Probability & Statistics

C. Caruvana, Ph.D.

Anno accademico MMXVIII
Contents

I Probability 1

1 Counting 3
  1.1 Binomial Coefficients .................................................. 7
  1.2 Poker Hands ............................................................... 11

2 Probability 13
  2.1 Sample Spaces and Events ............................................. 14
  2.2 The Probability of Events ............................................ 15
  2.3 Conditional Probability ................................................ 19
  2.4 Independent Events ..................................................... 22
  2.5 Bayes’ Theorem ............................................................ 24

3 Probability Distributions and Densities 29
  3.1 Random Variables ......................................................... 29
    3.1.1 Discrete Random Variables ..................................... 29
    3.1.2 Continuous Random Variables .................................. 31
  3.2 Probability Distributions .............................................. 31
  3.3 Probability Density Functions ....................................... 35
  3.4 Multivariate Distributions and Densities ....................... 38
    3.4.1 The Discrete Case ................................................ 38
    3.4.2 The Continuous Case ............................................. 41
    3.4.3 Beyond the Bivariate ............................................ 44
  3.5 Marginal Distributions and Densities ............................ 45
  3.6 Conditional Distributions and Densities ......................... 47

4 Expected Value and Moments 51
  4.1 Expected Value .......................................................... 51
  4.2 Moments, Variance, and Chebyshev’s Inequality ............... 57
  4.3 The Median ............................................................. 61
  4.4 Moment-generating Functions ...................................... 62
  4.5 Product Moments and Covariance .................................. 65
  4.6 The Weak Law of Large Numbers ................................. 71

5 Particular Probability Distributions and Densities 73
  5.1 Discrete Distributions ................................................ 73
    5.1.1 Uniform Distributions .......................................... 73
    5.1.2 Bernoulli Distributions ........................................ 73
    5.1.3 Binomial Distributions ......................................... 74
    5.1.4 Negative Binomial and Geometric Distributions .......... 75
    5.1.5 Hypergeometric Distributions ................................. 79
    5.1.6 Poisson Distributions ......................................... 80
    5.1.7 Multinomial Distributions ..................................... 83
    5.1.8 Multivariate Hypergeometric Distributions .............. 83
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>Continuous Densities</td>
<td>84</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Uniform Distributions</td>
<td>84</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Gamma, Exponential, and Chi-Square Distributions</td>
<td>85</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Beta Distributions</td>
<td>90</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Normal Distributions</td>
<td>95</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Bivariate Normal Distributions</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>Using the Standard Normal Distribution</td>
<td>105</td>
</tr>
<tr>
<td>6.1</td>
<td>Producing and Using $z$-tables</td>
<td>105</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Approximating Standard Normal Probability</td>
<td>105</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Approximating $z$ Given a Probability</td>
<td>106</td>
</tr>
<tr>
<td>6.2</td>
<td>Approximating the Binomial Distribution</td>
<td>107</td>
</tr>
<tr>
<td>6.3</td>
<td>Normal Score Plots</td>
<td>110</td>
</tr>
<tr>
<td>II</td>
<td>Statistics</td>
<td>113</td>
</tr>
<tr>
<td>7</td>
<td>Functions of Random Variables</td>
<td>115</td>
</tr>
<tr>
<td>7.1</td>
<td>Distribution Function Technique</td>
<td>115</td>
</tr>
<tr>
<td>7.2</td>
<td>Transformation Technique</td>
<td>116</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Single Variable</td>
<td>116</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Several Variables</td>
<td>119</td>
</tr>
<tr>
<td>7.3</td>
<td>Moment-generating Function Technique</td>
<td>121</td>
</tr>
<tr>
<td>8</td>
<td>Sampling</td>
<td>123</td>
</tr>
<tr>
<td>8.1</td>
<td>The Sample Mean</td>
<td>123</td>
</tr>
<tr>
<td>8.1.1</td>
<td>Finite Populations</td>
<td>128</td>
</tr>
<tr>
<td>8.2</td>
<td>A Note on Sample Variance</td>
<td>130</td>
</tr>
<tr>
<td>8.3</td>
<td>The Chi-Square Distribution</td>
<td>131</td>
</tr>
<tr>
<td>8.4</td>
<td>Student’s $t$-distribution</td>
<td>135</td>
</tr>
<tr>
<td>8.5</td>
<td>The Fisher-Snedecor Distribution</td>
<td>137</td>
</tr>
<tr>
<td>8.6</td>
<td>Order Statistics</td>
<td>138</td>
</tr>
<tr>
<td>8.6.1</td>
<td>The Sample Median</td>
<td>141</td>
</tr>
<tr>
<td>9</td>
<td>Point Estimation</td>
<td>143</td>
</tr>
<tr>
<td>9.1</td>
<td>Properties of Estimators</td>
<td>143</td>
</tr>
<tr>
<td>9.2</td>
<td>Producing Estimators</td>
<td>151</td>
</tr>
<tr>
<td>9.2.1</td>
<td>The Method of Moments</td>
<td>151</td>
</tr>
<tr>
<td>9.2.2</td>
<td>The Method of Maximum Likelihood</td>
<td>152</td>
</tr>
<tr>
<td>10</td>
<td>Interval Estimation</td>
<td>157</td>
</tr>
<tr>
<td>10.1</td>
<td>Population Means</td>
<td>157</td>
</tr>
<tr>
<td>10.1.1</td>
<td>With Known Variance</td>
<td>157</td>
</tr>
<tr>
<td>10.1.2</td>
<td>Using the Sample Variance</td>
<td>158</td>
</tr>
<tr>
<td>10.2</td>
<td>Differences between Means</td>
<td>159</td>
</tr>
<tr>
<td>10.2.1</td>
<td>With Known Variances</td>
<td>159</td>
</tr>
<tr>
<td>10.2.2</td>
<td>With Equal but Unknown Variances</td>
<td>159</td>
</tr>
<tr>
<td>10.3</td>
<td>Proportions</td>
<td>160</td>
</tr>
<tr>
<td>10.4</td>
<td>Differences between Proportions</td>
<td>161</td>
</tr>
<tr>
<td>10.5</td>
<td>Population Variances</td>
<td>161</td>
</tr>
<tr>
<td>10.6</td>
<td>Ratios of Variances</td>
<td>162</td>
</tr>
<tr>
<td>11</td>
<td>Hypothesis Testing</td>
<td>163</td>
</tr>
<tr>
<td>11.1</td>
<td>The Neyman-Pearson Lemma</td>
<td>164</td>
</tr>
<tr>
<td>11.2</td>
<td>Likelihood Ratio Tests</td>
<td>167</td>
</tr>
</tbody>
</table>
Part I

Probability
Chapter 1

Counting

“How do I love thee? Let me count the ways.”

date from Sonnet 43
by Elizabeth Barrett Browning

The Basic Principle of Counting. If there are \( n \) many ways to do something and \( m \) many ways to do something else, there are \( n \cdot m \) many ways of doing both. In general, if there are \( k \) many things and \( n_j, 1 \leq j \leq n \), many ways to do the \( j^{th} \) thing, the total number of ways to do all \( k \) things is

\[
n_1 \cdot n_2 \cdots n_k.
\]

Example 1. If I have 7 shirts and 5 pairs of pants in my closet, I have 35 different possible outfits in my closet.

Example 2. If I have a standard 6-sided die and a 20-sided die, there are 120 different outcomes to rolling both.

Example 3. Suppose you are taking a test consisting of 10 true/false questions. There are \( 2^{10} = 1024 \) ways of answering all of the questions on the test.

Recall the factorial operation: We can define it recursively with

- \( 0! = 1 \)
- For \( n \geq 0 \), \((n + 1)! = n \cdot n!\).

Alternatively,

\[
n! = 1 \cdot 2 \cdot 3 \cdots (n - 1) \cdot n = \prod_{j=1}^{n} j
\]

Definition 1. A permutation is a distinct arrangement of \( n \) different elements.

- There is only one permutation of a single object.
- There are only two permutations of the set of two objects. Let \( a \) and \( b \) be our distinct objects. Then the permutations are exactly \( ab \) and \( ba \).
- How many permutations are there of three objects \( a, b, \) and \( c \)? There are 6: \( abc, acb, bac, bca, cab, \) and \( cba \). We can see this more generally by noting that there are 3 different choices for the first item, 2 choices for the second, and 1 choice for the remaining member:

\[
3 \cdot 2 \cdot 1 = 6.
\]

Theorem 2. In general, there are \( n! \) permutations of \( n \) objects.
**Example 4.** Seven people are going to get in line to enter a shared ride. How many different ways can they line up?

There are $7! = 5040$ ways for them to line up.

**Example 5.** Seven people are going to get in line to enter a shared ride. How many different ways can they line up if two of them refuse to be next to each other?

The two who refuse to be next to each other can be next to each other in 6 ways: in spots $(1, 2), (2, 3), (3, 4), \ldots (6, 7)$. In each of those arrangements, person 1 or person 2 could be in front so there are 2 ways to arrange them. Moreover, there are $5! = 120$ ways of ordering the remaining persons. In total, there are

$$6 \cdot 2 \cdot 120 = 1440$$

ways for them to be next to each other. Therefore, there are $5040 - 1440 = 3600$ ways for the group to line up if a particular couple don’t want to be next to each other.

Another way to see this is to isolate one of the “problem” persons. There are $6! = 720$ ways to order the remaining persons and only 5 possible positions in line for the isolated person. That is, we count $5 \cdot 6! = 3600$ ways for the line to be formed.

In a more visual fashion, let our persons be $A, B, C, D, E, F,$ and $G$ where $A$ and $B$ refuse to be next to each other. Then there are $6!$ ways to arrange $B, C, D, E, F,$ and $G$ in a line. Next, we wish to assign a place to $A$ so pick a sample ordering of the remaining:

$$1 \quad C \quad \times \quad B \quad \times \quad F \quad \quad E \quad \quad G \quad \quad D \quad$$

That is, $A$ has only 5 choices to choose from.

But what if we just want to pick out a few things from a set of $n$ objects? For example, how many ways can we choose 2 things from a collection of 4 things, $a, b, c,$ and $d$? By the basic principle of counting, we know that there are $4 \cdot 3 = 12$ ways to pick two items from this list of four. We can represent this with a tree structure:

**Theorem 3.** In general, the number of ways to pick $r \leq n$ things (where order is taken into account) from a collection of $n$ things is

$$n^P_r = \frac{n!}{(n-r)!}$$

Let’s revisit the example where we were picking two things from four:
That is, \( ab \) and \( ba \) are getting counted as distinct choices. What if we wish to ignore order and treat \( ab \) as \( ba \)? In this particular example, we see 6 total options:

1. \( ab \) – \( ba \)
2. \( ac \) – \( ca \)
3. \( ad \) – \( da \)
4. \( bc \) – \( cb \)
5. \( bd \) – \( db \)
6. \( cd \) – \( dc \)

**Definition 4.** A **combination** is a choice of \( r \leq n \) things from \( n \) distinguishable objects without regard to the order of the selection.

**Theorem 5.** In general, there are

\[
\binom{n}{r} = \frac{n!}{(n-r)!r!} = \frac{n^P_r}{r!}
\]

ways to choose \( r \) things from a set of \( n \) ignoring the order in which they are chosen. That is, there are \( \binom{n}{r} \) many sub-collections of a set of \( n \) objects with size \( r \).

**Example 6.** In how many ways can we pick a primary color, a secondary color, and a tertiary color from a collection of 10 different colors? We can do it in \( 10^P_3 = 720 \) ways.

**Example 7.** There are \( \binom{10}{3} = 120 \) ways of choosing three colors from a collection of 10 different colors.

**Definition 6.** A **circular permutation** is one where the objects are arranged in a circle where two arrangements are considered the same if any object has the same object on their left and to their right.

