{ni}| < \infty$. If $a_n \to a$ then $\sum_{i=1}^{n} w_{ni}a_i \to a$ as $n \to \infty$.

Theorem A.6 Kronecker Lemma. If $\sum_{i=1}^{n} i^{-1}a_i \to a < \infty$ as $n \to \infty$ then $n^{-1}\sum_{i=1}^{n} a_i \to 0$ as $n \to \infty$.

A.3 Factorial

Factorials are widely used in probability formulae.

Definition A.8 For a positive integer n, the **factorial** $n!$ is the product of all integers between 1 and n:

$$n! = n \times (n-1) \times \cdots \times 1 = \prod_{i=1}^{n} i$$

Furthermore, $0! = 1$.

A simple recurrance property is $n! = n \times (n-1)!$

Definition A.9 For a positive integer n, the **double factorial** $n!!$ is the product of every second positive integer up to n:

$$n!! = n \times (n-2) \times (n-4) \times \cdots = \prod_{i=0}^{\lceil n/2 \rceil - 1} (n-2i)$$

Furthermore, $0!! = 1$.

For even n

$$n!! = \prod_{i=1}^{n/2} 2i.$$

For odd n

$$n!! = \prod_{i=1}^{(n+1)/2} (2i-1).$$

The double factorial satisfies the recurrance property $n!! = n \times (n-2)!!$

The double factorial can be written in terms of the factorial as follows. For even and odd integers we have:

$$(2m)!! = 2^m m!$$
$$(2m-1)!! = \frac{(2m)!}{2^m m!}$$

A.4 Exponential

An exponential is a function of the form a^x. We typically use the name exponential function to refer to the function $e^x = \exp(x)$ where e is the exponential constant $e \simeq 2.718...$.

Definition A.10 The **exponential function** is $e^x = \exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$

Theorem A.7 Properties of the exponential function

1. $e = e^0 = \exp(0) = \sum_{i=0}^{\infty} \frac{1}{i!}$
2. $\exp(x) = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$
3. $(e^a)^b = e^{ab}$
4. $e^{a+b} = e^a e^b$
5. $\exp(x)$ is strictly increasing on $\mathbb{R}$, everywhere positive, and convex.

A.5 Logarithm

In probability, statistics, and econometrics the term "logarithm" always refers to the natural logarithm, which is the inverse of the exponential function. We use the notation "log" rather than the less description notation "ln".

Definition A.11 The **logarithm** is the function on $(0,\infty)$ which satisfies $\exp(\log(x)) = x$, or equivalently $\log(\exp(x)) = x$.

Theorem A.8 Properties of the logarithm

1. $\log(ab) = \log(a) + \log(b)$
2. $\log(a^b) = b\log(a)$
3. $\log(1) = 0$
4. $\log(e) = 1$
5. $\log(x)$ is strictly increasing on $\mathbb{R}_+$ and concave.

A.6 Differentiation

Definition A.12 A function $f(x)$ is **continuous** at $x = c$ if for all $\epsilon > 0$ there is some $\delta > 0$ such that $\|x - c\| \le \delta$ implies $\|f(x) - f(c)\| \le \epsilon$.

Definition A.13 The **derivative** of $f(x)$, denoted $f'(x)$ or $\frac{d}{dx} f(x)$, is

$$\frac{d}{dx} f(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

A function $f(x)$ is **differentiable** if the derivative exists and is unique.

Definition A.14 The **partial derivative** of $f(x,y)$ with respect to x is

$$\frac{\partial}{\partial x} f(x,y) = \lim_{h\to 0} \frac{f(x+h,y) - f(x,y)}{h}.$$

