MAT 211: Mathematics for Business Analysis

Jenifer Boshes
Karin Saoub
Arizona State University
Department of Mathematics and Statistics

September 6, 2007
# Contents

1. **Derivative Review**

2. **Chapter 5 Functions of Several Variables**
   - 2.1 Section 5.1 Functions of Two Variables ............................................. 9
   - 2.2 Section 5.2 Surface & Distance ............................................................ 10
     - 2.2.1 The Distance Formula ................................................................. 11
     - 2.2.2 Spheres ....................................................................................... 11
   - 2.3 Section 5.3 Geometric Representation .................................................. 11
     - 2.3.1 Level Curves ............................................................................... 12
     - 2.3.2 Geometric Interpretations of Partial Derivatives ......................... 12
   - 2.4 Section 5.4 Partial Derivative with Two Variables ................................. 12
     - 2.4.1 Other Notation for Partial Derivatives ........................................ 13
     - 2.4.2 Limit Definition of Partial Derivatives ........................................... 13
     - 2.4.3 Higher-Order Partial Derivatives ................................................... 14

3. **Chapter 6 Multivariable Optimization**
   - 3.1 Section 6.1 Optimization in two variables ............................................. 15
     - 3.1.1 Finding maximums/minimums ..................................................... 16
   - 3.2 Section 6.2 Local Extreme Points ......................................................... 16
     - 3.2.1 Classifying stationary points ......................................................... 17
   - 3.3 Section 6.3 The Extreme Value Theorem ............................................. 17
     - 3.3.1 Theorem 6.3.1 – The Extreme Value Theorem ............................. 18
     - 3.3.2 How to find maxima and minima ................................................... 18
   - 3.4 Section 6.4 The Lagrange Multiplier Method ....................................... 18
     - 3.4.1 How to use the Lagrange Multiplier Method ................................ 19
   - 3.5 Section 6.7 Linear Programming .......................................................... 19
     - 3.5.1 Solving LPP’s .............................................................................. 20
## CONTENTS

### 4 Chapter 7 – Matrices

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Solving a System Using Elimination and Substitution</td>
<td>22</td>
</tr>
<tr>
<td>4.1.1 Dependent and Inconsistent Systems</td>
<td>22</td>
</tr>
<tr>
<td>4.2 Matrices and Matrix Operations</td>
<td>23</td>
</tr>
<tr>
<td>4.2.1 Rules for Matrix Addition and Scalar Multiplication</td>
<td>24</td>
</tr>
<tr>
<td>4.3 Matrix Multiplication</td>
<td>25</td>
</tr>
<tr>
<td>4.3.1 Rules for Matrix Multiplication</td>
<td>25</td>
</tr>
<tr>
<td>4.4 Elementary Row Operations</td>
<td>26</td>
</tr>
<tr>
<td>4.5 Determinants</td>
<td>27</td>
</tr>
<tr>
<td>4.5.1 Sarrus’s Rule</td>
<td>28</td>
</tr>
<tr>
<td>4.5.2 Expansion by Cofactors</td>
<td>29</td>
</tr>
<tr>
<td>4.6 Inverses</td>
<td>30</td>
</tr>
<tr>
<td>4.6.1 Finding the Inverse of a Matrix</td>
<td>30</td>
</tr>
<tr>
<td>4.6.2 Properties of the Inverse</td>
<td>32</td>
</tr>
<tr>
<td>4.7 Solving a System Using Matrices</td>
<td>32</td>
</tr>
<tr>
<td>4.7.1 Gaussian Elimination</td>
<td>32</td>
</tr>
<tr>
<td>4.7.2 Inverse of the Coefficient Matrix</td>
<td>33</td>
</tr>
<tr>
<td>4.7.3 Cramer’s Rule</td>
<td>34</td>
</tr>
<tr>
<td>4.8 Interpreting Solutions of Linear Equations</td>
<td>35</td>
</tr>
</tbody>
</table>

### 5 Chapter 8 – Probability

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Set Theory</td>
<td>37</td>
</tr>
<tr>
<td>5.1.1 Relationships between sets</td>
<td>37</td>
</tr>
<tr>
<td>5.1.2 Operations on sets</td>
<td>38</td>
</tr>
<tr>
<td>5.2 Counting</td>
<td>39</td>
</tr>
<tr>
<td>5.2.1 Permutations</td>
<td>40</td>
</tr>
<tr>
<td>5.2.2 Combinations</td>
<td>41</td>
</tr>
<tr>
<td>5.3 Assignment of Probabilities</td>
<td>42</td>
</tr>
<tr>
<td>5.3.1 Properties of Probability</td>
<td>42</td>
</tr>
<tr>
<td>5.3.2 Finding probabilities using counting techniques</td>
<td>45</td>
</tr>
<tr>
<td>5.4 Other types of probabilities</td>
<td>45</td>
</tr>
<tr>
<td>5.4.1 Conditional Probability</td>
<td>45</td>
</tr>
<tr>
<td>5.4.2 Independent Events</td>
<td>47</td>
</tr>
<tr>
<td>5.4.3 Bayes’ Theorem</td>
<td>47</td>
</tr>
</tbody>
</table>

### 6 Chapter 9 – Discrete & Continuous Distributions

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Discrete Random Variables</td>
<td>49</td>
</tr>
<tr>
<td>6.1.1 Properties of probability mass functions</td>
<td>50</td>
</tr>
<tr>
<td>6.1.2 Uniform Random Variables</td>
<td>51</td>
</tr>
<tr>
<td>6.1.3 Other examples...</td>
<td>51</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.1.4 Histograms</td>
<td>52</td>
</tr>
<tr>
<td>6.2 Mathematical Expectation</td>
<td>53</td>
</tr>
<tr>
<td>6.2.1 Examples</td>
<td>53</td>
</tr>
<tr>
<td>6.2.2 Expected Value</td>
<td>54</td>
</tr>
<tr>
<td>6.2.3 Properties of Expected Value</td>
<td>55</td>
</tr>
<tr>
<td>6.2.4 Mean, Variance, and Standard Deviation</td>
<td>56</td>
</tr>
<tr>
<td>6.3 Bernoulli Trials &amp; the Binomial Distribution</td>
<td>57</td>
</tr>
<tr>
<td>6.3.1 Bernoullis</td>
<td>57</td>
</tr>
<tr>
<td>6.3.2 Binomials</td>
<td>59</td>
</tr>
<tr>
<td>6.3.3 Examples</td>
<td>62</td>
</tr>
<tr>
<td>6.4 Integration Review</td>
<td>63</td>
</tr>
<tr>
<td>6.4.1 Basic Integrals &amp; General Rules</td>
<td>63</td>
</tr>
<tr>
<td>6.4.2 Indefinite Integrals, Definite Integrals and Improper Integrals</td>
<td>64</td>
</tr>
<tr>
<td>6.5 Continuous Random Variables</td>
<td>66</td>
</tr>
<tr>
<td>6.5.1 Properties of Continuous Random Variables</td>
<td>67</td>
</tr>
<tr>
<td>6.6 The Normal Distribution</td>
<td>73</td>
</tr>
<tr>
<td>6.6.1 Standard Normals</td>
<td>73</td>
</tr>
<tr>
<td>6.6.2 Normals with mean $\mu$ and standard deviation $\sigma$</td>
<td>75</td>
</tr>
</tbody>
</table>
Chapter 1

Derivative Review

- Power Rule: If $f(x) = ax^n$ then $f'(x) = anx^{n-1}$
- Exponential Rule: If $f(x) = e^x$ then $f'(x) = e^x$
- Natural Logarithm Rule: If $f(x) = \ln x$ then $f'(x) = \frac{1}{x}$
- Product Rule: If $f(x) = u \cdot v$ then $f'(x) = u' \cdot v + v \cdot u'$
- Quotient Rule: If $f(x) = \frac{u}{v}$ then $f'(x) = \frac{u' \cdot v - v' \cdot u}{v^2}$, where $u$ and $v$ are functions of $x$
- Examples:
Chapter 2

Chapter 5 Functions of Several Variables

Thus far we’ve considered functions of the form $y = f(x)$, where we pick one value of $x$ to get a corresponding $y$ value. This coordinate pair $(x, y)$ is now a point that we can graph on the $x – y$ plane.

- Ex:
  
  $$g(x) = x^2 + 2$$  \hspace{1cm} (2.1)

Recall that when we consider functions as in (2.1) we also define what is known as the domain, $\mathcal{D}$ of a function. Simply stated, the domain refers to the restrictions placed on the independent variables such that the function is defined. What is the domain of $g(x)$?

2.1 Section 5.1 Functions of Two Variables

Consider $z = f(x, y)$. This is a function of two variables and is a rule that assigns a specified number $f(x, y)$ to each point $(x, y)$ in the domain $\mathcal{D}$. Now we have $(x, y, z)$ as a point in 3-D space.

- Ex:
  
  $$f(x, y) = 3x^2 + xy - 9$$  \hspace{1cm} (2.2)

The function given in (2.2) is a surface in three-dimensional space. We can find points on the surface by picking values for $x$ and $y$ that are in $\mathcal{D}$.

- Evaluate $f(2, 2)$
• Evaluate \( f(0, 0) \)
• Evaluate \( f(0, 1) \)
• Evaluate \( f(a, b) \)

We still need to address the question: What is \( \mathcal{D} \)? The three main “problem” areas we look for are

1. What values of \( x \) and/or \( y \) would cause a 0 to be in the denominator? (If there is one)
2. What values of \( x \) and/or \( y \) would cause a negative number to appear under any square (or even) roots? (If there is one)
3. What values of \( x \) and/or \( y \) would cause us to take logarithms of negative numbers? (If there is one)

Back to (2.2) – We notice that we do not have to consider any of the “problem” areas since we do not have any fractions, roots, or logarithms. Hence, the domain of \( f \) is: \( \mathcal{D} = \{(x, y) \mid x, y \in \mathbb{R}\} \). This is read as “The domain is all coordinate pairs \((x, y)\) such that \(x\) and \(y\) are real numbers.”

Ex: Find the domain and evaluate each function at \((1, -1), (0, 1/2), (a, b)\)

\[ f(x, y) = \sqrt{x} - \frac{1}{y} \]
\[ g(x, y) = \sqrt{x - 2y} \]
\[ h(x, y) = x - \frac{1}{x-y} \]

2.2 Section 5.2 Surface & Distance

Recall from univariate calculus that we work with a function of the form \( y = f(x) \). Suppose that we have

\[ y = 2 - x \implies x + y = 2 \]

This is the set of all points \((x, y)\) such that \(x+y = 2\), as seen in GRAPH::

Similarly, a function of two variables can be represented as \( g(x, y, z) = c \). This is a point set in space, and the surface of \( g(x, y) \) is made of all coordinate triples \((x, y, z)\) such that \( g(x, y, z) = c \).

• Ex:
2.2.1 The Distance Formula

Recall: How do we find the distance between two points on the $x-y$ plane?

**Def:** The distance between $(x_1, y_1, z_1)$ and $(x_2, y_2, z_2)$ is given by:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (2.3)$$

2.2.2 Spheres

Recall: On the $x-y$ plane a circle is defined by the set of all points a certain distance ($r$) from another point. That is, for the circle centered at point $(s, t)$ with radius $r$, we could say that this is $C = \{(x, y) \mid d((x, y), (s, t)) = r\}$.

(How would we find the exact distance between $(x, y)$ and $(s, t)$? Hint: 2.3)

Well, in three-dimensional space, we do not consider circles, but rather the counterpart – spheres.

**Def:** For a point in space $(a, b, c)$ the sphere of radius $r$ centered at $(a, b, c)$ is the set of all points $(x, y, z)$ such that the distance between $(a, b, c)$ and $(x, y, z)$ equals $r$. Thus, using (2.3),

$$r = \sqrt{(x-a)^2 + (y-b)^2 + (z-c)^2} \implies$$

$$r^2 = (x-a)^2 + (y-b)^2 + (z-c)^2$$

2.3 Section 5.3 Geometric Representation

Recall: Any point in a plane can be represented by a pair of real numbers. Similarly, any point in space can be represented by a triplet of numbers.

Note that

- The equation $x = 0$ is satisfied by all points on the $y-z$ plane.
- The equation $y = 0$ is satisfied by all points on the $x-z$ plane.
- The equation $z = 0$ is satisfied by all points on the $x-y$ plane.

The three coordinate planes together can be divided into eight octants (not a dissimilar idea from the 4 quadrants that we are already familiar with in regards to the $x-y$ plane).

**Def:** Let $z = f(x, y)$ be defined over a domain $D$. The graph of the function $f$ is the set of all points $(x, y, f(x, y))$ in space obtained by evaluating every point $(x, y)$ in the domain.
2.3.1 Level Curves

Suppose we have some function \( z = f(x, y) \). The graph of the function in three-dimensional space can be visualized as being cut by horizontal planes parallel to the \( x - y \) plane. The intersection between each plane and the graph is projected onto the \( x - y \) plane and is known as the level curve at height \( c \) where \( c \) is the intersecting plane.

- Ex:

To get a better idea of what level curves are: Consider any 3D object. If you cut through the object parallel to the \( x - y \) plane (the ground), what would that cut look like from a bird’s eye view? Now cut again at a different height, still parallel to the \( x - y \) plane. What does the bird’s eye view look like now?

2.3.2 Geometric Interpretations of Partial Derivatives

Recall: What is a derivative? How do we interpret a derivative? We usually say that the derivative is simply the “slope of the tangent line”. We can extend our definition and interpretation of derivatives from one-variable calculus to two-variable calculus. First, however, we must consider what is known as a partial derivative. Consider any surface, and fix a point \( (x_0, y_0) \). The partial derivative \( f_x(x_0, y_0) \) is the slope of the tangent line to the curve at \( x = x_0 \). Another way to consider the partial derivative is: when holding \( y \) constant and moving in the positive \( x \) direction, what happens to the \( z \) values? Similarly, \( f_y(x_0, y_0) \) is the slope of the tangent line to the curve at \( y = y_0 \). Again, holding \( x \) constant and moving along in the positive \( y \) direction, what happens to the \( z \) values?

