Introduction to Probability and Statistics for Economists

by

Perry Shapiro

* ©Perry Shapiro – September 2005
Table of Content

Division 1

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>Division 1 Problem Set</td>
<td></td>
</tr>
<tr>
<td>Chapter 1</td>
<td>Probability</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Univariate Distribution</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Multivariate Distribution</td>
</tr>
</tbody>
</table>

Division 2

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division 2 Problem Set</td>
<td></td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Sampling</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Estimation</td>
</tr>
</tbody>
</table>
Introduction

Most students begin advanced studies with some knowledge of probability and statistics. An undergraduate degree in Economic and Finance require rudimentary understanding of basic probabilistic and statistical thinking. For those students with only a basic understanding of these subjects, these notes are meant to enhance your understanding in preparation for the rigors of modern Economics and Finance. Some students who enter advanced programs have studied probability and statistics in considerably more depth than is required for a standard undergraduate degree. For these students, the notes are intended to be a brief refresher of topics already mastered.

The treatment of most subjects in this book is different from those encountered elsewhere. In most text it is common to work from the general to the specific. For probability axioms are stated; properties deduced (usually through a set of rigorously proven theorems); illustrative examples presented followed and finally problems that further highlight the applicability of the abstract principles are offered to challenge and further enlighten the reader. In this text, the examples are put at the beginning of a section and, sometimes, solutions are proposed without first deriving the background principles. This back-forward approach is chosen to simulate the work of Economists. Most of us confront economic questions and then search for approaches to shed light on the issues. If we are lucky, we already have the technical tools to solve the problems that arise, but most often, particularly if we are engaged in original research, we are insufficiently equipped to make much progress. In that case we do what we can to understand the issues and then search for whatever technical tools will help us in the work of mathematicians, probabilist and statisticians. While the beauty of the mathematical abstractions may impress us, rigorous beauty is not our aim; instead we seek understanding of something fundamental about the economy or about the nature of economic behavior.

Even though these note stress economic applications the probabilistic and statistical subjects are treated with some rigor. It is my belief that with rigor come a deep understanding of the subject. While it is not our aim to prepare students for careers as a probabilist or statistician, the tools of both probability and statistics will become an integral part of the future professional life. While it is possible to use the tools with little knowledge of their workings (the black box approach has become increasingly available bundles in slick software packages), however, those who choose the 'whatever looks right' approach are unable to evaluate whether or not the appropriate analyses have been performed. This technician view leaves the user vulnerable to inappropriate application. For this reason, proofs, accompany most propositions in the texts. We urge the students to work at the proofs long enough so that they are understood. Attention paid to this part of the text will have its payoff in a deeper understanding and smarter use of the material.

The fundamental feature of the derivation of probability is an event and event space. Probability is a mapping from the event space, to be defined within the text, with the property that it is non-negative and that it sums to one. It is most familiarly understood, as it relates to our own experiences, as discrete values. These are subject to counting rules and probability computed as relative frequencies. Much insight can be
gleaned from the analysis of discrete probability measures. For this we concentrate on the binomial distribution. The events are represented by the outcome of a fixed number of coin flips. The set of all possible outcomes is the sample space. An event, for instance, head on the first flip of the coin, is a subset of the sample space. The set of all subsets is the event space. The entire events space is easily represented for a simple two coin flips, and many of the important properties are examined with this example.

Random variables, defined as functions from the sample space to the reals, are introduced and their distribution defined and discussed. Again, the concept is made concrete with reference to binary outcomes.

Following the material on discrete outcomes, is the more complicated subject of continuous ones. It is not surprising that there much in common that is exploited in the explanation. Rather than speak of events as discrete outcomes they are characterized as sets. Whereas the random variable in the discrete case are mappings from the sample space to the rationals in the unit interval from zero to one, they are now mapping from subsets of the sample space to subintervals of that same unit interval. It is a complicated and technical issue that probability measures cannot be attached to all event spaces. These problems are touched on here, so that the students are familiar with problems that may arise. Nonetheless, the coverage is not deep for there are issues that far exceed the ambitions of this text.

There is much that is done in the probability part of the text that anticipates the material on statistic that follows. This isn't too surprising since probability is fully integrated with statistics. While the part of statistics deals with much that is familiarly thought to be that subject, sampling, hypothesis testing and estimation, some of the issues, particularly those dealing with hypothesis testing, are integrated in the first part of the text. We stress, the testing of hypotheses, whether this be with traditional or Bayesian methods.

There is a rough division of the material. One of the divisions is probability, that is the topic of the first three chapters and the other division is on sampling and estimation. Whereas the second division most closely relates to subsequent, more advanced and detailed work on Econometrics, the first division relates to issues in Economic Theory as well as Econometrics. At the beginning of the first division, is a set of problems covering the topics of the entire division. A similar set of problems precede the last two chapters.
Division 1 Problems

The first six problems refer to the asset value example in the notes. The important aspect of that example is that there are only two possible year-to-year changes in value.

1.1) What are the potential asset values in a sixth year and how many pathways (from year 0) are there to the sixth year prices?
1.2) If the prevailing discount rate is 7.5%, what are the present values of the various fifth and sixth year asset values?
1.3) If the asset price either increases by $0.50 or decreases by $0.25 in any one year, what are the potential value pathways? For this example, prices cannot be negative. If the price drops to zero in any one period it remains at zero thereafter.
1.4) Compute the potential values of the asset if the year-to-year changes are either 10% increase or 5% decline after ten years.
1.5) Repeat the previous problem if the asset price can increase by $0.50 or decline by $0.25 every year.
1.6) How many paths are there to the different tenth year values in both the previous problems?

1.7) There are two famous problems in probability. It is arguable which one inspired the entire field. One of the problems was sent by Samuel Pepys to Issac Newton. Here is that problem:

   This sayd, and with great Truth & Respect; I goe on to tell you, that ye Bearer Mr. Smith is One I beare great goodwill to, noe less for what I personally know of his general Ingenuity, Industry, and Virtue, than for ye general Reputation he has in this Towne (inferior to none, but superiour to most) for his Maistry in the two Points of his Profession, namely, Faire-Writeing and Arithmetick, soe farr (principally) as is subervient to Accountantship. Now soe it is, That ye late Project (of which you cannot but have heard) of Mr Neale ye Groom-Porter his Lottery, has almost extinguish'd for some time all places of publick Conversation in this Towne, especially among Men of Numbers, every other Talk but what relates to ye Doctrine of determining between ye true proportions of the Hazards incident of this or that given Chance or Lot.

   The Question:
   A – has 6 dice in a Box, wth wch he is to fling a 6
   B – has in another Box 12 Dice, wth wch he is to fling 2 Sixes
   C – has in another Box 18 Dice, wth wch he is to fling 3 Sixes
   Q. whether B & C have not as easy a Taske as A, at even luck?

   How should Newton answer Pepys?

1.8) The second problem that competes for the foundational honor is one given by the Chevalier de Me're to Blaise Pascal. The problem is what is more likely:

   A) Rolling at least one six in four rolls of a single die
   B) Rolling at least one double six in 24 rolls of a pair of dice.
de Me're reasoned that
The chance of one six = 1/6
Average for four rolls 4*(1/6) = 2/3

The chance of a double six = 1/36
Average for twenty four rolls 24*(1/36) = 2/3.

The Chevalier was making bets based on his calculation and loosing a great deal of money. What is wrong with his reasoning?

1.9) There are 20 people in a room. What is the probability that at least two of them have the same birthday? There are n < 366 people in a room, what is the probability that at least two of them have the same birthday?

1.10) You have a dollar. A fair coin is flipped and if it comes up heads you will receive an additional dollar. If tails appears, you loose a dollar. You can stay in the game as long as you have at least one dollar. The coin will be flipped four times with the same payoff rules.

a) What is the probability that you must exit the game after the first flip?
b) What is the probability that you must exit the game after the second flip?
c) What is the probability that you must exit the game after the third flip?
d) What is the probability that you will be in the game for the fourth flip?
e) If you survive to the fourth flip, what is the expected value of your earnings?

1.11) There is a stone in square 0 (see below). You will draw from an urn that contains 3 red ball and 7 green ones. If you draw a red ball you move the stone one square to the left and if you draw a green ball you move it two squares to the right. You return the drawn ball to the urn after you move the stone.

| -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

Let $X_n$ be the number of the square where the stone lays after n draws from the urn.

a) What are the possible values for $X_n$ for n = 1,2,3 and 4?
b) What are the density function, mean and variance of $X_n$ for n = 1,2,3 and 4?
c) You are to be paid $Y_n=10X_n$ (negative amounts mean that you pay). What is the distribution (density) of $Y_n$ for n = 1,2,3 and 4? What are the expected value and variance of $Y_n$?
d) Suppose the stones are to be drawn one per year and the prevailing interest rate is 10%
   (i) What is the market value of the game?
(ii) What is the market value (before the first draw) to buy $Y_4$ for $50 after the 4th draw?

1.12) Use the same urn as in the previous problem and draw one ball at a time. If it is green you receive $1 and if it is red you pay $1. You begin with an initial amount of $1. You continue to draw balls until either you have no more money or you have drawn four balls. If you lose all of your money the drawing terminates.
   a) What is the probability that you are allowed to draw only one ball?
   b) What is the probability that you will make four draws?
   c) If you do make a fourth draw, what is the expected value of your earnings (total amount less initial $1)?
   d) What is the expected value of the game?

1.13) A game is proposed: a fair coin will be tossed in the air, if it lands heads the game stops and you receive nothing. If it lands tails you receive $1 and the coin is tossed again and again if heads appears the game stops and you get no more money but if the second toss is tails you are given another dollar. The game continues like that until the coin lands on heads. What is the probability that you end up with
   a) nothing?
   b) one dollar?
   c) two dollars?
   d) n dollars?
   e) at least one dollar?
   f) at least two dollars?
   g) at least n dollars?

1.14) The rate of interest (both borrowing and lending) is 5%. It is know that every year the value of the security will increase by 20% or decrease by 10%. Each of these events are equally likely every year.
   a) What is the value of the security today?
   b) What is today’s value of an option to buy the asset in two years for $120?
   c) What is today’s value of an option to buy the asset in two years for $100?

1.15) If the interest rate is 15% and we know that every year either a security increase in value by 30% or declines in value by 10%. What can we say about the (subjective) probability that a risk neutral investor assigns to an annual increase?

1.16) The rate of interest (both borrowing and lending) is 5%. It is know that every year the market price of a security valued at $100 today will increase by 20% or decrease by 10%. Each of these events is equally likely every year.
   a) What is the expected value of the security price at the end of the first year?
   b) What is today’s value of an option to buy the asset in two years for $120?
   c) What is today’s value of an option to buy the asset in two years for $100?
1.17) The beginning of the year market price of another security is $110 and the rate of interest is 10%. The probability that the price will increase by 40% any year is $p = 0.70$ and the probability that the price will fall by 60% is $q = 0.30$.

Part A: What is the expected value of the price of the security at the end of
   a) The first year?
   b) The second year?
   c) The third year?
   d) The $T^{th}$ year? $T \geq 1$.

Part B: What is the market value of an option to purchase the asset for $110 at the end of
   a) The first year?
   b) The second year?
   c) The third year?
   d) The $T^{th}$ year? $T \geq 1$

1.18) If the interest rate is 15% and we know that every year either a security increase in value by 30% or declines in value by 10%. What can we say about the (subjective) probability that a risk neutral investor assigns to an annual increase?

1.19) The power of a test is defined as the probability of rejecting the hypothesis (concluding that it is false). With this definition in mind consider the following test of an hypothesis:

   It is hypothesised that an urn contains an equal number of red and green balls. A random sample of size 10 is drawn, with replacement, from the urn. The test proposed is to reject the hypothesis if the number of balls drawn is either less than or equal to three or greater than or equal to seven. Evaluate the power of this test if the true proportion of red balls is 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 or 0.9.

1.20) You are setting air pollution standards for the country of Euphoria. Euphoria shares an air basin with the neighboring country of Wonderwhat. Recognizing the interdependence between the air quality choices of the two countries you formalize a treaty between Euphoria and Wonderwhat in which you both agree to limit your air pollution emissions to 5 each. If you both stick to the terms of the treaty the measured air pollution in the air basin will be the sum of what you both produce, 10, plus a random number $u$, random variable that can take on one of the following value $-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$ with probability $0.15 - (1/50)|u|$. If Wonderwhat does not abide by the treaty terms, its emissions will be 8 and the common level of air pollution will be $13 + u$. You decide to abide by the treaty.

   a) What are the possible levels of pollution in the common air basin.
   b) What are the probabilities of each level of pollution given that Wonderwhat abides by the treaty?
   c) What are the probabilities of each level of pollution if Wonderwhat does not abide by the treaty?
d) For each of the possible value of measured air pollution, what probability would you assign to Wonderwhat having abided by the treaty?

Suppose initially you believe there is a 0.75 probability that Wonderwhat will abide by the treaty terms.

e) After the agreement is signed you find air basin pollution is 9. How does this change your assessment of Wonderwhat's reliability?

f) Suppose you take a series of daily air pollution readings and they are as follows:

<table>
<thead>
<tr>
<th>Day</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

What probability do you assign to Wonderwhat's reliability after the fifth day?

g) Answer the same question for this set of observations

<table>
<thead>
<tr>
<th>Day</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

h) Answer the same question again for these observations

<table>
<thead>
<tr>
<th>Day</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
</tbody>
</table>

1.21) Remember Ann, she offers you the chance to gamble and she chooses the box from which the coin is taken. For two flips in a row the outcome is heads and you lose $1. Suppose the outcome is tails on the second flip and you choose to continue the gamble. The next flip is heads and you lose another dollar. Do you continue to a fourth try? If you do accept a fourth try and again heads is the outcome, do you accept an offer for a fifth try?

1.22) Ben appears as honest as his cousin Ann – a probability of $p = 0.9$ is assigned to his honesty. He has one coin that he asserts is fair and as proof of that he flips it ten times. Seven out of the ten tries come up heads. After this what probability do you assign (what is your posterior probability) to Ben's honesty?

1.23) I tell you that there are as many red balls as green ones in an urn. The alternative is that 80% of the balls are red. You consider me a shifty character and assign a probability of 0.2 to my honesty.

a) I draw out ten balls with replacement and there are an equal number of red and green balls. What is your new honesty probability assignment?
b) After you have re-evaluated the probability of honesty, I draw six more balls, and all but one is red. What is your new assessment of my honesty probability?

1.24) The Grungy Scum – a popular baroque-funk-rap-romance group – are to perform in Adelaide. As you know the Grungy-Scum is an enormously popular group, so it is strange that the promoter chose a venue with a capacity of 10. It is announced that the first 10 people in the queue will be admitted to the concert, all others will be excluded. Fans begin to arrive outside the door to the venue at 9:00AM as a Poisson process with expected arrival rate of \( \lambda = 8 \) per minute. You arrive at the venue at exactly 9:01AM. What is the probability that you will be admitted to the concert?

1.25) What are the (i) moment generating functions; (ii) expected values; (iii) variances and (iv) first, second and third moments of
   a) The Poisson distribution
   b) The uniform (both discrete and continuous) distribution
   c) The exponential distribution

1.26) Consider the random variable \( X \) with density function

\[
f_X(x) = \begin{cases} 
  k(1 + x); & x \in [-1,0) \\
  k(1 - x); & x \in [0,1] \\
  0, & \text{otherwise}
\end{cases}
\]

a) What is the value of \( k \)?

b) What is its moment generating function?

c) What are its mean, variance and third moment?

1.27) Provide a proof for the following:
   i) \( E(c) = c \) where \( c \) is a constant.
   ii) \( E(cg(X)) = cE(g(X)) \).
   iii) \( E(cg(X) + ah(X)) = cE(g(X)) + aE(h(X)) \), where \( a \) and \( c \) are constants and \( g(\ ) \) and \( h(\ ) \) are functions.
   iv) If \( g(x) \leq h(x) \) for all values of \( x \), the \( E(g(x)) \leq E(h(x)) \).

1.28) Demonstrate that the variance can be computed as \( V(X) = E(X^2) - \mu_X^2 \equiv \sigma_X^2 \) for both continuous and discrete distributions.

1.29) Provide a proof of Theorem 2.1 for discrete distributions.

1.30) Provided a proof of Theorem 2.2 (Jensen's Inequality) for convex functions.

1.31) Show that \( V(X) = E(X^2) - \mu_X^2 \equiv \sigma_X^2 \).
1.32) Let \( X \) be a random variable with expected value \( \mu \) and variance \( \sigma^2 = 1 \). Prove, or disprove the following statement:
"The probability is less than 0.25 that the absolute value of the difference between \( X \) and \( \mu \) is no smaller than 2."

1.33) You are attempting to estimate the expected value of a random variable \( X \) by computing the sample mean. The variance of \( X \) is equal to 1.
   a) How large a sample must you collect in order that the probability that the sample mean will be within 1 of the expected value is at least 0.90?
   b) You want the probability to be at least 0.90 that the sample mean is within 0.5 of the expected value. How large a sample must you collect?

1.34) Harry's preference for wealth is represented by the utility function
\[ U = W^{1/2} \]
where \( W \) is Harry's wealth. The value of Harry's wealth is $10,000, but with probability 0.20 he will lose $7,500. What is the most that Harry will pay for insurance that will fully compensate him for his loss? Explain this with reference to Jensen's inequality.

1.35) Mary's wealth is $40,000. With probability 0.1 she will lose $30,000. Mary's utility function is
\[ U = W^{1/2} \]
   a) What is Mary's expected utility?
   b) What amount, with certainty, would give Mary the same expected utility?
   c) What is the maximum Mary would pay to insure against the loss?
   d) An insurer offers her to fully insure her against the loss for $3,000. Does she accept the offer?
   e) There are millions of people exactly like Mary with the same preference functions, the same initial wealth and the same potential for loss. Analyse the market for insurance that fully insures against loss for $3,000.

1.36) Grace is endowed with wealth of $3,000,000. With probability 0.01 she will suffer a loss of $2,000,000. Grace's utility is equal to the square root of her wealth
\[ U = W^{1/2} \]
   a) What is her expected utility?
   b) Will she purchase insurance against the loss if it cost $36,000?
   c) Ten thousand exactly similar individuals are insured by one company and each pays $36,000 for the insurance. The probability that the revenues collected by the company exceed the payout for loss is smaller than 0.25. True or False?

1.37) \( X \), a random variable, can have one of five values: 1, 2, 3, 4 or 5. The probability density function for \( X \) is
\[ f_X(x) = A/2^x. \]
What is the value of \( A \)?
1.38) The following problems are to be done using Excel.
   1.38a) X is has a uniform density in the interval [-1,+1]. Plot the density function.
   1.38b) X is a normally distributed random variable with a zero mean. On the same graph plot the density function with variance 1, 0.25 and 4.
   1.38c) X is a random variable with a logistic density. It's cumulative density is \( F_X(x) = \frac{1}{1 + e^{-\beta x}} \). In this form X has mean zero and variance \( \frac{\pi^2}{3\beta^2} \).
     
     i) What is the density function of X?
     ii) What values of \( \beta \) give variance 1, 0.25 and 4?
     iii) For each of these \( \beta \) values plot the logistic density on the same graph.
     iv) For each of these \( \beta \) values plot the density on the same graph as the normal density with the same variance. What can you say about the comparison of the normal and logistic distributions?

1.39) These problems are simulations to be done using Excel.
   1.39a) X is a random variable with a uniform density over the [0,1] interval. With one-thousand draws plot the distribution of outcomes. Compare the plot with the graph of the "true" density function.
   1.39b) Y is a random variable with an exponential density with parameter \( \lambda = 2 \). Using one-thousand draws plot the distribution of the outcomes. Compare the plot with the graph of the "true" density function.
   1.39c) Z is a random variable with density 2(1 – x) for x in [0,1] interval and 0 otherwise. Use one-thousand draws and plot the distribution of the outcomes. Compare the plot with the graph of the "true" density function.

1.40) The following are experiments with the Central Limit Theorem. Remember the central limit theorem is about the limiting distribution of the normalized sample mean. These experiments, as the previous problems, are to be done using Excel. You should plot the distribution of sample means from samples of size 10, 100 and 1000 for each of the following distributions
   1.40a) Exponential density with parameter \( \lambda = 1 \).
   1.40b) Uniform density on the interval [0,1]
   1.40c) Bernoulli density with success probability \( p = 0.4 \)
   From inspection of the plots, evaluate whether or not the limiting distribution of the three densities is standard normal. If not why does the CLT fail in the particular case.

1.41) For the bivariate random variable with joint density \( f_{XY}(x,y) = 2x + 2y - 4xy \), compute the expected value and variance of each of the two random variables; the conditional expected value of each; the covariance and the correlation coefficient. Determine if they are independently distributed.

1.42) Prove the following 5 propositions:
1.42a) If X and Y are independent random variables with finite variances the
C(XY) = ρ(XY) = 0

1.42b) X is a random variable with finite variance and Y = a + bX. If b > 0 then
ρ(XY) = 1 and if b < 0 ρ(XY) = -1.

1.42c) If X and Y have finite variance then V(X + Y) = V(X) + V(Y) + 2C(XY)
and V(X − Y) = V(X) + V(Y) − 2C(XY).

1.42d) Let X₁ ... Xₘ and Y₁ ... Yₙ be random variables and C(XᵢYⱼ) exist for all
i and j. Let a₁ ...aₘ and b₁ ... bₙ be constants, then

\[ C(\sum_{i=1}^{m} a_i X_i \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j C(X_i Y_j). \]

1.42e) Let X₁ ... Xₘ be random variables each with a finite variance. Then

\[ V(\sum_{i=1}^{m} X_i) = \sum_{i=1}^{m} V(X_i) + 2\sum_{i<j} C(X_i X_j). \]

You can see from this that the variance of a sum of random variables is equal
to the sum of the variances if and only if their covariances (and thus their
correlation coefficients) are all zero.

1.43) Consider the following density function of two random variables X and Y:

\[ f_{XY}(-1,0) = 1/3 \quad f_{XY}(0,1) = 1/3 \quad f_{XY}(1,0) = 1/3 \quad f_{XY}(x, y) = 0 \text{ otherwise} \]

Evaluate the following two statements:

a) If two random variables are independently distributed their correlation
   coefficient is zero.

b) If the correlation coefficient for two random variables is zero, the random
   variables are independently distributed.

1.44) X and Y are normally distributed random variables both have expected value zero
and variance equal to 1. X and Y are independently distributed if and only if their
   correlation coefficient is zero.

1.45) The joint density function for the random variables X and Y is

\[ f_{XY}(x, y) = \frac{1}{20\pi} \exp\left(-\frac{(x-10)^2}{50}\right)\exp\left(-\frac{(y-100)^2}{8}\right). \]

a) Confirm the f_{XY} is a density function.
b) What are the means and variances of X and Y?

c) What is the covariance and correlation coefficient between X and Y?

d) Are X and Y independently distributed?

e) For X = 10, what are the mean and variance of Y?

f) What are the mean and variance of Y for X = 0 and X = 15?

g) What is the conditional density function for Y given X?

h) Draw the graph of the mean of Y as a function of X.

i) On that graph draw the density function of Y for X = -10, 0, 10.

j) What about the shape of the density for the various values of X?
Chapter 1

Probability
Asset Price

An asset with a value of $1 today will fluctuate in value from year to year. From one year to the next the asset price will either increase by 10% or decline by 5%. For any given year it is uncertain whether there will be an increase or a decline. All the possible prices for five years are illustrated in Figure 1. The possible paths are indicated by the arrows.

Figure 1
Asset Values Pathways for Five Years

Notice that after five years there are six possible prices for the asset. If there are five successive increases the asset price is $1.61 \((1.1^5)\) and if there is bad luck and there
are five successive down year the price is $0.77 (0.95^5). Intermediate values are the result of combinations of good and bad years. It is important to note that the final values depend only on the total number of good and bad years, not on the timing of successful and unsuccessful year. In fact if n is the number, out of t, of good years the asset value at the end of t years is, \( V_t = (1.1^n)(0.95^{t-n}) \). The intermediate values for \( t = 0,1,2,3,4,\) and 5 are illustrated in Figure 1. The possible terminal values (that is, the values after 5 years) and the paths to achieve those values are given in Table 1.