Theorem A.9 Chain Rule of Differentiation. For real-valued functions $f(x)$ and $g(x)$

$$\frac{d}{dx} f\left(g(x)\right) = f'\left(g(x)\right) g'(x).$$

Theorem A.10 Derivative Rule of Differentiation. For real-valued functions $u(x)$ and $v(x)$

$$\frac{d}{dx}\frac{u(x)}{v(x)} = \frac{v(x)u'(x) - u(x)v'(x)}{v(x)^2}.$$

Theorem A.11 Linearity of Differentiation

$$\frac{d}{dx}\left(ag(x) + bf(x)\right) = ag'(x) + bf'(x).$$

Theorem A.12 L'Hôpital's Rule. For real-valued functions $f(x)$ and $g(x)$ such that $\lim_{x\to c} f(x) = 0$, $\lim_{x\to c} g(x) = 0$, and $\lim_{x\to c} \frac{f(x)}{g(x)}$ exists, then

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \lim_{x\to c} \frac{f'(x)}{g'(x)}.$$

Theorem A.13 Common derivatives

1. $\frac{d}{dx} c = 0$
2. $\frac{d}{dx} x^a = ax^{a-1}$
3. $\frac{d}{dx} e^x = e^x$
4. $\frac{d}{dx} \log(x) = \frac{1}{x}$
5. $\frac{d}{dx} a^x = a^x \log(x)$.

A.7 Mean Value Theorem

Theorem A.14 Mean Value Theorem. If $f(x)$ is continuous on $[a,b]$ and differentiable on (a,b) then there exists a point $c \in (a,b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b-a}.$$

The mean value theorem is frequently used to write $f(b)$ as the sum of $f(a)$ and the product of the slope times the difference:

$$f(b) = f(a) + f'(c)\,(b-a)\,.$$

Theorem A.15 Taylor's Theorem. Let s be a positive integer. If $f(x)$ is s times differentiable at a, then there exists a function $r(x)$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(s)}(a)}{s!}(x-a)^s + r(x)$$

where

$$\lim_{x \to a} \frac{r(x)}{(x-a)^s} = 0.$$

The term $r(x)$ is called the **remainder**. The final equation of the theorem shows that the remainder is of smaller order than $(x-a)^s$.

Theorem A.16 Taylor's Theorem, Mean-Value Form. Let s be a positive integer. If $f^{(s-1)}(x)$ is continuous on $[a,b]$ and differentiable on (a,b), then there exists a point $c \in (a,b)$ such that

$$f(b) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(b-a)^2 + \cdots + \frac{f^{(s)}(c)}{s!}(b-a)^s.$$

Taylor's theorem is local in nature as it is an approximation at a specific point x. It shows that locally $f(x)$ can be approximated by an s-order polynomial.

Definition A.15 The **Taylor series expansion** of $f(x)$ at a is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

Definition A.16 The **Maclaurin series expansion** of $f(x)$ is the Taylor series at $a = 0$, thus

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k.$$

A necessary condition for a Taylor series expansion to exist is that $f(x)$ is infinitely differentiable at a, but this is not a sufficient condition. A function $f(x)$ which equals a convergent power series over an interval is called **analytic** in this interval.

A.8 Integration

Definition A.17 The **Riemann integral** of $f(x)$ over the interval $[a,b]$ is

$$\int_a^b f(x)dx = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} f\left(a + \frac{i}{N}(b-a)\right).$$

The sum on the right is the sum of the areas of the rectangles of width $(b-a)/N$ approximating $f(x)$.

Definition A.18 The **Riemann-Stieltjes** integral of $g(x)$ with respect to $f(x)$ over $[a,b]$ is

$$\int_a^b g(x)df(x) = \lim_{N \to \infty} \sum_{j=0}^{N-1} g\left(a + \frac{j}{N}(b-a)\right)\left(f\left(a + \frac{j+1}{N}(b-a)\right) - f\left(a + \frac{j}{N}(b-a)\right)\right).$$

The sum on the right is a weighted sum of the area of the rectangles weighted by the change in the function f.

Definition A.19 The function $f(x)$ is **integrable** on $\mathscr{X}$ if $\int_{\mathscr{X}} \left|f(x)\right| dx < \infty$.