2.4 Section 5.4 Partial Derivative with Two Variables

Recall: For \( y = f(x) \), \( f'(x) \) is a new function that measures the instantaneous rate of change of \( f(x) \).

- Ex: \( f(x) = 2x^3 - 3 \)

Similarly for \( z = g(x, y) \) we want to measure the rate of change for each of the independent variables.
Ex: $g(x) = 2x^3 + 3y^2 - 3$ \hspace{1cm} (2.4)

When we have a function of two variables we write the derivatives as: $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ to indicate the derivative of $z$ with respect to $x$ and the derivative of $z$ with respect to $y$, respectively. This is to indicate that we are only considering one independent variable changing while the other is held fixed.

What are $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ of (2.4)?

These expressions $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ are called the partial derivatives of $z$ with respect to $x$ and $y$. Note: $\frac{\partial z}{\partial x}$ is the rate of change of $f(x, y)$ with respect to $x$ when $y$ is held constant. Similarly for $\frac{\partial z}{\partial y}$.

- Ex:

2.4.1 Other Notation for Partial Derivatives

Note that it is common to use any of the following notation to refer to partial derivatives:

$$\frac{\partial z}{\partial x} = \frac{\partial f}{\partial x} = f_x(x, y) = \frac{\partial f(x, y)}{\partial x}$$

Similar notation holds for the partial derivative with respect to $y$.

Also note that just as we can evaluate functions of more than one variable at a specific point, we can also evaluate the derivative of functions of more than one variable at a specific point.

- Ex:

2.4.2 Limit Definition of Partial Derivatives

Recall that for a function of one variable, $y = f(x)$: Before you learned the “shortcut” method of how to take a derivative you learned that you were actually considering the limit. That is the derivative of $f$ is:

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \hspace{1cm} (2.5)$$
Similarly, for a function of two variables \( z = g(x, y) \), the derivatives of \( f \) with respect to \( x \) and \( y \) are:

\[
f_x(x, y) = \lim_{h \to 0} \frac{f(x + h, y) - f(x, y)}{h} \quad \text{and} \quad (2.6)
\]

\[
f_y(x, y) = \lim_{k \to 0} \frac{f(x, y + k) - f(x, y)}{k} \tag{2.7}
\]

That is, we consider a small change in one variable while holding the other one constant. If one of the limits does not exist (DNE) then we say that \( f \) is not differentiable with respect to that variable.

2.4.3 Higher-Order Partial Derivatives

Thus far we’ve considered \( \frac{\partial z}{\partial x} \) and \( \frac{\partial z}{\partial y} \), these are known as the first order partial derivatives. From these we can obtain the second order partial derivatives. These are:

\[
\frac{\partial}{\partial x} \left( \frac{\partial f}{\partial x} \right) = \frac{\partial^2 f}{\partial x^2}
\]

\[
\frac{\partial}{\partial y} \left( \frac{\partial f}{\partial x} \right) = \frac{\partial^2 f}{\partial y \partial x}
\]

\[
\frac{\partial}{\partial x} \left( \frac{\partial f}{\partial y} \right) = \frac{\partial^2 f}{\partial x \partial y}
\]

\[
\frac{\partial}{\partial y} \left( \frac{\partial f}{\partial y} \right) = \frac{\partial^2 f}{\partial y^2}
\]

It is important to note:

\[
\frac{\partial^2 f}{\partial x^2} \neq \left( \frac{\partial f}{\partial x} \right)^2
\]

• Ex:

Note that in most examples that you can come up with,

\[
\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}
\]

This is not just a coincidence, but rather these mixed second-order partial derivatives are equal when all second order partial derivatives are continuous. This is known as Young’s Theorem.
Chapter 3

Chapter 6 Multivariable Optimization

Recall from one-variable calculus that we spent much time finding extreme values of a function, that is the maximum or minimum of a function (if they exist). We can extend previous techniques to functions of more than variable in order to find these extreme values. Why do extreme values even interest us? From a business point of view, consider a company that mass produces two items: $x$ and $y$ and the cost of producing these items is given by a cost function $C(x, y)$, and the profit from selling the items is given by $P(x, y)$. It follows that a company would want to minimize $C$ and maximize $P$. It is such real world examples as this that motivates us to study the maximization/minimization techniques in chapter 6.

3.1 Section 6.1 Optimization in two variables

Suppose we have some function $z = f(x, y)$ which achieves its maximum somewhere on the graph. Call this point $(x_0, y_0)$. Fixing $y_0$, recall from one variable calculus that there is a specific $x = x_0$ such that at the point $(x_0, y_0)$, $f_x(x_0, y_0) = 0$. (Similarly there is some $y = y_0$ such that at $(x_0, y_0)$, $f_y(x_0, y_0) = 0$).

**Def:** A point at which both first order partial derivatives are equal to zero is called a stationary point. (Note: This is not unlike in one variable calculus when we take the first derivative and set it equal to zero to find a stationary or critical point).

**Theorem:** A differentiable function $z = f(x, y)$ achieves its maximum/minimum at an interior point $(x_0, y_0)$ if it is a stationary point. That is, if the point
satisfies:

\[ f_x(x, y) = 0 \text{ and } f_y(x, y) = 0 \]

### 3.1.1 Finding maximums/minimums

Recall that for functions of one variable, if \( f'(c) = 0 \) at some point \( c \in \mathcal{D} \) and \( f'(x) \geq 0 \) for \( x \leq c \) and \( f'(x) \leq 0 \) for \( x \geq c \), then \( c \) is a maximum.

**Def:** A set \( S \) in the \( x - y \) plane is convex if for every pair of points in \( S \) all points on the line segment between the points lie in \( S \).

**Theorem:** Suppose that \((x_0, y_0)\) is a stationary point for a twice continuously differentiable \((C^2)\) function in a convex set \( S \). Then:

1. If for every \((x, y)\) in \( S \) \( f_{xx}(x, y) \leq 0 \), \( f_{yy}(x, y) \leq 0 \) and \( f_{xx}(x, y) \cdot f_{yy}(x, y) - (f_{xy}(x, y))^2 \geq 0 \), then \((x_0, y_0)\) is a maximum.

2. If for every \((x, y)\) in \( S \) \( f_{xx}(x, y) \geq 0 \), \( f_{yy}(x, y) \geq 0 \) and \( f_{xx}(x, y) \cdot f_{yy}(x, y) - (f_{xy}(x, y))^2 \geq 0 \), then \((x_0, y_0)\) is a minimum.

- **Ex:**

### 3.2 Section 6.2 Local Extreme Points

**Def:** \((x_0, y_0)\) is a local maximum point of \( f \) in a set \( S \) if \( f(x, y) < f(x_0, y_0) \) for all \((x, y)\) near \((x_0, y_0)\). That is, if there is some positive number \( r > 0 \) such that for all points \((x, y)\) within the circle centered at \((x_0, y_0)\) with radius \( r \), \( f(x, y) < f(x_0, y_0) \), then \((x_0, y_0)\) is a local maximum. There is a similar definition for local minimum.

At a local extreme point, all first order partial derivatives equal 0. (Note, however, that a stationary point is not necessarily a local extreme point. For us, stationary points are just a place to start!)

**Def:** A stationary point \((x_0, y_0)\) which is neither a local maximum nor a local minimum is called a saddle point.

At a saddle point \((x_0, y_0)\) there exists points \((x, y)\) such that \( f(x, y) < f(x_0, y_0) \) and there exist other points \((x, y)\) such that \( f(x, y) > f(x_0, y_0) \)

- **Ex:**
3.2.1 Classifying stationary points

Stationary points can be classified as local maximum points, local minimum points, or saddle points. To decide which of these a stationary point is, we use the second derivative test for local extrema.

Theorem 6.2.1 Let \( f(x, y) \) be a function with continuous second order partial derivatives over some domain \( S \) and let \((x_0, y_0)\) be an interior point of \( S \) that is a stationary point of \( f \). Let \( A = f_{xx}(x_0, y_0) \), \( B = f_{xy}(x_0, y_0) \), and \( C = f_{yy}(x_0, y_0) \).

1. If \( A < 0 \) and \( AC - B^2 > 0 \) then \((x_0, y_0)\) is a local maximum.
2. If \( A > 0 \) and \( AC - B^2 > 0 \) then \((x_0, y_0)\) is a local minimum.
3. If \( AC - B^2 < 0 \) then \((x_0, y_0)\) is a saddle.
4. If \( AC - B^2 = 0 \) then \((x_0, y_0)\) is a local maximum or a local minimum or a saddle.

- Ex: Find the stationary points of \( f(x, y) \) and classify as local max/local min/saddle.

3.3 Section 6.3 The Extreme Value Theorem

Recall: In one variable calculus we learn the Extreme Value Theorem (Theorem 3.4.1) says that a function on a closed and bounded interval achieves its maximum and its minimum.

- Ex:

  **Def:** A point \((a, b)\) is called an interior point of a set \( S \) in the plan if there exists a disc centered at \((a, b)\) such that all points strictly inside the disc lie in \( S \).

  **Def:** A set is open if it only consists of interior points.

  **Def:** A point \((a, b)\) is a boundary point of a set \( S \) if every circle centered at \((a, b)\) has points inside \( S \) and outside \( S \).

  **Def:** A set that contains its boundary points is called a closed set.

- Ex:
In optimization problems, sets are defined by inequalities and boundary points occur when these inequalities are satisfied with equality.

**Def:** A set in a plane is **bounded** if the whole set is contained within a sufficiently large circle.

**Def:** A set that is closed and bounded is called **compact**.

- **Ex:**

3.3.1 **Theorem 6.3.1 – The Extreme Value Theorem**

If the function \( f(x, y) \) is continuous throughout a nonempty, closed, and bounded set \( S \) in the plane, then there exists a point \((a, b)\) in \( S \) where \( f \) has a minimum and a point \((c, d)\) in \( S \) where \( f \) has a maximum. That is:

\[
f(a, b) \leq f(x, y) \leq f(c, d) \quad \text{for all } (x, y) \in S
\]

(In other words: \( f \) achieves its maximum and minimum.)

3.3.2 **How to find maxima and minima**

Assume that \( f(x, y) \) is a differential function defined on a closed and bounded set \( S \).

1. Draw the region of interest (region of feasibility)
2. Find all stationary points of \( f \) in the interior of \( S \) and evaluate.
3. Find all points of intersection/corner points on the region of feasibility and evaluate.
4. Find the largest and smallest value of \( f \) on the boundary of \( S \).

- **Ex:**

3.4 **Section 6.4 The Lagrange Multiplier Method**

Consider maximizing a function \( f(x, y) \) when \( x \) and \( y \) must satisfy a complicated function. In other words we are asked something such as:

maximize \( f(x, y) \) subject to \( g(x, y) = c \)

This function \( g(x, y) \) may be considered as a contour on the \( x - y \) plane being projected onto our function \( z = f(x, y) \). It is along this contour, or path, that we are trying to maximize or minimize our function \( f \).
We introduce the Lagrange Multiplier $\lambda$, and the Lagrangian function, $\mathcal{L}$, as follows:

$$\mathcal{L}(x, y) = f(x, y) - \lambda(g(x, y) - c) \quad (3.1)$$

Notice that in (3.1) when the constraint $g(x, y) = c$ is satisfied, then $\mathcal{L}(x, y) = f(x, y).

### 3.4.1 How to use the Lagrange Multiplier Method

To find the only solutions of the problem:

$$\text{max(min) } f(x, y) \text{ subject to } g(x, y) = c$$

1. Write the Lagrangian function $\mathcal{L}(x, y) = f(x, y) - \lambda(g(x, y) - c)$, where $\lambda$ is a constant.
2. Differentiate $\mathcal{L}$ with respect to $x$ and $y$ and set these first order partial derivatives equal to 0.
3. Simultaneously solve the equations from previous step along with the constraint function to find $x$, $y$, and $\lambda$. (Usually try to solve for $\lambda$ first.)

- Ex:

### 3.5 Section 6.7 Linear Programming

A linear programming problem involves maximizing or minimizing a linear function of the form:

$$z = c_1x_1 + c_2x_2. \quad (3.2)$$

Equation (3.2) is known as the criterion or objective function. We want to max/min the objective function subject to a set of constraints of the form:

$$a_{11}x_1 + a_{12}x_2 \leq b_1 \quad (3.3)$$
$$a_{21}x_1 + a_{22}x_2 \leq b_2 \quad (3.4)$$

...$$
$$a_{m1}x_1 + a_{m2}x_2 \leq b_m \quad (3.5)$$
These are known as the inequality constraints.

In addition, we usually require (as seen in the context of an actual problem) the following constraints:

\[ x_1 \geq 0 \]

\[ x_2 \geq 0 \]

These are known as the nonnegativity constraints.

### 3.5.1 Solving LPP’s

- If an LPP has a solution, the solution **must** lie at an extreme point. This is because we are essentially trying to maximize (or minimize) the objective function on a plane. Examining the contours of the plane in question, we can see that all constraints will still be satisfied at a corner point (since the plane is in fact monotonic). Thus, we test corner points to maximize (minimize) the function.

- If the feasible region is unbounded, a finite optimal solution might not exist.

- To solve an LPP problem, carefully write out the objective function and all constraints.

- Ex:
Chapter 4

Chapter 7 – Matrices

Last chapter we saw systems of inequalities with linear programming, where there were many feasible solutions and we chose the maximum or minimum with respect to some objective function. Now we will discuss systems of linear equalities. Here is an example of a system of two variables and two equations:

\[
\begin{align*}
3x_1 + 2x_2 &= 12 \\
x_1 + 2x_2 &= 4
\end{align*}
\]

A solution to this system is an ordered pair of numbers \((s_1, s_2)\) which satisfies both equations. For example \((4, 0)\) is a solution to this system. Even though the ordered pair \((0, 2)\) satisfies the second equation it is not a solution to the system as it does not satisfy the first equation. The general notation for a system with \(m\) equations and \(n\) unknowns, or variables, is given by:

\[
\begin{align*}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
\vdots & \quad \vdots & \quad \vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{align*}
\]

Here \(a_{ij}, a_{12}, \ldots, a_{mn}\) are called the coefficients of the system, \(x_1, \ldots, x_n\) are the variables, and \(b_1, \ldots, b_m\) are called the constants. Note the order of the subscripts: \(a_{ij}\) is the coefficient in the \(i\)th equation of the \(j\)th variable \(x_j\).