<table>
<thead>
<tr>
<th>END</th>
<th>PATHWAYS</th>
<th>NUMBER UPS</th>
<th>VALUE ( V_5 )</th>
<th>NUMBER PATHWAYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>abcdjo abcijk abhijk aghijk abcdek</td>
<td>5</td>
<td>$1.61 = 1.1^5$</td>
<td>1</td>
</tr>
<tr>
<td>k</td>
<td>abcdjo abcijo abcino abhijo abhinr abhino abhmno aghino aghmno aglnno</td>
<td>4</td>
<td>$1.39 = (1.1^4)(0.95)$</td>
<td>5</td>
</tr>
<tr>
<td>o</td>
<td>abcdjo abcijo abcino abhijo abhinr abhino abhmno aghino aghmno aglnno</td>
<td>3</td>
<td>$1.20 = (1.1^3)(0.95^2)$</td>
<td>10</td>
</tr>
<tr>
<td>r</td>
<td>abcmr abhmnr abhmqr aghnq agmq aglnnr aglnq aglnq aglnq</td>
<td>2</td>
<td>$1.04 = (1.1^2)(0.95^3)$</td>
<td>10</td>
</tr>
<tr>
<td>t</td>
<td>abhmqt aghmqt aglnq aglnq aglnq</td>
<td>1</td>
<td>$0.90 = 1.1(0.95^4)$</td>
<td>5</td>
</tr>
<tr>
<td>u</td>
<td>aglnsu</td>
<td>0</td>
<td>$0.77 = 0.95^5$</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1
Asset Value Pathways for Five Years

There is only one pathway to the high value of $1.61 and one to the low of $0.77. There are however five pathways to the intermediate value of $1.39 and five to $1.04. There are twice as many, namely 10, ways to achieve $1.20 and 10 pathways to $1.04.

The example illustrates an important point. Namely, the asset at any given time depends only on the number of increasing years not on the path itself. For instance an year five value of $1.39 can be achieved with five distinct pathways, each with the common feature of four up years out of five. For instance for the path abhijk the years of increased value are 1,3,4 and 5 with a decline in year 2 and for the path abcedk the up years are 1,2,3,and 4 with a decline in year 5. Irrespective of the path with one down year out of five the asset value in year five is $1.39. This might lead us to consider a generalization of the asset value problem.

Another and more general way to view this issue is to consider the outcome of $1.39 year five value. Let's specify the years in a natural way 1 through 5. How many different numbering of the four up year (with the understanding that the unnumbered year
is down) are there with the five possible numbers. It is easy to attach numbers to the up years in all the paths that lead to $1.39 and find that there are exactly five distinct numbering: 1,2,3,4; 1,2,3,5; 1,2,4,5; 1,3,4,5 and 2,3,4,5. In a similar way we can count the number of distinct numbers for five, three, two, one and zero up years out of five. A generalization of the asset value problem leads to a greater understanding of a large number of probabilistic issues.

With the example we consider the possible numbering of k things (up years) out of n (five). A restatement of this is to consider a set of n (for the example n = 5) distinct (for the example the integers 1 through 5) objects and ask how many distinct k object subsets can assembled with the n objects (elements).

We start considering the generalization with a specific example, for instance, how many subsets with 3 different elements are there in a set with 5 different elements? To make the question more concrete let the set A = {1,2,3,4,5} be the successive integers from 1 to 5. Since there are so few, the subsets can be enumerated easily. Before laying out all the three elements subsets, recognize that a subset is defined by the inclusion, but not by the order, of the elements. So \( \{1,2,3\} \) is the same subset as \( \{2,3,1\} \) for instance. With this in mind, the subsets are \( \{1,2,3\} \{1,2,4\} \{1,2,5\} \{1,3,4\} \{1,3,5\} \{1,4,5\} \{2,3,4\} \{2,3,5\} \{2,4,5\} \{3,4,5\} \): there are ten subsets of size three in a set with five elements. While listing all the three subsets is easy for A with a small number of elements, it can be an enormous task for larger sets. Therefore, let's think how the question might have been answered without resorting to writing out the list.

Let \( x_3 \) be the answer to the posed question, namely the number of three elements subsets. In order to solve for \( x_3 \) first think how many ways a set of size three can be ordered. Use the set \( \{1,2,3\} \) as an example. The first element can be one of three values, namely 1, 2 or 3. Once the first element is chosen, there are two choices for the second. For instance if the first element is 1, the second can be 2 or 3. Finally, the choice of the first and second place leave only one choice for the third. Therefore there are \( 3 \times 2 \times 1 \) different orderings of the three numbers. A shorthand for expressing this product is \( 3! \). Since there are \( x_3 \) different size 3 subsets and for each there are \( 3! \) orderings, there are, therefore, \( 3! \times x_3 \) ordered collections of size three that can be formed from the elements of a set of size five.

There is an alternative way to calculate this same quantity by directly considering how many ordered collection with three elements can be formed from a set with five different elements. In this case there are five choices for the first element and once that is chosen there are four for the second. Three numbers are left for the third and final element of the ordered triple once the first and second are picked. There are \( 5 \times 4 \times 3 \) (60) ordered triples in a set of size five. There is another way to write this by observing that

\[
\frac{5 \times 4 \times 3}{2 \times 1} = \frac{5!}{2!} = \frac{5!}{(5-3)!}.
\]

There are two different ways to compute the number of ordered triples in a set of five and they must give the same value. The question about the number of subset (collection of numbers without regard to order) is answered by exploiting the equivalence of the two derivations:

\[
3! \times x_3 = \frac{5!}{(5-3)!} \tag{1.1}
\]
implies that the number of subset of size three in a set of size 5 is

$$x_3 = \frac{5!}{3!(5-3)!}. \quad (1.2)$$

$x_3$ is ten, as we found by writing-out the different subsets with three elements in a set of size 5.

Now extend this derivation to a set of arbitrary size $n$, $A = \{a_1 \cdots a_n\}$ and let $x_k$ be the number of $k$ element subsets of $A$. By reasoning similar to that for the subsets of size 3 described above, there are $k!$ different ordering for every $k$ element subset. Therefore there are $k!x_k$ ordered $k$-tuples in $A$. A second, and independent way to compute the number of ordered $k$-tuples is to consider that there are $n$ different elements for the first, $n-1$ for the second and finally $n-(k-1)$ for its final element. The number, defined $n_k$, is the product

$$n_k = n \cdot (n-1) \cdot (n-2) \cdots (n-(k-1)). \quad (1.3)$$

The number can be expressed in terms of factorials as

$$n_k = \frac{n!}{(n-k)!}. \quad (1.4)$$

Again, exploiting the equivalence between the two derivations

$$k!x_k = \frac{n!}{(n-k)!} = n_k. \quad (1.5)$$

The number of $k$-element subset in a set of size $n$ is

$$x_k = \frac{n!}{k!(n-k)!}. \quad (1.6)$$

Another, and commonly used, way to express (1.6) is as the number of (distinct) combinations of size $k$ that can be made from a collection of size $n$ (often expressed as the number of combinations of $n$ thing taken $k$ at a time). The common notation for this is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \quad (1.7)$$

The asset value example can be reformulated using the combinatorial formula just derived. The advantage is that it simplifies the counting of the pathways to the various terminal values: the number can be calculated without the need to list the pathways as is done in Table 1. In order to achieve $1.61$ it is necessary that there are 10% increases for
all five years. The number of orderings of the up years is \( \binom{5}{5} = 1 \). For $1.39 in five years there are \( \binom{5}{4} = 5 \). And for $1.20 there are \( \binom{5}{3} = 10 \).

This is a good departure from which to consider the question of chance. What are the probabilities of achieving each of the six possible terminal (year 5) values? Without further information it is reasonable to use the Principal of Insufficient Reason and assign an equal chance to every pathway (each pathway can be considered an event). That being the case, since there are 32 possible pathways (count them: \( 1 + 5 + 10 + 10 + 5 + 1 \)) each individual pathway is assigned a probability of \( 1/32 \) – its relative frequency. Each pathway is unique; achieving one excludes the possibility of experiencing any of the others. A reference short-hand is to label these pathways (events) as mutually exclusive. Since there are five distinct pathways to and asset value of $1.39 and $0.90 and there is only one unique pathway for asset value $1.61 and one for $0.77 it is reasonable to assign probabilities to the two former events that are five time those for the two latter values. By similar reasoning, probabilities ten times those assigned to the highest and lowest values are assigned to asset values $1.20 and $1.04. With the relative magnitudes tied down, the levels only remain to be established and this is achieved by requiring sum, over all possible values, to be one. The assignment that achieves this is \( P(V_5 = $1.61) = P(V_5 = $0.77) = 1/32; P(V_5 = $1.39) = P(V_5 = $0.90) = 5/32 \) and \( P(V_5 = $1.61) = P(V_5 = $0.77) = 10/32 \).

The probability logic of five period asset prices is generalized by using the combinatorial version. Let there be an arbitrary number of periods, \( n \) and let \( 0 \leq k \leq n \) be the number of periods in which there are year-to-year increases in value. The values at the final period if there are \( k \) up years is noted as \( V_n(k) = (1.1^k)(0.95^{n-k}) \). The number of pathways for achieving \( V_n(k) \) is \( \binom{n}{k} \). The total number of pathways from the initial period 0 to the final period \( n \) is \( \sum_{k=0}^{n} \binom{n}{k} \). By applying the logic of insufficient reason to we assign as probability of a value its relative frequency. Thus

\[
P(V_n = V_n(k)) = \frac{\binom{n}{k}}{\sum_{k=0}^{n} \binom{n}{k}}
\]

(1.7)

Combinatorial

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{by convention} \quad 0! = 1.
\]

Facts about combinatorial
1) \( \binom{n}{0} = \binom{n}{n} = 1 \)

2) \( \binom{n}{k} = \binom{n}{n-k} \)

3) \( \binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \)

Proof: \[
\binom{n}{k} + \binom{n}{k-1} = \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n+1-k)!} = \\
\frac{n!(n+1-k)}{k!(n+1-k)!} + \frac{n!k}{k!(n+1-k)!} = \\
\frac{n!}{k!(n+1-k)!}[n+1-k+k] = \\
\frac{n!(n+1)}{k!(n+1-k)!} = \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}.
\]

These are used to derive an important result about the nth power of a sum:

**Binomial Theorem**

Theorem: \( (a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \).

Proof.

Clearly it is true for \( n = 0 \) and \( n = 1 \).

Suppose it is true for \( n-1 \). Then

\[
(a + b)^n = (a + b) \sum_{k=0}^{n-1} \binom{n-1}{k} a^k b^{n-1-k}.
\]

The equation is rewritten
\[(a + b)^n = \sum_{k=0}^{n-1} \binom{n-1}{k} a^{k+1} b^{n-1-k} + \sum_{k=0}^{n-1} \binom{n-1}{k} a^k b^{n-k}. \]

A separate rewrite of each right hand side expression

\[
\sum_{k=0}^{n-1} \binom{n-1}{k} a^{k+1} b^{n-1-k} = \sum_{k=0}^{n-1} \binom{n-1}{k} a^k b^{n-k} + \binom{n-1}{n-1} b^n.
\]

and

\[
\sum_{k=0}^{n-1} \binom{n-1}{k} a^k b^{n-k} = \sum_{k=0}^{n-1} \binom{n-1}{k} a^k b^{n-k} + \binom{n-1}{0} b^n.
\]

Notice the equivalences

\[
\binom{n-1}{n-1} a^n = \binom{n}{n} a^n b^n
\]

and

\[
\binom{n-1}{0} b^n = \binom{n}{0} a^n b^n.
\]

With all of this information, the binomial expansion is

\[(a + b)^n = \sum_{k=1}^{n-1} \binom{n-1}{k-1} a^k b^{n-k-1} + \binom{n-1}{k} a^k b^{n-k} + \binom{n}{0} a^n b^0 + \binom{1}{n} a^n b^0. \]

It is known that

\[
\binom{n-1}{k-1} + \binom{n-1}{k} = \binom{n}{k}.
\]

With this the proof is complete because the binomial expansion is expressed

\[(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. ///
\]
The binomial theorem has a particularly interesting implication when \( a \) and \( b \) are both equal to one. This special case can be used to prove that the number of subsets of any set is \( 2^n \).

**Number of Subsets**

We know that the number of \( k \) element subsets in a set of size \( n \) is

\[
x_n = \binom{n}{k}.
\]

Therefore the total number of subsets (including the empty set \([k=0]\) and the entire set \([k=n]\)) is

\[
\sum_{k=0}^{n} x_n = \sum_{k=0}^{n} \binom{n}{k}
\]

The right hand term can be transformed by multiplying it by one in the following way

\[
\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} a^k b^{n-k} \quad \text{where } a = b = 1.
\]

From the binomial theorem

\[
\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a + b)^n = 2^n.
\]

The total number of subset of an \( n \) element set is \( 2^n \).

**Multinomial Theorem**

**Theorem:**

\[
\left( \sum_{i=1}^{n} a_i \right)^n = \sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \cdots \sum_{x_k=0}^{n-\sum_{i=1}^{k-1} x_i} \frac{n!}{k!} \prod_{i=1}^{k} a_i^{x_i}
\]

**Proof.**
Evaluate \( \left( \sum_{i=1}^{n} a_i \right)^n \) with a series of equalities and application of the Binomial Theorem

\[
\left( \sum_{i=1}^{k} a_i \right)^n = \left( a_1 + \sum_{i=2}^{n} a_i \right)^n = \sum_{x_1=0}^{n} \left( \begin{array}{c} n \\ x_1 \end{array} \right) a_1^{x_1} a_2^{x_2} + \sum_{i=3}^{n} \left( \begin{array}{c} n \\ x_1 \end{array} \right) a_1^{x_1} a_2^{x_2} + \sum_{i=4}^{n} \left( \begin{array}{c} n \\ x_1 \end{array} \right) a_1^{x_1} a_2^{x_2} \cdots a_i^{x_i} \cdots a_k^{x_k} = \\
\sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \cdots \sum_{x_k=0}^{n-x_1-x_2-\cdots-x_{k-1}} \left( \begin{array}{c} n \\ x_1 \end{array} \right) a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k} \\
\prod_{i=1}^{k} a_i^{x_i} /\!\!/k
\]

Probability

We all know, or at least think we know, what a probability is. If you were asked the probability of a coin flip that resulted in a heads facing up, you would say, with confidence, that the probability is 1/2. Similarly you would answer 1/6 to a question about the probability of three up facing dots on a thrown die. There are at least two good reasons for these responses. The first is that over many observations of coin flipping and die rolling, heads up happened approximately half of the coin flips and three dots up happened in approximately 1/6th of the die roles. The first is an experimental answer with the extension that the future duplicates the past. The second depends on pure thought. There are only two possible outcomes of the coin flip and there is no reason to think that one outcome is more likely than another, so 1/2 is the obvious probability assignment. There are six possible outcomes when rolling a die with no reason to think that one outcome is more likely than another. A probability of 1/6 reflects this. The assignment of equal probability to possible events, in the absence of information to the contrary, is called the 'Principle of Insufficient Reason'. The idea is attributed to Johann Bernoulli, a late 17th and early 18th century Swiss mathematician and father of Daniel Bernoulli. The Principle of Insufficient Reason states that if we have no reason to believe that one outcome is more likely than another, then all possible outcomes are equally likely. "Equal likeliness" is reflected in an equal assignment of a probability number.

While simple, coin flip and die roll probabilities have in common with all probability assignment two obvious characteristics. Probabilities are non-negative numbers between 0 and 1. The events described above – heads or tails and number of spots on a die – are disjoint, or mutually exclusive, events. The probabilities of all
disjoint events sum to 1. There are additional aspects that are common to and required of all probability assignments. These conditions are not as much a part of our experience as the coin flips and die rolls. They are easiest understood with specific examples. In this case the examples are focused on numbers in the [0,1] unit interval.

Start by first considering to what are we assigning probabilities. One might think about a four sided die, each side an equilateral triangle, with the numbers in Figure 1 printed, one to a side. The die is thrown in the air and it lands on a table. The outcome of interest is the side that lay flat against the table. The number printed on the side is noted. By the principle of insufficient reason the probability assigned to each event is 1/4.

A more formal development of probabilities starts with a consideration of the possible outcomes of an conceptual experiment. The collection of all possible outcomes is called the sample space. For the four-sided die example the sample space $S = \{0, 1/3, 2/3, 1\}$.

An event is a subset of the sample space. For instance, one event is that 1/3 lands face down. This is the subset $\{1/3\}$. Another event is a number greater than 1/3 lands face down. This is the subset $\{2/3, 1\}$. Another possibility is that anything can happen and that is the entire sample space, namely $S$, which is a subset of itself. The opposite of anything is that nothing happens, this is the empty set $\emptyset$ that is considered a subset of all sets. The collection of all events is the event space, which, for this experiment is denoted as $A$

$$A = \begin{bmatrix}
\{0\} & \{1/3\} & \{2/3\} & 1 \\
\{0, 1/3\} & \{0, 2/3\} & \{0, 1\} & \{1/3, 2/3\} \\
\{1/3, 1\} & \{2/3, 1\} & \{0, 1, 2/3\} & \{0, 1/3, 1\} \\
\{0, 2/3, 1\} & \{1/3, 2/3, 1\} & S & \emptyset
\end{bmatrix}$$

There are things to note about $A$. First the union of all events, namely $S$ itself, and the union of any arbitrary members of $A$, are members of $A$. Second, the intersection of one element of $A$ and any other is also in $A$ (if the two elements are disjoint (have no elements in common) their intersection is empty and thus contained in $A$). Third, the complement of any element of $A$, namely those events that are not the specific one (for
instance the complement of the event (singleton set) \{1/3\} is \{0, 2/3, 1\}) is also in A. It is said that the sample space is closed under union and intersection. Finally, there are sixteen elements of A as predicted since A is the collection of all subset of S, we have proven that there are \(2^4\) such subsets.

With the four-sided die, there are a small and countable number of outcomes. Consequently, there are only a small, and countable, number of events. However, for many common, and interesting problems, the sample space, and thus the event space, is not countable. For instance suppose the sample space is an interval on the closed unit interval \([0,1]\), represented in Figure 1.2 with subset L, M and N as half-open (on the right) intervals.

The set L, M and N fulfill the criteria of events, but so do the intervals \(L_1, L_2, M_1, M_2, N_1\) and \(N_2\) in Figure 1.3

In fact, the subdivision of the interval can continue and infinite number of times with the sub-intervals becoming smaller and smaller each iteration. Each small interval, so formed, constitutes and event of the original sample space \(S = [0,1]\). In fact the event space, \(A\), for this is the set of all half-open intervals (an infinity) in \(S\).

There are an uncountable infinity of such intervals and, thus, the sample and event spaces are infinite and uncountable as well. While, naturally, the elements of these sets cannot be enumerated, the analysis is restricted to event spaces that share certain properties. The following restrictions apply to both countable and uncountably infinite sets.

i) The sample space is an element of the event space: \(S \in A\)

ii) The complement of every event in the event space is, itself, contained in the event space: If \(A_1 \in A\) then \(A_1 \in A\).

iii) All infinite union of sets in the event space is a subset of \(A\)

\[A_1, A_2, \cdots, A_\infty \in A \Rightarrow \bigcup_{i=1}^{\infty} A_i \in A\]

These three conditions imply two more
iv) The event space contains the null set: $\emptyset \in A$

Proof:
By i) $S$ is in the event space and by (ii) its complement is also in the event space. But the complement of $S$ is $\emptyset$.

iv) All infinite intersections of subsets of $A$ are contained in $A$

$$A_1, A_2, \ldots A_{\infty} \in A \Rightarrow \bigcap_{i=1}^{\infty} A_i \in A$$

With the restrictions on the sample and event spaces, probability (abstract from physical or psychological experience) is given substance as a mathematical object: it is a mapping from the event space to the unit interval. A formal definition of a **Probability Function** $P[ ]$ is a function with domain $A$ (as defined above) and range $[0,1]$. A probability function satisfies the following three conditions:

P1) $P[A_i] \geq 0$ for all $A_i \in A$.

P2) $P[S] = 1$. The probability of a "sure thing" is equal to 1.

P3) If $A_1, \ldots$ is an infinite sequence of mutually exclusive events in $A$ ($A_i \cap A_j = \emptyset$ for $i \neq j$) and if $A_1 \cup A_2 \cup \cdots \cup A_1 \in A$ then $P[\bigcup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]$.

The proofs of the following propositions are problems for students:

A1.5 $P[\emptyset] = 0$.

A1.6 If $A_1 \cdots A_n$ are mutually exclusive events in $E$ then

$$P[\bigcup_{i=1}^{n} A_i] = \sum_{i=1}^{n} A_i$$

A1.7 If $A_i \in E$ then $P[\overline{A_i}] = 1 - P[A_i]$.

For the following, and all subsequent formulae, set designations typed without space between indicate intersection, e.g., $A_iA_j$ is equivalent to $A_i \cap A_j$.

A1.8 If $A_i$ and $A_j \in A$, then $P[A_i] = P[A_iA_j] + P[A_i\overline{A_j}]$ and

A1.9 Let \( B = \{ B_1, \ldots, B_n \} \) be a partition of the event space, \( A : A = \bigcup_{i=1}^{n} B_i, \ B_i \cap B_j = \emptyset \) for \( i \neq j \). Then \( P(A_i) = \sum_{j=1}^{n} P(A_i B_j) \) for all \( A_i \in A \).

Proof.

From 1.7, \( P(A_i) = P(A_i B_i) + P(A_i \overline{B}_i) \).

Since \( B_i B_j = \emptyset \) for \( i \neq j \), \( (A_i \overline{B}_i) \cap (A_i \overline{B}_j) = \emptyset \) and \( \overline{B}_i = \bigcup_{k=2}^{n} B_k \), then

\[
P(A_i \overline{B}_i) = \sum_{k=2}^{n} P(A_i B_k). \text{ The conclusion follows automatically.}\
\]

A1.10 (i) For every two events \( A_i \) and \( A_j \in E \),
\[
P[A_i \cup A_j] = P[A_i] + P[A_j] - P[A_i A_j]
\]

(ii) For events \( A_1 \ldots A_n \in A \),
\[
P[\bigcup_{i=1}^{n} A_i] = \sum_{j=1}^{n} P[A_j] - \sum_{i<j} A_i A_j + \sum_{i<j<k} P[A_i A_j A_k] - \cdots + (-1)^{n+1} P[A_1 A_2 \cdots A_n]
\]

A1.11 If \( A_i \) and \( A_j \in A \) and \( A_i \subset A_j \), then \( P[A_i] \leq P[A_j] \).

A1.12 Boole's Inequality  If \( A_1, A_2, \ldots, A_n \in A \), then \( P[\bigcup_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} P[A_i] \)

In Economic research most commonly don't research events themselves, but measured manifestations of them. For instance an individual's employment on one particular day can be thought of an experiment for which the outcomes will be either yes if employed or no if unemployed. We choose to quantify this by assigning a number to the particular outcome and a common way to do this is to make the assignment rule \( X = 1 \) if yes and \( X = 0 \) if no. The assignment rule, namely the function that assigns a real number to the outcome is called a Random Variable. A formal definition is as follows:

**A Random Variable** is a function \( X(\cdot) \) from the sample space to the reals \( X : S \to R \) such that
\[
X(s) = s \text{ for } s \in S.
\]

Since probabilities are unique mapping from the sample space to the real, the probabilities associated with the random variable are defined in a straightforward manner. As an illustration think of the four-sided die with the random variable the number on the down face. For \( X \) so defined
\[
P_X(0) = P_X(1/3) = P_X(2/3) = P_X(1) = 1/4.
\]
Read $P_X(x)$ as the probability that the random variable $X$ takes on the value $x$.