Theorem A.17 Linearity of Integration

$$\int_a^b \big(cg(x) + df(x)\big)\, dx = c\int_a^b g(x)dx + d\int_a^b f(x)dx.$$

Theorem A.18 Common Integrals

1. $\int x^a dx = \dfrac{1}{a+1}x^{a+1} + C$
2. $\int e^x dx = e^x + C$
3. $\int \dfrac{1}{x} dx = \log|x| + C$
4. $\int \log x = x\log x - x + C.$

Theorem A.19 First Fundamental Theorem of Calculus. Let $f(x)$ be a continuous real-valued function on $[a,b]$ and define $F(x) = \int_a^x f(t)\,dt$. Then $F(x)$ has derivative $F'(x) = f(x)$ for all $x \in (a,b)$.

Theorem A.20 Second Fundamental Theorem of Calculus. Let $f(x)$ be a real-valued function on $[a,b]$ and $F(x)$ an antiderivative satisfying $F'(x) = f(x)$. Then

$$\int_a^b f(x)dx = F(b) - F(a).$$

Theorem A.21 Integration by Parts. For real-valued functions $u(x)$ and $v(x)$

$$\int_a^b u(x)v'(x)dx = u(b)v(b) - u(a)v(a) - \int_a^b u'(x)v(x)dx.$$

It is often written compactly as

$$\int u\,dv = uv - \int v\,du.$$

Theorem A.22 Leibniz Rule. For real-valued functions $a(x)$, $b(x)$, and $f(x,t)$

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(x,t)\,dt = f\left(x,b(x)\right)\frac{d}{dx}b(x) - f\left(x,a(x)\right)\frac{d}{dx}a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,dt.$$

When a and b are constants it simplifies to

$$\frac{d}{dx}\int_a^b f(x,t)\,dt = \int_a^b \frac{\partial}{\partial x} f(x,t)\,dt.$$

Theorem A.23 Fubini's Theorem. If $f(x,y)$ is integrable then

$$\int_{\mathscr{Y}}\int_{\mathscr{X}} f(x,y)\,dx\,dy = \int_{\mathscr{X}}\int_{\mathscr{Y}} f(x,y)\,dy\,dx.$$

Theorem A.24 Fatou's Lemma. If f_n is a sequence of non-negative functions then

$$\int \liminf_{n\to\infty} f_n(x)dx \le \liminf_{n\to\infty} \int f_n(x)dx.$$

Theorem A.25 Monotone Convergence Theorem. If $f_n(x)$ is an increasing sequence of functions which converges pointwise to $f(x)$ then $\int f_n(x)dx \to \int f(x)dx$ as $n \to \infty$.

Theorem A.26 Dominated Convergence Theorem. If $f_n(x)$ is a sequence of functions which converges pointwise to $f(x)$ and $\left|f_n(x)\right| \le g(x)$ where $g(x)$ is integrable, then $f(x)$ is integrable and $\int f_n(x)dx \to \int f(x)dx$ as $n \to \infty$.

A.9 Gaussian Integral

Theorem A.27 $\int_{-\infty}^{\infty} \exp\left(-x^2\right) dx = \sqrt{\pi}$.

Proof:

$$\begin{aligned}
\left(\int_0^{\infty} \exp\left(-x^2\right) dx\right)^2 &= \int_0^{\infty} \exp\left(-x^2\right) dx \int_0^{\infty} \exp\left(-y^2\right) dy \\
&= \int_0^{\infty}\int_0^{\infty} \exp\left(-\left(x^2+y^2\right)\right) dxdy \\
&= \int_0^{\infty}\int_0^{\pi/2} r\exp\left(-r^2\right) d\theta dr \\
&= \frac{\pi}{2}\int_0^{\infty} r\exp\left(-r^2\right) dr \\
&= \frac{\pi}{4}.
\end{aligned}$$

The third equality is the key. It makes the change-of-variables to polar coordinates $x = r\cos\theta$ and $y = r\sin\theta$ so that $x^2 + y^2 = r^2$. The Jacobian of this transformation is r. The region of integration in the (x, y) units is the positive orthont (upper-right region), which corresponds to integrating θ from 0 to $\pi/2$ in polar coordinates. The final two equalities are simple integration. Taking the square root we obtain

$$\int_0^{\infty} \exp\left(-x^2\right) dx = \frac{\sqrt{\pi}}{2}.$$

Since the integrals over the positive and negative real line are identical we obtain the stated result.