A solution to this system is an ordered set or list of numbers \(s_1, s_2, \ldots, s_n\) that satisfies all the equations, with \(x_1 = s_1, x_2 = s_2, \ldots, x_n = s_n\). If the system has a unique solution, or exactly one value for each variable, then it is said to be independent. When the system has no solution it is said to be inconsistent. If the system has infinitely many solutions then it is said
to be **dependent**. This is caused by one of the equations being a linear combination of the other equations in the system.

### 4.1 Solving a System Using Elimination and Substitution

The elimination method uses linear combinations of the equations in the system to eliminate one or more variables. Consider the solution above. If we subtract the second equation from the first we obtain

\[
\begin{align*}
3x_1 + 2x_2 &= 12 \\
-x_1 + 2x_2 &= 4 \\
\hline
2x_1 &= 8
\end{align*}
\]

Thus \(x_1 = 4\). Plugging this value into either equation gives \(x_2 = 0\). So our system has the solution \((4,0)\). Since this gives a unique solution the system is independent.

In the substitution method we solve for one of the variables in terms of the others in one of the equations and plug this result into the other equations. This will allow us to solve for each variable. Again consider the example above. Solving for \(x_1\) in the second equation gives

\[x_1 = 4 - 2x_2\]

Plugging this into the first equation we have

\[3(4 - 2x_2) + 2x_2 = 12 - 4x_2 = 12\]

Thus \(x_2 = 0\) and \(x_1 = 4 - 2(0) = 4\), giving the same solution obtained above using elimination.

#### 4.1.1 Dependent and Inconsistent Systems

Here we explain how to identify when a system is inconsistent or dependent. Consider the following example

\[
\begin{align*}
3x_1 + 6x_2 &= 12 \\
x_1 + 2x_2 &= 5
\end{align*}
\]
Using elimination (add equation 1 to $-3$ times equation 2) we get $0 = -3$, which is a contradiction. Thus the system has no solution and is inconsistent. Note that we would get the same result if we used substitution. Next consider the following system.

\[
\begin{align*}
3x_1 + 6x_2 + 2x_3 &= 15 \\
x_1 + x_2 &= 4 \\
3x_2 + 2x_3 &= 3
\end{align*}
\]

First, let’s eliminate $x_1$ from the first equation using the second:

\[
\begin{align*}
3x_1 + 6x_2 + 2x_3 &= 15 \\
-3(x_1 + x_2 &= 4) \\
3x_2 + 2x_3 &= 3
\end{align*}
\]

Since this yields the third equation we know that this is a dependent system. To write out our solution set choose an arbitrary value for one of the variables, say $x_2 = t$. We then solve for the other variables in terms of $t$. This gives us a solution set of

\[
\begin{align*}
x_1 &= 4 - t \\
x_2 &= t \\
x_3 &= \frac{3}{2} - \frac{3}{2}t
\end{align*}
\]

### 4.2 Matrices and Matrix Operations

A **matrix** is simply a rectangular array of entries. When there are $m$ rows and $n$ columns we say the matrix has dimension $m \times n$. Generally we denote matrices by capital letters such as $A$, $B$, or $X$. In general an $m \times n$ matrix is of the form

\[
A = (a_{ij})_{m \times n} = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{pmatrix}
\]

Note that entry $a_{ij}$ denotes the element, or entry, in the $i$th row and $j$th column. If a matrix contains only one row then we say it is a **row vector** or **row matrix**. If it contains only one column then we say it is a **column**
vector or column matrix. If a matrix has the same number of rows and columns then we say it is a square matrix of size \( n \times n \). Also, if \( A \) is a square matrix the entries \( a_{11}, a_{22}, \ldots, a_{nn} \) which run from the top right corner to the bottom left are referred to as the main diagonal.

Examples: The following are matrices.

\[ A = \begin{pmatrix} 1 & 2 & 0 & -4 \\ 3 & -5 & 11 & -1 \\ 0 & 2 & 17 & 8 \end{pmatrix}, \quad B = \begin{pmatrix} 5 & 9 \\ -2 & 6 \end{pmatrix}, \quad C = \begin{pmatrix} -5 \\ 0 \\ 4 \\ 12 \end{pmatrix} \]

\( A \) is \( 3 \times 4 \), \( B \) is \( 2 \times 2 \), and \( C \) is \( 4 \times 1 \). Note that \( B \) is a square matrix and \( C \) is a column vector, with \( a_{14} = -4, b_{21} = -2, \) and \( c_{12} = 0 \). Also, the main diagonal of \( B \) has entries 5 and 6.

In order for two matrices \( A \) and \( B \) to be equal \(( A = B )\) they must have the same dimension and corresponding entries must be equal: \( a_{ij} = b_{ij} \). Addition of matrices is defined only for matrices of the same dimension and is done component-wise, that is if \( C = A + B \) then

\[ (c_{ij})_{m \times n} = (a_{ij} + b_{ij})_{m \times n} = (a_{ij})_{m \times n} + (b_{ij})_{m \times n} \]

. For example:

\[ \begin{pmatrix} 0 & 3 \\ -1 & 6 \end{pmatrix} + \begin{pmatrix} 2 & -4 \\ 5 & 10 \end{pmatrix} = \begin{pmatrix} 0 + 2 & 3 - 4 \\ -1 + 5 & 6 + 10 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 4 & 16 \end{pmatrix} \]

Also for any real number \( \alpha \), we define scalar multiplication as \( \alpha A = \alpha(a_{ij}) = (\alpha a_{ij})_{m \times n} \), that is we multiply each entry of the matrix by the real number. For example:

\[ 2 \begin{pmatrix} 0 & 3 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} 0 & 6 \\ -2 & 12 \end{pmatrix} \]

4.2.1 Rules for Matrix Addition and Scalar Multiplication

1. \((A + B) + C = A + (B + C)\)
2. \(A + B = B + A\)
3. \(A + 0 = A\)
4. $A + (-A) = 0$

5. $(\alpha + \beta)A = \alpha A + \beta A$

6. $\alpha(A + B) = \alpha A + \beta B$

### 4.3 Matrix Multiplication

Unlike the matrix operations defined so far, matrix multiplication is not as natural. For the product $AB$ to be defined $B$ must have as many rows as $A$ has columns. Let $A = (a_{ij})_{m \times n}$ and $B = (b_{ij})_{n \times p}$. Then the product $C = AB$ is an $m \times p$ matrix $C = (c_{ij})_{m \times p}$, where

$$c_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj} \quad (4.1)$$

Thus to find the $ij$th entry of the product of $A$ and $B$, we take the inner product of the $i$th row of $A$ and the $j$th column of $B$.

Example: Let $A = \begin{pmatrix} 1 & 1 \\ 3 & -5 \end{pmatrix}$ and $B = \begin{pmatrix} 3 & 1 & 0 \\ -1 & 6 & 7 \end{pmatrix}$. Then $A$ is $2 \times 2$ and $B$ is $2 \times 3$ and so $C = AB$ is defined as a $2 \times 3$ matrix. Here we show how to obtain the entry $c_{12}$ by using the first row of $A$ and the second column of $B$.

$$C = AB = \begin{pmatrix} 2 & \frac{(1)(1) + (1)(6)}{14} & 7 \\ -27 & -35 \end{pmatrix} = \begin{pmatrix} 2 & 7 & 7 \\ 14 & -27 & -35 \end{pmatrix}$$

Note that $BA$ is not defined since the number of rows of $A$ does not equal the number of columns of $B$. Even in cases where both $AB$ and $BA$ are defined, it is unlikely that they will be equal.

### 4.3.1 Rules for Matrix Multiplication

Here we assume that $A$, $B$, and $C$ are matrices whose dimension are such that the given operations are defined.

1. Associative Law: $(AB)C = A(BC)$

2. Left Distribution: $A(B + C) = AB + AC$

3. Right Distribution: $(A + B)C = AC + BC$
We can take powers of a matrix: $A^2 = A \cdot A$, $A^3 = A \cdot A \cdot A$, and $A^n = A \cdot A \cdot \cdots \cdot A$ (n times). Note: we do not simply take the power of each entry!

Example: Let $A = \begin{pmatrix} 1 & -2 \\ 1 & 1 \end{pmatrix}$

So $A^2 = \begin{pmatrix} -1 & -4 \\ 2 & -1 \end{pmatrix}$ and $A^3 = \begin{pmatrix} -5 & -2 \\ 1 & -5 \end{pmatrix}$

The identity matrix of order $n$ is the $n \times n$ matrix having 1’s along the diagonal and 0’s elsewhere:

$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ \quad I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

Let $A$ be any $m \times n$ matrix, $B$ any $n \times m$ matrix, and $C$ any $n \times n$ matrix. Then we have:

$AI_n = A \quad I_nB = B \quad CI_n = C = I_nC$

### 4.4 Elementary Row Operations

Often we want to “simplify” the information given in a matrix. However, we do not want to change the basic relationships that are expressed through the rows and columns of that matrix. To do this we are only permitted to use one of the following elementary row operations:

1. Switch rows: $R_i \leftrightarrow R_j$

2. Multiply a row by any nonzero constant: $R_i = \alpha r_i$

3. Add one row to any nonzero multiple of another row: $R_i = r_i + \alpha r_j$

We will use the elementary row operations to get a matrix into the following forms:

1. Row-Echelon Form (REF): the first nonzero entry of any row is a 1 and zeroes are below each of the leading 1’s. A few examples are given
below:

\[
\begin{pmatrix}
1 & 2 & 2 \\
0 & 0 & 1
\end{pmatrix}
\quad \begin{pmatrix}
1 & 2 & 2 & 15 \\
0 & 1 & 0 & 4
\end{pmatrix}
\quad \begin{pmatrix}
1 & 2 & 2 \\
0 & 0 & 1
\end{pmatrix}
\quad \begin{pmatrix}
1 & 2 & 3 \\
0 & 1 & 4
\end{pmatrix}
\quad \begin{pmatrix}
1 & 2 & 2 \\
0 & 1 & 3
\end{pmatrix}
\]

2. Reduced Row-Echelon Form (RREF): the first nonzero entry of any row is a 1 and zeroes are above and below each of the leading 1’s. A few examples are given below:

\[
\begin{pmatrix}
1 & 2 & 0 \\
0 & 0 & 1
\end{pmatrix}
\quad \begin{pmatrix}
1 & 0 & 0 & 15 \\
0 & 1 & 0 & 4
\end{pmatrix}
\quad \begin{pmatrix}
1 & 2 & 0 \\
0 & 0 & 1
\end{pmatrix}
\quad \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0
\end{pmatrix}
\quad \begin{pmatrix}
1 & 0 & 2 \\
0 & 1 & 3
\end{pmatrix}
\]

We will see the use of this later in Gaussian elimination and in finding the inverse of a matrix.

Example: Find the RREF of the matrix

\[
\begin{pmatrix}
2 & 2 & 5 \\
1 & 1 & 0 \\
1 & 1 & 3
\end{pmatrix}
\]

\[
\begin{pmatrix}
2 & 2 & 5 \\
1 & 1 & 0 \\
1 & 1 & 3
\end{pmatrix}
\quad \begin{pmatrix}
1 & 1 & 0 \\
2 & 2 & 5 \\
1 & 1 & 3
\end{pmatrix}
\quad \begin{pmatrix}
1 & 1 & 0 \\
0 & 0 & 5 \\
1 & 1 & 3
\end{pmatrix}
\quad \begin{pmatrix}
1 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{pmatrix}
\]

4.5 Determinants

The determinant of a square matrix \( A \) is a function that assigns a number, or scalar, to that matrix, denoted by \( \det(A) \) or \(|A|\). The formula for calculating the determinant of a \( 2 \times 2 \) matrix is shown below.
There are two methods for calculating the determinant of larger dimension matrices. The first is a shortcut that only works for $3 \times 3$ matrices, whereas the second can be used for square matrices of any dimension.