We might also think of the value of the random variable for the other elements of the event space. Consider the intersection of any two events in the sample space, for instance, $1/3 \cap 1$. Since the intersection is empty, its probability is 0. The random variable $X$ must also have probability zero for its mapping from this intersection to the reals

$$P_X(1/3 \text{ and } 1) = 0.$$ Consider a different subset of $S$, namely the union of $0$ and $2/3$. These are disjoint events (sets) and the probability of their union is $1/2$. Similarly, the probably of the random variable $X$ taking on the value of one of the two events is

$$P_X(0 \text{ or } 2/3) = P_X(0) + P_X(2/3) = 1/2.$$

The principle here is that as the probability of the union of disjoint events is the sum of the probabilities of the individual events. In terms of probabilities this means that the probability of one of two or more mutually exclusive values of a random variable is the sum of the individual probabilities.

Suppose, rather than events described by discrete values, the sample space is the entire closed unit interval $[0,1]$ and the events are inclusion in either two half-open interval $[0,1/3) \cup [1/3,2/3)$, or one closed interval $[2/3,1]$. Each subinterval is the same length and, again, the Principle of Insufficient Reason would have us assign equal probability to each event. Since the events are disjoint, the rule that individual probabilities sum to one, would indicate an assignment to each of $1/3$. As indicated in Figure 1.2, the intervals (events) are designated $L$, $M$ and $N$. By using these designation we could indicate the probability of the events as

$$P(L) = P(M) = P(N) = 1/3.$$ Another possibility is to use the random variable $X$ as previously defined, namely, the mapping from the value on the interval to the real line. In terms of the random variable, membership in a specific interval is indicated by values of $X$. For instance for the event $M$ the random variable takes on the values between the upper and lower bound of $M$. Thus, $P(M) = P_X(1/3 \leq x < 2/3)$. Similarly, $P(L) = P_X(0 \leq x < 1/3)$ and $P(N) = P_X(2/3 \leq x \leq 1)$.

These observations demonstrate additional properties of the sample and probabilities. As, with the case of discrete events the sample space is made up of the union of all its subsets. In this case the subsets $L$, $M$ and $N$ are disjoint, and their intersection is in the sample by virtue of the inclusion of the null set in all sets. Furthermore the complement of any subset is in the sample space as well. For instance the complement of the subset $L$ is the union of $M$ and $N$. Another aspect of the sample space and associated probabilities is that the intersection of all subsets of any subset of $S$

\footnote{The intervals are closed on the left indicated by [ and open on the right.}
is no more probable than the subset itself. This property translates into probabilities of the random variable. Think of this example that starts with subsets of $M$ defined as follows:

$$M_n = \left( \frac{1}{3^n}, \frac{2}{3^n}, \frac{1}{3^n} \right), \quad n = 1, 2, \ldots$$

As defined, each $M_n$ is a subset of the subsequent $M_{n+1} (M_n \subset M_{n+1})$ and all are subsets of $M$. The probability of each of the subsets is no greater than the probability of a higher numbered subset

$$P(M_n) \leq P(M_{n+1}).$$

Each of the subsets defined above is a subset of $M$ and each, in turn, is no more probable than $M$ itself. Furthermore the infinite union of all $M_n$ is no more probable than $M$

$$P(M_n) \leq P(M) \quad \text{and} \quad P(\bigcup_{n=1}^{\infty} M_n) \leq P(M).$$

**Uniform Distribution**

The first example of a probability function is the uniform distribution. For this distribution, and other that follow, it is useful to consider the example of an urn containing $M$ balls. The balls are numbered from 1 to $M$, $K$ of them are red and the others are green (the color of the balls has no bearing on the discussion of the uniform distribution, but will be employed for subsequent examples).

The thought experiment is to draw one ball, at random, from the urn and examine it number (only, its color is irrelevant for this example). Since all ball have the same chance of selection, the probability of drawing a given number, $k$, is the proportion of balls with the number $k$ (namely one out of $M$) in the urn. The random variable in this case is the number, and the probability of any number is

$$P[k] = \frac{1}{M} \quad \text{for all } k.$$  

This function is non negative for all number included in the urn and it sum over all $k$ is equal to one:

$$\sum_{k=1}^{M} P[k] = \sum_{k=1}^{M} \frac{1}{M} = 1$$  

(1.9)
**Bernoulli Distribution**

Employing the urn example, the random variable for the uniform distribution is the number of the ball selected. A Bernoulli random variable is its color. In the common Bernoulli experiment, the terminology used is that of success and failure. It is arbitrary what event (drawing a red ball or a green on) is labeled a success. Let a red ball indicate success and green failure. The probability, \( p \), of a success (drawing a red ball) is equal to the proportion of red ball in the urn:

\[
p = \frac{K}{M}.
\]  

(1.10)

The probability, \( q \), of a failure (drawing a green ball) is equal to the proportion of green balls in the urn:

\[
q = \frac{M - K}{M} = 1 - p.
\]  

(1.11)

Each probability is non-negative and the sum of the probabilities over the two possible outcomes (\( p + q \)) equals one.

**Binomial Distribution**

The binomial distribution is most easily introduced with reference to the urn with red and green balls. Suppose there are \( M \) balls in the urn of which \( K \) are red and \( M - K \) are green. In this conceptual experiment each ball has number (suppose without loss in generality that the balls with numbers 1 through \( K \) are red and those with number \( K+1 \) through \( M \) are green). We are going to draw \( n \) balls with replacement, one at a time, and each has the same chance (1/\( M \)) of being chosen on every draw. The sample space, \( S \), for this experiment, is the set of every possible outcome from \( n \) draws. For instance, if \( n \) is three and \( M \geq 7 \), two element of \( S \) are \( s_1 = \{ b_1, b_3, b_5 \} \) and \( s_2 = \{ b_7, b_7, b_7 \} \) (remember this is sampling with replacement. The number of elements in \( S \) is \( N(S) = M^n \). Every element, a subset of \( S \), has an equal chance of being selected.

The question posed is "what is the probability of choosing an \( s_i \in S \) that has exactly \( k \) red balls?" The answer, since each element of \( S \) is a likely to be chosen as any other, is the proportion of the \( M^n \) elements that have exactly \( k \) red balls.

For any draw (set) of \( n \) balls (elements) there are \( \binom{n}{k} \) subsets of size \( k \). Since the drawing is done with replacement, and there are \( K \) red balls in the urn, there are \( K^k \) different numberings of the red balls in the subset of size \( k \). For the remaining \( n-k \) elements in the sample there are \( (M-K)^{n-k} \) different numbering of white balls. Therefore, for any sample with exactly \( k \) red and \( n-k \) white balls, there are \( K^k (M-K)^{n-k} \) different ball numberings. Thus, there are \( \binom{n}{k} K^k (M-K)^{n-k} \) different,
and all equally likely, ways to find a sample of size \( n \) with exactly \( k \) red balls. The proportion of outcomes, out of the total equally likely, with a sample size of \( n \) is

\[
\binom{n}{k} \frac{K^k (M - K)^{n-k}}{M^n}.
\]

The expression can be rewritten

\[
\binom{n}{k} \left( \frac{K}{M} \right)^k \left( \frac{M - K}{M} \right)^{n-k}.
\]

The first fraction is the probability (proportion of red balls in the urn) of drawing a red ball from the urn and the second is the probability of drawing a green one. Thus, the probability of finding exactly \( k \) red balls in a sample of \( n \) is

\[
P[k] = \binom{n}{k} p^k q^{n-k}.
\]  (1.12)

The binomial probability function is clearly non-negative for all values of \( 0 \leq k \leq n \). It is simple to prove that it also sums to one over the same range:

\[
\sum_{k=0}^{n} P[k] = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}
\]  (1.13)

from the binomial theorem

\[
\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1.
\]  (1.14)

**Poisson Distribution**

On Highway 00X between Yarma and Zootsville, a distance of 100 miles, there are an average of 6 highway fatalities a year. The number of accidents is a random variable that might reasonably be modeled as a Bernoulli process with a probability of 1/2 (6/12) of an accident in any given month. A years worth of accident data can be thought of as twelve independent Bernoulli trials. With this view specify the number of accident in a year by \( A \) and the realization in any year as \( x \), the random variable \( A \) is binomially distribute with an \( n = 12 \)

\[
P(A = x) = \binom{n}{x} 0.5^x 0.5^{(n-x)}.
\]

Implicit in this subdivision and subsequent model is that there can be no more than one accident a month. But, of course, that is not the case; there is the potential for a much larger number. Perhaps a solution is to make finer subdivision, for example days rather than month. In this case \( n = 365 \) and the daily probability is 12/365. The expression for specific realizations of \( A \) is

\[
P(A = x) = \binom{n}{x} (6/365)^x (359/365)^{(n-x)}.
\]
This might be good, nonetheless, the issue of more than one occurrence per day remains. Perhaps if continue to make finer and finer subdivision, the time intervals will become so short that the chance of any more than one occurrence is so minute that it can be ignored. It is the limit of the process of finer subdivision from which the Poisson distribution is derived.

Let $\lambda$ be the expected (average) number of occurrences of a random event is a specified time period. For the highway fatality example $\lambda$ is 12 and represents the average number of fatalities in a year. Define

$$p_n \equiv \frac{\lambda}{n} \quad (1.15)$$

as the probability of an event occurrence is one of $n$ subdivisions. This definition makes sense only if the number of subdivisions is greater than $\lambda$. In this case the probability of $x$ occurrences when there are $n$ subdivisions is

$$P_n(A = x) = \binom{n}{x} p_n^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x}. \quad (1.16)$$

Define $n_x \equiv n(n-1)(n-2)\cdots(n+1-x)$ and rewrite 1.16

$$P_n(A = x) = \frac{\lambda^x}{x!} n_x^x (1 - \frac{\lambda}{n})^{-x} (1 - \frac{\lambda}{n})^n \quad (1.17)$$

The limit as $n$ becomes infinitely large of this probability is found by finding the limits of each of the multiplicative terms in 1.17:

$$\lim_{n \to \infty} \frac{n_x}{n^x} = 1$$

$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1$$

and

$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}. \quad ^2$$

Thus

$$\lim_{n \to \infty} P_n(A = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (1.18)$$

which is the Poisson density function.

---

^2 See Appendix A at the end of this chapter.
The Poisson distribution has application in both temporal and spatial analysis. For instance suppose a brokerage house, after many years of collecting data on incoming telephone calls, find that the average of the number of calls per hour is sixty. The calls arrive independently: the probability of one call is independent of the arrival of any other. If 60 is the true average the brokerage management needs to decide how many employees they need to answer telephones. In order to make this decision the managers want to know the probability distribution of calls per hour and call within any five minute period during the working day. Let X be the number of telephone calls per hour and Y the number within any five minute interval. If the telephone arrivals are a Poisson process the $\lambda$ for the hour interval is 60 and for the five minute interval is 5. The probabilities are

$$P_X(x) = \frac{60^x e^{-60}}{x!}$$
$$P_Y(y) = \frac{5^y e^{-5}}{y!}$$

<table>
<thead>
<tr>
<th>Telephone Calls</th>
<th>Five Minute</th>
<th>One Hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00673795</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.03368973</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.08422434</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.1403739</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.17546737</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.17546737</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.14622281</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.10444486</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.06527804</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.043348</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>0.046444</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>0.048889</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>0.050575</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>0.051432</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0.051432</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>0.050589</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>0.048957</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>0.046625</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>0.043711</td>
<td></td>
</tr>
</tbody>
</table>

A second example is a spatial one. Think of a flat plane square featureless plane with an area of 100 SqKM. It is divided into plats of 1 square KM each. It is know that there an average of 3 settlement per square kilometer on the plane. It is know that whether or not there is a settlement in one of the plats is independent of the location of other settlements. The Poisson probability can be used to predict the distribution of settlement sizes. The probability of settlement densities, X, is
\[ P_X(x) = \frac{3^x e^{-3}}{x!} \]

<table>
<thead>
<tr>
<th>Settlement Density X</th>
<th>Probability</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>0.15</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>0.22</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0.02</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
<td>1</td>
</tr>
</tbody>
</table>

**Conditional Probability**

In the previous section the focus is on the probability of particular events when drawing from the entire sample space. In this section we consider a partition of the sample space. The focus on the information (perhaps the precision of prediction) gained when we are told that the event is part of a particular partition of the sample space. To illustrate this consider two possible experiments of drawing from two urns. The experiments are labeled I and II.

I.

\[
\begin{align*}
&\text{A} \\
&3 \text{ Red} \\
&4 \text{ Green}
\end{align*}
\]

II.

\[
\begin{align*}
&\text{B} \\
&12 \text{ Red} \\
&16 \text{ Green}
\end{align*}
\]

In experiment I there are 35 balls, 15 Red and 20 Green, and in the second there are 13 balls, 4 Red and 9 Green.

In the following discussion subscripts on the probability function indicate the experiment and the terms in the parenthesis the event. For the entire example one ball is
drawn at random from either urn A or B in experiment I or urn A or C in experiment II.
The sample space for experiment I is
\[ S_I = \{(R,A), (R,B), (G,A), (G,B)\} \]
and for experiment II is
\[ S_{II} = \{(R,A), (R,C), (G,A), (G,C)\} \].
In the first experiment, the probability of drawing a red ball is
\[ P_I(R) = \frac{15}{35} = \frac{3}{7} \]
the probability of drawing a red ball that comes from urn A is
\[ P_I(R,A) = \frac{3}{35} \]
and one that comes from urn B is
\[ P_I(R,B) = \frac{12}{35}. \]
The probability of a ball being drawn from A, \[ P_I(A) = \frac{7}{35} \] (there are a total of 35 ball
and 7 of them are in A).
Similarly, find that \[ P_{II}(R) = \frac{4}{13}, P_{II}(R,A) = \frac{3}{13} \] and \[ P_{II}(R,C) = \frac{1}{13}. \] From this, if you
were asked for the probability of a red ball in experiment I you would answer \( \frac{3}{7} \) and in
experiment II your answer would be \( \frac{4}{13} \).

Suppose you are given some additional information: namely that the ball was
drawn from urn A. How would you revise your statement of the probability of red? In
both experiments, 3 out of the seven balls in urn A are red. Thus, in both experiments you
would assign a probability of \( \frac{3}{7} \) to a red if you know that the draw is from urn A.
However, there is one large difference between the assignments for the two experiments.
In experiment I, \( \frac{3}{7} \) is the probability assignment if the urn is unknown. In other words
there is nothing of value gained by the information about the urn. If urn B is announced
the probability of Red in that urn is the same as it is for the population as a whole. In the
case of experiment I learning about the urns does not change the probability of the ball
color. The events, ball color and urn, are said to be independent. A formal definition of
independence is given later.

On the other hand, in experiment II, ball color and urn are not independent. If the
urn choice is not known, the probability of Red is \( \frac{4}{13} \). However if it is known that the
ball is drawn from urn A, the probability of Red is \( \frac{3}{7} \) and if it is drawn from urn B \( \frac{1}{6} \) is
the chance of a Red. For experiment II there is value in knowing from which the ball is
taken. In this case color and urn are not independent.

Formal Definitions of Conditional Probability and Independence

Suppose there are two events X and Y and \( P(Y) > 0 \). The probability of X
conditional on Y (in the last example this would be the probability of Red conditional on
the urn from which the ball is drawn) is
\[ P(X \mid Y) = \frac{P(XY)}{P(Y)}. \quad (1.19) \]
This formula relates to the example above as follows. The probability of Red conditional on urn A
\[
P_1(R \mid A) = \frac{P_1(RA)}{P_1(A)} = \frac{3/35}{7/35}.
\]

The events X and Y are independent if knowing Y provides no additional information about the probability of X, namely, if the conditional probability of X given Y equals the total probability of X. Thus, X and Y are independent if, and only if
\[
P(X \mid Y) = P(X). \quad (1.20)
\]

Another criterion for independence is that the joint probability equals the product of the individual probabilities:
\[
P(XY) = P(X)P(Y). \quad (1.21)
\]

You should be able to show the equivalence of (1.20) and (1.21).

**Total Probability and Bayes' Theorem.**

Bayes' Theorem, a most important and useful result, is developed and follows immediately from the "Law of Total Probability"

1.12 Let \( B_1, \ldots, B_n \) be a partition of the sample space S and \( P(B_j) > 0 \) for all \( j \). Then for every \( A \in S \)
\[
P(A) = \sum_{j=1}^{n} P(B_j)P(A \mid B_j).
\]

Proof.  
Note that \( P(B_j)P(A \mid B_j) = P(A \cap B_j) \) and the conclusion follows from (A1.9).///

1.13 **Bayes' Theorem** Let \( B_1, \ldots, B_n \) be a partition of the sample space S and \( A \in S \) be any event such that \( P(A) > 0 \). Then for \( i = 1, \ldots, n \)
\[
P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)}. \quad (1.22)
\]
The proof is left to the student.

**Learning about Probability**

There are two points of view about probability. The one used in all that precedes is an objective view. The probability of an event is its relative frequency in the sample space: if a ball in drawn from an urn in \( K \) balls are red and \( M-K \) are green then the probability of red is \( K/M \). The second viewpoint might, with some reason, be labeled subjective. Probability is a measure of our certainty about an event, the higher the
probability, the more certain we are that the event will occur. It is natural, from the second viewpoint to consider learning and becoming more assured about the nature of the world around us. Bayes theorem suggests a rational way that uncertainty may be reduced with experience. An example is used to explain this.

There are only two types of people in the world honest ones (H), who always do what they say they will do and dishonest ones (D), who always do the opposite of what they say they will do. You are a bit of a gambler and always ready to play a fair game of chance (a fair game is one in which the probability of winning is equal to 0.5 the probability of loosing). In fact you enjoy gambling so much that you will play in an unfair game as long as you believe the probability of winning is 0.4 or higher.

You are approached by, Ann, a long-time and highly regarded acquaintance. You assess her as an honest person, in fact you assign a probability, \( P(H) = 0.90 \), that she is an honest person (and, of course \( P(D) = 1 - P(H) = 0.10 \)). She holds in her hand two boxes, one in each hand. You know that one box contains a fair coin, but the other contains a coin with two heads. Ann knows which is the fair box. Ann proposes the following gamble: she will choose the fair box take out the coin and flip it. If it comes up tails (t) she will pay you $1.00 and if it comes up heads (h) you will pay her $1.00. Do you accept the bet? Here are the considerations: if Ann is honest you have a 50% chance of winning – this is higher than your cut-off probability of 0.4 and so you would certainly accept. However, you are not certain of Ann's honesty. There is a probability of 0.10 that she is dishonest, in which case the probability of a win in zero. Applying the law of total probability (1.12) we find that the probability of a win

\[
P(W) = P(H)P(W|H) + (1 - P(H))P(W|D) = 0.9*0.5 + 0.1*0 = 0.45.
\]

With the calculation in hand, you accept the bet. Ann removes the coin from the box she selects and flips it – it comes up heads. She takes your dollar and then asks if you want to repeat the game with the same rules as before.

According to the decision rule given above, you will accept if you think the probability of winning is at least 0.4. Since \( P(W|D) = 0 \), you will continue playing as long you believe that \( P(H) \geq 0.80 \). The question is, how should the experience in the previous gamble affect you belief about the probability of honesty? Since the outcome was heads, what is the probability that Ann is honest? What is \( P(H|h) \)? This probability can be calculated using Bayes' theorem, namely

\[
P(H \mid h) = \frac{P(h \mid H)P(H)}{P(H)P(h \mid H) + P(D)P(h \mid D)}.
\]

In this equation, \( P(H) \) is called the prior probability and \( P(H|h) \) the posterior. For the problem at hand, the updated posterior probability of honesty is

\[
P(H \mid h) = \frac{0.5*0.9}{0.5*0.9 + 1*0.1} = 0.82.
\]
It appears that you still believe that the probability of an honest Ann is high enough to play the game one more time. You signal acceptance, and removes a coin, flips it and it comes up heads. What is your new belief about the probability of honesty?

**Monty Hall Problem**

There are three doors – an expensive prize behind one and nothing behind the other two. Without further information there is an equal, 1/3, probability that the prize is behind any specific door. A contestant chooses one door and may keep whatever is behind it: either a valuable prize or nothing.

![Diagram of three doors](image)

After the contestant has chosen (say door 1 \([D1]\)) Monty, who knows what is behind each door, opens one of the two remaining doors. Every time, there is nothing behind the door that Monty opens. The contestant is offered the chance to switch to \(D3\) or stay with \(D1\). Should the contestant switch?

The first thing to note is that Monty knows what is behind each door and he always opens a door to reveal no prize. Let \(M#\) designate Monty's choice. Suppose after \(D1\) is chosen he opens door 2. This is represented as \(M2\).

The contestants conundrum: either door 1 hides the prize – represented as \(D1+\) or it does not, \(D1-\); similarly either \(D3+\) or \(D3-\) (note the obvious if \(D1+\) then not \(D3+\)). The contestant's problem can be stated in terms of conditional probabilities, namely, what is the probability of \(D1+\) given \(M2\) compared with the probability of \(D3+\) given \(M2\)? In symbols

\[
P(D1+ | M2) \begin{cases} > & \text{if } M2 \text{ is opened} \\ < & \text{if } M2 \text{ is not opened} \\ \end{cases} P(D3+ | M2) .
\]

Bayes' Law can be used to compute the desired probabilities.

First determine that probabilities of \(M2\) given \(D1+\) and \(D3+\). It is reasonable to suppose that if the prize is behind door 1 that then door 2 and 3 have an equal chance of being Monty's selection

\[
P(M2 | D1+) = 1/2 .
\]

But if the prize is truly behind door 3, Monty will always choose door 2
P(M2 | D3+) = 1.

The probabilities of the joint events $D1 \cap M2$ and $D2 \cap M2$ are

$$P(D1 \cap M2) = P(M2 | D1+)P(D1+) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

and

$$P(D3 \cap M2) = P(M2 | D3+)P(D3+) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{3}.$$ 

Since $(D1 \cap M2)$ and $(D3 \cap M2)$ are mutually exclusive

$$P(M2) = P(D1 \cap M2) + P(D3 \cap M2) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}.$$ 

Therefore

$$P(D1+ | M2) = \frac{P(D1+ \cap M2)}{P(M2)} = \frac{1}{3}$$

$$P(D3+ | M2) = \frac{P(D3+ \cap M2)}{P(M2)} = \frac{2}{3}.$$ 

**Appendix B Poisson Density**

Let $x = \lambda/n$ and $f(x) = \ln(1-x)$, then the natural logarithm of $(1 - \frac{\lambda}{n})^n$ is $n f(x)$. A Taylor series approximation of $n f(x)$ around $x = 0$ is

$$n f(x) = n f(0) + nx f'(0) + nx^2 f''(0) + \cdots .$$

For $x = 0$, $f(x) = 0$. The first derivative $f'(x) = -\frac{1}{1 - \frac{\lambda}{n}}$ and thus $f(0) = -1$. It can be shown that all of the higher order derivative are equal to 1 when $x = 0$. Therefore substituting for $x$

$$n \ln(1 - \frac{\lambda}{n}) = -n\lambda/n + n(\lambda/n)^2 + \cdots$$
Therefore,

$$\lim_{n \to \infty} n \ln(1 - \lambda / n) = -\lambda$$

thus

$$\lim_{n \to \infty} (1 - \lambda / n)^n = e^{-\lambda}$$

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$
Chapter 2

Univariate Distribution
Univariate Distribution

A **Random Variable** is a function from the sample space to the real line. It is an assignment of a number to an event in the sample space. For instance, consider the sample space from the experiment of flipping a coin three times:

\[ S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}. \]

The number of heads is a random variable as it is a unique assignment of a real number to an outcome in the sample space – it is a function whose domain is the sample space and whose range is \( R \). The random variable, \( X \), in this case takes on one of four values

\[
X = \begin{cases}
0 \\
1 \\
2 \\
3
\end{cases}.
\]

A **Cumulative Density Function (CDF)** \( F(X) \) is a single valued mapping (namely a function), from the real numbers to the closed interval \([0, 1]\). It is defined as the probability that a random variable \( X \) is no greater than some specified value \( x \). This is written

\[ F_X(x) = P(X \leq x) \]

The subscript on \( F \) specifies the random variable. In the future, where there is no ambiguity about the variable, the subscript is dropped.