A.10 Gamma Function

Definition A.20 The **gamma function** for $x > 0$ is

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt.$$

The gamma function is not available in closed form. For computation numerical algorithms are used.

Theorem A.28 Properties of the gamma function

1. For positive integer n, $\Gamma(n) = (n-1)!$

2. For $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$.
3. $\int_0^\infty t^{x-1}\exp\left(-\beta t\right)dt = \beta^{-\alpha}\Gamma(x)$.
4. $\Gamma(1) = 1$.
5. $\Gamma(1/2) = \sqrt{\pi}$.
6. $\lim_{n\to\infty}\dfrac{\Gamma(n+x)}{\Gamma(n)\,n^x} = 1$.
7. **Legendre's Duplication Formula**: $\Gamma(x)\Gamma\left(x+\frac{1}{2}\right) = 2^{1-2x}\sqrt{\pi}\,\Gamma(2x)$.
8. **Stirling's Approximation**: $\Gamma(x+1) = \sqrt{2\pi x}\left(\dfrac{x}{e}\right)^x\left(1+O\left(\frac{1}{x}\right)\right)$ as $x\to\infty$.

Parts 1 and 2 can be shown by integration by parts. Part 3 can be shown by change-of-variables. Part 4 is an exponential integral. Part 5 can be shown by applying change of variables and the Gaussian integral. Proofs of the remaining properties are advanced and not provided.

A.11 Matrix Algebra

This is an abbreviated summary. For a more extensive review see Appendix A of *Econometrics*.

A **scalar** a is a single number. A **vector** a or $\boldsymbol{a}$ is a $k\times 1$ list of numbers, typically arranged in a column. We write this as

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

A **matrix** $\boldsymbol{A}$ is a $k\times r$ rectangular array of numbers, written as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

By convention a_{ij} refers to the element in the i^{th} row and j^{th} column of $\boldsymbol{A}$. If $r = 1$ then $\boldsymbol{A}$ is a column vector. If $k = 1$ then $\boldsymbol{A}$ is a row vector. If $r = k = 1$, then $\boldsymbol{A}$ is a scalar. Sometimes a matrix $\boldsymbol{A}$ is denoted by the symbol (a_{ij}).

The **transpose** of a matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}'$, $\boldsymbol{A}^\top$, or $\boldsymbol{A}^t$, is obtained by flipping the matrix on its diagonal. In most of the econometrics literature and this textbook we use $\boldsymbol{A}'$. In the mathematics literature $\boldsymbol{A}^\top$ is the convention. Thus

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Alternatively, letting $\boldsymbol{B} = \boldsymbol{A}'$, then $b_{ij} = a_{ji}$. Note that if $\boldsymbol{A}$ is $k\times r$, then $\boldsymbol{A}'$ is $r\times k$. If $\boldsymbol{a}$ is a $k\times 1$ vector, then $\boldsymbol{a}'$ is a $1\times k$ row vector.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $\boldsymbol{A} = \boldsymbol{A}'$, which requires $a_{ij} = a_{ji}$. A square matrix is **diagonal** if the off-diagonal elements are all zero, so that $a_{ij} = 0$ if $i \neq j$.

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. The $k \times k$ identity matrix is denoted as

$$\boldsymbol{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The **matrix sum** of two matrices of the same dimensions is

$$\boldsymbol{A} + \boldsymbol{B} = \left(a_{ij} + b_{ij}\right).$$

The product of a matrix $\boldsymbol{A}$ and scalar c is real is defined as

$$\boldsymbol{A}c = c\boldsymbol{A} = \left(a_{ij}c\right).$$

If $\boldsymbol{a}$ and $\boldsymbol{b}$ are both $k \times 1$, then their inner product is

$$\boldsymbol{a}'\boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = \sum_{j=1}^{k} a_j b_j.$$

Note that $\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a}$. We say that two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are **orthogonal** if $\boldsymbol{a}'\boldsymbol{b} = 0$.