To see how this way of arranging things, consider the following three equivalent circular permutations:

![Circular Permutations](image)

To see the importance of the distinction from regular permutations, notice that

1,2,3 \quad 3,1,2 \quad 2,3,1

would be considered different permutations but are equivalent as circular permutations. With this said, we can count the number of circular permutations for \( n \) objects.

**Theorem 7.** The number of circular permutations of \( n \) distinct objects is \( (n - 1)! \).

Now we consider the case when our collection of \( n \) objects consists of possibly indistinct objects. For example, in how many distinct ways can we reorder *mississippi*? Before we address the case for *mississippi*, let’s consider a simpler example: *seeds*. If we color code, we can identify different permutations that would be indistinguishable without the coloration:

- seeds
- seeds
- seeds

For even more clarity, for any permutation of *seed*, we can reorder the *e*’s present to get another indistinguishable permutation: e.g. *edesc* and *edesc*. Now, there are \( 4! = 24 \) ways to order 4 distinct objects and, for each of those permutations, there are \( 2! = 2 \) ways to reorder the *e*’s. This gives us \( 24/2 = 12 \) ways to reorder *seed*. 
For seeds, there are $5! = 120$ ways to permute 5 distinct objects and, for each of those permutations, there are $2! \cdot 2! = 4$ ways of reordering the $e$’s and $s$’s. Hence, there are $120/4 = 30$ ways to reorder seeds.

For mississippi, we count 1 $m$, 4 $i$’s, 4 $s$’s, and 2 $p$’s. There are 11 characters in mississippi. For any of the 11! permutations of 11 characters, there are $4!$ ways to rearrange the $i$’s, $4!$ ways to rearrange the $s$’s, and $2!$ ways to rearrange the $p$’s. Therefore there are 

$$\frac{11!}{4! \cdot 4! \cdot 2!} = 34650$$

different ways to rewrite mississippi.

**Theorem 8.** In general, if we have a set of $n$ items, only $k$ of which are distinct, and $n_j$ is the number of indistinct objects\(^1\) for $1 \leq j \leq k$, there are

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}$$

ways to order the $n$ objects.

**Definition 9.** A (finite) partition of a set $X$ is a collection $A_1, A_2, \ldots, A_n$ of subsets of $X$ so that $X = A_1 \cup A_2 \cup \cdots \cup A_n$ and $A_j \cap A_k = \emptyset$ for each $j \neq k$.

Given $n$ objects, we can partition them into $k$ groups, each consisting of $n_j$ objects for $1 \leq j \leq k$. Notice that $n = n_1 + n_2 + \cdots + n_k$. Now, how many different ways can we partition $n$ objects into $k$ groups where each group consists of $n_j$ objects for $1 \leq j \leq k$?

First, observe that

$$\sum_{j=1}^{k} n_j = n$$

by virtue of this being a partition. It turns out that there are

$$\binom{n}{n_1, n_2, \ldots, n_k}$$

ways to partition $n$ objects into $k$ groups where each group consists of $n_j$, $1 \leq j \leq k$, objects.

To illustrate this, let’s consider the situation where we have a class of 7 students where we wish to make groups consisting of 3, 2, and 2 students. Then consider the string

$$abc|de|fg$$

where the letters represent students and the $|$ represent the separators between the groups. Since we are just interested in the grouping, notice that

$$bac|ed|fg$$

is a different permutation of the string $abcdefg$ but the groups are the same. So, given any particular grouping expressed as a string, any permutation of a particular group doesn’t yield a new grouping. Hence, there are

$$\frac{7!}{3! \cdot 2! \cdot 2!} = \binom{7}{3, 2, 2}$$

ways to make these groups.

\(^1\)This means that $\sum_{j=1}^{k} n_j = n$


1.1 Binomial Coefficients

The following reflects a symmetry of choosing.

Theorem 10. For any positive integer \( n \) and \( 0 \leq r \leq n \),
\[
\binom{n}{r} = \binom{n}{n-r}.
\]

Proof. Note that
\[
\binom{n}{n-r} = \frac{n!}{(n-(n-r))!(n-r)!} = \frac{n!}{r!(n-r)!} = \binom{n}{r},
\]
the desired equality.

Theorem 11. For any positive integer \( n \geq 2 \) and \( 1 \leq r < n \),
\[
\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}.
\]

Proof. Observe that
\[
\binom{n-1}{r} + \binom{n-1}{r-1} = \frac{(n-1)!}{(n-1-r)!r!} + \frac{(n-1)!}{(n-r-1)!r!} = \frac{(n-1)!}{(n-r)!r!} + \frac{(n-1)!}{(n-r)!r!}\cdot \frac{n-r}{r} = \frac{n!}{(n-r)!r!} = \binom{n}{r},
\]
the desired equality.

In fact, this is seen easily using Pascal’s triangle:

\[
\begin{array}{cccccccccccc}
  & & & & & & & & 1 & & & & \\
  & & & & & & 1 & & 1 & & & & \\
  & & & & & 1 & & 2 & & 1 & & & \\
  & & & & 1 & & 3 & & 3 & & 1 & & \\
  & & & 1 & & 4 & & 6 & & 4 & & 1 & \\
  & & 1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
  & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 & & \\
\end{array}
\]
As an illustration to see how one builds Pascal’s triangle, notice that the number 1 is always on the extremities of each line. Then consider the more detailed diagram:

![Pascal's Triangle Diagram]

**Definition 12.** For a positive integer \( n \) and integer \( 0 \leq r \leq n \), we refer to \( \binom{n}{r} \) as a *binomial coefficient* due to the following result.

**Theorem 13 (Binomial Theorem).** For any positive integer \( n \),

\[
(x + y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}.
\]

**Proof.** First, we note that the rearrangement of the exponents results from commutative of addition or by replacing \( r \) with \( n - r \) and appealing to Theorem 11.

We prove this by induction on \( n \). For the base case, notice that

\[
(x + y)^1 = \binom{1}{0} x^1 y^0 + \binom{1}{1} x^0 y^1 = \sum_{r=0}^{1} \binom{1}{r} x^{n-r} y^r.
\]

So the result is true for \( n = 1 \).

Now suppose it holds up to \( n \geq 1 \) and observe that

\[
(x + y)^{n+1} = (x + y)(x + y)^n
\]

\[
= (x + y) \cdot \left[ \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r \right]
\]

\[
= x \cdot \left[ \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r \right] + y \cdot \left[ \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r \right]
\]

\[
= \left[ \sum_{r=0}^{n} \binom{n}{r} x^{n-r+1} y^r \right] + \left[ \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^{r+1} \right].
\]

Now, to determine the coefficient of \( x^{(n+1)-r} y^r \) in the expansion of \( (x + y)^{n+1} \), we consider cases.

1. Case \( r = 0 \): Notice that the sum on the right-hand side only contributes positive powers of \( y \) so the coefficient of \( x^{n+1} \) in the expansion of \( (x + y)^{n+1} \) is \( \binom{n}{0} = 1 = \binom{n+1}{1} \).

2. Case \( 0 < r < n + 1 \): The left-hand sum contributes \( \binom{n}{r} x^{n-r+1} y^r \) and the right-hand sum contributes

\[
\binom{n}{r-1} x^{n-(r-1)} y^{r-1+1} = \binom{n}{r-1} x^{n-r+1} y^r.
\]

Hence, by Theorem 13, the coefficient of \( x^{(n+1)-r} y^r \) in the expansion of \( (x + y)^{n+1} \) is

\[
\binom{n}{r} + \binom{n}{r-1} = \binom{n+1}{r}.
\]
3. Case \( r = n + 1 \): In this case, the left-hand side of the sum doesn’t contribute anything to the \( y^{n+1} \) term. But the right-hand sum contributes \( \binom{n}{n+1} = 1 = \binom{n+1}{n+1} \).

Thus, we see that
\[
(x + y)^{n+1} = \sum_{r=0}^{n+1} \binom{n+1}{r} x^{n+1-r} y^r
\]

and, by induction, we have completed the proof.

Another way to understand the truth of Theorem 13 is to consider the following:
\[
(x + y)^n = \underbrace{(x + y)(x + y) \cdots (x + y)}_{n \text{ times}}.
\]

Then to find the coefficient of \( x^{n-r} y^r \), one just needs to choose the \( r \) factors of the \((x + y)’s contributing the \( y’\)s. Since the order of choice doesn’t matter, there should be \( \binom{n}{r} \) ways to choose those factors.

For the next theorem, we adopt the convention that, for a positive integer \( n \) and \( r > n \), \( \binom{n}{r} = 0 \). To be sure, there is no way that we can pick \( r \) objects from \( n \) if \( r > n \).

**Theorem 14** (Shijie-Vandermonde’s Identity). For positive integers \( n, m \) and \( 0 \leq k \leq m + n \),
\[
\binom{m + n}{k} = \sum_{r=0}^{k} \binom{m}{r} \binom{n}{k-r}.
\]

**Proof.** First, consider the fact that \((1+y)^{m+n} = (1+y)^m (1+y)^n\). Then, the coefficient of \( y^k \) in the expansion of \((1+y)^{m+n}\) is, by Theorem 13,
\[
\binom{m + n}{k}.
\]

Also by Theorem 13, we compute
\[
(1+y)^m \cdot (1+y)^n = \left[ \sum_{r=0}^{m} \binom{m}{r} y^r \right] \cdot \left[ \sum_{r=0}^{n} \binom{n}{r} y^r \right].
\]

To get the coefficient of \( y^k \) in the expansion of \((1+y)^m \cdot (1+y)^n\), witness that we must consider all coefficients of \( y^{k-r} \) in the expansion of \((1+y)^m \) and \( y^r \) in the expansion of \((1+y)^n \) where \( 0 \leq r \leq k \). Behold that this is
\[
\sum_{r=0}^{k} \binom{m}{r} \binom{n}{k-r},
\]
finishing the proof.

As one may have realized earlier,
\[
\binom{n}{r} = \frac{n!}{(n-r)!r!} = \binom{n}{n-r, r}.
\]

This is suggestive of the following.

**Corollary 15.** For any multinomial \((x_1 + x_2 + \cdots + x_k)^n\), the coefficient of the term \(x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k}\) where \( r_1 + r_2 + \cdots + r_k = n \) is
\[
\binom{n}{r_1, r_2, \ldots, r_k}.
\]

**Proof.** We prove by induction on \( k \) and notice that the base case \( k = 2 \) is the result of Theorem 13. So suppose we’ve proved it up to \( k \geq 2 \) and consider
\[
(x_1 + x_2 + \cdots + x_k + x_{k+1})^n.
\]
Consider the fact that
\[(x_1 + x_2 + \cdots + x_k + x_{k+1})^n = ((x_1 + x_2 + \cdots + x_k + x_{k+1})^n = \sum_{r=0}^{\infty} \binom{n}{r} (x_1 + x_2 + \cdots + x_k)^r x_{k+1}^{n-r}.\]

To see it more vividly, we can rewrite the last sum in its expanded form:
\[(x_1 + x_2 + \cdots + x_k + x_{k+1})^n = \binom{n}{r} (x_1 + x_2 + \cdots + x_k)^r x_{k+1}^{n-r}.\]

Let \(r_1, r_2, \ldots, r_k, r_{k+1}\) be non-negative integers so that \(r_1 + r_2 + \cdots + r_k + r_{k+1} = n\). Let \(r = r_1 + r_2 + \cdots + r_k\) and notice that \(r_{k+1} = n - r\). By the above, we know that, for a term to contribute any coefficient to \(x_{k+1}^r\), it must occur at
\[\binom{n}{r} (x_1 + x_2 + \cdots + x_k)^r x_{k+1}^{n-r}.\]

By the inductive hypothesis, the coefficient of
\[x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k}\]
in the expansion of
\[(x_1 + x_2 + \cdots + x_k)^r\]
is
\[\binom{r}{r_1, r_2, \ldots, r_k}\]

Hence, the coefficient of
\[x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k} x_{k+1}^{r_{k+1}}\]
in the expansion of
\[\binom{n}{r} (x_1 + x_2 + \cdots + x_k)^r x_{k+1}^{n-r}\]
is
\[\binom{n}{r} \binom{r}{r_1, r_2, \ldots, r_k}.\]

Fortuitously,
\[\binom{n}{r} \binom{r}{r_1, r_2, \ldots, r_k} = \frac{n!}{r!(n-r)!} \cdot \frac{r!}{r_1!r_2! \cdots r_k!} = \frac{n!}{r_1!r_2! \cdots r_k!r_{k+1}!} = \binom{n}{r_1, r_2, \ldots, r_k, r_{k+1}}\]
concluding the proof.