As an example consider rolling two dice. The random variable is the difference between the number of dots on one die and the other. The outcome of this experiment is given in the following table in which the differences are listed.

<table>
<thead>
<tr>
<th>Die 2</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The entries in the table, the differences between the two dice, are the values of the random variable, \( X \). The values, \( x \), of \( X \) along with their probabilities and the value of the cumulative density function are
The CDF for X looks like this

\[
\begin{array}{|c|c|c|}
\hline
x & P(X=x) & F(x) \\
\hline
[-\infty,0) & 0 & 0 \\
0 & \frac{6}{36} & \frac{6}{36} \\
1 & \frac{10}{36} & \frac{16}{36} \\
2 & \frac{8}{36} & \frac{24}{36} \\
3 & \frac{6}{36} & \frac{30}{36} \\
4 & \frac{4}{36} & \frac{34}{36} \\
5 & \frac{2}{36} & 1 \\
\hline
\end{array}
\]

Properties of a CDF

1) \( F_X(-\infty) \equiv \lim_{x\to-\infty} F_X(x) = 0 \) and \( F_X(\infty) \equiv \lim_{x\to\infty} F_X(x) = 1 \).

2) \( F(\cdot) \) is a monotone, nondecreasing function, namely, \( F(a) \leq F(b) \) for \( a < b \).

3) \( F(\cdot) \) is continuous from the right (\( \lim_{0<h\to0} F(x+h) = F(x) \)).

Any function \( F(\cdot) \) satisfying 1, 2 and 3 with domain the real line and range \([0 \ 1]\) is a cumulative density function.

So far we have considered only discrete random variables. They are random variables that take on only a countable number of discrete values. The discrete density function is equivalent to the probability functions that were extensively discussed in the last chapter. Formally the density function of a discrete random variable \( X \) is

\( f_X(x) = P(X = x) \)
that is read: the density function of the random variable \( X \), \( f_X(\cdot) \), evaluated at the value \( x \) is equal to the probability that \( X \) equals \( x \). The cumulative density function for the discrete random variable is defined in the natural way:

\[
F_X(x) = \sum_{k=-\infty}^{x} f_X(k) = P(X \leq x).
\]

As previously defined, \( \sum_{x=-\infty}^{\infty} f_X(x) = 1 \).

Examples of discrete variable density function from the last chapter are

- **Uniform:** \( f(x) = \frac{1}{b-a} \) where \( a \) and \( b \) are integer and \( b > a \). The random variable lies within the closed interval \([a b]\) and every integer in that interval has equal probability of being drawn.

- **Bernoulli:** \( f(x) = p \) where \( x \) can take one of two values (0 or 1 for instance) and \( p \) is the probability of one of those values, 0, say, labeled a success (\( 1 - p = q \) is the probability of the other).

- **Binomial:** \( f(x) = \binom{n}{x} p^x q^{n-x} \) where \( x \) is the (integer) number of successes in \( n \) independent Bernoulli trials.

- **Poisson:** \( f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \) where \( x \) is the number of arrivals over a specified interval and \( \lambda \) is the expected rate of arrival.

A **continuous random variable** is one for which there exists a density function \( f_X(\cdot) \) such that the cumulative density function \( F_X(x) = \int_{-\infty}^{x} f_X(u)du \) for every real number \( x \). Since \( \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_X(u)du dx = \int_{-\infty}^{\infty} f_X(x)dx \), furthermore \( \int_{-\infty}^{\infty} F_X(x)dx = \int_{-\infty}^{\infty} f_X(x)dx = 1 \).

If \( X \) is a continuous random variable then \( P_X(x) = 0 \), that is, the probability of any single value of the random variable is zero. One can only contemplate the probability of a particular interval, as, for instance, the interval between the real numbers \( a \) and \( b \) (\( a < b \)).

\[
P_X(a \leq x \leq b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(u)du - \int_{-\infty}^{a} f_X(u)du = \int_{a}^{b} f_X(u)du.
\]

**Uniform Distribution**

A random variable \( X \) with a uniform distribution is contained in a continuous interval, \( X \in [a \ b] \) and its density function is
Note \( \int_{-\infty}^{\infty} f_X(u) \, du = \int_{-\infty}^{0} f(u) \, du + \int_{0}^{b} f(u) \, du + \int_{b}^{\infty} f(u) \, du = 0 + \int_{a}^{b} \frac{du}{b-a} + 0 = 1 \). The subscript noting the random variable is dropped when there is no ambiguity.

**Exponential Distribution**

A random variable \( X \) with an exponential distribution can take on any non-negative real number \( X \in [0, \infty) \). The density function is

\[
f_X(x) = \begin{cases} 
\frac{1}{b-a} & \text{for } x \in [a, b] \\
0 & \text{otherwise}
\end{cases}
\]

This is a density function: it is everywhere non-negative and the integral

\[
\int_{-\infty}^{\infty} f(u) \, du = \int_{-\infty}^{0} f(u) \, du + \int_{0}^{\infty} f(u) \, du = 0 + \int_{-\infty}^{\infty} \frac{du}{b-a} = 1.
\]

**Normal Distribution**

A random variable \( X \in (-\infty, \infty) \) is has a normal (gaussian) distribution if

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}.
\]
The normal density is non-negative over the entire real line, the range of the normally distributed random variable. It is a bit tricky to show that it fulfills the other requirement that the value of its integral, over its entire range, is equal to one. To do this, first define $y = \frac{x - \mu}{\sigma}$ and this means that $dy = \frac{dx}{\sigma}$. With this in mind the integral

$$
\frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy.
$$

Since the density function is positive over its entire range, proving that the square of the integral is equal to one is sufficient to prove that the integral itself is equal to one. The square is of the integral is

$$
\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{y^2+z^2}{2}} dy dz.
$$

Evaluating this integral is straightforward if it is transformed into polar coordinates. In order to do this first consider how to describe any $(y,z)$ pair in polar form:

It is apparent from the figure that the variables $y$ and $z$ can be expressed in polar form $y = r \cos \theta$ and $z = r \sin \theta$. Thus $y^2 + z^2 = r^2(\cos^2 \theta + \sin^2 \theta) = r^2$. The equivalent of the infinite range of both $y$ and $z$ in the limits of integration is $\theta \in [0,2\pi]$ and $r \in [0,\infty)$. Integration is simply the adding up of the infinity of infinitely small boxes into which any
area can be partitioned. In the original Cartesian coordinates of y and z these boxes are small rectangles. The shape, and thus the means of computing the area, is transformed with the change to polar coordinates. For an intuitive understanding of the nature of the transformation and its effect on the computation of area examine the following graph:

Integration in polar coordinates is adding up infinitely small curved boxes formed by small changes in the angle, $d\theta$, and the distance from the origin, $dr$. Each change in $\theta$ is $d\theta/2\pi$ of the total, thus the length of the arc formed by the change in the angle at a distance $r$ is $d\theta/2\pi$ of the circumference of a circle with radius $r$, that is the length of the arc is simply $rd\theta$. Thus the area of the shaded curved box is approximately $rdrd\theta$. The integration is transformed

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{y^2+z^2}{2}} \ dy \ dz = \frac{1}{2\pi} \int_{0}^{\infty} \int_{0}^{2\pi} e^{-r^2} \ r \ d\theta \ dr = 1.$$
**Expectation**

There are three definitions of expected value. For discrete random variable the expected value is

1) \[ E(X) = \sum_{i=1}^{n} x_i f_X(x_i) \equiv \mu_X; \]

for a continuous random variable the expected value is

2) \[ E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx \equiv \mu_X \]

and, for an arbitrary random variable the expected value is

3) \[ E(X) = \int_{0}^{\infty} [1 - F_X(x)] \, dx - \int_{-\infty}^{0} F_X(x) \, dx \equiv \mu_X. \]

The equivalence of 2 and 3 is demonstrated in the following two graphs.

**Figure 1**

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

**Figure 2**

\[ E(X) = \int_{0}^{\infty} [1 - F(x)] \, dx - \int_{0}^{\infty} F(x) \, dx \]
An alternative way to write 2 is

\[ (2') \quad E(X) = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{-\infty}^{0} x f(x) \, dx - \int_{0}^{\infty} x f(x) \, dx. \]

The integral is the summation of the rectangular boxes whose width is \( \Delta F \) (in figure 1 each of the \( \Delta F \)'s are the same) and whose height is \( x \), for very small \( \Delta F \) (\( \Delta F \to 0 \)). The integral \( \int_{-\infty}^{0} x f(x) \, dx \) is the summation of the boxes from \( x = 0 \) to \( -\infty \). It is the area between the CDF and 1 between 0 and \( -\infty \). Similarly, the \( \int_{0}^{\infty} x f(x) \, dx \) is the summation of the boxes from \( x = 0 \) to \( \infty \): the area under the CDF for that same interval.

The rectangular boxes for equation 3, displayed in figure 2, are constructed differently. In this case the boxes have width \( \Delta x \) (each of the \( \Delta x \)'s in figure 2 are the same width). For the first integral, \( \int_{0}^{\infty} [1 - F(x)] \, dx \), the height of the boxes is \( 1 - F(x) \).

The summation of these boxes for infinitesimally small \( \Delta x \) is the area between one and the CDF for \( x \) between 0 and \( \infty \). Similarly the integral \( \int_{-\infty}^{0} F(x) \, dx \) is the area under the CDF for \( x \) between 0 and \( -\infty \).

The expectation of any function \( g( \cdot ) \) of a random variable is defined

\[ 4) \quad E(g(X)) = \sum \quad \text{for a discrete distribution and} \]

\[ 5) \quad E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \, dx. \]

**Students should be able to prove the truth of the following:**

i) \( E(c) = c \) where \( c \) is a constant.

ii) \( E(cg(X)) = c E(g(X)) \).

iii) \( E(cg(X) + ah(X)) = c E(g(X)) + a E(h(X)) \), where \( a \) and \( c \) are constants and \( g( \cdot ) \) and \( h( \cdot ) \) are functions.

iv) If \( g(x) \leq h(x) \) for all values of \( x \), the \( E(g(x)) \leq E(h(x)) \).

**Variance**

As with expectation there are three ways to compute the variance of a distribution. It is one measure of the dispersion of a random variable. The three measures are

\[ 6) \quad V(X) = \sum \quad \text{for a discrete distribution and} \]

\[ 7) \quad V(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) \, dx \equiv \sigma_X^2 \]
for a continuous distribution. Another, and computationally handy, way to write the variance is

$$V(X) = E(X^2) - \mu_X^2 \equiv \sigma_X^2.$$  

**Students should prove the equivalence of this with the previous two definitions.**

There is a third way to compute the variance that is works for arbitrary (discrete, continuous or mixed) distributions, namely

$$V(X) = \int_{-\infty}^{\infty} x^2 [1 - F_X(x) + F_X(-x)] dx - \mu_X^2 \equiv \sigma_X^2.$$  

The equivalence of 8 and 9 is verified:

$$V(X) = E(X^2) - \mu_X^2 \equiv \sigma_X^2$$  

(a)

Employ 3, above to compute $E(X^2)$

$$E(X^2) = \int_{-\infty}^{\infty} [1 - F_X(x)] dx^2$$  

(b)

because $x^2 \geq 0$, $\int_{-\infty}^{\infty} F_X(x^2) dx^2 = 0$. Thus,

$$E(X^2) = \int_{-\infty}^{\infty} [1 - F_X(x)] dx^2 = \int_{0}^{\infty} 2x [1 - F_X(x^2)] dx.$$  

Further evaluation gives

$$F_X(x^2) = \int_{0}^{x^2} f_X(u^2) du^2 = \int_{-x}^{x} f_X(z) dz = F_X(x) - F_X(-x).$$

Substituting this expression (b) and then employing (a), shows the equivalence.

**Standard Deviation**

The standard deviation of a distribution is defined as the positive square root of the variance:

$$\sigma = \sqrt[4]{\sigma^2}.$$  

**Moment Generating Functions**

The moments of a distribution are defined as the expected value of the random variable raised to the appropriate power. For instance the first moment of the distribution
of the random variable $X$ is $E(X)$, the second moment is $E(X^2)$ and the $r^{th}$ moment is $E(X^r)$. The $r^{th}$ central moment, or, more descriptively the $r^{th}$ moment about the mean is $E((X - \mu)^r)$.

The moment generating function, defined as the expected value of the $e^{tX}$, is a powerful tool for computing the moments of a distribution. While it may not exist for some distributions, it is a unique identifier of those distributions for which it exits. The moment generating function of a distribution is either unique (namely it is unlike the moment generating function for any other distribution) or it does not exist. The formal definition is

$$m_x(t) = E(e^{tX}) = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx \text{ for continuous densities} \\
\sum_{i=1}^{n} e^{tx} f_x(x_i) \text{ for discrete densities} \end{cases}$$

The $r^{th}$ moment of a distribution is the $r^{th}$ derivative of its moment generating function evaluated at $t = 0$:

$$E(X^r) = \frac{\partial^r m(t)}{\partial t^r} \bigg|_{t=0}.$$ 

The truth of this is apparent for the first derivative of a continuous distribution

$$\frac{\partial m(t)}{\partial t} \bigg|_{t=0} = \int_{-\infty}^{\infty} xe^{tx} f(x) \, dx \bigg|_{t=0} = \int_{-\infty}^{\infty} xf(x) \, dx.$$ 

The subscript indicating the random variable will be dropped, as it is in the above expressions, when it is obvious what is the random variable. Students should show that the second and third moments are the second and third derivatives evaluated at $t = 0$. You should so this for discrete as well as continuous densities.

A few examples demonstrate the usefulness of computing the moments with the moment generating function. Consider the binomial distribution first:

$$m(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (e^t p)^x (1-p)^{n-x}$$

from the binomial theorem

$$m(t) = (e^t p + (1-p))^n.$$ 

The first partial derivative of $m(t)$ with respect to $t$ is
\[ \frac{\partial m(t)}{\partial t} = n(e^tp + (1-p))^{n-1}e^t \]

and

\[ E(X) = \frac{\partial m(t)}{\partial t} \bigg|_{t=0} = np. \]

The variance of the binomial is calculated as follows:

\[ V(X) = E(X^2) - E^2(X) = \frac{\partial^2 m(t)}{\partial t^2} \bigg|_{t=0} - \left( \frac{\partial m(t)}{\partial t} \bigg|_{t=0} \right)^2. \]

The second derivative of the moment generating function is

\[ \frac{\partial^2 m(t)}{\partial t^2} = n(n-1)[e^tp + (1-p)]^{n-2}e^{2t}p^2 + n[e^tp + (1-p)]^{n-1}e^t \]

thus, the second derivative evaluated at \( t = 0 \) is

\[ \frac{\partial^2 m(t)}{\partial t^2} \bigg|_{t=0} = (np)^2 + np(1-p). \]

It follows that the variance of the binomial distribution is

\[ V(X) = np(1-p). \]

As a second example, consider the normal density. The moment generating function

\[ m(t) = E(e^{tx}) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{tx} \left( \frac{1}{\sigma} \right)^2 dx = \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{t(x-\mu)} \left( \frac{1}{\sigma^2} \right)^2 dx = \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{t(x-\mu)} \left( \frac{1}{\sigma^2} \right)^2 dx = \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\frac{1}{2\sigma^2}[(x-\mu)^2-2\sigma^2t(x-\mu)+\sigma^4t^2]} dx = \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\frac{1}{2\sigma^2}[(x-\mu)^2-2\sigma^2t(x-\mu)+\sigma^4t^2]} dx = \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\frac{1}{2\sigma^2}[(x-\mu)^2-2\sigma^2t(x-\mu)+\sigma^4t^2]} dx. \]
The integral \[ \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-\mu-\sigma t)^2} \, dx = 1 \]. Therefore the moment generating function for the normal density is

\[ m(t) = e^{\frac{\mu + \sigma^2 t^2}{2}} \]

The expression is set-off in a box because of its importance. It will be used later as part of the proof of the Central Limit Theorem.

The first and second derivatives of \( m(t) \) with respect to \( t \) are

\[ m'(t) = (\mu + \sigma^2 t) e^{\frac{\mu + \sigma^2 t^2}{2}} \]

and

\[ m''(t) = [\sigma^2 + (\mu + \sigma^2 t)^2] e^{\frac{\mu + \sigma^2 t^2}{2}}. \]

Evaluate the two derivatives at \( t = 0 \)

\[ m'(0) = \mu \]

and

\[ m''(0) = \sigma^2 + \mu^2. \]

Thus, the mean and variance of the normal density are

\[ E(X) = \mu \text{ and } V(X) = \sigma^2 \]

If the normal density has a mean of zero and a variance of 1 (often called the standard normal density), then its moment generating function is

\[ m(t) = e^{\frac{t^2}{2}} \]

Remember this expression because it plays an important part in the proof the Central Limit Theorem.

---

3 This can be proven in the same way that the integral of the normal distribution is shown to equal 1. In fact this expression is the integral of a normal density function with expected value \( \mu + \sigma t \) and variance \( \sigma^2 \).
Chebyshev et.al.

The following are some important results about expectations and probability.

**Theorem 2.1**

Let $X$ be a random variable and $g(\cdot)$ a non-negative function with domain the real line; then

$$ P( g(x) \geq k ) \leq \frac{E( g(x) )}{k}. $$

Proof.

I will prove this for continuous densities, **students should prove the theorem for discrete densities**.

For continuous densities

$$ E(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx = \int_{x: g(x) \geq k} g(x)f(x)\,dx + \int_{x: g(x) < k} g(x)f(x)\,dx \geq \int_{x: g(x) \geq k} kf(x)\,dx = kP(g(x) \geq k). $$

The first inequality is a result of the non-negativity of $g(\cdot)$. //

**Theorem 2.2**  **Chebyshev's Inequality**

$$ P( |X - \mu_X| \geq r\sigma_X ) = P((X - \mu_X)^2 \geq r^2\sigma^2_X ) \leq \frac{1}{r^2} \text{ for } r > 0. $$

Proof.

Let $g(X) = (X - \mu_X)^2$, from which it follows that $E(g(x)) = \sigma^2_X$, and let $k = r^2\sigma^2_X$.

Applying Theorem 2.1

$$ P((X - \mu_X)^2 \geq r^2\sigma^2_X ) \leq \frac{\sigma^2_X}{r^2\sigma^2_X} = \frac{1}{r^2}. /// $$

The next result, called Jensen's Inequality, concerns the nature of the expected value of concave and convex functions. For that reason an explanation of convexity and concavity precedes the formal Theorem.

**Convexity**: A function $g(\cdot)$ is convex if, for every $x$, there exists a line, $l(x) = a + bx$, that goes through the point $(x, g(x))$ and lies on or under the graph of $g(\cdot)$.

**Concavity**: A function $g(\cdot)$ is concave if, for every $x$, there exists a line, $l(x) = a + bx$, that goes through the point $(x, g(x))$ and lies on or above the graph of $g(\cdot)$.  

55
**Theorem 2.3 (Jensen's Inequality):** Let $X$ be a random variable with expected value $E(X)$ and let $g(\cdot)$ be a \begin{align*} \text{convex function, then } E(g(X)) &\geq g(E(X)) \\ \text{concave function, then } E(g(X)) &\leq g(E(X)). \end{align*}

The following is a proof for a concave function, **students are responsible for proving the proposition in the case of a convex function.**

**Proof.** By concavity there exits a line $l(x) = a + bx$ such that $l(x) \geq g(x)$ and $l(E(x)) = g(E(x))$.

\[ \begin{align*} E(l(x)) &= a + bE(X) = l(E(X)) \\ g(E(X)) &= l(E(X)) = E(l(x)) \geq E(g(X)) \text{ because } l(x) \geq g(E(X)) \text{ for all } x./// \end{align*} \]

**Expected Utility and Insurance**
Jensen's inequality can be used to explain individuals' choices among risky alternative. In particular it is an useful tool to model the motives for purchasing insurance.

As a first step in explaining the insurance decision, consider an individual with wealth of $100,000 facing the possibility, with known probability, $p$, of a large loss of $90,000$. Suppose the individual's utility is a concave function of his wealth, for instance: 

$$U(W) = \ln W$$

(note that the line $1(W) = \ln(W - 1) + \frac{1}{W}$ lies everywhere above $U(W)$ except at $W$ where it is equal to $U(W)$ and this is true for all $W$). Jensen's inequality tells us that $E(U(W)) \leq U(E(W))$. In this case $E(U(W)) = (1-p)U(100,000) + pU(10,000)$ and $E(W) = (1-p)100,000 + p(10,000)$. If, for instance, $p = 0.1$, then $E(W) = 91,000$ and we know that a $90\%$ chance at $100,000$ and a $10\%$ chance at $10,000$ is worth less than $91,000$ to the individual. What this means is that the individual would be willing to accept less than $91,000$ if it were guaranteed with certainty, than to take a chance at sustaining a $90,000$ loss with probability $0.1$. The amount less than $91,000$ that he would be willing to accept is the amount he would be willing to pay as in insurance premium to guarantee that he would be fully compensated for his loss. In this, the maximum insurance premium he is willing to pay, say $C^*$, can be calculated

$$0.9\ln(100,000) + 0.1\ln(10,000) = \ln(100,000 - C^*)$$

or

$$C^* = 100,000 - 100,000^{0.9}10,000^{0.1} = 20,567.18.$$  

This means that the individual is indifferent between accepting the potential loss and having $79,433.82 with certainty, i.e., he would purchase any insurance policy for up to $C^*$ as long as it guaranteed him full compensation in case of loss.

The next step is to figure out how the insurance company does its financial calculations. For this we will use the Law of Large Numbers. Before stating the theorem it is necessary to define the sample mean. Many of the properties of the sample mean will be discussed, in detail, in later chapters, but here it suffices to define it as follows:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$  

It is the average value of the random variable $X$ found in a sample of size $n$, from the population, in which the sample values are $x_i$. The variance of a sample average is $\sigma^2/n$. For instance if the sample was a series of Bernoulli trials, with 1 representing success and 0 failure, the average would be the proportion of success in a sample of $n$. The variance of the sample proportion is $p(1-p)/n$. The Bernoulli trial is relevant to the insurance example, as will be explained later.

**Students should show that** $E(\bar{X}) = E(X)$ and $V(\bar{X}) = \sigma^2 / n$.

**Theorem 2.4 (Weak Law of Large Numbers (WLLN))**
Let \( f(. ) \) be a density with mean \( \mu \) and finite variance \( \sigma^2 \), and let \( \bar{X}_n \) be the mean of a sample of size \( n \) from \( f(. ) \). Let \( \varepsilon \) and \( \delta \) be any two specified numbers satisfying \( \varepsilon > 0 \) and \( 0 < \delta < 1 \). If \( n \) is any integer greater than \( \sigma^2 / (\varepsilon^2 \delta) \), then

\[
P(-\varepsilon < \bar{X}_n - \mu < \varepsilon) \geq 1 - \delta
\]

Proof:

\[
P(\varepsilon \geq k) \leq E(\varepsilon) / k \text{ from Theorem 2.1.}
\]

Thus

\[
-P(\varepsilon \geq k) \geq -E(\varepsilon) / k
\]

and

\[
1 - P(\varepsilon \geq k) \geq 1 - E(\varepsilon) / k
\]

but

\[
1 - P(\varepsilon \geq k) = P(\varepsilon < k).
\]

Thus

\[
P(\varepsilon < k) \geq 1 - E(\varepsilon) / k.
\]

Let \( g(x) = (X - \mu)^2 \) and \( k = \varepsilon^2 \) and substitute

\[
P((X - \mu)^2 < \varepsilon^2) \geq 1 - \frac{E(X - \mu)^2}{\varepsilon^2} = 1 - \frac{\sigma^2}{\varepsilon^2}.
\]

Define \( \delta = \frac{\sigma^2}{\varepsilon^2} \) and the theorem is proved.///

The law of large numbers suggests that the population mean can be estimated with arbitrary accuracy with sufficiently large observations. While there are innumerable examples in which this is applied, it is particularly relevant to the insurance example. It might be true that individuals wish to insure against potential loss, there would be no way to purchase that insurance unless there were those who profit by selling it.