If $\boldsymbol{A}$ is $k \times r$ and $\boldsymbol{B}$ is $r \times s$, so that the number of columns of $\boldsymbol{A}$ equals the number of rows of $\boldsymbol{B}$, we say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are **conformable**. In this event the matrix product $\boldsymbol{AB}$ is defined. Writing $\boldsymbol{A}$ as a set of row vectors and $\boldsymbol{B}$ as a set of column vectors (each of length r), then the **matrix product** is defined as

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_s \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_s \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_s \end{bmatrix}.$$

The **trace** of a $k \times k$ square matrix $\boldsymbol{A}$ is the sum of its diagonal elements

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} a_{ii}.$$

A useful property is

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}).$$

A square $k \times k$ matrix $\boldsymbol{A}$ is said to be **nonsingular** if there is no $k \times 1$ $\boldsymbol{c} \neq 0$ such that $\boldsymbol{Ac} = 0$.

If a square $k \times k$ matrix $\boldsymbol{A}$ is nonsingular then there exists a unique $k \times k$ matrix $\boldsymbol{A}^{-1}$ called the **inverse** of $\boldsymbol{A}$ which satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k.$$

For non-singular $\boldsymbol{A}$ and $\boldsymbol{C}$, some useful properties include

$$\left(\boldsymbol{A}^{-1}\right)' = \left(\boldsymbol{A}'\right)^{-1}$$
$$(\boldsymbol{AC})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{A}^{-1}.$$

A $k \times k$ real symmetric square matrix $\boldsymbol{A}$ is **positive semi-definite** if for all $\boldsymbol{c} \neq 0$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0$. This is written as $\boldsymbol{A} \geq 0$. A $k \times k$ real symmetric square matrix $\boldsymbol{A}$ is **positive definite** if for all $\boldsymbol{c} \neq 0$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} > 0$. This is written as $\boldsymbol{A} > 0$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are each $k \times k$, we write $\boldsymbol{A} \geq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B} \geq 0$. This means that the difference between $\boldsymbol{A}$ and $\boldsymbol{B}$ is positive semi-definite. Similarly we write $\boldsymbol{A} > \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B} > 0$.

Many students misinterpret "$\boldsymbol{A} > 0$" to mean that $\boldsymbol{A} > 0$ has non-zero elements. This is incorrect. The inequality applied to a matrix means that it is positive definite.

Some properties include the following:

1. If $\boldsymbol{A} = \boldsymbol{G}'\boldsymbol{B}\boldsymbol{G}$ with $\boldsymbol{B} \geq 0$ then $\boldsymbol{A} \geq 0$.
2. If $\boldsymbol{A} > 0$ then $\boldsymbol{A}$ is non-singular, $\boldsymbol{A}^{-1}$ exists, and $\boldsymbol{A}^{-1} > 0$.
3. If $\boldsymbol{A} > 0$ then $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}'$ where $\boldsymbol{C}$ is non-singular.

The **determinant**, written $\det \boldsymbol{A}$ or $|\boldsymbol{A}|$, of a square matrix $\boldsymbol{A}$ is a scalar measure of the transformation $\boldsymbol{A}\boldsymbol{x}$. The precise definition is technical. See Appendix A of *Econometrics* for details. Useful properties are:

1. $\det \boldsymbol{A} = 0$ if and only if $\boldsymbol{A}$ is singular.
2. $\det \boldsymbol{A}^{-1} = \dfrac{1}{\det \boldsymbol{A}}$.

References

[1] Amemiya, Takeshi (1994): *Introduction to Statistics and Econometrics*, Harvard University Press.

[2] Andrews, Donald W.K. (1994): "Empirical process methods in econometrics", in *Handbook of Econometrics, Volume 4*, ch 37. Robert F. Engle and Daniel L. McFadden, eds., 2247-2294, Elsevier.

[3] Ash, Robert B. (1972): *Real Analysis and Probability*, Academic Press.

[4] Billingsley, Patrick (1995): *Probability and Measure,* Third Edition, New York: Wiley.