Exercise 1. Chris forgot to study for an exam in his Metaontology class. If the exam has 20 multiple choice questions, each having 4 choices, in how many ways can he get a 70% or better on the exam?

Exercise 2. In how many ways can we give one book to each of 25 students where we have 12 of book A, 12 of book B, and 3 of book C?
1.2 Poker Hands

A standard deck of cards contains 52 cards. There are

- 13 ranks, in ascending order: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A
- and 4 suits: ♠, ♥, ♦, ♣

Definition 16. A hand is an unordered collection of 5 cards. There are special kinds of hands as well.

- A pair is a hand containing two cards of the same rank.
- A two pair is a hand containing two pairs.
- A three of a kind is a hand containing three cards of the same rank.
- A full house is a three of a kind and a pair.
- A four of a kind is a hand containing four cards of the same rank.
- A straight is a hand consisting of 5 consecutive ranks allowing the A to be consecutive to the 2.
- A flush is 5 cards of the same suit.
- A straight flush is a straight which is also a flush.

Exercise 3. How many distinct hands are there?

Example 8. How many ways can one get a straight flush?

First, we list the possible ranks in a straight:

\[ \rightarrow A, 2, 3, 4, 5 \]
\[ \rightarrow 2, 3, 4, 5, 6 \]
\[ \rightarrow 3, 4, 5, 6, 7 \]
\[ \rightarrow 4, 5, 6, 7, 8 \]
\[ \rightarrow 5, 6, 7, 8, 9 \]
\[ \rightarrow 6, 7, 8, 9, 10 \]
\[ \rightarrow 7, 8, 9, 10, J \]
\[ \rightarrow 8, 9, 10, J, Q \]
\[ \rightarrow 9, 10, J, Q, K \]
\[ \rightarrow 10, J, Q, K, A \]

So there are 10 ways of having a straight not accounting for suit values. Since a straight flush must be cards only from one suit, there are \(4 \cdot 10 = 40\) ways to get a straight flush.

Example 9. How many ways can one get a straight?

We already noted in Example 8 that there are 10 ways of getting the consecutive cards. Since there are four suits and we have 5 cards, there are \(4^5 = 1024\) ways to assign suit values to any hand. Therefore, there are \(10 \cdot 1024 = 10240\) different ways to have a straight.

If you wish to only count the straights which are not straight flushes, observe that there are \(10240 - 40 = 10200\) different straights that are not straight flushes.

Example 10. How many ways can one get a flush?

There are 13 cards in a suit so there are

\[ \binom{13}{5} = 1287 \]

ways to get a hand consisting of only one suit. There are 4 suits so there are \(4 \cdot 1287 = 5148\) ways of getting a flush.
**Example 11.** How many ways can one get a pair?

For any particular rank, there are $\binom{4}{2} = 6$ ways to pick two cards from that rank (this is already accounting for the suit). There are 13 different ranks. For the remaining three cards, since we just want to count the pairs, we need to make sure none of the rest of the three cards are of the same rank. So there are $\binom{12}{3} = 220$ different ways to pick 3 cards from the remaining card ranks. Moreover, there are $4^3 = 64$ ways to choose the suits on these last three. So there are

$$6 \cdot 13 \cdot 220 \cdot 64 = 1098240$$

different ways to get a pair. More concisely,

$$\binom{13}{1} \cdot \binom{4}{2} \cdot \binom{12}{3} \cdot 4^3 = 1098240.$$  

**Example 12.** How many ways can one get two pair?

For any particular rank, like we saw above, there are $\binom{4}{2} = 6$ ways to get a pair. There are $\binom{13}{2} = 78$ ways to choose the two ranks. The fifth card must be of the remaining 11 ranks and there are 4 ways to assign it a suit. Therefore, there are

$$6 \cdot 6 \cdot 78 \cdot 11 \cdot 4 = 123552$$

ways to get two pairs. More concisely,

$$\binom{13}{2} \cdot \binom{4}{2}^2 \cdot \binom{11}{1} \cdot (4^1)^3 = 123552.$$  

**Exercise 4.** How many ways can one get a three of a kind?

**Exercise 5.** How many ways can one get a full house?

**Exercise 6.** How many ways can one get a four of a kind?
Chapter 2

Probability

“Not probable — The barest Chance —
A smile too few — a word too much”
from *Not probable — The barest Chance* —
by Emily Dickinson

Probability can be thought of as the likelihood of a particular event occurring out of a total possible number of events. There are two common interpretations of probabilities.

**The Classical Approach.** If there are $n$ possible events, then the probability of a single event being the outcome is $1/n$. That is, each event is equally likely. Moreover, if $s \leq n$ events are chosen to be favorable, then the probability of a favorable outcome is $s/n$.

**Example 13.** Given a 6-sided die, the probability (from the classical point of view) of rolling a 1 or a 6 is

$$\frac{2}{6} = \frac{1}{3} \approx 33.33\%.$$ 

**The Relative Frequency Approach.** Suppose we are interested in finding the likelihood of an event $A$. Out of $n$ experiments where $A$ could have occurred, suppose that event $A$ actually occurred $s_n$ times. Then the idea is to let the probability of the event $A$ to be

$$P(A) = \lim_{n \to \infty} \frac{s_n}{n},$$

if the limit exists. As we can see, this interpretation is dependent upon experimentation and historical records. Also, it is physically impossible to actually conduct infinitely many experiments but, for many scenarios of interest, we can approximate $P(A)$ well enough with large $n$.

**Example 14.** Suppose a 6-sided die has been rolled 1 000 times and the number of rolls resulting in either 1 or 6 has been recorded to be 317. Then we see that the probability (according to the frequentist) of rolling a 1 or a 6 (based on this particular set of experiments) is

$$\approx \frac{317}{1000} = 31.7\%.$$ 

**Remark.** Although the difference between the Classical and Relative Frequency approaches seems to hinge on the *repeatability* of a particular experiment, it may be seen to be a bit more philosophically subtle. When one hears that there is a 30% change of rain, one is ready to accept the frequentist interpretation that, according to historical records, with similar weather patterns, rain tends to fall 30% of the time. It is hard to conceptualize the 30% rain chance as being calculated as the result of 3 events out of a total of 10 possible events.
On the other hand, when one rolls a die, it is easy to think a re-roll is an actual repetition of the roll experiment. The key here is that, although we actually cannot recreate the previous roll (many indiscernible and perhaps unknown forces could be at play), the roll of a die is similar enough to any other roll of a die that we treat them as identical experiments.

**Question.** Epiphania has been chillin’ in a ball pit for time immemorial. She has thrown a ball at the top of every hour since her arrival to the ball pit. You show up to the ball pit at 2:55pm. What is the probability that she will throw a ball at 3:00pm?

### 2.1 Sample Spaces and Events

**Definition 17.** By a sample space we mean a set $S$ consisting of all possible outcomes of an experiment. Each outcome in a sample space is called an element or a sample point.

**Example 15.** If I’m flipping a coin, the sample space can be represented with $S = \{H, T\}$. 

**Example 16.** If I’m rolling a 6-sided die, the sample space can be represented with $S = \{1, 2, 3, 4, 5, 6\}$. 

**Example 17.** If I’m flipping a coin until a heads appears, the sample space can be represented with

$$S = \{H, TH, TTH, TTTTH, TTTTT, \ldots\}.$$ 

So sample spaces need not be finite.

We usually categorize samples spaces in terms of their cardinality; i.e., the number of elements they contain. Recall that a set $A$ is said to be countable if there exists a bijective correspondence between $A$ and $\mathbb{N}$.

**Definition 18.** We say that a sample space $S$ is discrete if $S$ is finite or countable.

**Definition 19.** We say that a sample space $S$ is continuous if there is a bijective correspondence between $S$ and $\mathbb{R}$.

**Example 18.** Measuring the temperature in Yellowknife, Northwest Territories, Canada, can best be represented with a continuous sample space.

**Definition 20.** Let $S$ be a sample space. By an event, we mean a subset $E \subseteq S$.

**Example 19.** Suppose my sample space is $S = \{1, 2, 3, 4, 5, 6\}$, the outcomes of rolling a 6-sided die. Then I can define the event $E = \{2, 4, 6\}$, the ways in which I can roll an even number.

**Example 20.** Suppose I’m measuring the temperature in Yellowknife, Northwest Territories, Canada, in degrees Fahrenheit. Let the sample space be $S = \mathbb{R}$. I can define the event $E = (-\infty, 32]$, the set of outcomes where I measure a temperature less than or equal to 32°F.

**Example 21.** Give a sample space which represents the outcomes of rolling two dice. Then describe the event of rolling two dice that sum up to 7.

The sample space here can be represented by ordered pairs where the first number corresponds to the first die and the second number to the second:

$$S = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\
(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\
(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\
(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\
(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

The outcomes which result in a sum of 7 are

$$E = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}.$$
Definition 21. Given two events $A$ and $B$ of a sample space $S$, we say that $A$ and $B$ are **mutually exclusive** provided that $A \cap B = \emptyset$; i.e. that $A$ and $B$ have no sample points in common.

Example 22. Let the sample space be $S = \{ n \in \mathbb{N} : 1 \leq n \leq 20 \}$, $A$ be the event of getting an even number, and $B$ be the event of getting an odd number. Then $A$ and $B$ are mutually exclusive events.

Example 23. Let the sample space be $S = \{ n \in \mathbb{N} : 1 \leq n \leq 20 \}$, $A$ be the event of getting an even number, and $B$ be the event of getting a number divisible by 3. Then $A$ and $B$ are not mutually exclusive events. In fact, $6 \in A \cap B$.

### 2.2 The Probability of Events

The **Postulates of Probability**. Given a sample space $S$ and a real-valued function $P$ on subsets of $S$, then $P$ is a **probability measure** (on $S$) if all of the following hold:

(P1) For any event $E$, $P(E) \geq 0$.

(P2) $P(S) = 1$.

(P3) For any collection $A_1, A_2, \ldots$ of mutually exclusive events,

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$$

By the way we set things up, the sample space $S$ consists of all possible outcomes. So Postulate (P2) is capturing that fact; i.e., the probability that an outcome of the experiment is a sample point in $S$ is 100%.

We could rephrase Postulate (P1) in the following way: The probability $P(E)$ of any event $E$ must be non-negative.

**Theorem 22.** If $E$ is an event of a discrete sample space $S$ and $P$ is a probability measure on $S$, then

$$P(E) = \sum_{x \in E} P\{x\};$$

i.e., $P(E)$ is the sum of the probabilities of the individual outcomes which comprise $E$.

**Proof.** All sample points are distinct so $\{x\}$ and $\{y\}$ are mutually exclusive events for $x \neq y$. Notice that

$$E = \bigcup_{x \in E} \{x\}$$

so, by Postulate (P3),

$$P(E) = \sum_{x \in E} P\{x\},$$

the promised equality. \qed

**Corollary 23.** If $S$ is a finite sample space with cardinality $N$ where all outcomes are equally probable and $E$ is any event with cardinality $n$, then $P(E) = \frac{n}{N}$.