### 4.5.1 Sarrus’s Rule

If $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, write $a_{11} a_{22} a_{33} + a_{12} a_{23} a_{31} + a_{13} a_{21} a_{32}$

Multiply along the diagonals from top left to bottom right and take the sum. Subtract each of the “backwards diagonals” from top right to bottom left as shown below:

**Step 1:** First multiply along the diagonals from top left to bottom right and take the sum:

\[
\begin{align*}
    a_{11} & \rightarrow a_{12} & \rightarrow a_{13} & \rightarrow a_{11} & \rightarrow a_{12} \\
    a_{21} & \rightarrow a_{22} & \rightarrow a_{23} & \rightarrow a_{21} & \rightarrow a_{22} \\
    a_{31} & \rightarrow a_{32} & \rightarrow a_{33} & \rightarrow a_{31} & \rightarrow a_{32}
\end{align*}
\]

yields $(a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32})$

**Step 2:** Now multiply the “backwards diagonals” from top right to bottom left and take the sum:

\[
\begin{align*}
    a_{11} & \rightarrow a_{12} & \rightarrow a_{13} & \rightarrow a_{11} & \rightarrow a_{12} \\
    a_{21} & \rightarrow a_{22} & \rightarrow a_{23} & \rightarrow a_{21} & \rightarrow a_{22} \\
    a_{31} & \rightarrow a_{32} & \rightarrow a_{33} & \rightarrow a_{31} & \rightarrow a_{32}
\end{align*}
\]

yields $(a_{13}a_{22}a_{31} + a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33})$
**Step 3:** Subtract step 2 from step 1, yielding the determinant of \( A \):

\[
\text{det}(A) = (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) - (a_{13}a_{22}a_{31} + a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33})
\]

(4.3)

An example is shown below:

\[
A = \begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{pmatrix} \rightarrow \begin{pmatrix}
1 & 2 & 3 & 1 & 2 \\
4 & 5 & 6 & 4 & 5 \\
7 & 8 & 9 & 7 & 8
\end{pmatrix}
\]

\[
\text{det}(A) = (1 \cdot 5 \cdot 9 + 2 \cdot 6 \cdot 7 + 3 \cdot 4 \cdot 8) - (3 \cdot 5 \cdot 7 + 1 \cdot 6 \cdot 8 + 2 \cdot 4 \cdot 9)
\]

\[
= (45 + 84 + 96) - (105 + 48 + 72)
\]

\[
= 225 - 225
\]

\[
= 0
\]

### 4.5.2 Expansion by Cofactors

Here we choose a row or column to expand along and find the determinant of smaller matrices. It is usually best to choose the row or column that has the most zero entries. Here we show the process in general where we have chosen to expand along the first row.

\[
\text{det}(A) = \begin{vmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{vmatrix} = a_{11} \begin{vmatrix}
a_{22} & a_{23} \\
a_{32} & a_{33}
\end{vmatrix} - a_{12} \begin{vmatrix}
a_{21} & a_{23} \\
a_{31} & a_{33}
\end{vmatrix} + a_{13} \begin{vmatrix}
a_{21} & a_{22} \\
a_{31} & a_{32}
\end{vmatrix}
\]

The sign of \( a_{ij} \) can be found by summing the indices: if \( i + j \) is even then use \( a_{ij} \) and if \( i + j \) is odd then use \( -a_{ij} \). The smaller matrices are obtained by deleting the row and column that \( a_{ij} \) appears in. The example below shows how to find the determinant of \( A \) from the previous section by using expansion by cofactors along the first row:

\[
\text{det}(A) = \begin{vmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{vmatrix} = 1 \cdot \begin{vmatrix}
5 & 6 \\
8 & 9
\end{vmatrix} - 2 \cdot \begin{vmatrix}
4 & 6 \\
7 & 9
\end{vmatrix} + 3 \cdot \begin{vmatrix}
4 & 5 \\
7 & 8
\end{vmatrix}
\]

\[
= 1 \cdot (5 \cdot 9 - 8 \cdot 6) - 2 \cdot (4 \cdot 9 - 6 \cdot 7) + 3 \cdot (4 \cdot 8 - 5 \cdot 7)
\]

\[
= -3 + 12 - 9
\]

\[
= 0
\]
4.6 Inverses

Given any matrix $A$, we say $X$ is the inverse of $A$ if

$$AX = I =XA$$

and we say $A$ is invertible. Generally we denote the inverse of $A$ by $A^{-1}$. Note that only square matrices can have inverses and that $A$ and $A^{-1}$ must be of the same dimension. It is very easy to determine which matrices have an inverse:

$A$ has an inverse if and only if $A$ is square and $\det(A) \neq 0$.

4.6.1 Finding the Inverse of a Matrix

There are two main methods for finding the inverse of a matrix. The first uses the adjoint of a matrix and the second uses the RREF of the augmented matrix $(A|I)$.

General Formula for the Inverse

Let $A$ be any square matrix with $\det(A) \neq 0$. Then the formula for the unique inverse matrix is $A^{-1} = \frac{1}{\det(A)} \text{adj}(A)$, where

$$\text{adj}(A) = \begin{pmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{pmatrix}$$

and $C_{ij}$ is the cofactor obtained by removing the $i$th row and $j$th column. An example is shown below:

$$A = \begin{pmatrix} 3 & 6 & 2 \\ 1 & 1 & 0 \\ 0 & 3 & 3 \end{pmatrix}$$

So $\det(A) = -3$. The cofactors of $A$ are:

$$C_{11} = \begin{vmatrix} 1 & 0 \\ 3 & 3 \end{vmatrix} = 3 \quad C_{21} = -\begin{vmatrix} 6 & 2 \\ 3 & 3 \end{vmatrix} = -12 \quad C_{31} = \begin{vmatrix} 6 & 2 \\ 1 & 0 \end{vmatrix} = -2$$

$$C_{12} = -\begin{vmatrix} 1 & 0 \\ 0 & 3 \end{vmatrix} = -3 \quad C_{22} = \begin{vmatrix} 3 & 2 \\ 0 & 3 \end{vmatrix} = 9 \quad C_{32} = -\begin{vmatrix} 3 & 2 \\ 1 & 0 \end{vmatrix} = 2$$
\[ C_{13} = \begin{vmatrix} 1 & 1 \\ 0 & 3 \end{vmatrix} = 3 \quad C_{23} = -\begin{vmatrix} 3 & 6 \\ 0 & 3 \end{vmatrix} = -9 \quad C_{33} = \begin{vmatrix} 3 & 6 \end{vmatrix} = -3 \]

This gives the inverse matrix
\[
A^{-1} = \frac{1}{-3} \begin{pmatrix} 3 & -12 & -2 \\ -3 & 9 & 2 \\ 3 & -9 & -3 \end{pmatrix} = \begin{pmatrix} -1 & 4 & 2/3 \\ 1 & -3 & -2/3 \\ -1 & 3 & 1 \end{pmatrix}
\]

Note that when we do use this method for a \(2 \times 2\) matrix, a shortcut emerges:

\[
\text{if } A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ then } A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}
\]

Using RREF to find the inverse

Another method for finding the inverse uses elementary row operations on the augmented matrix \((A|I)\). If the RREF of this matrix is \((I|B)\) then \(B = A^{-1}\). Otherwise \(A^{-1}\) does not exist.

Example: Let \(A = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 1 \\ 0 & 1 & 3 \end{pmatrix}\). Then we can find the RREF of \((A|I)\) by the following set of elementary row operations:

\[
\begin{pmatrix} 1 & 0 & 0 & | & 1 & 0 & 0 \\ 2 & 1 & 1 & | & 0 & 1 & 0 \\ 0 & 1 & 3 & | & 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 = r_2 - 2r_1} \begin{pmatrix} 1 & 0 & 0 & | & 1 & 0 & 0 \\ 0 & 1 & 1 & | & -2 & 1 & 0 \\ 0 & 1 & 3 & | & 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_3 = r_3 - r_2} \begin{pmatrix} 1 & 0 & 0 & | & 1 & 0 & 0 \\ 0 & 1 & 1 & | & -2 & 1 & 0 \\ 0 & 0 & 2 & | & 2 & -1 & 1 \end{pmatrix}
\]

\[
\xrightarrow{R_3 = 1/2r_3} \begin{pmatrix} 1 & 0 & 0 & | & 1 & 0 & 0 \\ 0 & 1 & 1 & | & -2 & 1 & 0 \\ 0 & 0 & 1 & | & 1 & -1/2 & 1/2 \end{pmatrix} \xrightarrow{R_2 = r_2 - r_3} \begin{pmatrix} 1 & 0 & 0 & | & 1 & 0 & 0 \\ 0 & 1 & 0 & | & -3 & 3/2 & -1/2 \\ 0 & 0 & 1 & | & 1 & -1/2 & 1/2 \end{pmatrix}
\]

Thus \(A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 3/2 & -1/2 \\ 1 & -1/2 & 1/2 \end{pmatrix}\)(check this by multiplying \(A \cdot A^{-1}\) and \(A^{-1} \cdot A\))
4.6.2 Properties of the Inverse

Here we list some important properties of the inverse that may be useful. Let $A$ and $B$ be invertible $n \times n$ matrices. Then:

1. inverses are unique, i.e. $A^{-1}$ and $B^{-1}$ are unique.
2. $(A^{-1})^{-1} = A$
3. $(AB)^{-1} = B^{-1}A^{-1}$
4. $(A^t)^{-1} = (A^{-1})^t$
5. $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$ where $\alpha$ is any nonzero number

4.7 Solving a System Using Matrices

Our discussion of matrices has an important application to systems of equations. We will outline 3 different ways to solve a system through the use of matrix operations.

4.7.1 Gaussian Elimination

The process of Gaussian elimination has two parts. The first part reduces a given system to row-echelon form (REF). Remember that this is accomplished through the use of elementary row operations. The second step uses back-substitution to find the reduced row-echelon form (RREF) and from that the solution of the system.

First we need to explain how to get a system into matrix form. Any system of equations can be written in the form $AX = B$, where $A$ is the coefficient matrix, $X$ is the variable matrix, and $B$ is the matrix of constants. For example, given the system

\[
3x_1 + 6x_2 + 2x_3 = 15 \\
x_1 + x_2 = 4 \\
3x_2 + 3x_3 = 3
\]

we get the following matrices

\[
A = \begin{pmatrix} 3 & 6 & 2 \\ 1 & 1 & 0 \\ 0 & 3 & 3 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad B = \begin{pmatrix} 15 \\ 4 \\ 3 \end{pmatrix}
\]
For Gaussian elimination we wish to form the **augmented matrix** \((A|B)\).

From the example above we have

\[
\begin{pmatrix}
3 & 6 & 2 & 15 \\
1 & 1 & 0 & 4 \\
0 & 3 & 3 & 3
\end{pmatrix}
\]

To solve this system we are only permitted to use elementary row operations, as defined earlier, to get the augmented matrix into RREF. Once this is accomplished we can easily determine the solutions to the system of equations, if any exist. Examples are shown below.

1. \[
\begin{pmatrix}
1 & 0 & 0 & 3 \\
0 & 1 & 0 & -2 \\
0 & 0 & 1 & 0
\end{pmatrix}
\]
   This system has a unique solution of \(x_1 = 3\), \(x_2 = -2\), and \(x_3 = 0\). Thus it is independent.

2. \[
\begin{pmatrix}
1 & 0 & 1 & 3 \\
0 & 1 & 2 & -2 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]
   This system has no solution since the last row gives a contradiction \((0 \neq 1)\). Thus it is inconsistent.

3. \[
\begin{pmatrix}
1 & 0 & 1 & 3 \\
0 & 1 & 2 & -2 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]
   This system has an infinite number of solutions of the form \(x_1 = 3 - t\), \(x_2 = -2 - 2t\), and \(x_3 = t\). Thus it is dependent.

### 4.7.2 Inverse of the Coefficient Matrix

We stated above that any system of equations can be written in the form \(AX = B\), where \(A\) is the coefficient matrix, \(X\) is the variable matrix, and \(B\) is the matrix of constants. If \(A\) is a square matrix and \(\det(A) \neq 0\) then \(A^{-1}\) exists. So we have

\[X = IX = A^{-1}AX = A^{-1}B\]
Thus the system will have a unique solution.
From example, if
\[
A = \begin{pmatrix} 3 & 6 & 2 \\ 1 & 1 & 0 \\ 0 & 3 & 3 \end{pmatrix}
\]
then \( \det(A) = -3 \) and so \( A^{-1} \) exists. Using any of the techniques described earlier we find that
\[
A^{-1} = \begin{pmatrix} -1 & 4 & 2/3 \\ 1 & -3 & -2/3 \\ -1 & 3 & 1 \end{pmatrix}
\]
Thus we know the corresponding system has a unique solution and we find it as follows:
\[
X = A^{-1}B = \begin{pmatrix} -1 & 4 & 2/3 \\ 1 & -3 & -2/3 \\ -1 & 3 & 1 \end{pmatrix} \begin{pmatrix} 15 \\ 4 \\ 3 \end{pmatrix}
\]
\[
= \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}
\]
Thus \( x_1 = 3, \ x_2 = 1, \) and \( x_3 = 0. \)

4.7.3 Cramer’s Rule
As above we will represent a system of equations as \( AX = B. \) If \( A \) is an invertible square matrix then Cramer’s Rule states that
\[
x_i = \frac{\det(A_i)}{\det(A)}
\]
where \( A_i \) is the matrix formed by replacing the \( i \)th column of \( A \) by the column vector \( B. \) For example, consider the following system
\[
3x_1 + 2x_2 = 12 \\
x_1 + 2x_2 = 4
\]
We first check that \( A = \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix} \) is invertible: \( \det(A) = 3 \cdot 2 - 2 \cdot 1 = 4. \)
Next we find \( A_1 \) and \( A_2: \)
\[
A_1 = \begin{pmatrix} 12 & 2 \\ 4 & 2 \end{pmatrix} \quad A_2 = \begin{pmatrix} 3 & 12 \\ 1 & 4 \end{pmatrix}
\]
Thus there is a unique solution:

\[ x_1 = \frac{\det A_1}{\det A} = \frac{16}{4} = 4 \]
\[ x_2 = \frac{\det A_2}{\det A} = \frac{0}{4} = 0 \]

### 4.8 Interpreting Solutions of Linear Equations

Consider the solutions throughout this section. In the last example (also Example 1 in section 4.1), we found a solution of \((4,0)\). What does this solution mean, however? Since this coordinate pair satisfies both of our linear equations, this actually represents the point on the graph that these two lines intersect.

This means that when we solve a system of linear equations (regardless of whether it’s 2 equations or 10 equations) we are looking for the point at which all of the lines (or planes) meet.

In the second example in section (4.1.1) we encountered a system that had no solution. If a solution implies that the lines intersect, then no solution implies that the lines never intersect. The only way this can happen is if we have parallel lines (or planes).

Finally, the third type of solution was when a system had infinitely many solutions. The reason for this is because of our \(n\) equations, at least two of them were actually the same equation. Hence, they intersect in infinitely many places.

We can also relate this to the previous discussion on determinants. Recall that a determinant is nonzero if and only if a coefficient matrix \(A\) has an inverse. If \(A^{-1}\) exists, then a solution exists to a system of linear equations.

Hence, in summary:

<table>
<thead>
<tr>
<th>System is...</th>
<th>it has ___ solution(s)</th>
<th>the lines (planes)...</th>
<th>determinant is...</th>
</tr>
</thead>
<tbody>
<tr>
<td>independent</td>
<td>1 unique</td>
<td>intersect exactly once</td>
<td>nonzero</td>
</tr>
<tr>
<td>dependent</td>
<td>infinitely many</td>
<td>are the same line</td>
<td>0</td>
</tr>
<tr>
<td>inconsistent</td>
<td>no</td>
<td>are parallel</td>
<td>0</td>
</tr>
</tbody>
</table>
Chapter 5

Chapter 8 – Probability

5.1 Set Theory

5.1.1 Relationships between sets

A set is a collection of distinct objects. The objects are referred to as elements of a set. We denote a set in two common ways:

- \( B = \{0,1,2,3,4,5,6,7,8,9\} \). This is known as roster notation - when all the elements are listed out.
- \( B = \{x : x \text{ is a digit}\} \). This is known as set-builder notation.