The potential market is made up of ten million identical people with identical preference functions (\( \ln(W) \)), initial wealth of $100,000 and independent probability (0.1) of losing $90,000. Suppose there are ten companies each insuring one million against the loss. The insured are randomly distributed among the companies. The potential payout loss for each company ranges from 0, if no customer is damaged, to $90 billion, if all customers are damaged. With this range of potential loss, what premium will be charged?

To answer this question it is useful to think of this as sampling one million people, from a very large population and counting the number who are damaged. Assume that the sampling is done with replacement so that each draw from the population is independent of all others.\(^4\) The sample average, namely the number damaged divided by one million, is the proportion of those insured to whom the

---

\(^4\) This is not strictly accurate because the drawing cannot be done with replacement. Nonetheless, the error is small because of the large population from which the sample is drawn.
insurance company must pay $90,000. The Law of Large Numbers can be employed to put bounds on the probability that the proportion of damaged deviates significantly from the population probability.

The variance of the sample proportion is \( p(1-p)/n \). Thus, if the sample proportion is designated \( \bar{p} \),

\[
P(-\epsilon < \bar{p} - 0.1 < \epsilon) \geq 1 - \frac{0.09}{\epsilon^2 \times 10^6}.
\]

Suppose the insurance company want to know the probability that the proportion of their damaged insured is in the range 0.09 to 0.11 (\( \epsilon = 0.01 \))

\[
P(-0.01 < \bar{p} - 0.1 < 0.01) \geq 1 - \frac{0.09}{.0001 \times 10^6} = 0.991.
\]

The insurers are virtually certain that they will be making no more than 110,000 (1.1 X 10^9) payouts. With probability of greater than .99 the insurers will pay no more than $9.9 billion in damage claims. As calculated above, insurers are willing to pay in excess of $20,000 to insure against the loss. If a company charged that amount, it revenue would be $20 billion. There is a probability of .99 that the companies profits would be no less than $10.1 billion.\(^5\)

\(^5\) Shit Happens – Do not forget that there is a small (but not zero) probability that the insurers loss is catastrophic. Insurers, while gambling at very favorable odds, are, nonetheless, known to lose everything (go bankrupt).
Chapter 3

Multivariate Distribution
Economists are often interested in the behavior of more than one random variable. While it is interesting to contemplate the probability of winning at dice, or suffering a large insurance loss, we are most often contemplating, for instance, the relationship between investment and gross domestic product; income, prices and demand for goods and services and labor supply, wages, education and family background. Most economic problems focus on the behavior and interrelationship of many random variables. In order to study these issues it is necessary to understand multivariate distributions – distribution of vectors rather than single random variables.

As a beginning, consider a simple problem like the number of dots on each of two dice – one green and the other red. The dice are both fair, so the probability of any number on each is 1/6. Let X be the number of dots showing on the green die and Y the number of dots showing on the red one. The potential joint values of the random variables X and Y, written as ordered pairs \((x,y)\), are

\[(1,1) \quad (1,2) \quad (1,3) \quad (1,4) \quad (1,5) \quad (1,6)\]
\[(2,1) \quad (2,2) \quad (2,3) \quad (2,4) \quad (2,5) \quad (2,6)\]
\[(3,1) \quad (3,2) \quad (3,3) \quad (3,4) \quad (3,5) \quad (3,6)\]
\[(4,1) \quad (4,2) \quad (4,3) \quad (4,4) \quad (4,5) \quad (4,6)\]
\[(5,1) \quad (5,2) \quad (5,3) \quad (5,4) \quad (5,5) \quad (5,6)\]
\[(6,1) \quad (6,2) \quad (6,3) \quad (6,4) \quad (6,5) \quad (6,6)\]

There are 36 equally probable outcomes, \(f(x,y) = 1/36\) for all feasible X,Y pairs. For pedagogical purposes the probabilities of the X,Y outcomes are given in the following table

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>g(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>h(Y)</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td></td>
</tr>
</tbody>
</table>

The entries in the interior of the table represent the **joint density** (joint probability in the case of discrete random variables) of the X and Y. The entries in the last row and column are the **marginal densities**. Since, by construction, the joint events \((X,Y)\) are mutually exclusive, the marginal densities are the row and column sums of the joint densities. The random variables are distributed independently if

\[f(x,y) = g(X)h(Y).\]

That is, if the joint density is the product of the two marginal densities.
From the point of view of an economic researcher engaged in empirical studies, if two variables are independently distributed, nothing can be learned about one from knowledge of the other. In the case above, the researcher would know nothing more about the probability of a six, say, on the red die, given that a green one was three. Suppose that demand for oranges and personal income are both independently distributed random variables. Then the prediction about whether or not a person will purchase an orange is unaffected by knowledge of her income.

Now, consider a slight variation on the above example. In this case X is as it was before – the number of dots showing on the green die. But, this time Y is the sum of the dots showing on the green and red. The sample space for this is

$$(1,2) (1,3) (1,4) (1,5) (1,6) (1,7)$$

$$(2,3) (2,4) (2,5) (2,6) (2,7) (2,8)$$

$$(3,4) (3,5) (3,6) (3,7) (3,8) (3,9)$$

$$(4,5) (4,6) (4,7) (4,8) (4,9) (4,10)$$

$$(5,6) (5,7) (5,8) (5,9) (5,10) (5,11)$$

$$(6,7) (6,8) (6,9) (6,10) (6,11) (6,12)$$

The joint density of every x,y pair is, as before 1/36. However, the table of X Y outcomes is different than before, as illustrated in the following table

<table>
<thead>
<tr>
<th>Y</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>g(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
<td></td>
</tr>
</tbody>
</table>

The joint density does not meet the criterion for independence: They are not the product of the marginal densities. In terms relevant to the economic researcher, knowledge of X gives information about the probability of Y. For instance if we know X = 2, then we know that Y = 2 as well as 9 through 12 are impossible (probability = 0). We also know that the probability is 1/6 for any Y between, and including 3 and 8.

**Multinomial Distribution**

The multinomial distribution is a generalization of the binomial distribution. Here we consider drawing from a population in which there are k+1 different possible outcomes. Think, perhaps, of an urn with M balls of which there are k different colors (M_1 white balls, M_2 red balls, M_3 green balls …M_k yellow balls) each ball, in addition to a color is labeled with a unique number, for instance, 1,2 …, M.

Suppose a sample of size n is drawn with replacement. The random variables, X_i (i = 1, …, k), are the number of type i balls drawn. We will distinguish the designation of
the random variable from the actual outcome that is labeled with a lower case letter. As described, \( x_i \) is a specific experimental value for the random variable \( X_i \). The question that suggests the form of the multinomial distribution is what is the probability of a specific experimental outcome, namely, what is \( P(X_1 = x_1 \cdots X_k = x_k) \).

Because there are \( M_i \) type \( i \) balls the number of labelings (that is the number of different number combinations) of the \( x_i \) drawn is \( M_i^{x_i} \). Thus the total number of labelings is \((M_1^{x_1})(M_2^{x_2})\cdots(M_{k-1}^{x_{k-1}})(M - \sum_{i=1}^{k-1} M_i)\). For each of the \( M_i^{x_i} \) labelings for type one balls, there are \( \binom{n}{x_1} \) different ways (that is the number of the draw) that the specific numbers can be picked. Then for type two, given that \( x_1 \) of the positions are taken, there are \((n - x_1)\) positions left. Therefore for each of the \( M_2^{x_2} \) labeling for type two balls there are \( \binom{n-x_1}{x_2} \) different ways that those numbers can be picked. Finally for the \( k+1 \) types there are \( \binom{n-\sum_{i=1}^{k-1} x_i}{x_{k-2}} \) different ways to draw those numbers. Since the probability of picking a type \( i \) ball on any draw is \( p_i = M_i/M \), the probability of picking one specific ordering of type \( i \) balls is \( p_i^{x_i} \). Therefore the probability of a specific count of balls is

\[
P(X_1 = x_1 \cdots X_k = x_k) = \binom{n}{x_1} \binom{n-x_1}{x_2} \cdots \binom{n-\sum_{i=1}^{k-1} x_i}{x_i} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}
\]

Students should show that this expression is equivalent to

\[
P(X_1 = x_1 \cdots X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \prod_{i=1}^{k} p_i^{x_i}.
\]

The density function is clearly non-negative and applying the Multinomial Theorem proves that it sums to one.
Continuous Distributions

In the examples so far, I have been careful to label the distribution functions as densities rather than probabilities. For discrete variables, these two are the same. However to continuous ones they are not, as you know already. As an example consider two random variables X and Y with uniform distributions on the 0,1 interval. This means that the value of the density function is the same for every point in the unit square.

The value of the joint density function for the two random variables is

$$f_{XY}(x, y) = \begin{cases} 
1 & \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\
0 & \text{otherwise}
\end{cases}.$$  

$f_{XY}$ satisfies one condition for a density, namely it is strictly non-negative. Further more it integral over X and Y $\int_0^1 \int_0^1 dydx$ is equal to 1. The marginal densities, defined similarly as in the discrete case are $f_X(\cdot) = \int_0^1 f_{XY}(x, y)dy$ and $f_Y(\cdot) = \int_0^1 f_{XY}(x, y)dx$. In the case of the joint uniform distribution given above $f_X(x) = f_Y(y) = 1$ for all $x, y \in [0,1]$.

The condition for independence is the same for the continuous as it is for the discrete variable: X and Y are independently distributed if and only if $f_{XY}(x, y) = f_X(x)f_Y(y)$. This test is passed for the joint uniform distribution given above. The distribution passes the intuitive test for independence as well: nothing is learned about the probability of any particular range the random variable X (Y) from knowledge of the value (or range) of Y (X). Each random variable is uniformly distributed on the unit interval, independent of the value of the other.

---

6 For the discrete random variables the marginal density is the sum rather than the integral.
Since the $X$ and $Y$ are independently distributed, it is expected that their conditional distributions are the same as their marginal distribution. Not surprisingly, the conditional distributions are defined as they are for the discrete distributions:

\[
\begin{align*}
   f_{X|Y}(x | y) &= \frac{f_{XY}(x, y)}{f_Y(y)} \\
   f_{Y|X}(y | x) &= \frac{f_{XY}(x, y)}{f_X(x)}.
\end{align*}
\]

For this example the value of the conditional distribution of $X$ given $Y$ and that of $Y$ given $X$ are the same, namely

\[
   f_{X|Y}(x | y) = 1 = f_{Y|X}(y | x).
\]

A second example, taken from Rice, page 79, derives a non-independently distributed joint density function and provides a simple illustration how the knowledge of one variable provides information about the other. The random variables are the $X$ and $Y$ coordinates of points in a circle with its center at the origin and a radius of 1. The area of such a circle is $\pi$. Thus the density function for a point chosen at random from the circle is

\[
f_{XY}(x, y) = \begin{cases} 
\frac{1}{\pi} & \text{if } x, y \in [0,1] \text{ and } x^2 + y^2 \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

The marginal density, of, for instance, $X$ is found by integrating $f_{XY}(x, y)$ over $Y$:

\[
f_X(x) = \frac{1}{\pi} \int_{\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy = \frac{2}{\pi} \sqrt{1-x^2} \text{ for } -1 \leq x \leq 1.
\]

It then follows that the conditional of $y$ on $x$ is

\[
f_{Y|X}(y | x) = \frac{1}{2\sqrt{1-x^2}}.
\]

It is clear from the conditional density that $X$ and $Y$ are not independently distributed. Knowledge of one gives significant information about the probable value of the other. For instance, suppose you know that $x = 0.5$ then it is known that $y$ must be value in the
interval from 0.75 to 0. If, on the other hand, $x = 0.75$ the possible value of $y$ must lie in the smaller interval from $9/16$ to 0.

As another example consider the following function

$$f_{XY}(x, y) = \begin{cases} 
ke^{\lambda(x+y)} & \text{for } -1 \leq x \leq 1 \text{ and } -1 \leq y \leq 1 \\
0 & \text{otherwise}
\end{cases}$$

The function is non-negative, so it is necessary only to set the constant, $k$, to a value such that

$$\int_{-1}^{1} \int_{-1}^{1} f_{XY}(x, y) dy dx = 1 = k \int_{-1}^{1} \int_{-1}^{1} e^{\lambda(x+y)} dy dx.$$

Students should be able to show that for $f_{XY}(\cdot)$ to be a density function the constant $k = \frac{\lambda^2 e^{2\lambda}}{(1 - e^{2\lambda})^2}$. Student should also find the marginal, as well as the conditional, densities for the two random variables. Are $X$ and $Y$ independently distributed?

The final, and most important example of a multivariate distribution is the bivariate normal density. The density function itself is a complicated expression. It is sufficiently important that the student should make an attempt to remember its form.

$$f_{XY}(x, y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y}\right)\right).$$

Because of its complexity, it is convenient to use the following shorthand:

$$\phi_1 \equiv \frac{1}{\sqrt{2\pi} \sigma_X \sqrt{1-\rho^2}} \quad ; \quad \phi_2 \equiv \frac{1}{\sqrt{2\pi} \sigma_Y} \quad ; \quad u \equiv \frac{x-\mu_X}{\sigma_X} \quad \text{and} \quad v \equiv \frac{y-\mu_Y}{\sigma_Y}.$$ The integral of the joint density function is

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_1 \phi_2 \exp\left(-\frac{1}{2(1-\rho^2)} [u^2 - 2\rho uv + v^2]\right) dy dx.$$

By adding and subtracting $\rho^2 v^2$ to the expression in the exponent

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_1 \phi_2 \exp\left(-\frac{1}{2(1-\rho^2)} [u^2 - 2\rho uv + \rho^2 v^2 + (1-\rho^2)v^2]\right) dy dx =$$
\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ \phi_1 \exp - \frac{1}{2(1-\rho^2)} (u - \rho v)^2 \right] \phi_2 \exp - \frac{v^2}{2} \, dx \, dy .
\]

Since \( x \) appears only in the first term within the square brackets, the integral expression can be rewritten as

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ \phi_1 \exp - \frac{1}{2(1-\rho^2)} (u - \rho v)^2 \right] dx \phi_2 \exp - \frac{v^2}{2} \, dy
\]

The first square bracketed term \( \left[ \phi_1 \exp - \frac{1}{2(1-\rho^2)} (u - \rho v)^2 \right] \) is the density function of a normal random variable whose values are \( x \) with expected value \( \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y) \) and variance \( \sigma^2 (1 - \rho^2) \). Since the integral is over the full range of \( x \)

\[
\int_{-\infty}^{\infty} \phi_1 \exp - \frac{1}{2(1-\rho^2)} (u - \rho v)^2 \, dx = 1.
\]

Then the full integral can be evaluated as

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[ \phi_1 \exp - \frac{1}{2(1-\rho^2)} (u - \rho v)^2 \right] dx \phi_2 \exp - \frac{v^2}{2} \, dy = \int_{-\infty}^{\infty} \phi_2 \exp - \frac{v^2}{2} \, dy
\]

that is recognized as the integral of a normal density function and is itself equal to 1. Therefore

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy \, dx = 1
\]

as was to be shown.///

**Expectations and Moments**

The expectation and variance of multivariate joint density functions are defined much the same as they are for univariate ones. For the discrete densities they are

\[ E(X) = \sum_x \sum_y x f_{XY}(x, y) \]

and
\[
V(X) = \sum_{x} \sum_{y} (x - E(X))^2 f_{XY}(x, y).
\]

And, for continuous case there are similar definitions

\[
E(X) = \iint_{xy} x f_{XY}(x, y) \, dy \, dx
\]

and

\[
V(X) = \iint_{xy} (x - E(X))^2 f_{XY}(x, y) \, dy \, dx \equiv \sigma_X^2.
\]

Students should be able to show that for both the discrete and continuous cases the variance can be written \( V(X) = E(X^2) - E(X)^2 \).

The covariance of a joint density is a measure of the association between the random variables. It is defined

\[
C(XY) = \sum_{x} \sum_{y} (x - E(X))(y - E(Y)) f_{XY}(x, y)
\]

for the discrete case and

\[
C(XY) = \iint_{xy} (x - E(X))(y - E(Y)) f_{XY}(x, y) \, dy \, dx
\]

for the continuous case.

Students should be able to show that \( C(XY) = E(XY) - E(X)E(Y) \) for both discrete and continuous distributions. You should also show that \( C(XY) = 0 \) if the random variables are independently distributed.

**Conditional Expectation and Variance**

In this section the conditional expectation and variance defined and some of their properties examined. For expositional ease, the discussion is limited to the bivariate case: two random variables X and Y. All of the results and examples are applicable to the general multivariate distribution, as long as all of the relevant moments exist. The existence and finiteness of both the first and second moments are assumed.

The conditional expectation of Y for a given value of X is defined
The following theorem that gives a useful property of conditional expectations is proved for the continuous case, **student should prove it for discrete densities**.

**Theorem 3.1**

\[ E_Y(Y) = E_X(E_{Y|X}(y|x)). \]

Proof.

\[ E_Y(y) = \int y f_Y(y) dy = \int y \int f_{XY}(x, y) dx dy = \int f_X(x) \int y f_{Y|X}(y | X = x) dy dx = E_X(E_{Y|X}(Y | X)). \]

**Theorem 3.2**

\[ V(Y) = V(E(Y|X)) + E(V(Y|X)) \]

Proof.

\[ V(Y|X) = E(Y^2|X) - E^2(Y|X). \]

Therefore,

\[ E(V(Y|X)) = E(Y^2) - E[E^2(Y|X)]. \] *(This is a straightforward extension of the Theorem 3.1, student should show that this step is correct).*

\[ V(E(Y|X)) = E[E^2(Y|X)] - E^2[E(Y|X)]. \]

Therefore

\[ V(E(Y|X)) + E(V(Y|X)) = E[E^2(Y|X)] - E^2[E(Y|X)] + E(Y^2) - E[E^2(Y|X)] = E(Y^2) - E(E(Y))E(E(Y)) = E(Y^2) - E^2(Y) = V(Y)./// \]

**Example – The Bivariate Normal**

The density function of the bivariate normal is
Students should show that the conditional density is as follows:

\[
f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_Y^2 \sqrt{1 - \rho^2}}} \exp\left( -\frac{1}{2(1 - \rho^2)} \left( \frac{(y - \mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} \right) \right)
\]

This is the density function of a normally distributed random variable with

\[
E(Y | X) = \mu_Y + \frac{\sigma_Y}{\sigma_X} \rho(x - \mu_X)
\]

and variance

\[
V(Y | X) = (1 - \rho^2)\sigma_Y^2.
\]

**Moment Generating Function**

The moment generating function for multivariate distributions is similar to those in the univariate case. For \(n\) random variables it is

\[
m_{X_1X_2\ldots X_n}(t) = E(e^{t_1X_1 + t_2X_2 + \ldots + t_nX_n}) = E(\prod_{i=1}^{n} e^{t_iX_i}).
\]

An important consequence of the last expression is that if the \(X\)'s are independent random variables, the joint moment generating function factors in a most useful way. The joint moment generating function is the product of the individual moment generating functions.

Now consider the sum of independent random variables

\[
Z = \sum_{i=1}^{n} X_i
\]

The moment generating function for the random variable \(Z\) is
\[ m_Z(t) = E(e^{t \sum X_i}) = E(\prod_{i=1}^{n} e^{tX_i}) = \prod_{i=1}^{n} e^{tX_i} = \prod_{i=1}^{n} m_{X_i}(t) \]

The next to the last step is a consequence of independence from which the claimed result follows.

There is an additional consequence of independence, namely, that if two, or more, random variables (or vector of random variables) are independently distributed, then the moment generating function of functions of these random variables are also independently distributed. This will be shown in two steps for continuous densities student should extend the result to discrete densities.

**Theorem 3.3**

Let \( X = (X_1, \ldots, X_k) \) and \( Y = (Y_1, \ldots, Y_l) \) be independent random (vectors) variables. Let \( g(X) \) and \( h(Y) \) be functions of these random variables. Then the moment generating function

\[ m_{g(X)h(Y)}(s, t) = m_{g(X)}(s) m_{h(Y)}(t). \]

Proof.

\[ m_{g(X)} m_{h(Y)}(s, t) = \iint_{xy} e^{sg(x)} e^{th(y)} f_{XY}(x, y) dx dy \]

By independence

\[ m_{g(X)} m_{h(Y)}(s, t) = \int_{x} e^{sg(x)} f_{X}(x) dx \int_{y} e^{th(y)} f_{Y}(y) dy. \]

**The moments of a multivariate density**

The moments of the distribution, as in the univariate case, the partial derivatives \( m(t) \) evaluated at \( t = 0 \).

\[ \frac{\partial m(t)}{\partial t_j} = E(X_j e^{\Sigma_t X_i}) \quad \frac{\partial^2 m(t)}{\partial t_j^2} = E(X_j^2 e^{\Sigma_t X_i}) \quad \text{and} \quad \frac{\partial^2 m(t)}{\partial t_k \partial t_j} = E(X_k X_j e^{\Sigma_t X_i}) \]

thus

\[ \frac{\partial m(0)}{\partial t_j} = E(X_j) \quad \frac{\partial^2 m(0)}{\partial t_j^2} = E(X_j^2) \quad \text{and} \quad \frac{\partial^2 m(0)}{\partial t_k \partial t_j} = E(X_k X_j e). \]
We will derive, and demonstrate the computation of some of the moments, for both the multinomial and bivariate normal densities.