**Proof.** By Theorem 22,

$$P(E) = \sum_{x \in E} P\{x\}.$$ 

Since all outcomes are equally probable,

$$P\{x\} = \frac{1}{N}.$$
Ergo,

\[ P(E) = \sum_{x \in E} P(\{x\}) = \sum_{x \in E} \frac{1}{N} = \frac{n}{N}, \]

and we’re done.

In the remainder of this section we will collect facts that follow readily from the Postulates (P1) – (P3). Recall that, given \( E \subseteq S \), then the complement of \( E \) is

\[ E^c = \{ x \in S : x \notin E \}. \]

**Theorem 24.** For any event \( E \), \( P(E^c) = 1 - P(E) \).

**Proof.** Of course, we know that \( E \cap E^c = \emptyset \) and \( E \cup E^c = S \). Since \( E \cap E^c = \emptyset \), we have

\[
\begin{align*}
1 &= P(S) & \text{by (P2)} \\
   &= P(E \cup E^c) \\
   &= P(E) + P(E^c) & \text{by (P3)}
\end{align*}
\]

Then we see that \( 1 - P(E) = P(E^c) \), the desired result. \hfill \Box

**Theorem 25.** \( P(\emptyset) = 0 \).

**Proof.** Clearly, \( (\emptyset)^c = S \) so, by Theorem 24, we see that

\[
\begin{align*}
1 &= P(S) & \text{by (P2)} \\
   &= 1 - P(\emptyset), & \text{by Theorem 24}
\end{align*}
\]

Hence, \( P(\emptyset) = 0 \). \hfill \Box

**Theorem 26.** If \( A \) and \( B \) are two events in a sample space with \( A \subseteq B \), then \( P(A) \leq P(B) \).

**Proof.** Notice that

\[ B = (B \cap A^c) \cup (B \cap A) = (B \cap A^c) \cup A \]

since \( A \subseteq B \). Moreover, \( B \cap A^c \) and \( A \) are mutually exclusive. Then, by Postulate (P3),

\[ P(B) = P(B \cap A^c) + P(A). \]

By Postulate (P1), \( P(B \cap A^c) \geq 0 \) so we have that

\[ P(B) = P(B \cap A^c) + P(A) \geq P(A), \]

the desired conclusion. \hfill \Box

**Theorem 27.** For any event \( E \), \( 0 \leq P(E) \leq 1 \).

**Proof.** The fact that \( 0 \leq P(E) \) is Postulate (P1). Then, since \( E \subseteq S \), combining Theorem 26 and Postulate (P2), we see that

\[ P(E) \leq P(S) = 1, \]

finishing the proof. \hfill \Box

**Theorem 28.** For any two events \( A \) and \( B \),

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B). \]
2.2. THE PROBABILITY OF EVENTS

Proof. Note that
\[ A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B) \]
and that all three sets \( A \cap B^c, A \cap B, \) and \( A^c \cap B \) are mutually exclusive. So, by Postulate (P3),
\[ P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B). \]

Now, \( A = (A \cap B^c) \cup (A \cap B) \) and both \( A \cap B^c \) and \( A \cap B \) are mutually exclusive so
\[ P(A) = P(A \cap B^c) + P(A \cap B). \]

Similarly, \( B = (B \cap A^c) \cup (A \cap B) \) and both \( B \cap A^c \) and \( A \cap B \) are mutually exclusive so
\[ P(B) = P(B \cap A^c) + P(A \cap B). \]

That is,
\[ P(A \cap B^c) = P(A) - P(A \cap B) \]
and
\[ P(A^c \cap B) = P(B) - P(A \cap B). \]

Putting it all together,
\[ P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \]
\[ = P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B) \]
\[ = P(A) + P(B) - P(A \cap B), \]
the coveted conclusion. \qed

Theorem 29. For any three events \( A, B, \) and \( C, \)
\[ P(A \cup B \cup C) = P(A) + P(B) + P(C) \]
\[ -P(A \cap B) - P(A \cap C) - P(B \cap C) \]
\[ +P(A \cap B \cap C). \]

Proof. We can employ Theorem 28 to see that
\[ P(A \cup B \cup C) = P(A \cup B) + P(C) - P((A \cup B) \cap C) \]
\[ = P(A) + P(B) - P(A \cap B) + P(C) - P((A \cap C) \cup (B \cap C)). \]

Then we calculate
\[ P((A \cap C) \cup (B \cap C)) = P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C)) \]
\[ = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C). \]

Putting it together,
\[ P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P((A \cap C) \cup (B \cap C)) \]
\[ = P(A) + P(B) + P(C) - P(A \cap B) \]
\[ -P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \]
\[ = P(A) + P(B) + P(C) \]
\[ -P(A \cap B) - P(A \cap C) - P(B \cap C) \]
\[ +P(A \cap B \cap C), \]
the intended result. \qed
To help visualize consider the Venn diagram:

If this were a counting exercise and we imagine simply counting the things in $A$, the things in $B$ and the things in $C$, then we would have counted the things in $A \cap B$, $A \cap C$, and $B \cap C$ twice while counting the things in $A \cap B \cap C$ three times. To adjust for counting twice, we need to subtract the extra counts in $A \cap B$, $A \cap C$, and $B \cap C$. After this adjustment, we’ve over adjusted the $A \cap B \cap C$; namely, we have counted them thrice and removed them thrice. So we simply need include $A \cap B \cap C$ again.

**Exercise 7.** As we saw in Exercise 1, Chris forgot to study for an exam in his Metaontology class. Recall that the exam has 20 multiple choice questions, each having 4 choices. What is the probability that he scores 70% or better on the exam?

**Exercise 8.** Recall the material from Section 1.2. You are dealt a Poker hand from a standard deck of cards. Determine the probability that you are dealt

(a) a straight flush.
(b) a straight.
(c) a flush.
(d) a pair.
(e) two pair.
(f) a three of a kind.
(g) a full house.
(h) a four of a kind.

**Exercise 9.** There is a 67% probability of meeting Jemima, a 43% probability of meeting Ben, and a 19% probability of meeting both at The Rice Pancake Convention. Find the probability of meeting Jemimia or Ben at the convention.

**Exercise 10.** There is a

- 24% probability that Earth will be visited by Dassians
- 37% probability that Earth will be visited by Learyians
- 41% probability that Earth will be visited by Wilsonians
- 12% probability that Earth will be visited by both Dassians and Learyians
- 9% probability that Earth will be visited by both Dassians and Wilsonians
- 7% probability that Earth will be visited by both Learyians and Wilsonians
- 2% probability that Earth will be visited by all 3.

Find the probability that Earth will be visited by Dassians, Learyians, or Wilsonians.
2.3 Conditional Probability

Conditional probability comes into play when one wishes to calculate the probability of an event relative to another event in a larger sample space. For example, if $S$ is the sample space of all flowers at The Flowing Florist but one wishes to determine the probability that a randomly selected tulip is yellow, one would need to restrict attention to the set of tulips. Here, we discuss how to do just that.

**Definition 30.** Let $S$ be a sample space and $A$ and $B$ be events. Then, as long as $P(B) \neq 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

is the conditional probability of $A$ relative to $B$, read “the probability of $A$ given $B$”.

**Theorem 31.** Let $S$ be a sample space and $A$ and $B$ be events where $P(B) \neq 0$. Then

$$P(A \cap B) = P(A|B) \cdot P(B).$$

**Example 24.** Suppose The Flowing Florist has 560 flowers, 70 of which are tulips, and that 12 of the tulips are yellow. Calculate the following probabilities:

- that a randomly selected flower is a tulip,
- that a randomly selected flower is a yellow tulip, and
- that a randomly selected tulip is yellow.

The probability of randomly selecting a tulip is $\frac{70}{560}$.

There are 12 yellow tulips so the probability of randomly selecting a yellow tulip is $\frac{12}{560}$.

The probability of randomly selecting a yellow tulip from the tulips is $\frac{12}{70}$.

Let $T$ represent the event of selecting a tulip and $Y$ represent the event of selecting a yellow tulip. Notice that $Y \subseteq T$ so $T \cap Y = Y$. Moreover, by definition,

$$P(Y|T) = \frac{P(Y \cap T)}{P(T)} = \frac{P(Y)}{P(T)} = \frac{12}{560} \cdot \frac{560}{70} = \frac{12}{70}.$$

**Example 25.** Jack Example tells you that he has exactly one sibling. What is the probability that he has a brother?

Working under the convention that siblings are categorized as sister or brother, the sample space of two siblings in a family can be seen to be

$$SS \quad SB \quad BS \quad BB$$
To see this clearly, think in terms of chronological birth; i.e., older/younger sibling. Given that Jack is a brother, the relative sample space is

\[
\begin{array}{ccc}
SB & BS & BB \\
\end{array}
\]

So the probability that Jack has a brother should be \( \frac{1}{3} \). Is that surprising? Let’s consider the following events:

- \( T \): of two siblings, both are brothers
- \( A \): of two siblings, there is at least one brother

From our original sample space,

\[
\begin{array}{ccc}
SS & SB & BS & BB \\
\end{array}
\]

we see that \( P(T) = \frac{1}{4} \) and \( P(A) = \frac{3}{4} \). Now, the probability that Jack has a brother is really

\[
P(T|A) = \frac{P(T \cap A)}{P(A)} = \frac{P(T)}{P(A)} = \frac{1}{4} \cdot \frac{4}{3} = \frac{1}{3}.
\]

Exercise 11. In a diplomatic meeting with the Queen of Toratopia, she informs you that she has exactly one sibling.

(a) What is the probability that she has a brother?

(b) If she is the oldest sibling, what is the probability that she has a brother?

Example 26. A museum curator knows that the probability Gustav Klimt finishes a commissioned piece in time for a gallery opening is 0.87. The probability that piece will be ready in time for the opening and delivered to the museum is 0.78. What is the probability that the painting will be delivered on time given that it is ready on time?

Let \( R \) denote the event of the painting being ready on time and \( D \) denote the event of it being delivered on time. Using the definition of conditional probability, we compute

\[
P(D|R) = \frac{P(R \cap D)}{P(R)} = \frac{0.78}{0.87} \approx 0.8966,
\]

the probability that the painting will be delivered on time given that it is ready on time.

Example 27. A quality control employee at BiC is to randomly select 2 ballpoint pens from a batch of 72, 5 of which are defective. What is the probability that both pens selected are defective?

Let \( A \) be the event of the first pen selected being defective and \( B \) be the event of the second pen being defective. The probability of \( A \) is

\[
P(A) = \frac{5}{72}.
\]

Having selected a pen from the group already and putting it aside, there are only 71 pens left to choose from. Assuming that the first pen pulled was defective, the probability of the next pen being defective is

\[
P(B|A) = \frac{4}{71}.
\]

So, the probability of both selected pens being defective is

\[
P(A \cap B) = P(B|A) \cdot P(A) = \frac{4}{71} \cdot \frac{5}{72} = \frac{5}{1278}.
\]

Alternatively, we can also approach this example using combinatorial methods. Notice that there are \( \binom{5}{2} \) ways of picking two of the defective pens and \( \binom{72}{2} \) different possible samples. Then check that

\[
\frac{\binom{5}{2}}{\binom{72}{2}} = \frac{5}{1278}.
\]
2.3. CONDITIONAL PROBABILITY

Example 28. An aspiring artist wishes to incorporate randomness into her work to express what she experiences as randomness in the universe. To do this, she has a box of 43 colored pencils, 7 of which are black. In her first minimalist piece, she will color the interior of a rhombus with a randomly selected color. Then she will color the exterior of the rhombus with a randomly selected color.

(a) If she replaces the color she used in the first step, find the probability that she will use black twice.

(b) If she doesn’t replace the first color choice, find the probability that she will use a non-black color for the rhombus and black for the exterior.