Elements of a set are never repeated. That is, we would not say that a set \( B = \{1,0,1,4\} \), but rather we would say that the set \( B = \{1,0,4\} \). Also notice that it does not matter how we arrange the elements of the set. \( B \) could also be written as \( B = \{0,1,4\} \) or \( B = \{4,0,1\} \). In fact, there are 6 different ways to list the elements of set \( B \). (We will learn where the 6 different arrangements come from later!)

A set is well defined if there is a clear whether or not an element belongs in the set. For example, we could define a set to be the set of all Tom Cruise movies. If you name any movie, it is clear whether or not it belongs inside of this set. This is well defined. Now consider the set of all good Tom Cruise movies. Whether or not an element belongs to this set is not as clear – you might think that Mission Impossible is a good Tom Cruise movie and someone else may not agree. This is not a well defined set.
It is convenient for us to define the empty set or null set, denoted {} or ∅. This is the set that has no elements in it. For example, we could define a set \( C = \{ x : x \text{ is a state that has 4 senators} \} \). Since there are no such elements that exist in this set (namely, there are no states that have more than 2 senators), set \( C \) is an empty set.

Now we will discuss the algebra of sets:

Let \( A \) and \( B \) be two sets. Then we say that \( A \) is equal to \( B \), or \( A = B \), if and only if \( A \) and \( B \) have the same elements. Again, note that the elements of the two sets do not have to be in the same order, as we can reorder the elements of a set without changing the set itself. If two sets do not have the same exact elements, we say that they are not equal or \( A \neq B \).

Let \( A \) and \( B \) be two sets. Then we say that \( A \) is a subset of, or \( A \) is contained in \( B \), denoted \( A \subseteq B \), if every element of \( A \) is also in \( B \). If \( A \) is not a subset of \( B \) (if \( A \) contains elements that are not in \( B \)) then we write this as \( A \nsubseteq B \). Note that technically, every set is a subject of itself, so for any set \( A \), \( A \subseteq A \). In addition, the null set is a subset of every set. So for any set \( B \), \( \emptyset \subseteq B \).

The Universal Set, \( U \) is defined as the set consisting of all elements under consideration. The need to define the universal set will become more apparent as we consider operations on sets.

### 5.1.2 Operations on sets

A Venn diagram is a way to display relationships between sets. We will use Venns when considering relationships between sets.

Let \( A \) and \( B \) be any two sets. The union of \( A \) and \( B \) or \( A \cup B \), is defined to be the set consisting of all elements in \( A \) or \( B \) or both. Note that “or” in this sense means \( A \) or \( B \) or both. This is known as the “inclusive or”. An element only needs to be in at least one of the sets in order to be a part of the union.

Let \( A \) and \( B \) be any two sets. The intersection of \( A \) and \( B \) or \( A \cap B \), is defined as the set consisting of all elements that are in both \( A \) and \( B \). These are the common elements of the two sets.

If two sets \( A \) and \( B \) have no common elements, that is: \( A \cap B = \{\} \), then
**A** and **B** are disjoint sets or mutually exclusive sets.

If **A** is any set, then the complement of **A**, denoted **A’** or \( \bar{A} \), consists of all elements in the universal set, \( U \) that are not in **A**. Note than \( U’ = \emptyset \) and likewise \( \emptyset’ = U \).

If **A** is a set, then \( c(A) \) denotes the number of elements in **A**. (Many other texts will denote the cardinality, or number of elements in a set **A** by: \( n(A) \) or \( N(A) \).)

The Counting Formula establishes the number of elements in the union of 2 sets. (Note that if two sets are disjoint, since there is no intersection, the cardinality of their union is simply the sum of the cardinality of each set. See [5.2].)

\[
c(A \cup B) = c(A) + c(B) - c(A \cap B) \quad (5.1)
\]

\[
c(A \cup B) = c(A) + c(B) \text{ (when } A \cap B = \{\}) \quad (5.2)
\]

### 5.2 Counting

Suppose you have to pick a new ATM code, a code consisting of 4 digits. How many different ways are there to pick a new code? We can consider this problem by looking at each digit as an “experiment” or “task”. That is, choosing digit 1 is a task, then choosing digit 2 is a separate task, etc.

How many ways can each of these “tasks” be accomplished? This leads us to The multiplication principle.

If an experiment **E**\( _1 \) has \( n_1 \) outcomes and for each of these experiment **E**\( _2 \) has \( n_2 \) outcomes, then the combined experiment \( E_1E_2 \) has \( n_1n_2 \) outcomes.

The above math-speak simply says that the total number of possible outcomes of a series of decisions is found by multiplying the number of choices in each decision. This total number of outcomes is easiest found using a tree diagram (when possible).

Again consider the ATM example. Applying the multiplication principle, and considering choosing each digit to be a separate experiment, we see
that there are $10 \cdot 10 \cdot 10 \cdot 10 = 10000$ (10 choices for each digit) different ways to choose a code.

5.2.1 Permutations

When we count the number of elements in a set, or the number of ways an event can occur, we tend to count the permutations or the combinations, depending on the situation. These two types of problems will eventually lead us into probability. First, however, it is useful to introduce the following:

**Def:** The symbol $n!$, read as “$n$ factorial” is defined for integers greater than or equal to 1 as:

$$n! = n \cdot (n - 1) \cdot (n - 2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$$  \hspace{1cm} (5.3)

We also define

$$0! = 1,$$  \hspace{1cm} (5.4)

and note that:

$$n! = n(n - 1)!$$  \hspace{1cm} (5.5)

Simply stated, $n!$ is the product of all positive integers less than or equal to $n$.

Using factorials essentially develops a shorthand way to evaluate expressions, and rewriting a factorial as in (5.5) will prove to be quite useful.

**Def:** A permutation is an ordered arrangement of $r$ objects from $n$ objects (note: $r \leq n$) and is denoted $nP_r$.

There are essentially three types of permutations as follows:

- The $n$ objects are distinct, repetition is allowed.

  EX: The abbreviations for departments at ASU are 3 letters long with repetition allowed. How many different abbreviations are possible?

- The $n$ objects are distinct, repetition is not allowed.
EX: Now consider ASU department abbreviations, without allowing for repetition.

- The $n$ objects are not distinct, and all are used in an arrangement.

When dealing with permutations, it is key to remember that ORDER MATTERS. Consider this: If I were to give away 3 cash prizes $100, $500, $1000, in that order, you would care in which order you were chosen for a prize, correct? This is a permutation. (This is different than if I were to give away 3 cash prizes, all the same amount, and you clearly don’t care in which order you receive a prize, as long as you receive one.)

We have two useful formulas when considering permutations.

\[
\text{Number of permutations} = n^r \tag{5.6}
\]

and

\[
nPr = \frac{n!}{(n-r)!} \tag{5.7}
\]

Equation (5.6) is used for distinct objects when repetition is allowed, while (5.7) is used for distinct objects when repetition is not allowed. The difference between these types becomes most clear through many many examples.

5.2.2 Combinations

Recall that permutations are used when order matters. This is unlike combinations. We consider combinations when order DOES NOT matter. Consider a 5-card poker hand. If you are dealt 4 aces and a K – do you care if you receive the K then the aces? Or 2 aces, a K, then 2 more aces? Or first the 4 aces then the K? No, you don’t care because in the end you still have four of a kind.

More formally, a combination is an arrangement without regard to order of $r$ objects selected from $n$ distinct objects (note: $r \leq n$) without repetition and is denoted $nC_r$ or

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5.8}
\]
Note that the difference between (5.7) and (5.8) is in the denominator. The addition of \( r! \) in the denominator essentially “unorders” the number of permutations.

Just as with permutations, the way to calculate the number of ways an event can occur in regards to combinations becomes more clear with lots and lots of examples.

### 5.3 Assignment of Probabilities

What is a probability?

We experience probabilities in every day life. What is the chance the it will rain today? What is the probability of winning the lottery? What is the likelihood of correctly guessing every question on a multiple choice exam? We will now discuss how to use set theory in conjunction with permutations and combinations to calculate probabilities.

An experiment is the process by which an outcome is obtained. The set of all outcomes of an experiment is known as the sample space, denoted \( S \). Any subset of the sample space \( S \) is known as an event.

#### 5.3.1 Properties of Probability

Suppose a sample space has \( n \) outcomes, \( S = \{e_1, e_2, ..., e_n\} \). To each of these outcomes, we assign a real number, \( P(e) \) called the probability of \( e \) such that

\[
P(e_i) \geq 0 \quad \forall \quad i = 1, ..., n \quad (5.9)
\]

and

\[
\sum_{i=1}^{n} P(e_i) = 1 \quad (5.10)
\]

Note what (5.9) and (5.10) are saying: probabilities are \textbf{ALWAYS} non-negative and the sum of the probabilities of events in one individual experiment \textbf{MUST} sum to 1.
The probability of any event is a simply a measure of the likelihood that the event will occur. The probability of an event \( E \) is denoted as \( p(E) \).

Consider a trivial example such as tossing a coin. The sample space consisting of all possible outcomes is: \( S = \{H, T\} \). That is, there are two outcomes – heads, or tails. Assuming a non-weighted coin, there is an equal chance of either of these things occurring, so \( p(H) = p(T) = \frac{1}{2} \). Notice that each of these probabilities are nonnegative in accordance with (5.9) and as they are the only outcomes, the sum of their probabilities equals 1, in accordance with (5.10).

In an experiment, outcomes are said to be equally likely outcomes when the same probability is assigned to each outcome of the sample space, \( S \). (Consider the coin example or the experiment of tossing a non-weighted die.) Using this, we can formally define how to find a probability:

Let a sample space \( S = \{e_1, e_2, ..., e_n\} \). Suppose that each outcome in \( S \) is equally likely to occur. Let event \( E \) be so that \( E \subseteq S \). If \( E \) contains \( m \) of the \( n \) outcomes of \( S \), then we define the probability of \( E \) as:

\[
p(E) = \frac{m}{n} \text{ or equivalently } p(E) = \frac{c(E)}{c(S)} \tag{5.11}
\]

Equation (5.11) explains as to why (5.9) is always true. Since \( c(E) \geq 0 \) and \( c(S) > 0 \), the probability of any event must always be nonnegative. That is, since an event cannot have negative cardinality, its probability cannot be negative.

**Probabilities & Odds**

**Def:** The odds for any event \( E \) are defined as:

\[
c(E) : c(\overline{E}) \tag{5.12}
\]

and the odds against any event \( E \) are defined as:

\[
c(\overline{E}) : c(E) \tag{5.13}
\]
Note the relationship between the assignment of odds of an event $E$ as in (5.12) and the assignment of a probability of an event $E$ as in (5.11), i.e. $c(E) + c(\overline{E}) = c(S)$.

More on probabilities...

Def: If event $E = \emptyset$, then $E$ is an impossible event, and

$$p(E) = 0. \hspace{1cm} (5.14)$$

If event $E$ is the union of $r$ disjoint events, $E = \{e_1\} \cup \{e_2\} \cup \ldots \cup \{e_r\}$, then $p(E) = p(e_1) + p(e_2) + \ldots + p(e_r)$.

Def: Two or more events of a sample space are called mutually exclusive events if and only if they have no elements in common. Mutually exclusive events are also called disjoint sets.

If two events are in fact mutually exclusive, the probability of their union can easily be found as follows:

$$p(E \cup F) = p(E) + p(F) \hspace{1cm} (5.15)$$

Note how this is different than the general additive rule for probability for any two sets:

$$p(E \cup F) = p(E) + p(F) - p(E \cap F) \hspace{1cm} (5.16)$$

Let’s think about the relationship between (5.15) and (5.16). Recall: If events $E$ and $F$ are in fact mutually exclusive, or disjoint, then we have that $E \cap F = \emptyset$. Also recall from (5.14) that the probability of the empty set is zero. Hence, (5.16) can also be used for mutually exclusive events, and in fact (5.15) is just a special case of the general additive rule!

Properties of the Probability of an Event

There are a few basic properties of the probability for any event:

- $0 \leq p(E) \leq 1 \forall$ events $E \subseteq S$.
- $p(\emptyset) = 0$ and $p(S) = 1$
• \( p(E \cup F) = p(E) + p(F) - p(E \cap F) \) for any two events \( E \) and \( F \).

• \( p(E) = 1 - p(E') \)

Notice what this last property is saying: Since in fact the union of an event \( E \) and its complement \( E' \) make up the entire sample space \( S \), we have that:

\[
\begin{align*}
E \cup E' &= S \\
p(E) + p(E') &= p(S) \\
\frac{c(E)}{c(S)} + \frac{c(E)}{c(S)} &= \frac{c(S)}{c(S)} \\
&= 1
\end{align*}
\]

5.3.2 Finding probabilities using counting techniques

Recall: For any event \( E \), the probability that the event occurs can be found by \( \frac{c(E)}{c(S)} \). How do we find \( c(E) \)? We use counting techniques as in Section 5.2. Once you are able to clearly define the ways an event can occur and you find the cardinality, simple divide it by the total number of ways the event can happen.

See in class examples.

5.4 Other types of probabilities

5.4.1 Conditional Probability

We oftentimes restrict our sample space based upon previous knowledge. Instead of computing a probability in relation to everything that could happen, we instead compute the probability of an event relative to the occurrence of another event in our sample space.

As an example, suppose you poll a sample of ASU students regarding whether or not they watch “Lost”, denoted \( L \). You compile the data as follows:
If a person were chosen at random, the probability that the person watches “Lost” is found by
\[
\frac{c(L)}{c(S)} = \frac{56}{134}.
\]
However, if you had additional information, such as the person is a female, you could use this information to get a more precise probability. In this case, we would say that the probability the person watches “Lost” given or conditioned upon the fact that she is a female, denoted as \[ p(L | F) = \frac{31}{77}. \] Where did this probability come from? Consider the following more formal derivation of conditional probability:

\[
p(A | B) = \frac{p(A \cap B)}{p(B)}
\]

(5.17)

\[
= \frac{c(A \cap B)}{c(B)}
\]

(5.18)

\[
= \frac{c(A \cap B)}{c(B)}, \ p(B) > 0
\]

(5.19)

The requirement that \( p(B) > 0 \) implies that \( B \) is NOT the empty set. This is a fair requirement since in fact we are using the fact that the event \( B \) has happened, hence \( B \) cannot be the empty set.