[5] Billingsley, Patrick (1999): *Convergence of Probability Measure,* Second Edition, New York: Wiley.

[6] Bock, Mary Ellen (1975): "Minimax estimators of the mean of a multivariate normal distribution," *The Annals of Statistics*, 3, 209-218.

[7] Casella, George and Roger L. Berger (2002): *Statistical Inference*, Second Edition, Duxbury Press.

[8] Efron, Bradley (2010): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction,* Cambridge University Press.

[9] Epanechnikov, V. I. (1969): "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Application*, 14, 153-158.

[10] Gallant, A. Ronald (1997): *An Introduction to Econometric Theory*, Princeton University Press.

[11] Gosset, William S. (a.k.a. "Student") (1908): "The probable error of a mean," *Biometrika*, 6, 1-25.

[12] Hansen, Bruce E. (2022): *Econometrics*, Princeton University Press, forthcoming.

[13] Hansen, Bruce E. (2021): "A modern Gauss-Markov theorem," *Econometrica*, forthcoming.

[14] Hodges Joseph L. and Erich L. Lehmann (1956): "The efficiency of some nonparametric competitors of the t-test," *Annals of Mathematical Statistics*, 27, 324-335.

[15] Hogg, Robert V. and Allen T. Craig (1995): *Introduction to Mathematical Statistics*, Fifth Edition, Prentice Hall.

[16] Hogg, Robert V. and Elliot A. Tanis (1997): *Probability and Statistical Inference*, Fifth Edition, Prentice Hall.

[17] James, W. and Charles M. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361-380.

[18] Jones, M. C. and S. J. Sheather (1991): "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives," *Statistics and Probability Letters*, 11, 511-514.

[19] Koop, Gary, Dale J. Poirier, Justin L. Tobias (2007): *Bayesian Econometric Methods*, Cambridge University Press.

[20] Lehmann, Erich L. and George Casella (1998): *Theory of Point Estimation*, Second Edition, Springer.

[21] Lehmann, Erich L. and Joseph P. Romano (2005): *Testing Statistical Hypotheses*, Third Edition, Springer.

[22] Li, Qi and Jeffrey Racine (2007): *Nonparametric Econometrics*.

[23] Linton, Oliver (2017): *Probability, Statistics, and Econometrics*, Academic Press.

[24] Mann, Henry B. and Abraham Wald (1943): "On stochastic limit and order relationships," *The Annals of Mathematical Statistics* 14, 217-736.

[25] Marron, James S. and Matt P. Wand (1992): "Exact mean integrated squared error," *The Annals of Statistics*, 20, 712-226.

[26] Pagan, Adrian and Aman Ullah (1999): *Nonparametric Econometrics*, Cambridge University Press.

[27] Parzen, Emanuel (1962): "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, 33, 1065-1076.

[28] Pollard, David (1990): *Empirical Processes: Theory and Applications*, Institute of Mathematical Statistics.

[29] Ramanathan, Ramu (1993): *Statistical Methods in Econometrics*, Academic Press.

[30] Rosenblatt, Murrey (1956): "Remarks on some non-parametric estimates of a density function," *Annals of Mathematical Statistics*, 27, 832-837.

[31] Rudin, Walter (1976): *Principles of Mathematical Analysis*, Third Edition, McGraw-Hill.

[32] Rudin, Walter (1987): *Real and Complex Analysis*, Third Edition, McGraw-Hill.

[33] Scott, David W. (1992): *Multivariate Density Estimation*, Wiley.

[34] Shao, Jun (2003): *Mathematical Statistics*, Second Edition, Springer.

[35] Sheather, Simon J. and M. C. Jones (1991): "A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

[36] Silverman, Bernard W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[37] van der Vaart, Aad W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[38] van der Vaart, Aad W. and Jon A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

[39] Wasserman, Larry (2006): *All of Nonparametric Statistics*, New York: Springer.

[40] White, Halbert (1982): "Instrumental variables regression with independent observations," *Econometrica*, 50, 483-499.

[41] White, Halbert (1984): *Asymptotic Theory for Econometricians,* Academic Press.