- Let \(I\) be the event of picking black for the interior and \(E\) be the event of picking black for the exterior. Notice that \(P(I) = \frac{7}{43}\).

Now, given that the first color was black and that she replaced the first color, we calculate

\[
P(E|I) = \frac{7}{43}.
\]

Then

\[
P(I \cap E) = P(E|I) \cdot P(I) = \frac{7}{43} \cdot \frac{7}{43} = \frac{49}{1849}.
\]

- Let \(N\) be the event that she picks a non-black color for the interior and \(B\) be the event that she picks black for the exterior. Notice that \(P(N) = \frac{43 - 7}{43} = \frac{36}{43}\).

Given that the first color was non-black, we calculate

\[
P(B|N) = \frac{7}{42}.
\]

Then

\[
P(N \cap B) = P(B|N) \cdot P(N) = \frac{7}{42} \cdot \frac{36}{43} = \frac{6}{43}.
\]

We can extend Theorem 31 in the following way.

**Theorem 32.** Let \(S\) be a sample space and \(A, B,\) and \(C\) be events so that \(P(A \cap B) \neq 0\). Then

\[
P(A \cap B \cap C) = P(C|A \cap B) \cdot P(B|A) \cdot P(A).
\]

**Proof.** Notice that

\[
P(A \cap B \cap C) = P(C \cap (B \cap A)) = P(C|B \cap A) \cdot P(B \cap A) = P(C|A \cap B) \cdot P(B|A) \cdot P(A),
\]

the desired end.

Following inductively, one could obtain many more generalizations.

Example 29. A shipment of 30 smartphones, 5 of which are defective, arrive at WeKonekt. If 3 phones are to be randomly selected for testing (in succession without replacement), find the probability that all 3 phones are defective.

Let \(A\) be the event that the first phone is defective, \(B\) be the event that the second phone is defective, and \(C\) be the event that the third phone is defective. Notice that

- \(P(A) = \frac{5}{30}\)
- \(P(B|A) = \frac{4}{29}\)
22

CHAPTER 2. PROBABILITY

• \( P(C | A \cap B) = \frac{3}{28} \)

Then, we use

\[
P(A \cap B \cap C) = P(C | A \cap B) \cdot P(B | A) \cdot P(A) \]
\[
= P(A) \cdot P(B | A) \cdot P(C | A \cap B) \]
\[
= \frac{5}{30} \cdot \frac{4}{29} \cdot \frac{3}{28} \]
\[
= \frac{1}{406}.
\]

2.4 Independent Events

Recall from Example 28 (a), the probability of getting the color black in the second selection wasn’t affected at all by the first selection. This inspires the notion of independence in probability. Formally, let \( A \) and \( B \) be events. The idea is that \( A \) and \( B \) are independent if \( P(A | B) = P(A) \) and \( P(B | A) = P(B) \). Using Theorem 31, assuming \( A \) and \( B \) are independent, we obtain

\[
P(A \cap B) = P(A) \cdot P(B | A) \]
\[
= P(A) \cdot P(B).
\]

Definition 33. Two events \( A \) and \( B \) are said to be independent if

\[
P(A \cap B) = P(A) \cdot P(B).
\]

Otherwise, \( A \) and \( B \) are said to be dependent.

Example 30. A coin is tossed three times. Let \( A \) be the event that the first two outcomes are heads, \( B \) be the event that the third outcome is a tails, and \( C \) be the event that exactly two tails occur. Show that

(a) \( A \) and \( B \) are independent, and

(b) \( B \) and \( C \) are dependent.

Consider the sample space

\[
S = \{ \text{TTT, HHT, THT, TTH, HHT, HTH, THH, HHH} \}
\]

Then

\[
A = \{ \text{HHT, HHH} \} \implies P(A) = \frac{2}{8} = \frac{1}{4}
\]
\[
B = \{ \text{TTT, HHT, THT, HHT} \} \implies P(B) = \frac{4}{8} = \frac{1}{2}
\]
\[
C = \{ \text{HTT, THT, TTH} \} \implies P(C) = \frac{3}{8}
\]

and

\[
A \cap B = \{ \text{HHT} \} \implies P(A \cap B) = \frac{1}{8}
\]
\[
B \cap C = \{ \text{HTT, THT} \} \implies P(B \cap C) = \frac{2}{8} = \frac{1}{4}
\]

From this we see that

\[
P(A) \cdot P(B) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} = P(A \cap B),
\]

establishing that \( A \) and \( B \) are independent. On the other hand,

\[
P(B) \cdot P(C) = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16} \neq \frac{1}{4} = P(B \cap C),
\]

establishing that \( B \) and \( C \) are dependent.
Example 31. Working under the convention that siblings are categorized as sister or brother, suppose a family has two children. Let $A$ be the event that there is a sister and $B$ be the event that there is a brother. Show that $A$ and $B$ are dependent.

Like before, the sample space can be seen to be

\[
\begin{array}{cc}
SS & SB \\
BS & BB \\
\end{array}
\]

Now, $P(A) = \frac{3}{4}$ and $P(B) = \frac{3}{4}$ whereas

\[
P(A \cap B) = \frac{1}{2} \neq \frac{9}{16} = \frac{3}{4} \cdot \frac{3}{4} = P(A) \cdot P(B).
\]

Exercise 12. Let $S = \{a, b, c\}$ where each outcome is equally probable. Then let $A = \{a, b\}$ and $B = \{b, c\}$. Show that $A$ and $B$ are dependent.

Theorem 34. If $A$ and $B$ are independent, $A$ and $B^c$ are independent.

Proof. Note that $A = (A \cap B) \cup (A \cap B^c)$ and that $A \cap B$ and $A \cap B^c$ are mutually exclusive. It follows that

\[
P(A) = P((A \cap B) \cup (A \cap B^c)) = P(A \cap B) + P(A \cap B^c) = P(A) \cdot P(B)
\]

which provides

\[
P(A \cap B^c) = P(A) - P(A) \cdot P(B) = P(A)(1 - P(B)) = P(A) \cdot P(B^c).
\]

Therefore, $A$ and $B^c$ are independent.

Definition 35. Three events $A$, $B$, and $C$ are said to be independent provided that all of the following hold:

- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A \cap C) = P(A) \cdot P(C)$
- $P(B \cap C) = P(B) \cdot P(C)$
- $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

Example 32. Consider the sample space

\[
\{a, b, c, z\}
\]

where each event is equally likely. Let $A = \{a, z\}$, $B = \{b, z\}$, and $C = \{c, z\}$. Show that

(a) $A$ and $B$ are independent,
(b) $A$ and $C$ are independent,
(c) $B$ and $C$ are independent, but
(d) $A$, $B$, and $C$ are dependent.
First, we organize the facts:

\[ \{z\} = A \cap B = A \cap C = B \cap C = A \cap B \cap C \]

\[ \implies \frac{1}{4} = P(A \cap B) = P(A \cap C) = P(B \cap C) = P(A \cap B \cap C) \]

Also,

\[ \frac{1}{2} = P(A) = P(B) = P(C). \]

Immediately, we see that

- \( P(A \cap B) = P(A) \cdot P(B) \)
- \( P(A \cap C) = P(A) \cdot P(C) \)
- \( P(B \cap C) = P(B) \cdot P(C) \)

Alas,

\[ P(A) \cdot P(B) \cdot P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \neq \frac{1}{4} = P(A \cap B \cap C). \]

**Definition 36.** More generally, given \( n \) events \( E_1, E_2, \ldots, E_n \), we say that they are independent provided that, for any \( 2 \leq m \leq n \) and selection

\[ 1 \leq k_1 < k_2 < \cdots < k_{m-1} < k_m \leq n, \]

we have that

\[ P \left( \bigcap_{j=1}^{m} E_{k_j} \right) = \prod_{j=1}^{m} P(E_{k_j}). \]

In words, they are independent if the probability of the intersection of \( m \) of these events is equal to the product of their respective probabilities.

The usefulness of independence is displayed in the following example.

**Example 33.** A 6-sided die is rolled eight times and each roll is independent from the others. Find the probability of getting six 5’s followed by two other numbers.

We calculate the probability to be

\[ \left( \frac{1}{6} \right)^6 \cdot \left( \frac{5}{6} \right)^2 = \frac{25}{1679616}. \]

**2.5 Bayes’ Theorem**

**Theorem 37.** If the events \( B_1, B_2, \ldots, B_n \) form a partition of the sample space \( S \) and \( P(B_j) \neq 0 \) for each \( 1 \leq j \leq n \), then for any event \( A \),

\[ P(A) = \sum_{j=1}^{n} P(A|B_j) \cdot P(B_j). \]

**Proof.** By the fact that \( B_1, B_2, \ldots, B_n \) form a partition of the sample space \( S \), It follows that

\[ A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n) \]

and we know that \( B_j \cap B_k = \emptyset \) whenever \( j \neq k \). Therefore,

\[ P(A) = \sum_{j=1}^{n} P(A \cap B_j) = \sum_{j=1}^{n} P(A|B_j) \cdot P(B_j), \]

the desired conclusion.
Example 34. Tu Swole is making a new music video but the production may be delayed due to uncertain availability of a select car. The probabilities are

- 0.43 that the Rolls-Royce Phantom will be available,
- 0.91 that the video will be completed on time with the Rolls-Royce Phantom, and
- 0.17 that the video will be completed on time without the Rolls-Royce Phantom.

What is the probability that the video will be completed on time?

Let $V$ be the event that the video is completed on time and $R$ be the event that the Rolls-Royce Phantom is available for the video. Immediately, $P(R) = 0.43$, $P(V|R) = 0.91$, and $P(V|R^c) = 0.17$. Then, we can apply Theorem 37 to compute

$$P(V) = P(V|R) \cdot P(R) + P(V|R^c) \cdot P(R^c)$$

$$= 0.91 \cdot 0.43 + 0.17 \cdot 0.57$$

$$= 0.4882.$$ 

Example 35. A pretentious film critic likes to decorate her house with prints. She gets 63% of her prints from Guggenheim, 21% of her prints from the MoMA, and 16% of her prints from Tate. If 23% of prints from Guggenheim, 30% of prints from MoMA, and 7% of prints from Tate are avant-garde, what is the probability that her next print will be avant-garde?

Let $A$ be the event that her next print is avant-garde, $G$ be the event her print comes from Guggenheim, $M$ be the event her print comes from MoMA, and $T$ be the event her print comes from Tate. Using Theorem 37, we calculate

$$P(A) = P(A|G) \cdot P(G) + P(A|M) \cdot P(M) + P(A|T) \cdot P(T)$$

$$= 0.23 \cdot 0.63 + 0.3 \cdot 0.21 + 0.07 \cdot 0.16$$

$$= 0.2191.$$ 

Theorem 38 (Bayes’ Theorem). If the events $B_1, B_2, \ldots, B_n$ form a partition of the sample space $S$ and $P(B_j) \neq 0$ for each $1 \leq j \leq n$, then for any event $A$ with $P(A) \neq 0$,

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{j=1}^{n} P(A|B_j) \cdot P(B_j)}$$

Proof. First, note that

$$P(B_k|A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A|B_k) \cdot P(B_k)}{P(A)}.$$ 

The rest follows from Theorem 37. 

Example 36. Recall our film critic from Example 35. Assuming she just received a new avant-garde print, find the probability that it came from Tate.

We are asked to find $P(T|A)$. By Bayes’ Theorem, we compute

$$P(T|A) = \frac{P(A|T) \cdot P(T)}{P(A|G) \cdot P(G) + P(A|M) \cdot P(M) + P(A|T) \cdot P(T)}$$

$$= \frac{0.07 \cdot 0.16}{0.2191}$$

$$\approx 0.05112.$$ 

When testing for certain phenomena like drug usage or having a particular disease, some amount of error is possible. For example, a test could claim that you haven’t used steroids when you actually have used them. This would be known as a false negative. The following chart outlines the possibilities.
Example 37. We have discovered that aliens have implanted nanomachines in 3% of the human population and a new test has an accuracy rate of 98%. That is, it correctly identifies a person with nanomachines 98% of the time and it correctly identifies someone without nanomachines 98% of the time.