The two ways of calculating conditional probabilities in (5.17) and (5.19) both refer to the same basic idea: we have taken a subset of the entire sample space and are thus limiting ourselves to a sample space relative to information we already have.

The multiplication rule

The probability that two events \( A \) and \( B \) both occur, known as the multiplication rule is given as:

\[
cell
cell
cell
\]

<table>
<thead>
<tr>
<th></th>
<th>( L )</th>
<th>( L' )</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>25</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>Female</td>
<td>31</td>
<td>46</td>
<td>77</td>
</tr>
<tr>
<td>total</td>
<td>56</td>
<td>78</td>
<td>134</td>
</tr>
</tbody>
</table>

If a person were chosen at random, the probability that the person watches “Lost” is found by \( \frac{c(L)}{c(S)} = \frac{56}{134} \). However, if you had additional information, such as the person is a female, you could use this information to get a more precise probability. In this case, we would say that the probability the person watches “Lost” given or conditioned upon the fact that she is a female, denoted as \( p(L | F) = \frac{31}{77} \). Where did this probability come from? Consider the following more formal derivation of conditional probability:

\[
p(A | B) = \frac{p(A \cap B)}{p(B)}
\]

(5.17)

\[
= \frac{c(A \cap B)}{c(B)}
\]

(5.18)

\[
= \frac{c(A \cap B)}{c(B)}, \ p(B) > 0
\]

(5.19)

The requirement that \( p(B) > 0 \) implies that \( B \) is NOT the empty set. This is a fair requirement since in fact we are using the fact that the event \( B \) has happened, hence \( B \) cannot be the empty set.

The two ways of calculating conditional probabilities in (5.17) and (5.19) both refer to the same basic idea: we have taken a subset of the entire sample space and are thus limiting ourselves to a sample space relative to information we already have.

The multiplication rule

The probability that two events \( A \) and \( B \) both occur, known as the multiplication rule is given as:
\[ p(A \cap B) = p(A)p(B | A) = p(B)p(A | B) \] (5.20)

### 5.4.2 Independent Events

Recall that in the previous section, knowing that an event has *already* occurred changed the calculated probability of another event occurring. These were *dependent* probabilities. In this section we discover that for certain pairs of events, the occurrence of one may or may not change the probability of the occurrence of the other.

**Def:** Events \( A \) and \( B \) are independent if and only if

\[ p(A \cap B) = p(A)p(B) \] (5.21)

**Theorem:** If two events \( A \) and \( B \) are independent, then the following pairs of events are also independent:

- \( A \) and \( B' \)
- \( A' \) and \( B \)
- \( A' \) and \( B' \)

### 5.4.3 Bayes’ Theorem

First recall that if two events \( A \) and \( B \) are mutually exclusive events, then \( A \cap B = \emptyset \). Given a sample space \( S \), we say that subsets of \( S \), \( B_1, B_2, \ldots, B_n \) constitute a partition of \( S \) if:

- \( S = B_1 \cup B_2 \cup \ldots \cup B_n \)
- \( B_i \cap B_j = \emptyset, \ i \neq j \) and
- \( p(B_i) > 0 \ \forall \ i \)

What does the above mean? It says that a group of events partition the sample space if the union of the events make up the entire sample space, the intersection of any two events is the empty set, and the probability of any one of the events occurring is nonzero.
Once we can partition a sample space as indicated above, we might want to discuss other events that may happen in relation to this partition.

For example, suppose we can partition the entire population of ASU students into freshmen, sophomores, juniors and seniors. The union of these groups make up the undergraduate population and the intersection of any two of these is the empty set. Since there is at least one person in each of the groups (really there are thousands in this case) then then probability for each group occurring is nonzero. Thus all three of the above are fulfilled and we have thus partitioned our sample space. Now suppose we are curious regarding whether or not students drive to school. We might gather information about each one of the groups and what percentage of each group drives. In addition, we can collect information regarding what percentage of the overall population is a freshman, is a sophomore, etc... Thus, We can discuss probabilities now such as

- What is the probability that a student drives to school?
- What is the probability that a student is a junior given that he drives to school?

If we want to answer questions like these, we need to use Bayes’ Theorem.

Given that a group of events do make up a partition of the sample space, we can derive Bayes’ Theorem and we arrive at the following formula:

\[
p(B_k \mid A) = \frac{p(B_k \cap A)}{p(A)} = \frac{p(B_k)p(A \mid B_k)}{\sum_{i=1}^{n} p(B_i)p(A \mid B_i)} \quad (5.22)
\]
Chapter 6

Chapter 9 – Discrete & Continuous Distributions

6.1 Discrete Random Variables

We have thus far been discussing sample spaces and the associated probabilities of the outcomes in the sample space. We will now consider defining functions that map from the sample space to the real numbers.

Consider the following example: As an experiment we flip a coin three times. Using the multiplication principle we know ahead of time that there are $2\cdot2\cdot2 = 8$ possible outcomes from this experiment. We can list the sample space as follows: $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$. Suppose we are interested in the number of heads, not where they occur in the experiment. We can then define $X$ to be a function defined on $S$ such that $X(e_i) = x$, where $e_i \in S$ and $x$ equals either 0, 1, 2, or 3, depending on how many “H’s” appear in $e_i$. Thus, $X$ has domain $S$ and range $\{0, 1, 2, 3\}$. (Recall: What is a domain? What is a range?)

Now consider another example where we take a multiple choice exam with possible answers a, b, c so $S = \{a, b, c\}$. Now define a function $Y$ such that $Y(a) = 0$, $Y(b) = 5$, $Y(c) = 8$. Then $Y$ is a real valued function that has the elements of $S$ as its domain and the set $\{0, 5, 8\}$ as its range.

Notice that in both of the above examples, the outcomes of the experiments are not numbers. However, we often want to represent these outcomes as numbers for one reason or another. As above all we’ve done is taken non-
numeric outcomes and assigned to them a real number.

**Def:** For an experiment with sample space $S$, a function $X$ that assigns to each element $s$ in $S$ one and only one real number $X(s) = x$ is called a random variable. The space of $X$ is the set of real numbers $\{x \mid X(s) = x, s \in S\}$.

Simply stated - a random variable is a function that takes every element of a sample space and assigns to it a real number.

**Def:** Let $X$ denote a random variable with space $S$ and $S$ contains either a finite number of points or the points of $S$ can be put into a one to one correspondence with the positive integers. (That is, you can literally count the elements of $S$ even if there are infinitely many of them, you can still list them). Then the set $S$ is called a discrete space and the random variable $X$ is called a discrete random variable. Examples of discrete random variables include the number of students in a class, the number of people at a concert, the number of CD’s that a person owns, the number of pizzas a pizza place delivers on a Saturday night.

For a discrete random variable, the probability of $X$ taking on a specific value $x$, $p(X = x)$, is denoted as $f(x)$. This function $f(x)$ is called the probability mass function or pmf. Note: Though we are going to discuss the probability that a random variable takes on specific values, how probabilities are calculated is no different than it was before.

### 6.1.1 Properties of probability mass functions

1. Since $f(x) = p(X = x)$ for $x \in S$ $f(x) > 0$.
2. $\sum f(x) = 1$ for $x \in S$
3. If $A \subset S$, then $p(X \in A) = \sum_{x \in A} f(x)$

Note what the first item above is saying. Since we are simply redefining a probability as a function, namely $f(x)$, any output from this function must be positive for any element in the sample space since we previously defined all probabilities to be positive. (or nonnegative) Why are excluding the probability equaling 0? Recall how we are defining the space a random variable is defined on. We are saying that for elements inside the sample
space, the probability of any of those outcomes occurring is positive. Inversely, for elements not in the sample space, the probability occurring is zero since these outcomes cannot occur. Be careful!!! These are probabilities so they are still NEVER negative!

6.1.2 Uniform Random Variables

**Def:** When a probability mass function is constant over a space, we say that the random variable associated with the space has a uniform distribution.

For example, consider picking a random digit. The sample space under consideration is $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Define the random variable $X(s) = s$. That is, our function takes an element from the space $S$ and assigns it a real number, namely itself. (This is the identity function). Each of the elements inside of $S$ has the same probability of occurring, so we say that $p(X = x) = f(x) = 1/10$, where $x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$. This is often written as follows:

$$p(X = x) = f(x) = \begin{cases} 
\frac{1}{10} & x = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 \\
0 & \text{otherwise}
\end{cases}$$

The above example is that of a uniform random variable since the probability mass function is constant over the space.

6.1.3 Other examples...

Consider an experiment of rolling 2 dice. Define the random variable $X$ to be the sum of the two dice.

- **What values can $X$ take on?**

  When we roll 2 dice we can see a sum of 2,3,4,5,6,7,8,9,10,11, or 12.

- **What are the probabilities of each of the above possible sums?**

- **Write the pmf of $X$.**

  Using the appropriate notation we can succinctly write the pmf as follows:
\[
p(X = x) = f(x) = \begin{cases} 
\frac{1}{36} & x = 2, 12 \\
\frac{2}{36} & x = 3, 11 \\
\frac{3}{36} & x = 4, 10 \\
\frac{4}{36} & x = 5, 9 \\
\frac{5}{36} & x = 6, 8 \\
\frac{6}{36} & x = 7 \\
0 & \text{otherwise}
\end{cases}
\]

Note: the above pmf is a valid pmf since for all values \( x \), \( f(x) > 0 \) and in addition, if we sum over the probabilities of the points in the sample space, \( \sum_{x \in S} f(x) = 1 \).

Also note: a pmf yields the probability that our random variable \( X \) takes on any number – this is why we carefully include the “otherwise” category.

### 6.1.4 Histograms

Once we create a pmf, we can depict it graphically. One way of doing this is by drawing a probability histogram. This is a graphical representation of the pmf that has a rectangle of height \( f(x) \) and base length of 1 centered at each \( x \), for all \( x \in S \).

The following is the histogram for the sum of two dice example:
6.2 Mathematical Expectation

6.2.1 Examples

As an experiment, consider what happens if you flip a coin and if it shows H, you win $2, but if it shows T, you lose $1. What is your expected winnings? If you play the game once, you expect to win $2 for every $1 that you lose. However, each of these monetary value is associated with a specific probability, namely $\frac{1}{2}$. That is, your expected winnings becomes:

$$2 \left( \frac{1}{2} \right) - 1 \left( \frac{1}{2} \right) = \frac{1}{2}$$

Now consider a game where there are 6 containers in front of you. 3 contain salt, 2 contain pepper and 1 contains sugar. Your friend randomly picks one. If he picks salt, you owe him $1, if he picks pepper, you owe him $3, but if he picks sugar, he gives you $4. How much should your friend expect to win?

First consider that the probability that your friend chooses salt if $\frac{1}{2}$ (Why?), the probability of choosing pepper is $\frac{1}{3}$ (Why?) and the probability of choosing sugar is $\frac{1}{6}$ (Why?). Choosing any of these also has an expected payout for your friend, namely $1, $3, and -$4, respectively. Thus for any one round of the game, your friend can expect to win $1 half the time, $3 one third of the time, and -$4 one sixth of the time, or

$$1 \left( \frac{1}{2} \right) + 3 \left( \frac{1}{3} \right) - 4 \left( \frac{1}{6} \right) = \frac{5}{6} \approx 0.83.$$ 

Notice that in both of these examples the expected winnings is not one of the actual payments of the experiment, but is instead more of a “weighted average”. This is what we refer to as mathematical expectation.

We can now define a random variable $X$ for example 2, where $X$ takes on the potential payments from this game. Thus we could define the pmf as follows:
\begin{align*}
p(X = x) &= f(x) = \begin{cases} \frac{1}{2} & x = 1 \text{ (salt)} \\ \frac{1}{3} & x = 3 \text{ (pepper)} \\ \frac{1}{6} & x = -4 \text{ (sugar)} \\ 0 & \text{otherwise} \end{cases} \\
(\text{As always, verify that the above is a valid pmf!})
\end{align*}

\section*{6.2.2 Expected Value}

**Def:** If \( f(x) \) is the pmf of a random variable \( X \) with sample space \( S \) and if

\[
E[X] = \sum_{x \in S} xf(x) \quad (6.1)
\]

exists, then this sum is called the mathematical expectation or long term average or expected value of the random variable and is denoted as \( E[X] \).

\( E[X] \) is just a weighted mean of the values that \( X \) can assume. It is the theoretical value of what would happen if the experiment were performed over and over again. The expected value can be positive, negative, or zero and **DOES NOT** have to be one of the points of the sample space!

An alternative (yet quite useful!) way to calculate \( (6.1) \) is as follows:

\[
E[u(X)] = \sum_{x \in S} u(x)f(x) \quad (6.2)
\]

Note: \( (6.2) \) is simply saying that if you are interested in the expected value of a function of your random variable, you still essentially use \( (6.1) \), but you must apply the function in question to each point \( x \) in your sample space. Be careful: **DO NOT** alter the probabilities that \( x \) occurs!!! Consider the following example:

**Examples**

Roll a standard die. Let \( X \) be the face of the die that shows. Find the expected value.

Using formula \( (6.1) \):

\[
E[X] = 1 \left( \frac{1}{6} \right) + 2 \left( \frac{1}{6} \right) + 3 \left( \frac{1}{6} \right) + 4 \left( \frac{1}{6} \right) + 5 \left( \frac{1}{6} \right) + 6 \left( \frac{1}{6} \right) = 3.5
\]
Again, notice that the expected value is not one of the points in the sample space. However, the idea is that if you were to roll a single die over and over again, recording the number showing every time, and then you were to average the number, it would theoretically be (or be very close to) the number 3.5.

Now find $E[X^2]$.

We are now considering a function of our random variable, so we will use (6.2). That is, in the right-hand side of the formula we will substitute the values of $x^2$. Thus:

$$E[X^2] = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}$$

Notice: the probabilities did not change, just the values of $x$ since we substituted $x^2$ values.