**Multinomial Density**

\[
m(t) = \sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \ldots \sum_{x_k=0}^{n-\sum_{i=1}^{k-1} x_i} \frac{n!}{\prod_{i=1}^{k} x_i} \left( \prod_{j=1}^{k} x_j \right) = \sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \ldots \sum_{x_k=0}^{n-\sum_{i=1}^{k-1} x_i} \frac{n!}{\prod_{i=1}^{k} x_i} \left( \prod_{j=1}^{k} x_j \right) \cdot \]

From the Multinomial Theorem, the moment generating function for the multinomial distribution is

\[
m(t) = \left( \sum_{i=1}^{k} e^{t_i} p_i \right)^n.
\]

The relevant partial derivatives are

\[
\frac{\partial m(t)}{\partial t_i} = n e^{t_i} p_i \left( \sum_{j=1}^{k} e^{t_j} p_j \right)^{n-1};
\]

\[
\frac{\partial^2 m(t)}{\partial t_i^2} = n e^{t_i} p_i \left( \sum_{j=1}^{k} e^{t_j} p_j \right)^{n-1} \cdot n(n-1)(e^{t_i} p_i)^2 \left( \sum_{j=1}^{k} e^{t_j} p_j \right)^{-2}
\]

and

\[
\frac{\partial m(t)}{\partial t_i \partial t_j} = n(n-1) e^{t_i} e^{t_j} p_i p_j \left( \sum_{j=1}^{k} e^{t_j} p_j \right)^{n-2}.
\]

The first two moments and the cross-product moments are found as
\[ E(X_i) = \frac{\partial m(t)}{\partial t} |_{t=0} = np_i ; \]
\[ E(X_i^2) = \frac{\partial^2 m(t)}{\partial t^2} |_{t=0} = np_i + n(n-1)p_i^2 \]

and

\[ E(X_i X_j) = \frac{\partial^2 m(t)}{\partial t_i \partial t_j} |_{t=0} = n(n-1)p_i p_j . \]

The variance and covariance are

\[ V(X_i) = E(X_i^2) - E^2(X_i) = np_i (1 - p_i) \]

and

\[ C(X_i X_j) = E(X_i X_j) - E(X_i)E(X_j) = -np_i p_j . \]

**Bivariate Normal Density**

The two random variables are \( X \) and \( Y \) for this example and the moment generating function is

\[ m(t_1 t_2) = E(e^{(t_1 X + t_2 Y)}) = e^{(t_1 \mu_X + t_2 \mu_Y)} E(e^{(t_1(X-\mu_X) + t_2(Y-\mu_Y))}) \]

Define \( u = \frac{x - \mu_X}{\sigma_X} \) and \( v = \frac{y - \mu_Y}{\sigma_Y} \). The moment generating function is then

\[ m(t_1 t_2) = \frac{e^{(t_1 \mu_X + t_2 \mu_Y)}}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{[t_1 \sigma_X u + t_2 \sigma_Y v - \frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)]} \, du \, dv \]

With considerable manipulation, the moment generating function for the bivariate normal density is

\[ m(t_1 t_2) = e^{[t_1 \mu_X + t_2 \mu_Y + \frac{1}{2}(t_1^2 \sigma_X^2 + 2\rho t_1 t_2 \sigma_X \sigma_Y + t_2^2 \sigma_Y^2)]} \]
The first and second moment for X are found by evaluating the first and second partial derivatives with respect to $t$

$$\frac{\partial m(t_1 t_2)}{\partial t_1} = (\mu_X + t_1\sigma_X^2 + \rho t_2 \sigma_X \sigma_Y ) m(t_1 t_2)$$

and

$$\frac{\partial^2 m(t_1 t_2)}{\partial t_1^2} = [\sigma_X^2 + \mu_X^2 + (t_1 \sigma_X^2 + \rho t_2 \sigma_X \sigma_Y )\mu_X + (t_1 \sigma_X^2 + \rho t_2 \sigma_X \sigma_Y )^2]m(t_1 t_2).$$

Note that $m(00) = 1$, then the first and second partial derivative evaluated at $t_1 = t_2 = 0$ are

$$\frac{\partial m(t_1 t_2)}{\partial t_1} |_{t=0} = \mu_X$$

and

$$\frac{\partial^2 m(t_1 t_2)}{\partial t_1^2} |_{t=0} = \sigma_X^2 + \mu_X^2.$$

Thus

$$E(X) = \mu_X$$

and

$$V(X) = E(X^2) - E^2(X) = \sigma_X^2.$$ 

We find the covariance of X and Y by first finding the cross-partial derivative

$$\frac{\partial m(t_1 t_2)}{\partial t_1} = [(\mu_X + t_1\sigma_X^2 + \rho t_2 \sigma_X \sigma_Y ) (\mu_Y + t_2 \sigma_Y^2 + \rho t_2 \sigma_X \sigma_Y ) + \rho \sigma_X \sigma_Y ] m(t_1 t_2)$$

and evaluating at $t = 0$

$$\frac{\partial m(t_1 t_2)}{\partial t_1} |_{t=0} = \mu_X \mu_Y + \rho \sigma_X \sigma_Y.$$

The covariance is

$$C(XY) = \rho \sigma_X \sigma_Y.$$
It is immediate from this that the correlation coefficient is

$$
\rho = \frac{C(XY)}{\sigma_X \sigma_Y}.
$$

**Theorem 3.4  Cauchy-Swartz Inequality**

Let \(U\) and \(V\) have finite second moments, then

$$
E^2(UV) = |E(UV)|^2 \leq E(U^2)E(V^2),
$$

with equality if and only if \(P(V = cU) = 1\) for some constant \(c\).

**Proof:**

The inequality certainly holds for \(E(U^2)\) or \(E(V^2)\) = 0 since that would mean that Either (or both) are constants equal to 0. In this case, of course \(E(UV) = 0\). Furthermore the inequality certainly holds in the case where either \(E(U^2)\) or \(E(V^2)\) or both are infinite.

For the intermediate case let \(a\) and \(b\) be two real numbers. It is then the case that

$$
0 \leq E[(aU + bV)^2] = a^2E(U^2) + b^2E(V^2) + 2abE(UV) \quad (A)
$$

and

$$
0 \leq E[(aU - bV)^2] = a^2E(U^2) + b^2E(V^2) - 2abE(UV). \quad (B)
$$

Choose \(a = \sqrt{E(V^2)}\) and \(b = \sqrt{E(U^2)}\). In this case the two inequalities can be written

$$
0 \leq 2\sqrt{E(U^2)E(V^2)}[\sqrt{E(U^2)E(V^2)} \pm E(UV)] \quad (C)
$$

Therefore

$$
0 \leq \sqrt{E(U^2)E(V^2)} \pm E(UV). \quad (D)
$$

Since the inequality must hold for either + or - , the first part of the theorem follows.

For (D) to hold with equality, the term on the left of the equal sign in (B) must equal zero. Since that is the expectation of a square, the difference between \(ax\) and \(by\) must equal zero for all values of \(x\) and \(y\), namely \(P(aU - bV = 0) = 1\). For that to be true then \(V = \frac{a}{b}U\) which is the second part of the theorem for \(c = a/b.\\)

If we let \(U = X - E(X)\) and \(V = Y - E(Y)\), then from the Cauchy-Schwartz inequality
\[-\sqrt{\sigma_X^2 \sigma_Y^2} \leq C(XY) \leq \sqrt{\sigma_X^2 \sigma_Y^2}\]

Using this inequality and dividing through by the square root of the product of the variances, we get a natural definition and range of the correlation coefficient, \( \rho \), namely

\[-1 \leq \rho(XY) \equiv \frac{C(XY)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \equiv \rho(XY) \leq 1.\]
Chapter 4

Sampling
Sampling

Most often we do not know the distribution of the phenomena of interest. Suppose you are hired by Sunkist orange growers to find the demand functions for oranges. Since you have studied Economics you know that the demand is a function of the income of the consumer, the price of oranges and all sorts of other things that determine personal preferences. Sunkist is very interested in the results of your study and have given you sufficient funds for a survey of consumers in which you will discover all sorts of things – their income, the oranges price charged by their market and a bunch of other stuff about them. You are clever enough to design the survey so that your sample of consumers is truly random and each observation, i.e., orange purchase decision, can be treated as independent of all others. Buyer choice and the other 'explanatory' variables are random variables. If you are able to survey the entire population you could discover the exact nature of the distribution, i.e., the density function for each random variable. It is unfortunate that the cost of a universal sample is prohibitive so you must contend with a limited sample and then discern the underlying population densities.

The next chapter will focus on the estimation issues directly; this chapter deals with what we can know from sampling data about the population distributions. First random samples, sample statistics and sample moments are defined and then there properties elaborated.

Definitions

Random Sample

Think of the draws from a population as random variables where $X_1$ is the outcome first draw, $X_2$ the outcome on the second draw, etc., and $X_n$ is the outcome on the final $n$th draw. The joint density function of these random variables is $f_{X_1X_2\cdots X_n}(\cdot)$. This is a random sample if the X's are independently distributed, namely, if

$$f_{X_1X_2\cdots X_n}(x_1x_2\cdots x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

where $f(\cdot)$ is the common density function.

Statistic

A statistic is a function of observable random variables, which is itself an observable random variable (it does not contain any unknown parameters).

Sample Mean or Sample Average

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$
One of the properties of the sample average is that it expected value is equal to the expected value (mean) of the population density from which the sample is drawn:

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu. \tag{7}$$

Any sample statistic whose expected value is equal to an underlying density parameter is called an **unbiased** estimate of that parameter. Therefore the sample average is an unbiased estimate of the population mean.

Unbiasedness is a property of all sample averages, no matter the density from which the sample is drawn (conditional, of course on the existence of the mean of the underlying density). An important special case applies to a random sample drawn from a normal distribution. As the subsequent Central Limit Theorem shows, with large enough sample sizes most distributions can be approximated by the normal.

**Theorem 4.1.**

The sample average of a random sample of size $n$, $\bar{X}_n$, from a normal density, is distributed normal with mean $\mu$ and variance $\sigma^2$.

Proof.

The moment generating function of the sample average is

$$m_{\bar{X}}(t) = E(e^{\frac{t}{n} \sum X_i}) = E(e^{\sum_{i=1}^{n} \frac{t}{n} X_i}) = \prod_{i=1}^{n} E(e^{\frac{t}{n} X_i})$$

the last step is a consequence of the independence.

Because each of the $X_i$ are normally distributed, the moment generating function, written as a function of $t/n$ is

$$m_{\bar{X}}(t/n) = \prod_{i=1}^{n} e^{\frac{t}{n} \mu + \frac{\sigma^2 t^2}{2n^2}} = e^{\frac{t}{n} \mu + \frac{\sigma^2 t^2}{2n}} = e^{\mu \frac{t}{n} + \frac{\sigma^2 t^2}{2n}}$$

The final term is the moment generating function of a normal density with mean $\mu$ and variance $\sigma^2/n$.\footnote{If the expected value of a sample statistic is equal to a parameter of the density function the sample statistic is an unbiased estimator of the density parameter. There is more about this property in the chapter on estimation.}
The variance of the sample average (it has a variance since it is a random variable) is one-nth the variance of the population density:

\[ V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right). \]

Because this is a random sample, the outcomes from each observation are independent, the pairwise covariances are all 0 and the variance of the sum is the sum of the variances. Thus,

\[ V(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} V(X_i) = \frac{1}{n} \sum_{i=1}^{n} \sigma_X^2 = \frac{1}{n} \sigma_X^2. \]

**Sample Variance**

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \]

The sample variance is the sum of squares of the deviation from the sample mean, divided by \( n-1 \) rather than \( n \) as might have been expected. The reason for this is that the expected value of the sample variance, as defined, is equal to the variance of the underlying density function:

\[ E(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} E(X_i - \bar{X})^2 = \sigma^2. \]

In order to show this note that \( (X_i - \bar{X})^2 = (X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) \).

It follows that \( E(X_i - \bar{X})^2 = \sigma^2 + \frac{1}{n} \sigma^2 - 2C(X_i, \bar{X}) \). The only term that need further evaluation is the covariance between the two random variables: the outcome of observation \( i \) and the average of all observations. The average is the \( 1/n \) times the sum of the outcomes on all observations. Since the observation outcomes are independent, the covariances are all zero except for the covariance between an observation outcome and itself. That covariance \( C(X_i, X_i) = E(X_i - \mu)(X_i - \mu) = \sigma^2 \). Thus \( C(X_i, \bar{X}) = \frac{1}{n} \sigma^2 \) and

\[ E(X_i - \bar{X})^2 = \sigma^2 - \frac{1}{n} \sigma^2. \]

It follows that \( E(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} E(X_i - \bar{X})^2 = \sigma^2 \).  

\[ ^8 S^2 \text{ is an unbiased estimator of } \sigma^2. \]
Central Limit Theorem

The sample mean has a remarkable property: no matter the distribution from which the sample is drawn, as long as it has a finite variance, the distribution of the sample mean is approximately normal for very large sample size. The phrase used is that the asymptotic distribution of the sample mean is normal.

As an aid to understanding the proof of the central limit theorem two mathematical facts precede its formal statement and proof.

The first mathematical fact is the evaluation of a function by a Taylor Series:

A differential function can be expressed as a polynomial

\[ f(x) = f(a) + (x - a)f'(a) + \frac{(x - a)^2}{2!}f''(a) + \cdots + R^n \]

where

\[ R^n = \frac{f^n(a + \theta(x - a))}{n!}(x - a)^n \quad 0 \leq \theta \leq 1. \]

Consider the function with the graph

With \( a = 0 \) and \( R^1 = [f'(\theta x)/1!]x \), \( f(x) = f(0) + R^1 \).

The second mathematical fact is \( \lim_{n \to \infty} (1 + \frac{t^2}{2n})^n = e^{t^2/2} \).

**Theorem 6.4 (Central Limit Theorem):**

Let \( X_1, \ldots, X_n \) be a random sample from a distribution with mean \( \mu \) and positive variance \( \sigma^2 \). Assume that the moment generating function for the distribution, \( M(t) \) exists for \( -h \leq t \leq h \)
Then the random variable \( Y_n = \left( \sum_{i=1}^{n} X_i - n\mu \right) / \sigma \sqrt{n} = \sqrt{n}(\bar{X} - \mu) / \sigma \) has a limiting distribution which is normal with mean zero and variance 1 \( N(0,1) \).

Proof:

The moment generating function for \( Y_n \) is

\[
M_{Y_n}(t, n) = E[\exp(t \left( \sum_{i=1}^{n} X_i - n\mu \right) / \sigma \sqrt{n})]
\]

because the random variables are independently and identically distributed (iid) the moment generating function can be rewritten

\[
M_{Y_n}(t, n) = \left( E[\exp(t \frac{X_i - \mu}{\sigma \sqrt{n}})] \right)^n
\]

this can be written as the mean deviate form of the moment generating function, namely

\[
M_{Y_n}(t, n) = [m(t \frac{n}{\sigma \sqrt{n}})]^n
\]

By Taylor’s Theorem

\[
[m(t \frac{n}{\sigma \sqrt{n}})]^n = [m(0) + m'(0) \frac{t \sigma \sqrt{n}}{\sigma \sqrt{n}} + m''(\xi) \frac{t^2}{2n\sigma^2}]^n
\]

where \( 0 \leq \xi \leq \frac{t}{\sigma \sqrt{n}} \).

It is known that \( m(0) = 1 \) and \( m'(0) = 0 \). Use this and add and subtract \( \frac{t^2\sigma^2}{2n\sigma^2} \) and rewrite the moment generating function

\[
[m(t \frac{n}{\sigma \sqrt{n}})]^n = [1 + \frac{t^2}{2n} + (m''(\xi) - \sigma^2) \frac{t^2}{2n\sigma^2}]^n
\]

Since \( \frac{t}{\sigma \sqrt{n}} \to 0 \) as \( n \to \infty \) then \( \xi \to 0 \) as \( n \to \infty \). It follows that
\[
\lim_{n \to \infty} \left[ \frac{t}{\sigma \sqrt{n}} \right]^n = \lim_{n \to \infty} E[1 + \frac{t^2}{2n}] = E(\exp \frac{t^2}{2}).
\]

The final term is the moment generating function for the N(0,1) distribution. ///

The central limit theorem plays an important role in most economic applications. It is unlikely that we know the 'true' distribution of the random variables we study. Nonetheless, we invoke the central limit theorem and assert that the random variables are normally distributed if the number of observations is sufficiently large. The usefulness of this large sample approximation will become obvious in later study of Econometrics; however, the following demonstrates how the approximation approaches the true distribution as the sample size increases.

You are sampling, with replacement, from an urn that contains half red balls and half green ones (p = 0.5). You want to know the probability that the sample proportion of red balls is within 0.05 of the true population proportion. If \( \hat{p} \) is the sample proportion, the question can be posed: what is \( P(|\hat{p} - p| \leq 0.05) \)? For the half/half proportion the probability statement reads \( P(0.45 \leq \hat{p} \leq 0.55) \). Define \( k_1 = 0.45n \) and \( k_2 = 0.55n \), as long as \( k_1 \) and \( k_2 \) are integers, the probability of finding a sample proportion within the range specified is

\[
P(0.45 \leq \hat{p} \leq 0.55) = \sum_{k=k_1}^{k_2} \binom{n}{k} \left( \frac{1}{2} \right)^k \left( \frac{1}{2} \right)^{n-k}.
\]

If either one \( k \) or both are not integers, the number can be rounded to the nearest integer. The desired probability is difficult to compute, especially for large \( n \). The central limit theorem suggests an alternative.

Since the sample proportion is approximately normally distributed for large values of \( n \), that is

\[
\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{pq}}
\]

is distributed approximately normal with expected value of 0 and variance of 1. The question about the probability can be answered approximately by reformulating the question to

\[
P(-0.05 \leq \hat{p} - p \leq 0.05) = P(-0.05(\sqrt{n} / \sqrt{pq}) \leq z \leq 0.05(\sqrt{n} / \sqrt{pq}))
\]

where \( z \) is distributed normal with mean zero and variance 1, write \( z \sim N(0,1) \).

In order to show how good an approximation is the normal density to the binomial I calculated the true probability and the normal approximation. These calculations are present in the following table.
Note that the approximate probabilities are closer to the true ones for larger sample sizes. The difference between the true value and the approximate does not appear to approach zero monotonically. Rounding to induce integer values of k introduces small errors in the calculation.

An aside You should notice that the question posed is what is the probability that the sample proportion is within a certain value of the true proportion? This means that with repeated sampling, what proportion of the experiments will yield a sample proportion that is within 0.05 of the true value. The answer to that question gives the confidence level of the interval (in this case p +/- 0.05). The interval itself is labeled the confidence interval.

Chi-Squared ($\chi^2$) Distribution

The $\chi^2$ distribution is used to test hypotheses about the population variance among other thing. The sum of squares of standard normal variable are distributed as $\chi^2$. A first step in the derivation of the distribution is the Gamma function. That function is

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx .$$

Integrating by parts the gamma function is

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx = -x^{k-1}e^{-x}\big|_0^\infty + \int_0^\infty (k-1)x^{k-2}e^{-x}dx .$$
In other words
\[
\Gamma(k) = (k - 1)\Gamma(k - 1).
\]

Note that for \( k = 1 \) \( \Gamma(1) = 1 \), furthermore \( \Gamma(2) = 1\Gamma(1) \), \( \Gamma(3) = 2\Gamma(2) = 2^1\Gamma(1) \) and so forth. So \( \Gamma(n) = (n-1)! \) for integer values of \( n \). For non-integer values, the case is a bit more complicated.\(^9\)

The Chi-squared distribution incorporates the gamma function as follows: If \( X \) is distributed Chi-square with \( k \) degrees of freedom its density function is

\[
f_X(x) = \frac{1}{\Gamma(k/2)} (1/2)^{k/2} x^{k-1} e^{-x/2}.
\]

The moment generating function for this distribution is

\[
m(t) = E(e^{tx}) = \int_0^\infty \frac{1}{\Gamma(k/2)} (1/2)^{k/2} x^{k-1} e^{-(1/2)x} \, dx.
\]

Let \( \lambda = 1/2 - t \), then

\[
m(t) = \left(\frac{1}{2\lambda}\right)^{k/2} \frac{1}{\Gamma(k/2)} \int_0^\infty \frac{k}{\lambda^2} x^{k-1} e^{-\lambda x} \, dx.
\]

\(^9\) Of particular importance is the gamma function when \( k = 0.5 \). Let \( \lambda = 1 \)

\[
\Gamma(1/2) = \int_0^\infty e^{-y^2/2} \, dy.
\]

\( \Gamma(1/2) \) is evaluated by letting \( y = \sqrt{2x} \), (since \( x \in [0,\infty] \), \( y \in [-\infty, +\infty] \) then \( dy = \frac{1}{\sqrt{x}} \) and the integral is

\[
\Gamma(1/2) = \int_0^\infty e^{-y^2/2} \, dy = \int_{-\infty}^\infty e^{-y^2} \, dy = \sqrt{2\pi}.
\]

The last integral resembles the integral of the standard normal density. Now consider an arbitrary fraction with \( n \) in the numerator and 2 in the denominator \( (n/2) \). The gamma function evaluated at \( n/2 \), \( \Gamma(n/2) \) as has been derived above

\[
\Gamma(n/2) = \left(\frac{n}{2}\right)(\frac{n}{2} - 1)(\frac{n}{2} - 2) \cdots (\frac{n}{2} - m)\Gamma\left(\frac{1}{2}\right)\sqrt{\pi}
\]

where \( m = (n-1)/2 \).
The integral is $\Gamma(k/2)$, thus

$$m(t) = \left(\frac{1}{2\lambda}\right)^{k/2} = \left(\frac{1}{1-2t}\right)^{k/2}.$$

The first and second derivatives of the moment generating function with respect to $t$ are

$$m'(t) = k(1-2t)^{-(k+1)}$$

and

$$m''(t) = 2k(k+1)(1-2t)^{-(k+2)}.$$

This means that the expected value of a chi-squared random variable is

$$E(x) = m'(0) = k$$

and the variance is

$$V(X) = E(X^2) - E^2(X) = 2k.$$  

As the next theorem shows, the chi-squared distribution comes up naturally as the sum of squares of standard normal variables.

**Theorem 4.3:**
If the random variables $X_i$, $i = 1 \ldots k$ are normally and independently distributed with mean $\mu_i$ and variances $\sigma^2_i$, then

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

has a chi-squared distribution with $k$ degrees of freedom. This is written

$$U \sim \chi^2_k.$$  

Proof.

Let $Z_i = \left(\frac{X_i - \mu_i}{\sigma}\right)$; $Z_i \sim N(0,1)$.

The moment generating function for $U$ is

$$m_U(t) = E(e^{tU}) = E(e^{t \sum_{i=1}^{k} Z_i^2}) = E\left(\prod_{i=1}^{k} e^{tZ_i^2}\right) = \prod_{i=1}^{k} E(e^{tZ_i^2})$$

the last step follows from independence.
Since \( e^{tZ_i^2} \) is a function of standard normal variables

\[
E(e^{tZ_i^2}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tZ_i^2} e^{-\frac{1}{2}Z_i^2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)Z_i^2} dz = 
\]

\[
\frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)Z_i^2} dz
\]

The expression within the square brackets is recognized as the integral of a normal density with mean 0 and variance \( 1/(1-2t) \). Thus the value of the expression in the square bracket is equal to 1.

The moment generating function for \( U \) is

\[
m_U(t) = \left( \frac{1}{\sqrt{1-2t}} \right)^k = \left( \frac{1}{1-2t} \right)^{k/2}
\]

that has been shown, previously to be the moment generating function of the chi-square distribution with \( k \) degrees of freedom.

Next we show that the sum of squares of the deviations of standard normal random variables from the sample mean is distributed chi-square with \( n - 1 \) degrees of freedom.

**Theorem 4.4.**

If \( Z_1, \ldots, Z_n \) is a random sample from a standard normal distribution

(i) \( \bar{Z} \) has a normal distribution with mean 0 and variance \( 1/n \)

(ii) \( \bar{Z} \) and \( \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \) are independent

(iii) \( \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \) has a chi-squared distribution with \( n-1 \) degrees of freedom.

Proof.

(i) This is a special case of, and follows immediately from, Theorem 4.1.
(ii) We first show that \( \bar{Z} \) and the vector \((Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})\) are independent. Since \( \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \) is a function of the vector, the assertion follows.

Let \( z_i = Z_i - \bar{Z} \) and let \( z \) and \( t \) be the vectors \( z = (z_1, \ldots, z_n) \) and \( t = (t_1, \ldots, t_n) \). The moment generating function

\[
m_{\bar{Z}}(s, t) = E(e^{s\bar{Z} + \sum_{i=1}^{n} t_i z_i}).
\]

Examine

\[
\sum_{i=1}^{n} t_i z_i = \sum_{i=1}^{n} t_i Z_i - n\bar{Z}t.
\]

Then

\[
s\bar{Z} + \sum_{i=1}^{n} t_i z_i = \sum_{i=1}^{n} \left( \frac{s}{n} + (t_i - \bar{t}) \right) Z_i = \sum_{i=1}^{n} a_i
\]

where

\[
a_i = \frac{s}{n} + (t_i - \bar{t});
\]

and

\[
\sum_{i=1}^{n} a_i = s
\]

\[
\sum_{i=1}^{n} a_i^2 = \frac{s^2}{n} + \sum_{i=1}^{n} (t_i - \bar{t})^2.
\]

Using this the moment generating function can be written as a function of the \( Z_i \)'s alone

\[
m_{\bar{Z}}(s, t) = m_{Z_1 \cdots Z_n}(a_1 \cdots a_n)
\]

but, by virtue of the independence of the \( Z \)’s.
Because the $Z$'s are standard normal random variables, the moment generating function is

$$m_Z(a) = \prod_{i=1}^{n} e^{\left(\frac{\mu a_i + \sigma^2 a_i^2}{2}\right)} = e^{\left(\frac{\sum a_i + \sigma^2 \sum a_i^2}{2}\right)}$$

Expanding the final expression yields

$$m_Z(s, t) = e^{\left(\mu s + \frac{\sigma^2 s^2}{2n}\right)} e^{\frac{\sigma^2}{2} \left(\sum (t_i - \tilde{t})^2\right)} = m_Z(s) m_Z(t).$$

Thus, the independence of $\bar{Z}$ and $(Z - \bar{Z})$ is proved. It immediately follows that $\bar{Z}$ and $(Z - \bar{Z})^2$ are independent as well.////

**Chi-Square**

**Theorem 4.4**

If a random sample $X_1 \ldots X_k$ is iid $N(\mu, \sigma^2)$ then

$$(k-1)S^2 / \sigma^2 \sim \chi^2(k-1)$$

Proof.