(a) If you have tested positive for nanomachines, find the probability that you actually have nanomachines.

(b) If you have tested positive for nanomachines, find the probability that you don’t actually have nanomachines.

(c) If you have tested negative for nanomachines, find the probability that you actually don’t have nanomachines.

(d) If you have tested negative for nanomachines, find the probability that you actually do have nanomachines.

Let $H$ be the event of having nanomachines and $T$ be the event of testing positive for nanomachines.

• We’ll approach (a) in three different ways.

  ◦ By Bayes’ Theorem,

  \[
  P(H|T) = \frac{P(T|H) \cdot P(H)}{P(T|H) \cdot P(H) + P(T|H^c) \cdot P(H^c)}
  \]

  \[
  = \frac{0.98 \cdot 0.03}{0.98 \cdot 0.03 + 0.02 \cdot 0.97}
  \]

  \[
  \approx 0.6025
  \]

  ◦ Imagine that the population is 10000. Then 300 have nanomachines. Of those 300 with nanomachines, 98% of them will test positively; i.e., 294 of them will be correctly identified. Of those 9700 without nanomachines, 98% of them will test negatively; i.e., 9506 of them will be correctly identified. Let’s organize this in a table:

<table>
<thead>
<tr>
<th>actually positive</th>
<th>294</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>actually negative</td>
<td>194</td>
<td>9506</td>
</tr>
</tbody>
</table>

So there are 488 people which test positive, 294 of those actually have it. Hence, the probability that you actually have nanomachines given you tested positive for them is $\frac{294}{488} \approx 0.6025$.

  ◦ In a tree diagram:
From here, we see that $0.0294 + 0.0194 = 0.0488$ tested positive, $0.0294$ of those are correctly identified so 
\[
\frac{0.0294}{0.0488} \approx 0.6025
\]
would actually have nanomachines given that they tested positive for them.

- By Bayes’ Theorem,
\[
P(H^c|T) = \frac{P(T|H^c) \cdot P(H^c)}{P(T|H^c) \cdot P(H^c) + P(T|H) \cdot P(H)}
= \frac{0.02 \cdot 0.97}{0.02 \cdot 0.97 + 0.98 \cdot 0.03}
\approx 0.3975.
\]

- By Bayes’ Theorem,
\[
P(H^c|T^c) = \frac{P(T^{c}|H^c) \cdot P(H^c)}{P(T^{c}|H^c) \cdot P(H) + P(T^{c}|H) \cdot P(H)}
= \frac{0.02 \cdot 0.03 + 0.98 \cdot 0.97}{0.02 \cdot 0.03 + 0.98 \cdot 0.97}
\approx 0.9994.
\]

- By Bayes’ Theorem,
\[
P(H|T^c) = \frac{P(T^{c}|H) \cdot P(H)}{P(T^{c}|H) \cdot P(H) + P(T^{c}|H^c) \cdot P(H^c)}
= \frac{0.02 \cdot 0.03}{0.02 \cdot 0.03 + 0.98 \cdot 0.97}
\approx 0.00063.
\]
Chapter 3

Probability Distributions and Densities

3.1 Random Variables

Definition 39. If $S$ is a (not necessarily discrete) sample space with a probability measure, then any real-valued function $X$ defined on $S$ is called a random variable. If the range of $X$ is discrete, then we say that $X$ is a discrete random variable. If $S$ is a continuous sample space and the range of $X$ is also a continuous space, $X$ is said to be a continuous random variable. By convention, random variables will always be denoted with a capital letter and the values which they take on will be denoted with lower case letters.

Remark. Despite the nomenclature, a continuous random variable need not be a continuous function supposing the sample space is $\mathbb{R}$ or so.

3.1.1 Discrete Random Variables

Example 38. Let $S$ be the set of ordered pairs with integer entries ranging between 1 and 6; the space of two 6-sided dice rolls. Then define $X$ by the rule $X(a, b) = a + b$, the sum of the rolls. Then $X$ is a discrete random variable.

Example 39. Consider the sample space corresponding to flipping a coin 5 times. Define $X$ to be the number of heads appearing. Then $X$ is a random variable.

When using random variables, their values also determine events. Using set notation along with the fact that $X$ is a function, we can write

$$\{a \in S : X(a) = x\},$$

a subset of the sample space $S$ corresponding to all outcomes where $X$ is equal to $x$. This would be the event that $X = x$. We will write $P(X = x)$ to refer to the probability that the random variable $X$ is equal to the value $x$. As alluded to in Definition 39, the convention is to write the values which the random variable assumes in the lower case.

When rolling two dice as in Example 38 where the random variable is the sum of the rolls, we can talk about the probability that the sum is greater than 9. We would write this as $P(X \geq 9)$.

Example 40. Consider again the space of two 6-sided dice rolls with the random variable defined by summing the two rolls. Find $P(X = 7)$.

We will list out the entire sample space, the probability of each event, and the value of the random variable:
CHAPTER 3. PROBABILITY DISTRIBUTIONS AND DENSITIES

<table>
<thead>
<tr>
<th>outcome</th>
<th>prob.</th>
<th>x</th>
<th>outcome</th>
<th>prob.</th>
<th>x</th>
<th>outcome</th>
<th>prob.</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1)</td>
<td>1/36</td>
<td>2</td>
<td>(2, 1)</td>
<td>1/36</td>
<td>3</td>
<td>(3, 1)</td>
<td>1/36</td>
<td>4</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>1/36</td>
<td>3</td>
<td>(2, 2)</td>
<td>1/36</td>
<td>4</td>
<td>(3, 2)</td>
<td>1/36</td>
<td>5</td>
</tr>
<tr>
<td>(1, 3)</td>
<td>1/36</td>
<td>4</td>
<td>(2, 3)</td>
<td>1/36</td>
<td>5</td>
<td>(3, 3)</td>
<td>1/36</td>
<td>6</td>
</tr>
<tr>
<td>(1, 4)</td>
<td>1/36</td>
<td>5</td>
<td>(2, 4)</td>
<td>1/36</td>
<td>6</td>
<td>(3, 4)</td>
<td>1/36</td>
<td>7</td>
</tr>
<tr>
<td>(1, 5)</td>
<td>1/36</td>
<td>6</td>
<td>(2, 5)</td>
<td>1/36</td>
<td>7</td>
<td>(3, 5)</td>
<td>1/36</td>
<td>8</td>
</tr>
<tr>
<td>(1, 6)</td>
<td>1/36</td>
<td>7</td>
<td>(2, 6)</td>
<td>1/36</td>
<td>8</td>
<td>(3, 6)</td>
<td>1/36</td>
<td>9</td>
</tr>
<tr>
<td>(4, 1)</td>
<td>1/36</td>
<td>5</td>
<td>(5, 1)</td>
<td>1/36</td>
<td>6</td>
<td>(6, 1)</td>
<td>1/36</td>
<td>7</td>
</tr>
<tr>
<td>(4, 2)</td>
<td>1/36</td>
<td>6</td>
<td>(5, 2)</td>
<td>1/36</td>
<td>7</td>
<td>(6, 2)</td>
<td>1/36</td>
<td>8</td>
</tr>
<tr>
<td>(4, 3)</td>
<td>1/36</td>
<td>7</td>
<td>(5, 3)</td>
<td>1/36</td>
<td>8</td>
<td>(6, 3)</td>
<td>1/36</td>
<td>9</td>
</tr>
<tr>
<td>(4, 4)</td>
<td>1/36</td>
<td>8</td>
<td>(5, 4)</td>
<td>1/36</td>
<td>9</td>
<td>(6, 4)</td>
<td>1/36</td>
<td>10</td>
</tr>
<tr>
<td>(4, 5)</td>
<td>1/36</td>
<td>9</td>
<td>(5, 5)</td>
<td>1/36</td>
<td>10</td>
<td>(6, 5)</td>
<td>1/36</td>
<td>11</td>
</tr>
<tr>
<td>(4, 6)</td>
<td>1/36</td>
<td>10</td>
<td>(5, 6)</td>
<td>1/36</td>
<td>11</td>
<td>(6, 6)</td>
<td>1/36</td>
<td>12</td>
</tr>
</tbody>
</table>

So we see that $P(X = 7) = \frac{6}{36} = \frac{1}{6}$.

Random variables can even be seen as more robust than sample spaces. What we mean here is that any discrete sample space can be translated into a random variable without any loss of information. More explicitly, given a discrete sample space $S$, we can enumerate the outcomes and such an enumeration is a random variable.

**Example 41.** Consider the sample space $S = \{SS, SB, BS, BB\}$ and the random variable $X$ defined as

<table>
<thead>
<tr>
<th>outcome</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>1</td>
</tr>
<tr>
<td>SB</td>
<td>2</td>
</tr>
<tr>
<td>BS</td>
<td>3</td>
</tr>
<tr>
<td>BB</td>
<td>4</td>
</tr>
</tbody>
</table>

Then every outcome has a unique identifier in terms of the random variable. Moreover, we could calculate the probability of the event $A = \{SB, BS\}$ by computing $P(2 \leq X \leq 3)$.

**Example 42.** A box contains 12 red balls and 7 green balls. We are to randomly select three balls from the box successively. Let $X$ be the random variable counting the number of green balls. Describe the sample space, find the probability of each outcome, and summarize this information along with the random variable values. Using the table, calculate $P(X = 2)$ and $P(X < 2)$.

Using the tools of conditional probability we calculate

<table>
<thead>
<tr>
<th>outcome</th>
<th>probability</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRR</td>
<td>(12/19)(11/18)(10/17) = 220/969</td>
<td>0</td>
</tr>
<tr>
<td>RRG</td>
<td>(12/19)(11/18)(7/17) = 154/969</td>
<td>1</td>
</tr>
<tr>
<td>RGR</td>
<td>(12/19)(7/18)(11/17) = 154/969</td>
<td>1</td>
</tr>
<tr>
<td>GRR</td>
<td>(7/19)(12/18)(11/17) = 154/969</td>
<td>1</td>
</tr>
<tr>
<td>RGG</td>
<td>(12/19)(7/18)(6/17) = 28/323</td>
<td>2</td>
</tr>
<tr>
<td>GRG</td>
<td>(7/19)(12/18)(6/17) = 28/323</td>
<td>2</td>
</tr>
<tr>
<td>GGR</td>
<td>(7/19)(6/18)(12/17) = 28/323</td>
<td>2</td>
</tr>
<tr>
<td>GGG</td>
<td>(7/19)(6/18)(5/17) = 35/969</td>
<td>3</td>
</tr>
</tbody>
</table>

Then,

$$P(X = 2) = \frac{28}{323} + \frac{28}{323} + \frac{28}{323} = \frac{84}{323}$$
3.2. PROBABILITY DISTRIBUTIONS

and

\[ P(X < 2) = \frac{220}{969} + \frac{154}{969} + \frac{154}{969} + \frac{154}{969} = \frac{682}{969}. \]

3.1.2 Continuous Random Variables

Hitherto, we’ve focused on discrete sample spaces out of convenience. Nevertheless, continuous sample spaces are also important.

Example 43. Consider the depth of a lake which varies over time. Using drainage methods, suppose the lake can always be maintained below a depth of 500 meters. Then our sample space can be modeled with the set \( S = [0, 500] \) and we can define the random variable \( X \) to be the depth measurement. If we define

\[ P(a \leq X \leq b) = \frac{b - a}{500}, \]

we can build a probability measure on \( S \) though there are some subtleties with regards to what subsets can be coherently called events exceeding the scope of this course. In any case, by taking smaller intervals, given any value \( x \), we can see that \( P(X = x) = 0 \) even though any measurement of the depth of the lake would produce a particular value. Intuitively, if all particular outcomes could be considered to be equally likely, there are too many outcomes to allow any of them to be positive. With that said, notice that \( X = x \) is not an impossible event even though it is a probability zero event.