Now consider rolling a die where the faces are 2, 4, 6, 8, 10, and 12. (Each still with equal probabilities). Find the expected value of this die.

Using formula (6.1) and the fact that each side still has equal probability, namely $\frac{1}{6}$:

$$E[X] = 2\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) + 8\left(\frac{1}{6}\right) + 10\left(\frac{1}{6}\right) + 12\left(\frac{1}{6}\right) = \frac{7}{1}$$

6.2.3 Properties of Expected Value

Theorem: When it exists, the expected value $E$ has the following properties:

1. If $c$ is a constant, then

$$E[c] = c.$$
2. If $c$ is a constant and $u$ is a function, then
\[
E[cu(X)] = cE[u(X)].
\]

3. If $c_1$ and $c_2$ are constants and $u_1$ and $u_2$ are functions, then
\[
E[c_1u_1(X) + c_2u_2(X)] = c_1E[u_1(X)] + c_2E[u_2(X)].
\]

### 6.2.4 Mean, Variance, and Standard Deviation

Using the properties of expected value we can define the following mathematical concepts:

Def: For a pmf, define the **mean** or average of a random variable $X$ by
\[
\mu = E[X].
\]

Def: For a random variable $X$, we define the **variance** as the average spread of the values of $X$ as
\[
\sigma^2 = \text{Var}[X] = E[X^2] - \mu^2 = \sum_{x \in S} x^2 f(x) - \mu^2
\] (6.3)

We further define the **standard deviation** of a random variable by
\[
\sigma = \sqrt{\sigma^2} = \sqrt{\text{Var}[X]}
\] (6.4)

The variance (and by extension the standard deviation) give us a sense of how spread out the values of $X$ are. The more spread out a set of data is, the large the values of $\sigma$ and $\sigma^2$ will be.

Example: Define the pmf of a random variable $X$ as follows:
\[
p(X = x) = f(x) = \begin{cases} 
\frac{1-x}{6} & x = -2, -1, 0, 1 \\
0 & \text{otherwise}
\end{cases}
\]

Verify that this is a valid pmf and also verify the following:

1. $p(X < 0) = \frac{5}{6}$
2. $p(X \leq 0) = 1$
3. $p(X \geq -1) = \frac{1}{2}$
4. $\mu = E[X] = -\frac{4}{3}$

5. $\sigma^2 = \text{Var}[X] = \frac{7}{3}$

6. $\sigma = \text{SD}[X] = \sqrt{\frac{7}{3}}$

7. $E[(X + 4)^2] = \frac{23}{3}$

6.3 Bernoulli Trials & the Binomial Distribution

6.3.1 Bernoulli

Def: A Bernoulli experiment is an experiment in which the outcome can be classified in one of two mutually exclusive ways.

ex: success/failure

life/death

true/false

male/female

etc...

We say that a sequence of Bernoulli trials occurs when a Bernoulli experiment is performed several independent times where the probability of success, denoted $p$, remains constant from trial to trial.

Ex:

- The probability a person has a (noncontagious) disease is $\frac{1}{100}$.
- The percentage chance a student has for passing a standardized test is 80%
- You flip a weighted coin and the probability of seeing T is $\frac{1}{5}$

In each of these above examples, there are two outcomes.

- A person has the disease or doesn’t.
• A student passes or doesn’t.

• You see a T or H.

In each of these above examples, there is a constant probability of success from trial to trial and the trials are independent.

• \( p = \frac{1}{100} \), where “success” means a person has the disease. One person having it doesn’t mean the next person does.

• \( p = .8 \), where “success” means a student passes. One student passing doesn’t affect the next one.

• \( p = \frac{1}{5} \), where “success” means flipping T. A T on this turn doesn’t affect the chance of it occurring on the next one.

We can define the random variable \( X \) associated with a Bernoulli trial as follows:

\[
X(\text{success}) = 1 \\
X(\text{failure}) = 0
\]

and thus the pmf is:

\[
f(x) = p(X = x) = p^x (1 - p)^{1-x}, \ x = 0, 1
\]  

(6.5)

To apply (6.5) consider example three above. The probability that the random variable \( X = 1 \), that is, the probability that you see a T is:

\[
f(1) = p(X = 1) = \frac{1}{5} \left( 1 - \frac{1}{5} \right)^{1-1}
\]

The Mean and Variance of a Bernoulli

Recall: We’ve already talked about the meaning of the expected value of a random variable as well as the variance of a random variable using formulas (6.1) and (6.3), respectively.
The nice thing about identifying a random variable with a known distribution (such as a Bernoulli distribution) is that the formulas for the expected value and variance are much simpler than the formulas we used before. Let’s derive the mean and variance for a Bernoulli as follows:

\[
\mu = \mathbb{E}[X] = \sum_{x \in S} x f(x) = 0 \cdot p^0 (1-p)^{1-0} + 1 \cdot p^1 (1-p)^{1-1} = p
\]

and

\[
\sigma^2 = \text{Var}[X] = \mathbb{E}[X^2] - \mu^2 = \sum_{x \in S} x^2 f(x) - \mu^2 = 0^2 \cdot p^0 (1-p)^{1-0} + 1^2 \cdot p^1 (1-p)^{1-1} - p^2 = p - p^2 = p(1-p)
\]

This derivation of variance also leads us to the formula for the standard deviation for a Bernoulli random variable, namely:

\[
\text{SD}[X] = \sqrt{p(1-p)} \quad (6.6)
\]

### 6.3.2 Binomials

To develop the binomial distribution we will first consider the following experiment which is a Bernoulli trial. (Why is it Bernoulli?)

Ex: Between your house and ASU there are 7 stoplights. Suppose that the probability of any one stoplight being green is 0.6 and each stoplight is independent. What is the probability that on your morning commute to school you hit exactly 5 green lights?
First note what we are asking: We are treating each stoplight as a Bernoulli trial (is the light green or isn’t it?) with constant probability of success 0.6 and by assumption each light is independent of the next. (That is, the probability of the light at McClintock and Apache being green doesn’t change the probability that the light at McClintock and University is also green.) We treat each light as its own Bernoulli trial but we are interested in the total number of successes (in this case, 5) not when they appear. (Is it the first five? 1,3,5,6, and 7? The last five? We don’t care, as long as exactly 5 are green when we get there!)

For each light that you get to along the way you are facing a Bernoulli trial: will the light be green (with constant probability of success 0.6) or won’t it?

Define the random variable X to be the number of “successes” (in this case “a green light”) in n (in our case, 7) Bernoulli trials. The possible values for X are 0, 1, 2, . . . , n (in our case, 0,1,2,3,4,5,6,7). The number of successes and failures must add up to n since that’s how many trials there are, so if there are x successes, there must be n − x “failures” (not green lights). The number of ways to choose the x positions for the x successes is a combination (again, which give of the seven lights will be green?), namely the combination \( \binom{n}{r} \) or \( \left( \frac{n}{r} \right) \).

Since the trials (by assumption) are independent and the probability of success and failure remain constant from trial to trial, the pmf for this new random variable X can be derived from (6.5). This pmf is:

\[
f(x) = p(X = x) = \binom{n}{r} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n \quad (6.7)
\]

So, what is the probability that on your way to school you hit exactly 5 green lights?

\[
f(5) = p(X = 5) = \binom{7}{5} 0.6^5(1 - 0.6)^{7-5} = 0.26127
\]

In (6.7) we have derived the binomial distribution. A random variable with a binomial distribution has the following properties:

• A Bernoulli experiment (success/failure) is performed n times.
• Each trial is independent.

• The probability of success \( p \) and probability of failure \( 1 - p \) are constant from trial to trial.

• A binomially is often denoted by \( b(n, p) \) where \( n \) is the number of trials and \( p \) is the constant probability of success.

• We could derive formulas for the mean and variance, similar as in (6.6) and (6.6). Instead of formally deriving these, we will just state these here:

The mean and variance for a Binomial

The mean and variance for a binomially distributed random variable \( X \) distributed as \( b(n, p) \) is:

\[
E[X] = np
\]

and

\[
\text{Var}[X] = np(1 - p)
\]

Caution: These above formulas can only be used to mean the expected value and variance if you know that the random variable is a binomial random variable. If this is not true, you must use the formulas given in (6.1) and (6.3).

As always, the standard deviation is found by taking the square root of the variance.

Back to our stoplight example...

1. What is the probability that at least 6 lights are green?

This is \( p(X \geq 6) = p(X = 6) + p(X = 7) = f(6) + f(7) = 0.1586304 \) (verify this!)
2. What is the probability that at most 5 lights are green?

This is

\[
p(X \leq 5) = p(X = 0) + p(X = 1) + \ldots + p(X = 4) + p(X = 5) \\
= 1 - p(X \geq 6) \\
= 1 - 0.1586304 \\
= 0.8413696
\]

(verify this!)

Notice how we have progressed from defining probabilities like \( p(X = x) \)

to defining accumulated probabilities such as \( p(X \leq x) \).

This expression

\[
F(X) = p(X \leq x)
\]

(instead of \( f(x) \)) is known as the cumulative distribution function (or cdf) of \( X \). We will address this idea later in more detail.

### 6.3.3 Examples...

Suppose a salesman has a 45% chance of making a sale to each customer that comes through his door. In one day 25 people come into his shop. Let \( X \) denote the number of sales made that day.

1. What is the probability that he makes exactly 23 sales?

\[
p(X = 23) = 0.0000009588
\]

(verify this!)

2. What is the probability that he makes at least 23 sales?

\[
p(X \geq 23) = 0.000001026
\]

(verify this!)

3. What is the probability that he makes at most 22 sales?

\[
p(X \leq 22) = 0.999998974
\]

(verify this!)
4. What is the probability that he makes between 18 and 21 sales inclusive?

\[ p(18 \leq X \leq 21) = 0.005823 \text{ (verify this!)} \]

5. What is the expected number of sales that he makes?

\[ E[X] = 11.25 \text{ (verify this!)} \]

6. What is the variance of the number of sales?

\[ \text{Var}[X] = 6.1875 \text{ (verify this!)} \]

7. What is the standard deviation of the number of sales made?

\[ \text{SD}[X] = 2.4875 \text{ (verify this!)} \]

6.4 Integration Review

6.4.1 Basic Integrals & General Rules

The follow are just some basic rules with which you should already be familiar. For integration practice problems, refer to our first homework assignment!

\[
\int ax^n dx = \frac{a}{n+1}x^{n+1} + C
\]

\[
\int e^x dx = e^x + C
\]

\[
\int e^{ax} dx = \frac{1}{a}e^{ax} + C
\]

\[
\int \frac{1}{x} dx = \ln x + C
\]

\[
\int af(x)dx = a \int f(x)dx
\]
\[ \int [f(x) + g(x)] \, dx = \int f(x) \, dx + \int g(x) \, dx \]

\[ \int_a^b f(x) \, dx = - \int_b^a f(x) \, dx \]

\[ \int_a^a f(x) \, dx = 0 \]

\[ \int_a^b f(x) \, dx = \int_a^c f(x) \, dx + \int_c^b f(x) \, dx \]

\[ \int f(x)g'(x) \, dx = f(x)g(x) - \int f'(x)g(x) \, dx \]

\[ \int_a^\infty f(x) \, dx = \lim_{b \to \infty} \int_a^b f(x) \, dx \]

\[ \int_{-\infty}^b f(x) \, dx = \lim_{a \to -\infty} \int_a^b f(x) \, dx \]

**6.4.2 Indefinite Integrals, Definite Integrals and Improper Integrals**

**Indefinite Integrals**

When we say *indefinite* integral we refer to integrals without limits. Specifically, if we use the following notation:

\[ F'(x) = f(x) \implies \int f(x) \, dx = F(x) + C \]

then we say that \( f(x) \) has an *indefinite* integral since its antiderivative \( F(x) + C \) is not just one function but rather an entire *class* of functions, all having the same derivative, namely \( f(x) \). In other words, there are many functions that have the same derivative. Consider the following:

\[
\begin{align*}
F_1(x) &= x^2 + 7 \\
F_2(x) &= x^2 - \pi \\
F_3(x) &= x^2
\end{align*}
\]

What is the (common) derivative \( f(x) \) of each of these function? Key word: *common*. 
Definite Integrals

When we talk about a definite integral, we are probably referring to the area under the curve. That is, we’re asking what is the area between two regions on our coordinate plane.

The rules for integration of a definite integral are exactly the same as the rules for integration of an indefinite integral. The only difference is the addition of the constant $C$ at the end of our integration process. Instead of adding an arbitrary (possibly zero) constant at the end, we have the limits of integration when dealing with a definite integral. That is, if we’re integrating a function as follows:

$$\int_{a}^{b} f(x)\,dx$$

we first integrate to find the function $F(x)$ and then evaluate between our limit points:

$$\int_{a}^{b} f(x)\,dx = F(b)|_{a} = F(b) - F(a)$$

This above concept is very important as we will soon be connecting this idea with the aforementioned cumulative distribution function as in (6.8).

Improper Integrals

Consider the following integral:

$$\int_{0}^{\infty} xe^{-cx^2}\,dx$$

Using methods of substitution we can easily find the integral of this function.

Letting $u = -cx^2$ and $du = -2cx\,dx$, we arrive at the following:

$$\int_{0}^{\infty} xe^{-cx^2}\,dx = \frac{1}{-2c} \int_{x=0}^{\infty} e^{u}\,du$$

$$= \frac{1}{-2c} e^{-cx^2} \bigg|_{0}^{\infty}$$

Consider this last line, however. Normally as in (6.10) we normally just plug in the value of the upper limit and subtract the function evaluated at
the lower limit. However, what does it mean to “plug in” infinity? Actually, it doesn’t really mean a whole lot. Instead we consider taking a limit.

Integration over an infinite region If the limit of the above integral exists (that is, if the “area under the curve” is a finite number) then we can continue the above integration process as follows.