By Theorem 4.3 $\frac{1}{\sigma^2} \sum_{i=1}^{k} (X_i - \mu)^2 \sim \chi^2(k)$.

Recall that $\sum_{i=1}^{k} (X_i - \bar{X}) = 0$, then

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{k} (X_i - \bar{X})^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

The expression has the form

$$W = U + V.$$

By Theorem 4.4 $U$ and $V$ are independent therefore the moment generating function is
\[ m_w(t) = m_U(t)m_V(t). \]

Thus,

\[ m_U(t) = m_W(t)/m_V(t). \]

Since \( W \) is chi-square with \( k \) degrees of freedom and \( V \) is chi-square with 1 degree of freedom the moment generating function of \( U \) is

\[
m_U(t) = \frac{(1 - 2t)^{k/2}}{(1 - 2t)^{1/2}} = (1 - 2t)^{k/2}. \]

Therefore

\[
U = \frac{1}{\sigma^2} \sum_{i=1}^{k} (X_i - \bar{X})^2 = (n - 1)S^2 / \sigma^2 \sim \chi^2 (k - 1). ///
\]
Problem Set IV

1) MGB (4-265)

Balls are drawn with replacement from an urn containing one white and two black balls. Let \( X = 0 \) for the white ball and \( X = 1 \) for the black. For samples \( X_1 \ldots X_9 \) of size 9, what is the joint distribution of the observations? The distribution of the sum of the observations? What are the expected value of the sample mean and sample variance.

2) MGB (7-266)

Use Chebyshev inequality to find how many times a fair coin must be tossed in order that the probability will be at least 0.90 that the sample mean will lie between 0.4 and 0.6.

3) MGB (8-266)

If a population has \( \sigma = 2 \) and \( \bar{X} \) is the mean of samples of size 100, find the limits between which \( \bar{X} - \mu \) will lie with probability 0.90. Use both Chebyshev inequality and the central limit theorem. Why do the two results differ?

4) You must decide whether or not a particular coin is fair (equal probability of heads or tails). You are going to flip it 100 times and record the number of heads. You have decided to reject the hypothesis that the coin is fair if the proportion of heads is smaller than 0.45 or larger than 0.55.

   a) What is the power of this test if the coin is biased in favor of heads such that the probability of heads is 0.75 rather than 0.5?

   b) Compute the power of the test for the following probability of heads: 0.10, 0.15, 0.20, 0.25, ..., 0.45, 0.50, 0.55, ..., 0.85, 0.90.

   c) Plot the calculated power values against the potential true probabilities.

   d) Construct a test with a 0.05 probability of rejecting the hypothesis that the coin is fair.

      i) What is the confidence interval for this test
      ii) Plot the power of the test for the same probability values as in (c).

5) A random sample of 100, \( X_1 \ldots X_{100} \), is drawn from a normal distribution with unknown mean \( \mu \) and known variance \( \sigma^2 = 1 \).

   Construct a test with a level of significance \( \alpha = 0.05 \) of the hypothesis that the mean of the distribution is zero

   \( H_0 : \mu = 0 \).
a) What is the confidence interval of your test?

b) Compute and plot the power of your test for true values of $\mu$ of -3, -2, -1, 0, 1, 2, 3.

c) Answer (a) and (b) for $\alpha = 0.10$

d) Suppose the following sample is collected. Do you reject $H_0$?

\[
\begin{array}{cccccc}
-2.567 & 1.578 & 0.504 & -0.431 & 0.532 \\
1.212 & -0.268 & 2.641 & 1.559 & 0.220 \\
-0.853 & 0.809 & 1.133 & 2.704 & 1.758 \\
1.478 & 0.803 & 1.356 & 0.312 & 1.636 \\
-2.201 & -0.270 & 1.448 & 2.459 & 1.393 \\
1.175 & 0.472 & 1.138 & 0.927 & 0.673 \\
0.518 & 2.683 & 1.250 & 0.892 & 3.251 \\
0.471 & 1.211 & 1.017 & 0.260 & 1.724 \\
0.342 & 1.502 & 0.856 & 1.993 & -0.085 \\
1.649 & 2.083 & 0.520 & -0.773 & 0.292 \\
2.999 & 1.125 & -0.606 & -0.825 & 1.850 \\
-0.102 & 0.733 & 3.168 & 0.208 & 1.074 \\
-0.229 & 1.360 & -1.068 & 1.683 & -0.412 \\
1.516 & 1.015 & 1.015 & -0.756 & 2.419 \\
1.464 & 0.683 & 1.788 & 0.747 & 1.420 \\
1.494 & 0.507 & 1.226 & 0.650 & 0.885 \\
0.906 & 1.334 & 2.465 & 1.712 & 0.502 \\
2.186 & 0.405 & 0.525 & 0.498 & 2.276 \\
2.452 & 2.035 & 2.504 & 2.855 & 0.807 \\
1.110 & 2.112 & 1.771 & -0.893 & 0.540 \\
\end{array}
\]

6) You are going to take a sample of 20 observations from a normal distribution with unknown mean and variance. Construct a test with a 10% level of significance for the hypothesis that the variance of the distribution is 16.

a) What is the confidence interval?

b) What is the power of this test if the true variance is 9?

c) Compute and plot the power of the test against variances from 1 to 36.

d) Apply the constructed test to the following observation

\[
\begin{array}{cccccc}
1.3243102 \\
1.9474245 \\
0.37705607 \\
-1.4851785 \\
5.3004505 \\
-2.5721721 \\
-1.3243102 \\
0.37682326 \\
-1.0784376 \\
3.6760503 \\
-1.2343861 \\
\end{array}
\]
- 0.094780
- 2.825576
- 2.441675
2.182597
- 0.43844
- 1.0834
3.8646
  4.4461
-0.9915
-2.2802
Chapter 5

Estimation
Estimation

Most empirical studies in Economics begin with a model. The model is fit to existing data and various propositions are tested with the resulting outcome. It is common to treat estimation and testing as separate subjects. However, as in previous chapters, testing is integrated with estimation because they are employed together in most studies.

This chapter focuses on the best way to estimate underlying distribution parameters, commonly designated as the vector $\theta$. For the binomial distribution $\theta$ is a single element vector

$$\theta = (p)$$

and for the normal it has two elements

$$\theta = (\mu, \sigma^2).$$

The true value of $\theta$ is unknowable, our mission is to find the 'best' way to estimate it given the knowledge we have: namely, our observations of the real world. There is no perfect estimation technique, because it is the nature of all estimates to be wrong. What we will be looking for are estimators that have desirable properties.

Properties of Estimators

Sufficiency

If you choose an estimator it is seems appropriate that it should be based on all the information available (from the sample) about the underlying parameters. If it does not do so, then there is probably another estimator that can be found that tells us more about the true parameter values. What is said is that one desirable property of an estimator is that it be a function of sufficient statistics.

A statistic is any real valued function of observed values of the underlying random variable(s). A statistic is a sufficient statistic if the density function of the random sample, conditional on the value of $\hat{\theta}$ does not depend on the true value, $\theta$, itself. Formally:
A statistic, \( T = T(X_1 \cdots X_n) \), is a sufficient statistic if
\[
f_{X_1 \cdots X_n | T}(x_1 \cdots x_n ; \theta | T) = \frac{f_{X_1 \cdots X_n}(x_1 \cdots x_n ; T; \theta)}{f_T(T; \theta)} = g_{X_1 \cdots X_n | T}(x_1 \cdots x_n | T).
\]

An example of an estimator that is a sufficient statistic is average number of successes, i.e., \( x = \bar{x} \), in a Bernoulli experiment with \( n \) observation. The estimator is
\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i.
\]
For this distribution the statistic \( k = \sum_{i=1}^{n} x_i \) is a sufficient statistic.

The density of the experimental outcome conditional on \( \hat{p} \) is
\[
f_{X_1 \cdots X_n | k}(x_1 \cdots x_n ; k; p) = \frac{p^k (1-p)^{n-k}}{\binom{n}{k}} = \frac{1}{\binom{n}{k}}
\]
that does not depend on \( p \).

It is often difficult to determine whether or not a particular estimator is a sufficient statistic. However, sufficiency is assure if, and only if, the sample density function factors in a particular way:

**Theorem 5.1 Factorization Theorem**

*The statistic \( T(X_1 \cdots X_n) \) is a sufficient statistic if and only if the density function factors as*
\[
f_{X_1 \cdots X_n}(x_1 \cdots x_n ; \theta) = g_T(T(x_1 \cdots x_n); \theta) h(x_1 \cdots x_n).
\]

**Proof.**

The proof is for discrete densities and follows Rice page 281-2.

First assume that the joint density function can be factored as indicated. In order to simplify the notation let \( X \) be the vector \((X_1 \cdots X_n)\), \( x \) be the vector of realized values \((x_1 \cdots x_n)\) and \( \tau \) be a particular value of the statistic. Then
\[
P(T = \tau) = \sum_{T(x) = \tau} P(X = x) = g(\tau; \theta) \sum_{T(x) = \tau} h(x).
\]
Therefore
\[
P(X = x | T = \tau) = \frac{P(X = x, T = \tau)}{P(T = \tau)} = \frac{h(x)}{\sum_{T(x) = \tau} h(x)}.
\]
To prove the second part of the theorem, suppose that the conditional distribution of $X$ given $T$ is independent of $\theta$. Let

$$g(\tau; \theta) = P(T = \tau; \theta) \text{ and } h(x) = P(X = x \mid T = \tau).$$

The joint density function is then

$$f_{X_1 \cdots X_n}(x_1 \cdots x_n; \theta) = g_\theta(\hat{\theta}(x_1 \cdots x_n); \theta)h(x_1 \cdots x_n).$$

Examples of Sufficient Statistics.

1) Bernoulli Density

$$f_X(x, \theta) = \theta^x(1 - \theta)^{1-x} \quad X = \begin{cases} 1 \text{ with probability } \theta \\ 0 \text{ with probability } 1 - \theta \end{cases}$$

$$f_{X_1 \cdots X_n}(x_1 \cdots x_n; \theta) = \theta^{\sum x_i} (1 - \theta)^{1-\sum x_i}$$

The estimator $T = \sum_{i=1}^n x_i$ is a sufficient statistic. To see this let

$$g(\hat{\theta}; \theta) = \theta^T (1 - \theta)^{1-T} \text{ and } h(x) = 1,$$

then

$$f_{X_1 \cdots X_n}(x_1 \cdots x_n; \theta) = g_\theta(\hat{\theta}(x_1 \cdots x_n); \theta)h(x_1 \cdots x_n).$$

2) Normal Density

In this case $\theta$ is a vector $(\mu, \sigma^2)$ and while Theorem 3.1 is stated for $\theta$ as a single quantity, it extends to the case of a vector of parameters. The density function is

$$f_X(x, \theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right).$$

The density function for the random sample of size $n$ is

$$f_{X_1 \cdots X_n}(x_1 \cdots x_n; \theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \sum (x_i - \mu)^2\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \left(\sum x_i^2 - 2\mu\sum x_i + \mu^2\right)\right)$$
With \( T = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2) \) let \( g(T, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left( \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + \mu^2 \right) \right) \) and \( h(x) = 1 \). Again the sample density function can be written

\[
f_{X_1 \cdots X_n}(x_1 \cdots x_n; \theta) = g_T(T(x_1 \cdots x_n); \theta)h(x_1 \cdots x_n).
\]

**Unbiasedness**

Certainly, a desirable property of an estimator is that 'on average' it gives the correct value for the parameters of interest. This is a vague idea, but consider taking the average of the estimates from a large number of random samples. For more specificity, think about repeatedly drawing a sample of \( n = 100 \) and computing the parameter estimates, using a particular estimator, from each sample and then averaging the estimates. As the number of samples gets large, one would like the average values to get close to the 'true' parameter values. If, in fact, the average values converge to the true values, the estimator is said to be unbiased. More formally:

\[
\hat{\theta} = \hat{\theta}(x_1 \cdots x_n) \text{ is an unbiased estimator of } \theta \text{ if and only if } E(\hat{\theta}) = \theta.
\]

The concept of unbiasedness has already been used in the discussion of the sample average and variance. It was established that

\[
E\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \mu \text{ and } E\left( \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) = \sigma^2.
\]

Whether or not the estimator is unbiased does not depend on the sample size to which it is applied\(^\text{10}\). For this reason unbiasedness is called a small sample property of an estimator. There are many estimators that are biased (not unbiased), but approach the true parameter values for large sample. Estimators with this characteristic are consistent.

**Consistency**

An estimator is consistent if the probability if its application yields estimates that are arbitrarily close to the true value. If we let \( \hat{\theta}_n \) be the estimator of \( \theta \) with sample size \( n \), consistency is defined

\[
\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0 \text{ for all } \varepsilon > 0.
\]

\(^\text{10}\) Even though the variance of the estimator is sample size dependent
The term on the right is called the probability limit with the shorthand notation

$$\text{plim}(\hat{\theta}_n) = \theta.$$  

A related, but not necessarily the same property is asymptotic unbiasedness. An estimator is asymptotically unbiased if its expected value approaches the true parameters as the sample size gets large. Formally:

$$\lim_{n \to \infty} E(\hat{\theta}_n) = \theta.$$  

**Efficiency**

Suppose there are two estimators $\hat{\theta}$ and $\tilde{\theta}$ both with the same expected value (for simplicity let the expected value be the true parameter). Which estimator should be chosen if their variances are not the same (let $\text{V}(\hat{\theta}) < \text{V}(\tilde{\theta})$). Reference to the normal distribution is helpful for this, so let's assume that the estimators are normally distributed with variances $\sigma^2 < \tilde{\sigma}^2$. The distributions of the estimators are represented in the following figure.

A larger fraction of the distribution mass of $\hat{\theta}$ is concentrated close to the true value than it is for $\tilde{\theta}$. In imprecise terms the lower the variance of the estimator the higher the probability of being close to the true value. So, everything else the same, a low variance estimator is preferred to a high variance one.

In the discussion of sufficiency, it was suggested that estimators should be based on all the information about the distribution parameters. If they are not, there are other estimators that tell us more about the true parameter values. This idea is made operational by suggesting that estimators with low variance "tell us more" about $\theta$ than those with higher variance because there is a higher probability that the estimates are
close to \( \theta \). The variance of an unbiased estimator that is a function of sufficient statistics is no larger than the variance of any other unbiased estimator. The following theorem—a somewhat less general version of the Rao-Blackwell theorem—makes this idea formally.

**Theorem 5.2**

Let \( \tilde{\theta} \) be a finite variance estimator of \( \theta \). Let \( T \) be a sufficient statistic for \( \theta \) and let \( \hat{\theta} = E(\tilde{\theta} \mid T) \). Then, for all \( \theta \),

\[
V(\hat{\theta}) \leq V(\tilde{\theta})
\]

with equality if and only if \( \tilde{\theta} \) is a function of \( T \) as well.

Proof.

From Theorem 3.2

\[
V(\tilde{\theta}) = V(E(\tilde{\theta} \mid T)) + E(V(\tilde{\theta} \mid T)) = V(\tilde{\theta}) + E(V(\tilde{\theta} \mid T)).
\]

Since the variance is strictly non-negative, \( V(\hat{\theta}) < V(\tilde{\theta}) \) unless \( V(\tilde{\theta} \mid T) = 0 \) and this will be true only if \( \tilde{\theta} \) is a function of \( T \). /// **Important:** Suppose both \( \hat{\theta} \) and \( \tilde{\theta} \) are unbiased.

**Maximum Likelihood Estimation**

There are a number of estimators—regression and generalize method of moments estimators are commonly employed. Each are worthy of concentrated study, however they are ignored in these lecture. Instead we will concentrate on maximum likelihood estimators with the examination of alternatives left for subsequent study.

Maximum likelihood estimation (MLE) conceptually turns the density function on its head. We derived the density function as a description of the distribution of random variables (the vector \( X \)) with given parameter values (\( \theta \)'s). For MLE the density function, in this case called a **likelihood function**, is the distribution function for the estimators (\( \hat{\theta} \)'s) given the observed sample values (\( x \)'s). The maximum likelihood estimators are the parameters that maximize the probability, ex ante (before the observations are made), of the particular sample.

Consider a sample of size \( n \) of independent and identically distributed (iid) random variables \( X \). The sample is \( x = (x_1 \ldots x_n) \). With the parameters of the distribution \( \theta \), the sample density function, or likelihood function is
\[ \Lambda(x, \theta) = f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta). \]

An alternative, and shorthand, way to write this is

\[ \Lambda(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta). \]

The ML estimators are the parameter values that maximize the value of the likelihood function. For discrete random variables ML estimators are those that maximize the ex ante probability of observing the particular sample values found. For continuous random variables the ML estimators maximize the ex ante probability of getting a sample in the neighborhood of the observed one. For more precision consider the probability that the sample will be within the interval \([x - \varepsilon, x + \varepsilon]\) where \(x\) is the vector of observed values. That probability is

\[ P(x - \varepsilon \leq X \leq x + \varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} \prod_{i=1}^{n} f(x_i, \theta) dx. \]

The value of \(\theta\) that maximizes that probability is the one that maximizes the likelihood function.

The maximum likelihood estimator, \(\hat{\theta}\), is defined

\[ \hat{\theta} = \arg\max_{\theta} \Lambda(x, \theta). \]

Often it is convenient to work with a linear, rather than multiplicative function. For that reason the log transformation of the likelihood function (the log-likelihood function)

\[ L(x, \theta) = \ln \Lambda(x, \theta) = \sum_{i=1}^{n} \ln f(x_i, \theta) \]

is used. Because \(L(\ )\) is a monotonically increasing function of \(\Lambda\) the same estimator that maximizes \(\Lambda(x, \theta)\) maximizes \(L(x, \theta)\) as well

\[ \hat{\theta} = \arg\max_{\theta} \Lambda(x, \theta) = \arg\max_{\theta} L(x, \theta). \]

Three examples will clarify the nature of maximum likelihood estimation. The first, and simplest, is the estimator for the probability in a Bernoulli distribution. Consider a trial of \(n\) observation, \(k\) of which are 1 and \(n-k\) 0. The likelihood function for the sample is

\[ \Lambda(x, \theta) = \theta^k (1 - \theta)^{n-k}. \]
The log-likelihood is
\[ L(x, \theta) = k \ln \theta + (n - k)\ln(1 - \theta). \]

Since the log-likelihood function is differentiable, as long as it is concave, its maximum is found by setting its first derivative, with respect to \( \theta \), equal to 0:
\[ \frac{\partial L(x, \theta)}{\partial \theta} = \frac{k}{\theta} - \frac{n - k}{1 - \theta} = 0 \Rightarrow \hat{\theta} = \frac{k}{n}. \]

To insure that this is a maximum check the second derivative
\[ \frac{\partial^2 L(x, \theta)}{\partial \theta^2} = -\frac{k}{\theta^2} - \frac{n - k}{(1 - \theta)^2} < 0. \]

The second example is the normal density, for which there are two parameter, the mean and variance. The likelihood function is
\[ \Lambda(x, \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2} \]
and the log-likelihood is
\[ L(x, \theta) = -n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \]

The first partial derivatives of the log-likelihood with respect to \( \mu \) and \( \sigma \) are
\[ \frac{\partial L(x, \theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \]
and
\[ \frac{\partial L(x, \theta)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2. \]

Set both derivatives equal to zero and solve for the maximum likelihood estimators for the mean and variance
\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

and

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 . \]

Notice that the mle for the mean is unbiased, but the estimator for the variance is not.

The final example is the most common Economics model, used for empirical studies of almost all economic phenomena. It is the simple linear model with a normally distributed error

\[ y_i = \beta x_i + \epsilon_i \]

where \( \epsilon_i \sim N(0, \sigma^2) \), \( \mathbb{E}(\epsilon_i \epsilon_j) = 0 \) for \( i \neq j \), and \( \mathbb{E}(\epsilon_i x_j) = 0 \) \( \forall i, j \). This says that the error term, \( \epsilon \) is normally distributed with expected value 0 and variance \( \sigma^2 \). Furthermore the error in every observation \( \epsilon_i \) is uncorrelated with the error \( \epsilon_j \) of any other observation. In addition the \( x \)'s and \( \epsilon \)'s are uncorrelated. With these conditions, the dependent variable \( y_i \) is normally distributed with expected value \( \beta x_i \) and variance \( \sigma^2 \) and the log-likelihood function for \( n \) observations is

\[ L(y, x, \beta) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 . \]

The partial derivative of the log-likelihood with respect to \( \beta \) is

\[ \frac{\partial L(y, x, \beta)}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)x_i . \]

Set this derivative equal to zero to find the maximum likelihood estimator of \( \beta \)

\[ \hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} . \]

**Properties of Maximum Likelihood Estimators**

The following discussion relies heavily, and often mimics that in Rice pp261-5. While there is an attempt at rigor throughout many important technical details are left out. The material is presented to given the student an insight into the essential nature of
the proofs. More rigorous treatments of the asymptotic results await later Econometric study. The demonstrations are all presented for single parameter problems. The results all carry forward, with considerable notational complexity, to a finite vector of parameters.

The first property of the MLE is that it is a function of only the sufficient statistics that can be computed from the sample.

**Theorem 5.3**

*Maximum likelihood estimators depend on the sample through a (set of) sufficient statistic(s).*

Proof.
The proof is for a single estimator and single sufficient statistic. The theorem is true for a vector of estimators and a set of sufficient statistics. The extension of the proof is easy and left to the interested student.

From the Factorization Theorem (5.1) the likelihood function

\[ \Lambda(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta) \]

can be factored

\[ \prod_{i=1}^{n} f(x_i, \theta) = g(T(x_1 \cdots x_n), \theta) h(x_1 \cdots x_n), \]

where \( T(x_1 \cdots x_n) \) is a sufficient statistic.

Therefore

\[ \max_{\theta} \Lambda(x, \theta) = h(x_1 \cdots x_n) \max_{\theta} g(T(x_1 \cdots x_n, \theta). /// \]

This is an important idea. Taken in conjunction with Theorem 5.2, it implies that if the MLE is unbiased, there is no other unbiased estimator with a smaller variance.

Another important and useful property of maximum likelihood estimation is that the MLE of functions of density parameters are the same functions of the MLE of the parameters. This is handy if one is interested in the value of the functions because it is often easier to find the MLE of the parameters than the functions. A formal statement of this is
Theorem 5.4  Invariance of Maximum Likelihood Estimators

Let \( \hat{\theta} = \hat{\theta}(x) \) be a maximum likelihood estimator of \( \theta \) in the density \( f(x, \theta) \). A maximum likelihood estimator of any function \( v = v(\theta) \) is \( v(\hat{\theta}) \).

Proof.

Define \( M(v, x) \) to be the largest value of the likelihood function over which \( v(\theta) = v \):

\[
M(v, x) = \max_{\theta : v(\theta) = v} L(\theta, x) \leq \max_{\theta} L(\theta, x)
\]

It is important for students to understand the reason for the final inequality.

But,

\[
\max_{\theta} L(\theta, x) = L(\hat{\theta}, x) = \max_{\{\theta : v(\theta) = v(\hat{\theta})\}} L(\theta, x) = M(v(\hat{\theta}), x).
\]

Students should be able to explain this equality.