3.2 Probability Distributions

Definition 40. If \( X \) is a discrete random variable, the real-valued function \( f \) defined on the range of \( X \) by the rule \( f(x) = P(X = x) \) is called the probability distribution of \( X \).

Abstractly, the probability distribution of \( X \) exists for any discrete random variable \( X \) though it is much more useful if one were to have a formula.

Example 44. Recall Example 38 where we roll two dice and the random variable is the sum of the rolls. Using the table provided in Example 40, we collect the values of \( X \) along with their respective probabilities:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( P(X = x) )</th>
<th>( x )</th>
<th>( P(X = x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \frac{1}{36} )</td>
<td>8</td>
<td>( \frac{5}{36} )</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{2}{36} )</td>
<td>9</td>
<td>( \frac{4}{36} )</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{3}{36} )</td>
<td>10</td>
<td>( \frac{3}{36} )</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{4}{36} )</td>
<td>11</td>
<td>( \frac{2}{36} )</td>
</tr>
<tr>
<td>6</td>
<td>( \frac{5}{36} )</td>
<td>12</td>
<td>( \frac{1}{36} )</td>
</tr>
<tr>
<td>7</td>
<td>( \frac{6}{36} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then the probability distribution \( f \) of \( X \) is given by

\[ f(x) = \frac{6 - |x - 7|}{36}. \]

To obtain the formula, one could identify the common denominator of 36 and examine a plot of the numerators:
This should lead one directly to the formula $6 - |x - 7|$. 

**Example 45.** Find a formula for the probability distribution of the total number of heads appearing in a sequence of 8 coin tosses.

There are $2^8 = 256$ total outcomes. There are $\binom{8}{x}$ many ways to get exactly $x$ heads. Therefore, the probability distribution is given by

$$f(x) = \frac{\binom{8}{x}}{256}.$$ 

But what about situations like the one in Example 42? The probability distribution there is expressed in the table

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>220/969</td>
</tr>
<tr>
<td>1</td>
<td>462/969</td>
</tr>
<tr>
<td>2</td>
<td>252/969</td>
</tr>
<tr>
<td>3</td>
<td>35/969</td>
</tr>
</tbody>
</table>

**Exercise 13.** Find a formula $f(x)$ in terms of non-negative integers $x$ so that

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>220/969</td>
</tr>
<tr>
<td>1</td>
<td>462/969</td>
</tr>
<tr>
<td>2</td>
<td>252/969</td>
</tr>
<tr>
<td>3</td>
<td>35/969</td>
</tr>
</tbody>
</table>

Notice that, if we treat the range $R$ of a discrete random variable as a sample space, then the probability distribution is a probability measure on $R$. Hence, a translation of the Postulates of Probability yields the following.

**Theorem 41.** A function $f(x)$ can serve as the probability distribution of a discrete random variable $X$ if and only if the following two conditions hold:

- $f(x) \geq 0$ for all values $x$
- $\sum_x f(x) = 1$ where the summation ranges over all points of the domain of $f$.

We can also express probability distributions graphically. Recall the probability distribution in Exercise 13 and consider the following histogram:
As a means of foreshadowing, the area of each rectangle is equal to the probability the random variable equals the number at the base since we set the base length of each rectangle to be 1. This should remind one of Riemann sums and, in turn, integration. More on that later.

**Definition 42.** If $X$ is a discrete random variable, the function $F(x)$ defined by

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

for $x \in \mathbb{R}$ where $f(t)$ is the probability distribution of $X$ is called the *cumulative distribution*, or the *distribution function*, of $X$.

**Theorem 43.** Suppose $F(x)$ is a cumulative distribution of a discrete random variable $X$. Then the following three conditions hold:

- $F$ is monotonically increasing; i.e., if $a \leq b$, then $F(a) \leq F(b)$

- $\lim_{x \to -\infty} F(x) = 0$

- $\lim_{x \to \infty} F(x) = 1$

Recall the greatest integer function $\lfloor x \rfloor$ which returns the largest integer $n$ so that $n \leq x$.

**Example 46.** Again, revisit Example 38 where we roll two dice and the random variable is the sum of both rolls. Discuss the cumulative distribution.

Relying on the probability distribution formula obtained in Example 44, the cumulative distribution is given by

$$F(x) = \sum_{t=2}^{\lfloor x \rfloor} \frac{6 - |t - 7|}{36}.$$ 

Graphically,
As a piece-wise defined function,

$$F(x) = \begin{cases} 
0, & x < 2; \\
1/36, & 2 \leq x < 3; \\
3/36, & 3 \leq x < 4; \\
6/36, & 4 \leq x < 5; \\
10/36, & 5 \leq x < 6; \\
15/36, & 6 \leq x < 7; \\
21/36, & 7 \leq x < 8; \\
26/36, & 8 \leq x < 9; \\
30/36, & 9 \leq x < 10; \\
33/36, & 10 \leq x < 11; \\
35/36, & 11 \leq x < 12; \\
1, & 12 \leq x
\end{cases}$$

which we also obtain from the table in Example 44.

What we’ve seen is how to get the cumulative distribution given the probability distribution. As long as $X$ is a finite random variable, we can recover the probability distribution from the cumulative distribution.

**Theorem 44.** If the range of a random variable $X$ consists of the values $x_1 < x_2 < \cdots < x_n$ and $F$ is the cumulative distribution of $X$, then $f(x_1) = F(x_1)$ and, for $2 \leq j \leq n$,

$$f(x_j) = F(x_j) - F(x_{j-1})$$

where $f$ is the probability distribution of $X$.

**Example 47.** Given the cumulative distribution

$$F(x) = \begin{cases} 
0, & x < 0; \\
0.3, & 0 \leq x < 1; \\
0.75, & 1 \leq x < 2; \\
0.95, & 2 \leq x < 3; \\
1, & 3 \leq x
\end{cases}$$
of $X$, find the probability distribution of $X$.

We can infer that the random variable $X$ assumes the values 0, 1, 2, and 3. Then

- $f(0) = 0.3$,
- $f(1) = 0.75 - 0.3 = 0.45$,
- $f(2) = 0.95 - 0.75 = 0.2$, and
- $f(3) = 1 - 0.95 = 0.05$.

Remark. Although we can recover information about the probability distribution from the cumulative distribution, we have no license to infer anything about the original sample space.

### 3.3 Probability Density Functions

Recall a similar scenario to the one in Example 43 where we were measuring the depth of a lake. Let’s establish the sample space to be the interval $[-10, 10]$ where 0 corresponds to the desired lake level, 450 meters. The negative values represent how many meters below 450 the lake is actually measured at and the positive values represent how many meters above 450 the lake is actually measured at. If we decide to round off our measurements to the nearest meter, our sample space is now a discrete one where the sample space is

$$
\{d \in \mathbb{Z} : -10 \leq d \leq 10\}.
$$

We can refine our measurements even more, say to the nearest centimeter. This would give us more information about the depth of the lake, yet still be a discrete sample space. Consider the two potential probability histograms according to our discrete approximations:
The more we refine our measurements, the closer we get to the continuous “distribution”. As alluded to before, this approximation process should remind one of Riemann sums.

**Definition 45.** A function \( f(x) \) defined on \( \mathbb{R} \) is called a **probability density function** of the continuous random variable \( X \) if and only if

\[
P(a \leq X \leq b) = \int_a^b f(x) \, dx
\]

for all real numbers \( a \) and \( b \) with \( a \leq b \).

**Remark.** The reason we introduce a different term here, in contrast to Definition 40, the discrete analog, is because the value \( f(a) \) is not the value \( P(X = a) \). In fact, in agreement with the comments made in Example 43, we have that \( P(X = a) = 0 \) for any real value \( a \).

**Theorem 46.** If \( X \) is a continuous random variable and \( a \) and \( b \) are real numbers with \( a \leq b \), then

\[
P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b).
\]

As an analog to Theorem 41, we obtain the following.

**Theorem 47.** A non-negative function \( f(x) \) can serve as a probability density of a continuous random variable \( X \) if and only if

\[
\int_{-\infty}^{\infty} f(x) \, dx = 1.
\]

**Example 48.** Suppose \( X \) has the probability density

\[
f(x) = \begin{cases} k \cdot e^{-5x}, & x > 0; \\ 0, & x \leq 0 \end{cases}
\]

Find \( k \) and \( P(2 \leq X \leq 4) \).

First, we compute

\[
\int_{-\infty}^{\infty} f(x) \, dx = \int_0^{\infty} k \cdot e^{-5x} \, dx
\]

\[
= k \cdot \lim_{t \to \infty} \left[ -\frac{e^{-5x}}{5} \right]_0^t
\]

\[
= k \cdot \left( \frac{1}{5} + \lim_{t \to \infty} -\frac{e^{-5t}}{5} \right)
\]

\[
= \frac{k}{5}.
\]

To find the desired \( k \),

\[
1 = \int_{-\infty}^{\infty} f(x) \, dx = \frac{k}{5} \implies k = 5.
\]

Now, for \( P(2 \leq X \leq 4) \), we compute

\[
\int_2^4 5e^{-5x} \, dx = -e^{-5x} \bigg|_2^4
\]

\[
= -e^{-20} + e^{-10}
\]

\[
= \frac{e^{10} - 1}{e^{20}}.
\]

The following is completely analogous to Definition 42.
3.3. PROBABILITY DENSITY FUNCTIONS

**Definition 48.** If $X$ is a continuous random variable with a probability density function $f(x)$, then the cumulative distribution function, or simply distribution function, of $X$ is given by

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

Relying on the Fundamental Theorem of Calculus, we obtain

**Theorem 49.** If $f(x)$ is the probability density of a continuous random variable $X$ and $F(x)$ is the cumulative distribution function of $X$, then

$$P(a \leq X \leq b) = F(b) - F(a)$$

for real $a \leq b$ and

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

where the derivative exists.

**Example 49.** Find the cumulative distribution function of the random variable $X$ in Example 48. Then use it to evaluate $P(2 \leq X \leq 5)$.

We calculate

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

$$= \int_{0}^{x} 5e^{-5t} \, dt$$

$$= -e^{-5t}\bigg|_{0}^{x}$$

$$= -e^{-5x} + 1$$

$$= 1 - e^{-5x}.$$ 

To evaluate $P(2 \leq X \leq 5)$, we use

$$F(5) - F(2) = (1 - e^{-25}) - (1 - e^{-10}) = \frac{e^{15} - 1}{e^{25}}.$$

**Example 50.** Suppose a continuous random variable $X$ has a cumulative distribution function

$$F(x) = \begin{cases} 
0, & x < 0; \\
 x, & 0 \leq x \leq 1; \\
1, & 1 < x
\end{cases}$$

Find the probability density function.

We can apply Theorem 49 to see that

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

whenever the derivative exists. So we obtain

$$f(x) = \begin{cases} 
0, & x < 0; \\
1, & 0 \leq x \leq 1; \\
0, & 1 < x
\end{cases}$$

the probability distribution.
Example 51. Nothing in Definition 45 required the probability density \( f(x) \) to be continuous. As by Theorem 47, the function

\[
f(x) = \begin{cases} 
0, & x \leq 0; \\
1/4, & 0 < x \leq 1; \\
-\frac{1}{6} \cdot x + 1, & 1 < x < 2; \\
0, & 2 \leq x 
\end{cases}
\]

satisfies

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

and thus is a fine probability density.