Instead of (13), consider:

\[ \int_{0}^{\infty} x e^{-cx^2} \, dx = \lim_{b \to \infty} \int_{0}^{b} x e^{-cx^2} \, dx \]

\[ = \lim_{b \to \infty} \frac{1}{-2c} \int_{x=0}^{b} e^{x} \, du \]

\[ = \lim_{b \to \infty} \frac{1}{-2c} e^{-cx^2} \bigg|_{0}^{b} \]

\[ = \lim_{b \to \infty} \left[ \frac{1}{-2c} \left( e^{-c(b)^2} - e^{-c(0)^2} \right) \right] \]

\[ = \lim_{b \to \infty} \left[ \frac{1}{-2c} \left( \frac{1}{e^{b^2}} - 1 \right) \right] \]

\[ = \lim_{b \to \infty} \left[ \frac{1}{-2c} (0 - 1) \right] \]

\[ = \frac{1}{2c} \]

Notice what we have done: we have integrated over an infinite region and gotten a finite area! In this case, our improper integral (6.11) is said to converge.

This idea of being able to integrate over an infinite region is very important in manipulating probabilities, and more specifically in dealing with continuous random variables.

6.5 Continuous Random Variables

Def: A continuous random variable can assume any value within a given range.

Consider this example: Suppose as an experiment you throw a dart at a number line such that it always lands on the interval [1,3]:
Notice that since this is the real number line, there are an infinite of numbers on which the dart can land (if we hand the means to measure something exactly, without rounding at any point.) That is, the cardinality of the sample space (where the sample space is the numbers on which the dart can land) is: \( c(S) = \infty \). Using this, consider:

What is the probability that the dart lands exactly on 1.5? Using what we know about probabilities:

\[
p(X = 1.5) = \frac{c(X = 2)}{c(S)} = \frac{1}{\infty} \to 0
\]

Since this value is infinitely small, we treat it as zero. In other words, the probability that the dart lands exactly on 1.5 is zero. This may be intuitive in that sense that since there are an infinite number of places that the dart could land, the chance of landing on a very specific number is near impossible. Hence, this probability is zero.

With continuous random variables, the probability of any particular value is zero. (How is this different than with discrete random variables?)

### 6.5.1 Properties of Continuous Random Variables

**Def** A probability density function (or pdf) of a random variable \( X \) over space \( S \) is an integrable function \( f(x) \) such that:

- \( f(x) \geq 0, \forall x \in S \)
- \( \int_S f(x)dx = 1 \)
- \( p(a \leq X \leq b) = \int_a^b f(x)dx \)
Let’s comment on the main differences between the above and what we learned previously about discrete random variables in (6.1.1).

Since we have an infinite number of possible outcomes, we can no longer simply sum up the probabilities of events occurring. Instead, we integrate over the region. Remember, that’s what an integral is! It’s the area between two points. Thus, we when we discuss continuous r.v.’s, we discuss the probability that a random observation falls between two points.

In addition, as before, the sum of our probabilities must equal 1. That is why when we integrate over our entire space, it must integrate to 1.

**Example of a Continuous R.V.**

Let the random variable \( X \) be the height in inches of 2-year olds. Suppose a reasonable model for \( X \) is the pdf:

\[
f(x) = \frac{1}{10} e^{-\frac{x}{10}}, \quad 0 \leq x \leq \infty
\]

First notice that for all \( x \) in our region of interest, \( f(x) \geq 0 \). In addition,

\[
\int_{-\infty}^{\infty} f(x) \, dx = \int_{-\infty}^{0} f(x) \, dx + \int_{0}^{\infty} f(x) \, dx
\]

\[
= \int_{-\infty}^{0} 0 \, dx + \int_{0}^{\infty} \frac{1}{10} e^{-\frac{x}{10}} \, dx
\]

\[
= \lim_{b \to \infty} \int_{0}^{b} \frac{1}{10} e^{-\frac{x}{10}} \, dx
\]

\[
= \lim_{b \to \infty} -e^{-\frac{x}{10}} \bigg|_{0}^{b}
\]

\[
= 0 - (-1)
\]

\[
= 1
\]

That is, when we integrate over our region, the integral equals 1.

Suppose further that we are interested in the probability that a randomly chosen 2 year old is at least 20” tall. That is, we are asking \( p(X \geq 20) \). Using the fact that we now integrate to find probabilities:
\[ p(X \geq 20) = \int_{20}^{\infty} \frac{1}{10} e^{-\frac{x}{10}} dx \]
\[ = \ldots \]
\[ = 0.135335 \]

The actual graph of this function is:

Think about what this picture is telling us: As \( x \) gets bigger, the probability that a randomly chosen child is taller than \( x \) goes down. This makes sense, right? Children are not infinitely tall.

**Other examples:**

Let \( X \) have the pdf:
\[ f(x) = \frac{1}{9} x^2, \quad x \in [0, 3] \]

a) Verify that \( f(x) \) is a pdf (Hint: Integrate to 1)

b) What is the probability that \( (X) \) is between 0 and 1?

Find \( c \) such that \( f(x) \) is a pdf on the interval between 0 and 2, \( f(x) = c(x + x^2) \) (Try this out! \( c \) should equal \( \frac{3}{14} \)).
Significance Levels & Critical Values

It is often of interest to consider what is the probability that a random variable is greater than or equal to some observed value. For example, reconsider the heights of 2 year olds – say that I randomly choose one of these children and he is 30” tall. (yes, this is a tall 2 year old.) I am interested in what is the probability that a peer is taller than he is? This probability is known as the significance level.

The probability that my random variable (heights of 2 year olds) is greater than an observed value (say, 30”) is calculated as:

\[ p(X \geq 30) = \int_{30}^{\infty} \frac{1}{10} e^{-\frac{x}{10}} dx = 0.049787 \]

This value of 0.049787 is the significance value of this experiment.

Now instead suppose that I have a probability in mind. That is, I want to take the tallest 10% of my sample of 2 year olds and further study them. I need to find the cut off value that separates the shortest 90% from the tallest 10%. This is called the critical value that corresponds to a probability of interest that we usually denote as \( \alpha \) (alpha).

Suppose I want to study the tallest 10% of 2 year olds. What is the corresponding critical value? We are now looking at:

\[ p(X \geq c) = \int_{c}^{\infty} \frac{1}{10} e^{-\frac{x}{10}} dx = 0.10 \]

Using the same techniques as before we simply carry out the integration on the left hand of the equation and set it equal to 0.10. Try this! You should get a critical value of \( c \approx 23.02 \)

The Cumulative Distribution Function

CDF’s & Discrete RV’s  In section (6.3.2) we first defined the cumulative distribution function as in (6.8). Again, the cdf is:

\[ F(X) = \sum_{x \in X} p(X \leq x) \]  

This is the probability that our random variable \( X \) takes on a value \( x \) plus the probability that \( X \) takes on the value of everything less than \( x \).
Example: Let $X$ have the pmf:

$$f(x) = \begin{cases} 
\frac{1}{8} & x = 2 \\
\frac{1}{3} & x = 3 \\
\frac{2}{8} & x = 4 \\
\frac{3}{8} & x = 5 \\
\end{cases}$$

We can now find the cdf by breaking our region up into pieces and considering the accumulated probabilities over each of those regions.

Thus, the cdf is:

$$F(x) = \begin{cases} 
0 & x < 2 \\
\frac{1}{8} & 2 \leq x < 3 \\
\frac{2}{8} & 3 \leq x < 4 \\
\frac{3}{8} & 4 \leq x < 5 \\
1 & 5 \leq x \\
\end{cases}$$

with corresponding graph:

Notice than when dealing with cdf’s of discrete random variables, the graph of the cdf is a step function. This is because probability is only accumulated at the points in the sample space. Everywhere else, the probability added to the previous probabilities is zero.

**CDF’s & Continuous RV’s**

**Def:** For a continuous random variable $X$ the cumulative distribution function $F$ is given by:

$$F(X) = p(X \leq x) = \int_{-\infty}^{x} f(t)dt \quad (6.15)$$
The interpretation of the cumulative distribution function for a continuous random variable is the same – \( F(x) \) is the probability that \( X \) is less than or equal to \( x \).

Consider \( X \) with pdf

\[
f(x) = \frac{2}{3}x, \quad -1 \leq x \leq 2
\]

Find the cdf of \( X \):

Using (6.15),

\[
F(x) = \int_{-\infty}^{x} f(t)dt
\]

\[
= \int_{-\infty}^{-1} f(t)dt + \int_{-1}^{x} f(t)dt
\]

\[
= \int_{-\infty}^{-1} 0dt + \int_{-1}^{x} \frac{2}{3}tdt
\]

\[
= \frac{1}{3}t^2 \bigg|_{-1}^{x}
\]

\[
= \frac{1}{3} \left[ x^2 - (-1)^2 \right]
\]

\[
= \frac{1}{3} \left[ x^2 - 1 \right]
\]

Hence, the cdf \( F(x) \) can be written:

\[
F(x) = \begin{cases} 
0 & x < -1 \\
\frac{1}{3} \left[ x^2 - 1 \right] & -1 \leq x < 2 \\
1 & 2 \leq x
\end{cases}
\]

Why is this useful? Suppose you wanted to know the probability associated with a critical value (6.5.1), thus you’re asking \( p(X \leq c) \). Well, if you have the cdf, instead of having to integrate, simply just evaluate \( F(c) \).

For example, what is \( p(X \leq 1.5) \)? By definition, \( F(1.5) = p(X \leq 1.5) \), which by our cdf listed above, is \( \frac{1}{3} \left[ x^2 - 1 \right] = \frac{1}{3} \left[ 1.5^2 - 1 \right] \approx .4167 \)
The mean and variance of a continuous random variable

Def: The mean of a continuous random variable $X$ is given by:

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

(6.16)

Def: The variance of a continuous random variable $X$ is given by:

$$
\sigma^2 = \text{Var}[X] = \mathbb{E}[X^2] - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2
$$

(6.17)

Notice that these formulas are quite similar to the formulas for the mean and variance of a discrete random variable as given in (6.1) and (6.3). The primary difference is that when dealing with continuous random variables, we integrate over the region of interest instead of sum.

As always, the standard deviation is the square root of the variance:

$$
\sigma = \sqrt{\sigma^2} = \sqrt{\text{Var}[X]}
$$

(6.18)

6.6 The Normal Distribution

6.6.1 Standard Normals

Def: A continuous random variable $X$ has a standard normal distribution if its pdf is given by:

$$
f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}
$$

(6.19)

The pdf has a standard normal curve, often called a bell curve and looks like:
Notice that this curve is symmetric about zero, and thus the mean is zero. In other words, \( \mu = \mathbb{E}[X] = 0 \), where \( X \) is a standard normal random variable. In addition, \( \sigma^2 = \text{Var}[X] = 1 \rightarrow \sigma = \text{SD}[X] = 1 \).

Finding probabilities associated with the normal distribution is very computationally difficult as we are dealing with the pdf given in (6.19) which cannot be integrated easily. Instead, the areas of interest are approximated using numerical techniques.

A random variable that is known to be distributed with a mean of 0 and a standard deviation of 1 is called a standard normal and is denoted as \( Z \sim N(0, 1) \). This notation means that \( Z \) is a standard normal with mean 0 and standard deviation 1.

To find probabilities associated with a \( Z \), we refer to what are called z-tables as in this.

To use a standard normal table, consider the following examples:

Ex: Assume \( Z \) is a normally distributed r.v. with \( \mu = 0 \) and \( \sigma = 1 \). Find

1. \( p(Z \leq 1.27) \)

First locate the row that has 1.2 in it. Then locate the column that has 0.07 in it. Where this row and column meet is the probability that the random variable is less than 1.27. Namely, \( p(Z \leq 1.27) = 0.8980 \)

The probabilities in the z-table give the cumulative probabilities – that is, the probability that the random variable is less than or equal to the value of interest. Using part (a), try the following:

2. \( p(Z \leq -2.91) = 0.0018 \)

3. \( p(Z > 0.21) \)

Note that we can actually re-write this using the complement. We need to re-write this using the complement since, again, the z-table gives cumulative probabilities. Also notice that the distribution is symmetric. Hence:
\[ p(Z > 0.21) = 1 - p(Z \leq 0.21) = 1 - 0.5793 = 0.4207 \text{ AND } p(Z > 0.21) = p(Z \leq -0.21) = 0.4207 \]

4. \( p(-1.77 \leq Z \leq 2.68) \)

Consider this picture to investigate the probability in question:

Since the question asks for the probability between -1.77 and 2.68, we are interested in the area that is red only. Thus, we find the \( p(Z \leq 2.68) \) and find the \( p(Z \leq -1.77) \) then find the difference between these two areas. Hence,

\[ p(-1.77 \leq Z \leq 2.68) = p(Z \leq 2.68) - p(Z \leq -1.77) = 0.9963 - 0.0384 = 0.9579. \]

6.6.2 Normals with mean \( \mu \) and standard deviation \( \sigma \)

A normally distributed variable does not have to have a mean of 0 and standard deviation of 1. Instead, consider a continuous random variable \( X \) with mean \( \mu \) and standard deviation \( \sigma \). This can be transformed back into a standard normal by subtracting the mean and dividing by the standard deviation.
That is, if \( X \sim N(\mu, \sigma) \), (if \( X \) is distributed normally with mean \( \mu \) and std deviation \( \sigma \)), then

\[
Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \tag{6.20}
\]

(6.20) is referring to as standardizing a normal random variable. \( Z \) is known as a z-score and it tells us how many standard deviations away a specific \( x \) value lies from the mean. We standardize data in order to easily compare data that are on different scales.

When we do the standardization process, we can now use the z-table to compute probabilities.

ex: Let \( X \) be normally distributed with \( \mu = 12 \) and \( \sigma = 3 \).

1.

\[
p(X \leq 9) = p(X - 12 \leq 9 - 12) = p\left(\frac{X - 12}{3} \leq \frac{9 - 12}{3}\right) = p(Z \leq -1) = 0.1587
\]

2.

\[
p(X > 7) = p(X - 12 > 7 - 12) = \left(\frac{X - 12}{3} > \frac{7 - 12}{3}\right) = p(Z > -\frac{5}{3}) = 1 - p(Z \leq -\frac{5}{3}) = 1 - 0.0475 = 0.9525
\]