The final step is the conclusion

\[
M(v(\hat{\theta}), x) \geq M(v, x).
\]

Consistency, Efficiency and Asymptotic Normality

Let \( \hat{\theta} = \hat{\theta}(x_1 \cdots x_n) \) be an estimator of \( \theta \), and assume the five 'regularity conditions':

i) \[
\frac{\partial \ln f(x, \theta)}{\partial \theta} \text{ exists for all } x \text{ and all } \theta;
\]

ii) \[
\frac{\partial}{\partial \theta} \left[ \cdots \int \prod_{i=1}^{n} f(x_i, \theta) dx_1 \cdots dx_i \right] = \int \cdots \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i, \theta) dx_1 \cdots dx_n;
\]

iii) \[
\frac{\partial^2}{\partial \theta^2} \left[ \cdots \int \prod_{i=1}^{n} f(x_i, \theta) dx_1 \cdots dx_i \right] = \int \cdots \frac{\partial^2}{\partial \theta^2} \prod_{i=1}^{n} f(x_i, \theta) dx_1 \cdots dx_n;
\]

iv) \[
\frac{\partial}{\partial \theta} \left[ \cdots \int \hat{\theta}(x_1 \cdots x_n) \prod_{i=1}^{n} f(x_i, \theta) dx_1 \cdots dx_i \right] = \int \cdots \frac{\partial}{\partial \theta} \hat{\theta}(x_1 \cdots x_n) \prod_{i=1}^{n} f(x_i, \theta) dx_1 \cdots dx_n.
\]
\( v) \quad 0 < E_\theta \left[ \left( \frac{\partial \ln f(x, \theta)}{\partial \theta} \right)^2 \right] < \infty. \)

**Information**

The concept of Fisher information is introduced before beginning the exposition of the properties of the maximum likelihood estimators. Let \( \rho(\theta) = \ln f(x, \theta) \) and the first and second derivatives be \( \rho'(x, \theta) \) and \( \rho''(x, \theta) \) respectively. Then the information, denoted at \( I(\theta) \) is defined as the expected value, with respect to \( \theta \), of the square of the first derivative

\[ I(\theta) = E_\theta ([\rho'(x, \theta)]^2). \]

**Lemma 5.1**

\[ E_\theta (\rho'(x, \theta)) = 0. \]

**Proof.**

\[ E_\theta (\rho'(x, \hat{\theta})) = \int \frac{\partial f(x, \theta)}{\partial \theta} f(x, \theta) dx = \int \frac{\partial f(x, \theta)}{\partial \theta} dx = 0. \]

The final equality to zero because \( \int f(x, \theta) dx = 1 \) and \( \frac{\partial}{\partial \theta} \int f(x, \theta) dx = \int \frac{\partial}{\partial \theta} f(x, \theta) dx \).

**Lemma 5.2**

\[ I(\theta) = E_\theta (\rho''(x, \theta)) = V(\rho'(x, \theta)). \]

**Proof.**

Since \( \int f(x, \theta) dx = 1 \) then \( \frac{\partial}{\partial \theta} \int f(x, \theta) dx = 0 \). Notice \( \frac{\partial f(x, \theta)}{\partial \theta} = \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) \).

Apply these two facts with condition (ii) to find

\[ \int \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) dx = 0. \]

Therefore, from (iii)
\[
\int \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) dx = \int \frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} f(x, \theta) dx + \int \left( \frac{\partial \ln f(x, \theta)}{\partial \theta} \right)^2 f(x, \theta) dx = 0.
\]

Thus,

\[
\int \left( \frac{\partial \ln f(x, \theta)}{\partial \theta} \right)^2 f(x, \theta) dx = - \int \frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} f(x, \theta) dx
\]

\[
E([\rho'(\theta)]^2) = I(\theta) = - E(\rho''(\theta)).
\]

However, since \(E(\rho'(\theta)) = 0\), \(V(\rho'(\theta)) = I(\theta).///

**Theorem 5.5**

*Under conditions i – iii the MLE from a random sample is consistent.*

**Proof.**

The sample average of the log-likelihood function is

\[
L_n(x, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ln f(x_i, \hat{\theta}).
\]

From the Weak Law of Large Numbers it is known that

\[
p\lim_{n \to \infty} L_n(x, \hat{\theta}) = E(\ln f(X, \hat{\theta})).
\]

In order to understand this you must keep in mind that
since the density \( f(\ ) \) is a function of the random variable \( X \), it is itself a random variable. The expected value of the density is assumed to be finite for all \( X \). The convergence means that the likelihood function converges (probabilistically) to a non stochastic function. Figure A illustrates the convergence of a stochastic log likelihood function to a one that is fixed. The remainder of the proof will show that the fixed function achieves a maximum at the true parameter value.

The first part has show that the log likelihood function converges in probability to 
\[
E(\log f(X, \hat{\theta})) .
\]
To continue the proof define 
\[
Z = \frac{f(X, \hat{\theta})}{f(X, \theta)} .
\]
Since \( \log(\ ) \) is a concave function it follows from Jensen's inequality that
\[
E(\log(Z)) \leq \log E(Z)
\]
or
\[
\int f(x, \hat{\theta}) f(x, \theta) dx \leq \log \int f(x, \hat{\theta}) f(x, \theta) dx = \log(1) = 0
\]
Since, as shown above, log likelihood function converges (in probability) to its expected value, this implies that 
\[
L(\theta) \leq L(\theta^*) .
\]

**Theorem 5.6**

*The probability distribution of \( \sqrt{nI(\theta)}(\hat{\theta} - \theta) \) approaches \( N(0,1) \) as \( n \to \infty \).*

**Proof.**

Take a Taylor first order Taylor series approximation of the first derivative of the log-likelihood function (evaluated at the MLE \( \theta \))

\[
L'(x, \hat{\theta}) = L'(x, \theta) + (\hat{\theta} - \theta)L''(x, \theta) + R
\]

where \( R \) is the remained that disappears as the sample size gets very large. For that reason it will be ignored for henceforth. It then follows that

\[
(\hat{\theta} - \theta) \approx -\frac{L'(\theta)}{L''(\theta)} .
\]

A simple rewriting is
\[
\sqrt{n}(\hat{\theta} - \theta) = -\frac{1}{\sqrt{n}} \frac{L'(\theta)}{\frac{1}{n} L''(\theta)}.
\]

Take the limit of this expression

\[
\lim_{n \to \infty} \sqrt{n}(\hat{\theta} - \theta) = -\lim_{n \to \infty} \frac{1}{\sqrt{n}} L'(\theta) \lim_{n \to \infty} \frac{1}{n} L''(\theta).
\]

By the Law of Large Numbers denominator approaches the expected value of \(L''(\theta)\), namely \(I(\theta)\). Thus for large \(n\)

\[
\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \frac{L'(\theta)}{I(\theta)}.
\]

By Lemma 5.2 \(E(L'(\theta)) = 0\) and thus

\[
E(\sqrt{n}(\theta - \hat{\theta})) = 0.
\]

This shows that the maximum likelihood estimator is **asymptotically unbiased**.

Apply Lemma 5.2 to find the variance

\[
V(\sqrt{n}(\theta - \hat{\theta})) = \frac{V(\frac{1}{\sqrt{n}} L'(\theta))}{I^2(\theta)} = \frac{I(\theta)}{I^2(\theta)} = \frac{1}{I(\theta)}.
\]

It follows that

\[
V(\hat{\theta} - \theta) = \frac{1}{nI(\theta)}.
\]

Since \(L'(\theta)\) is the sum of functions of identically and independently distributed, iid, random variables, it is, itself the sum of iid random variables with expected value 0 and variance \(I(\theta)\). Therefore, from Theorem 4.2 (Central Limit Theorem)

\[
(\hat{\theta} - \theta)\sqrt{nI(\theta)} = \frac{L'(x, \theta)}{\sqrt{nI(\theta)}} \to N(0,1) \text{ for large } n.
\]
Therefore \((\hat{\theta} - \theta)\sqrt{nI(\theta)}\) is asymptotically standard normal.///

**Efficiency of MLE**

We use one theorem to demonstrate the efficiency of the maximum likelihood estimators. As before, the theorem is stated, and proved, for one parameter. Nonetheless, the result is true for a vector of parameters.

**Theorem 5.7 (Cramer-Rao Inequality)**

Let \(T = T(x_1 \ldots x_n)\) be any (asymptotically) unbiased estimator of \(\theta\). Then under the regularity conditions (i) – (v) given above

\[
V(T) \geq 1/nI(\theta).
\]

Proof.

Because \(T(\cdot)\) is and unbiased estimator

\[
\theta = \int \ldots \int T(x_1 \ldots x_n) \prod_{i=1}^{n} f(x_i) \, dx_1 \ldots dx_n
\]

Then

\[
1 = \frac{\partial}{\partial \theta} = \int \ldots \int \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i) \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i) \, dx_1 \ldots dx_n.
\]

Recall that

\[
\int \ldots \int \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i) \, dx_1 \ldots dx_n = 0.
\]

Thus

\[
1 = \int \ldots \int \left[ T(x_1 \ldots x_n) - \theta \right] \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i, \theta) \, dx_1 \ldots dx_n = E \left[ \frac{\partial}{\partial \theta} \log f(x_i, \theta) \right].
\]

It follows immediately that
Apply theorem 3.4 (The Cauchy-Swartz Inequality)

\[
E \left[T(x_1 \cdots x_n) - \theta \right]^{n} \sum_{i=1}^{n} \frac{\partial \log f(x_i, \theta)}{\partial \theta}\right]^2 = 1.
\]

The first term on the right hand side of the inequality is the variance of T and the second term is equal to nI(\theta). With this the inequality follows:

\[
V(T) \leq 1 / nI(\theta).
\]

An important conclusion follows from this inequality. In words it say that the variance of \( \hat{\theta} \), the maximum likelihood estimator, is no larger than the variance of any other unbiased estimator. The conclusion is that under the regularity conditions specified, the MLE is an efficient estimator.

**Likelihood Ratio**

Suppose you know that the parameter \( \theta \) is one of two values \( \theta_0 \) or \( \theta_1 \). How would you choose between them, based on a sample of \( n \) observations, \( x = x_1 \cdots x_n \)? It makes sense to choose the value that would have assigned the highest ex-ante probability to the sample. In other word that parameter value that makes the likelihood, for the sample, the largest. In other words

\[
\text{choose } \begin{cases} 
\theta_0 & \text{if } \Lambda(\theta_0, x) > \Lambda(\theta_1, x) \\
\theta_1 & \text{if } \Lambda(\theta_0, x) < \Lambda(\theta_1, x)
\end{cases}
\]

This is the reasoning that is the basis for the likelihood ratio test. In the simplistic question stated above, the criterion can be specified as the ratio of the likelihood functions

\[
\text{choose } \begin{cases} 
\theta_0 & \text{if } \Lambda(\theta_0, x) / \Lambda(\theta_1, x) \geq 1 \\
\theta_1 & \text{if } \Lambda(\theta_0, x) / \Lambda(\theta_1, x) < 1
\end{cases}
\]

For more realistic problems, the tests are somewhat more complicated, but the idea behind it remains the same.

A problem that mirrors the one confronted in many Economic studies focuses on the significance of posited relationships. Many economic theories predict that certain variables are related (often linearly). The models are made operational and susceptible to
empirical verification with the specification of particular relationships. For instance, consider a naive multiplier model in which national income, Y, depends on government spending, G. The simplest such relationship is linear in the following form

$$Y_t = \alpha + \beta G_t + \varepsilon_t,$$

where t, rather than i, subscript is used because most macro-economic studies use time series rather than cross-section samples. A specific distribution is specified for the random error, \(\varepsilon\) and from that the log-likelihood function follows. The parameter of interest is \(\beta\). To hold that there is a relationship between income and government spending is to hypothesize that \(\beta\) is not zero.

A sensible way to test this is to compare the maximum value of the likelihood function when \(\beta\) is restricted to be zero with the maximum value when \(\beta\) is unrestricted. Of course, the latter value will be larger than the former, but if it is not sufficiently bigger it means that there is very little predictive value in presuming that \(\beta\) is other than zero. In that case we are inclined to favor the proposition (called an hypothesis) that there is no relationship between government spending and national income (they are independently distributed). If the latter value is significantly larger (what is significant is to be determined), then we are inclined to reject the hypothesis that \(\beta\) is zero. This idea of the relative sizes of the likelihood functions, evaluated at their respective maxima, is formalized in the likelihood ratio test.

The likelihood ratio \(\lambda\), for this problem, is defined as

$$\lambda = \frac{\sup_{\alpha, \beta=0} \Lambda(\alpha, \beta, Y, G)}{\sup_{\alpha, \beta} \Lambda(\alpha, \beta, Y, G)}.$$

This is the ratio of the largest value of the likelihood function (or the limit of the likelihood function) with the (in this case) one restriction that \(\beta = 0\), to the unrestricted largest value. For a more specific example, suppose the random error \(\varepsilon\) is normally distributed with \(\mu = 0\) and variance \(\sigma^2\). The unrestricted likelihood function is

$$L(\alpha, \beta, x) = \frac{-T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (Y_t - \alpha - \beta G_t)^2.$$

The restricted likelihood function is

\[^{11}\text{The supremum (sup) is used, rather than maximum (max) because maximization could be over an open set for which the maximum does not exist.}\]
Let the maximum likelihood estimators of $\alpha$ and $\beta$ be $a$, $b$ and $S^2$ respectively. **Students should derive $a$, $b$ and $S^2$ for both the unrestricted and restricted cases. From this the students should derive $\lambda$.**

Examining the size of $\lambda$, the larger it is the more likely is it that $\beta = 0$. This is a qualitative criterion. There is one that is rather more precise because the asymptotic distribution of a simple function of $\lambda$, namely twice the negative of its natural logarithm, is distributed $\chi^2(r)$ (chi-square with $r$ degrees of freedom) where $r$ is the number of restriction. In the case just examined $r = 1$. But for larger econometric models, ones in which there many right hand variable and, thus, many $\beta$'s, $r$ is a much larger number. Formally, the likelihood ratio statistic is

$$-2\log \lambda = -2 [L(a,0,x) - L(a,b,x)] \sim \chi^2(1).$$

As a detailed example of the use of the likelihood ratio test consider taking two samples of size $n_1$ and $n_2$ (total sample size of $n = n_1 + n_2$). Both are drawn from a normal distribution, but it is unknown whether or not each is from he same, or different distribution. The assumption that imposes the fewest restriction (call this the unrestricted form) is one that allows for differences in both the means ($\mu_1$ and $\mu_2$) and the variances ($\sigma_1^2$ and $\sigma_2^2$). The most restrictive is the one that requires the equality of both mean and variance ($\mu_1 = \mu_2 = \mu$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$). There are two intermediate restrictive assumptions – one is that the means are the same, but the variances may be different ($\mu_1 = \mu_2 = \mu$ and $\sigma_1^2 \neq \sigma_2^2$) and the other is that the means may be different but the variances are the same ($\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$). The possibilities are stated as four different hypotheses:

- $H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$
- $H_1 : \mu_1 = \mu_2 = \mu$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$
- $H_2 : \mu_1 = \mu_2 = \mu$ and $\sigma_1^2 = \sigma_2^2$
- $H_3 : \mu_1 = \mu_2$ and $\sigma_1^2 \neq \sigma_2^2 = \sigma^2$. 
The process of computing the likelihood ratio statistic begins with specifying the log likelihood functions and then computing the maximum likelihood estimators under the different hypotheses. As a notational simplification define \( p = -n \log 2\pi \). The log likelihood functions are

\[
L_0 = p - n_1 \log \sigma_1^2 - n_2 \log \sigma_2^2 - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n_1} (x_{1i} - \mu_1)^2 - \frac{1}{2\sigma_2^2} \sum_{i=1}^{n_2} (x_{2i} - \mu_2)^2
\]

\[
L_1 = p - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
L_2 = p - n_1 \log \sigma_1^2 - n_2 \log \sigma_2^2 - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n_1} (x_{1i} - \mu)^2 - \frac{1}{2\sigma_2^2} \sum_{i=1}^{n_2} (x_{2i} - \mu)^2
\]

\[
L_3 = p - n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\].

Students should verify that the following are the maximum likelihood estimators.

The MLE's for the means are designated by m's and for the variances by \( S^2 \).

0: \( m_1 = \bar{X}_1 \quad m_2 = \bar{X}_2 \quad S_1^2 = \sum_{i=1}^{n_1} (X_{1i} - m_1)^2 \) and \( S_2^2 = \sum_{i=1}^{n_2} (X_{2i} - m_2)^2 \).

1: \( m = \bar{X} \quad S^2 = \sum_{i=1}^{n} (X_i - m)^2 \)

2: \( S_1^2 = \sum_{i=1}^{n_1} (X_{1i} - m_1)^2 ; \quad S_2^2 = \sum_{i=1}^{n_2} (X_{2i} - m_2)^2 \quad m = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 S_1^2 + n_2 S_2^2} \)

3: \( m_1 = \bar{X}_1 \quad m_2 = \bar{X}_2 \quad S^2 = \sum_{i=1}^{n_1} (X_{1i} - m_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - m_2)^2 \).

The following two columns are random samples drawn from a standard normal distributions (each has mean 0 and variance 1).

\[
\begin{array}{c}
x_1 \\
x_2
\end{array}
\]
The maximum likelihood estimates for the four hypotheses are

0: \[ m_1 = 0.1582 \quad m_2 = 0.2691 \quad S_1^2 = 0.8354 \quad S_2^2 = 0.7509 \]

1: \[ m = 0.2137 \quad S^2 = 0.7962 \]

2: \[ S_1^2 = 0.8387 \quad S_2^2 = 0.7536 \quad m = 0.2166 \]

3: \[ m_1 = 0.1582 \quad m_2 = 0.2690 \quad S^2 = 0.7962. \]

With these, the maximum values of the likelihood functions for the various hypotheses are

L0 = -84.1878

L1 = -84.3993

L2 = -84.3425

L3 = -84.3221.
Notice, as expect, the unconstrained likelihood value is the highest.

As a comparison with samples drawn from the same density, this final example are two samples, the first drawn from a normal distribution with mean 1 and variance 1 and the second from a normal distribution with mean 0 and variance 4.

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0753937</td>
<td>2.4622331</td>
</tr>
<tr>
<td>0.95607622</td>
<td>-0.86549733</td>
</tr>
<tr>
<td>0.76060697</td>
<td>1.4008768</td>
</tr>
<tr>
<td>-0.66638678</td>
<td>0.38146732</td>
</tr>
<tr>
<td>-0.1824259</td>
<td>-0.7042369</td>
</tr>
<tr>
<td>0.56012146</td>
<td>-0.100416</td>
</tr>
<tr>
<td>2.6831778</td>
<td>-1.2232932</td>
</tr>
<tr>
<td>-0.28783090</td>
<td>1.4380826</td>
</tr>
<tr>
<td>1.0213592</td>
<td>-1.5809552</td>
</tr>
<tr>
<td>-0.1365249</td>
<td>-2.6698730</td>
</tr>
<tr>
<td>0.9987508</td>
<td>1.4376902</td>
</tr>
<tr>
<td>0.568770</td>
<td>-0.39583039</td>
</tr>
<tr>
<td>1.361485</td>
<td>1.1855396</td>
</tr>
<tr>
<td>1.780677</td>
<td>-1.5387181</td>
</tr>
<tr>
<td>0.803384</td>
<td>0.15343129</td>
</tr>
<tr>
<td>1.99619</td>
<td>2.3337503</td>
</tr>
<tr>
<td>2.18802</td>
<td>1.0665248</td>
</tr>
<tr>
<td>-0.39965</td>
<td>-1.9472835</td>
</tr>
<tr>
<td>-0.666942</td>
<td>1.3429603</td>
</tr>
<tr>
<td>1.597150</td>
<td>1.4380906</td>
</tr>
</tbody>
</table>

0: \( m_1 = 0.8005 \) \( m_2 = 0.1807 \) \( S_1^2 = 0.8942 \) \( S_2^2 = 2.1116 \)

1: \( m = 0.4906 \) \( S^2 = 1.5990 \)

2: \( S_1^2 = 0.9257 \) \( S_2^2 = 2.3073 \) \( m = 0.6231 \)

3: \( m_1 = 0.8005 \) \( m_2 = 1.807 \) \( S^2 = 3.006 \).

With these, the maximum values of the likelihood functions for the various hypotheses are

\( L0 = -106.2283 \)

\( L1 = -112.2896 \)

\( L2 = -108.6932 \)

\( L3 = -111.0882 \).
Problem Set V

1) (MGB 1-362)
An urn contains black and white balls. A sample of size n is drawn with replacement. What is the maximum likelihood estimator of the ratio R of black to white balls?

Suppose that one draws balls one by one with replacement until a black ball appears. Let X be the number of draws required (not counting the last draw). This operation is repeated n times to obtain a sample $X_1 \ldots X_n$. What is the maximum likelihood estimator based on this sample?

2) (MGB 4-362)
Two samples of size $n_1$ and $n_2$ are drawn from two normally distributed populations with means and variances $\mu_1, \sigma_1^2$ and $\mu_2, \sigma_2^2$ respectively. What is the maximum likelihood estimator of $\theta = \mu_1 - \mu_2$? If the sample size $n = n_1 + n_2$ is fixed, how should the n observations be divided between the two populations in order to minimize the variance of that estimator?

3) (MGB 5-363)
A sample of size n is drawn from each of four normal populations, all of which have the same variance $\sigma^2$. The mean of the four populations are $a + b + c$, $a + b - c$, $a - b + c$ and $a - b - c$. What are the maximum likelihood estimators of a, b, c and $\sigma^2$?

4) (MGB 11-363)
Let $X_1\ldots X_n$ be a random sample from some density with mean $\mu$ and variance $\sigma^2$.

a) Show that $\sum_{i=1}^{n} a_i X_i$ is an unbiased estimator of $\mu$ for any set of known constants (a's) that sum to one.

b) Assume the a's sum to one and show that the variance of the estimator is minimized for $a_i = 1/n$ for all i.

5) Individual's value placed on a particular commodity is a linear function of a number of measurable variables (these variables represent aspects of the individual, e.g., income, age, education, and the product, e.g., amount spent on advertising, it rating in consumer report) and a normally distributed random error. The value function is
\[ V_i = \sum \beta_i X_i + \sum \gamma_j Z_j + \epsilon_i ; \epsilon_i \sim N(0, \sigma^2) \] for all i.
The X's represent the product characteristics and the Z's the personal ones.

While the X's are observed, neither the V's nor the \( \epsilon \)'s are. What is observed, however, is the price at which the consumer is offered the product, \( p_i \), and whether or not it is purchased. We know that the product is purchased if and only if the value is at least a great as the price: \( V_i \geq p_i \). We wish to estimate the value of the \( \beta \)'s and \( \gamma \)'s.

a) What is the likelihood function?

b) Devise a test for the hypothesis that the personal characteristics do not affect individual value.

6) This problem is similar to the last in that the magnitude of value is unobserved. The value function is simpler in that it depends on one variable X and a random term \( u \) as follows

\[ V_i = \beta X_i + u_i. \]

In this case the random variable \( u \), is drawn from a uniform distribution with density

\[ f(u) = \begin{cases} 1 & \text{for } u_i \in [0,1] \\ 0 & \text{otherwise} \end{cases}. \]

We observe a purchase if the individual's value is positive and no purchase if it is negative. For the following observations compute the maximum likelihood estimator of \( \beta \) and test the hypothesis that it equals 0.

<table>
<thead>
<tr>
<th>X</th>
<th>Purchase (1 = buy  0 = not buy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.078</td>
<td>1</td>
</tr>
<tr>
<td>-2.388</td>
<td>0</td>
</tr>
<tr>
<td>-1.530</td>
<td>0</td>
</tr>
<tr>
<td>0.189</td>
<td>1</td>
</tr>
<tr>
<td>0.444</td>
<td>1</td>
</tr>
<tr>
<td>-0.275</td>
<td>1</td>
</tr>
<tr>
<td>-1.013</td>
<td>0</td>
</tr>
<tr>
<td>0.619</td>
<td>1</td>
</tr>
<tr>
<td>-0.435</td>
<td>1</td>
</tr>
<tr>
<td>1.937</td>
<td>1</td>
</tr>
</tbody>
</table>

You will need to write a program to numerically approximate the maximum likelihood estimates.