Example 52. We can also combine Theorems 44 and 49 to generalize cumulative distributions to be monotonically increasing functions so that \( P(a \leq X \leq b) = F(b) - F(a) \) and open the door to discontinuous cumulative distributions. Then the function

\[
F(x) = \begin{cases} 
0, & x < 0; \\
x^2, & 0 \leq x < \frac{1}{2}; \\
x, & \frac{1}{2} \leq x < 1; \\
1, & 1 \leq x
\end{cases}
\]

is a cumulative distribution, say for some random variable \( X \). The added complication here is that

\[
P(X = 1/2) = F(1/2) - \lim_{t \to \frac{1}{2}^+} F(t) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.
\]

With that said, we will concern ourselves with continuous cumulative distribution functions when dealing with continuous random variables.

Exercise 14. Show that the function

\[
f(x) = \exp \left(\frac{-x^2}{\sqrt{\pi}}\right) = \frac{e^{-x^2}}{\sqrt{\pi}}
\]

serves as a probability density.

### 3.4 Multivariate Distributions and Densities

In this section, we entertain scenarios where potentially multiple random variables defined on a single sample space are being considered. Given a sample space \( S \),

- the **univariate** case is where only one random variable is in consideration,
- the **bivariate** case is where two random variables defined on \( S \) are being considered, and
- the **multivariate** case is where finitely many random variables defined on \( S \) are being considered.

#### 3.4.1 The Discrete Case

**Notation.** If \( X \) and \( Y \) are discrete random variables, we write \( P(X = x, Y = y) \) to mean the probability that \( X = x \) and \( Y = y \). Notice that this is the intersection of the events \( X = x \) and \( Y = y \).
Example 53. Consider the space of two 6-sided dice rolls where $X$ is the sum of the rolls and $Y$ is the product of the rolls. The following table summarizes this information.

<table>
<thead>
<tr>
<th>outcome</th>
<th>$x$</th>
<th>$y$</th>
<th>outcome</th>
<th>$x$</th>
<th>$y$</th>
<th>outcome</th>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>2</td>
<td>1</td>
<td>(2,1)</td>
<td>3</td>
<td>2</td>
<td>(3,1)</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>(1,2)</td>
<td>3</td>
<td>2</td>
<td>(2,2)</td>
<td>4</td>
<td>4</td>
<td>(3,2)</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>(1,3)</td>
<td>4</td>
<td>3</td>
<td>(2,3)</td>
<td>5</td>
<td>6</td>
<td>(3,3)</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>(1,4)</td>
<td>5</td>
<td>4</td>
<td>(2,4)</td>
<td>6</td>
<td>8</td>
<td>(3,4)</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>(1,5)</td>
<td>6</td>
<td>5</td>
<td>(2,5)</td>
<td>7</td>
<td>10</td>
<td>(3,5)</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>(1,6)</td>
<td>7</td>
<td>6</td>
<td>(2,6)</td>
<td>8</td>
<td>12</td>
<td>(3,6)</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>(4,1)</td>
<td>5</td>
<td>4</td>
<td>(5,1)</td>
<td>6</td>
<td>5</td>
<td>(6,1)</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>(4,2)</td>
<td>6</td>
<td>8</td>
<td>(5,2)</td>
<td>7</td>
<td>10</td>
<td>(6,2)</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>(4,3)</td>
<td>7</td>
<td>12</td>
<td>(5,3)</td>
<td>8</td>
<td>15</td>
<td>(6,3)</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>(4,4)</td>
<td>8</td>
<td>16</td>
<td>(5,4)</td>
<td>9</td>
<td>20</td>
<td>(6,4)</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>(4,5)</td>
<td>9</td>
<td>20</td>
<td>(5,5)</td>
<td>10</td>
<td>25</td>
<td>(6,5)</td>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td>(4,6)</td>
<td>10</td>
<td>24</td>
<td>(5,6)</td>
<td>11</td>
<td>30</td>
<td>(6,6)</td>
<td>12</td>
<td>36</td>
</tr>
</tbody>
</table>

Example 54. Suppose a box contains 12 red markers, 7 green markers, and 3 blue markers. Three markers are randomly selected from the box, $X$ is the number of red markers, and $Y$ is the number of green markers. Describe this bivariate distribution.

Note that the possibilities for values of $X$ and $Y$ are

\[(0,0), (1,0), (2,0), (3,0), (0,1), (1,1), (2,1), (0,2), (1,2), (0,3).\]

There are $\binom{22}{3} = 1540$ total samples possible.
To find the probability of $(x, y)$ where $0 \leq x \leq 3$, $0 \leq y \leq 3$, and $x + y \leq 3$, notice that there are

\[\begin{align*}
&\text{• } \binom{12}{x} \text{ ways to choose } x \text{ red markers}, \\
&\text{• } \binom{7}{y} \text{ ways to choose } y \text{ green markers, and} \\
&\text{• } \binom{3}{3-(x+y)} \text{ to fill the remaining sample with blue markers.}
\end{align*}\]

Then we obtain the formula

\[P(X = x, Y = y) = \frac{\binom{12}{x} \binom{7}{y} \binom{3}{3-(x+y)}}{1540}.\]

Now, we summarize the information in the following table:

\[
\begin{array}{c|cccc}
   & x & 0 & 1 & 2 & 3 \\
\hline
   y &   & \frac{1}{1540} & \frac{9}{385} & \frac{9}{77} & \frac{1}{7} \\
   1 & \frac{3}{220} & \frac{9}{55} & \frac{3}{11} &   \\
   2 & \frac{9}{220} & \frac{9}{55} &   &   \\
   3 & \frac{1}{33} &   &   &   \\
\end{array}
\]
**Definition 50.** If \(X\) and \(Y\) are discrete random variables over a sample space, the function given by
\[
f(x, y) = P(X = x, Y = y)
\]
for each pair of values \((x, y)\) in the range of \(X\) and \(Y\) is called the joint probability distribution of \(X\) and \(Y\).

**Theorem 51.** A bivariate function \(f(x, y)\) can serve as the joint probability distribution of a pair of discrete random variables \(X\) and \(Y\) if and only if the following two properties hold:
- \(f(x, y) \geq 0\) for all pairs \((x, y)\) in the domain and
- \(\sum_{(x, y)} f(x, y) = 1\) where the summation ranges over all pairs \((x, y)\) in the domain.

**Example 55.** Determine the value of \(k\) for which the function \(f(x, y) = kxy^2\), \(x = 1, 2, 3, 4, \quad y = 1, 2, 3\), \(x + y \leq 5\), can serve as a joint probability distribution.

Consider the table

\[
\begin{array}{|c|cccc|}
\hline
x & 1 & 2 & 3 & 4 \\
\hline
y & 1 & k & 2k & 3k & 4k \\
2 & 4k & 8k & 12k & \\
3 & 9k & 18k & \\
\hline
\end{array}
\]

listing the values of \(f(x, y)\). For \(f(x, y)\) to serve as a joint probability distribution, we need
\[
1 = k + 2k + 3k + 4k + 8k + 12k + 9k + 18k = 61k.
\]
Therefore, \(k = \frac{1}{61}\).

**Definition 52.** If \(X\) and \(Y\) are discrete random variables, the function
\[
F(x, y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s, t)
\]
defined for all real values of \(x\) and \(y\) where \(f(s, t)\) is the joint probability distribution of \(X\) and \(Y\) at \((s, t)\) is called the joint cumulative distribution, or simply the joint distribution, of \(X\) and \(Y\).

**Example 56.** Refer back to Example 53 and use the table to verify that \(F(7, 10) = \frac{19}{36}\).

Given two random variables \(X\) and \(Y\) over a sample space \(S\), we can form new random variables via algebraic expressions involving \(X\) and \(Y\); e.g. \(X + Y\).

**Example 57.** Refer to Example 54 and calculate \(P(X + Y \leq 2)\).

To satisfy \(X + Y \leq 2\), we need the pairs \((0, 0), (0, 1), (0, 2), (1, 0), (1, 1),\) and \((2, 0)\). Using the probability distribution
\[
f(x, y) = \frac{\binom{12}{x} \binom{7}{y} \binom{3}{x+y}}{1540},
\]
we compute
\[
P(X + Y \leq 2) = f(0, 0) + f(0, 1) + f(0, 2) + f(1, 0) + f(1, 1) + f(2, 0)
\]
\[
= \frac{1}{1540} + \frac{3}{220} + \frac{9}{385} + \frac{9}{55} + \frac{9}{70}
\]
\[
= \frac{571}{1540}.
\]
3.4.2 The Continuous Case

Definition 53. A bivariate function $f(x, y)$ defined on the plane is called a joint density function of the continuous random variables $X$ and $Y$ if

$$P[(X, Y) \in A] = \int_A f(x, y) \, dx \, dy$$

for any region $A$ of the plane.

Example 58. Given the joint probability density

$$f(x, y) = \begin{cases} \frac{xy + x^2}{17}, & 0 < x < 2, \ 0 < y < 3; \\ 0, & \text{else} \end{cases}$$

of two continuous random variables $X$ and $Y$, find $P[(X, Y) \in A]$ where $A$ is the region

$$A = \{(x, y) : 0 < x < 1, 1 < y < 2\}.$$ 

First, consider the following graph

where the gray region is where $f(x, y)$ takes on positive values and the red region is the region $A$. Now, to find the desired probability, we calculate

$$P[(X, Y) \in A] = P(0 < X < 1, 1 < Y < 2)$$

$$= \int_A f(x, y) \, dx \, dy$$

$$= \int_1^2 \int_0^1 \frac{xy + x^2}{17} \, dx \, dy$$

$$= \int_1^2 \left[ \frac{x^2y}{34} + \frac{x^3}{51} \right]_{x=0}^{x=1} \, dy$$

$$= \int_1^2 \left[ \frac{y}{34} + \frac{1}{51} \right] \, dy$$

$$= \frac{y^2}{68} + \frac{y}{51} \bigg|_1^2$$

$$= \frac{4}{68} + \frac{2}{51} - \frac{1}{68} - \frac{1}{51}$$

$$= \frac{13}{204}.$$
**Theorem 54.** A non-negative bivariate function $f(x, y)$ can serve as a joint probability density for a pair of continuous random variables $X$ and $Y$ if and only if

$$
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1.
$$

**Example 59.** Find the value of $k$ for which

$$
f(x, y) = \begin{cases} 
  kxy, & 0 < x < 1, \ 0 < y < 1, \ x^2 + y^2 < 1; \\
  0, & \text{else}
\end{cases}
$$
can serve as a joint probability density.

First, notice that the function $f(x, y)$ takes on positive values only on the region

We can integrate against $y$ first. In that case, we are integrating along the lines that look like

and the dashed line is given by the relation $y = \sqrt{1-x^2}$.

To ensure that $f(x, y)$ is a joint probability density, we need

$$
1 = \int_0^1 \int_0^{\sqrt{1-x^2}} kxy \, dy \, dx \\
= k \cdot \left[ \int_0^1 xy^2 \sqrt{1-x^2} \, dy \right] \\
= k \cdot \left[ \int_0^1 \frac{x^2 - x^4}{2} \, dx \right] \\
= k \cdot \left[ \frac{x^3}{3} - \frac{x^5}{10} \right]_0^1 \\
= \frac{k}{8}
$$

Therefore, $k = 8$.

**Definition 55.** For continuous random variables $X$ and $Y$, the function given by

$$
F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t) \, ds \, dt
$$

for all real values $x$ and $y$ where $f(s, y)$ is the joint probability density of $X$ and $Y$ is called the joint cumulative distribution of $X$ and $Y$. 
Analogously to the univariate case, partial differentiation leads us to the relationship

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

where $F(x, y)$ is the joint cumulative distribution and $f(x, y)$ is the joint density and the derivative exists.

**Example 60.** Refer to Example 58 and find $F(x, y)$.

Immediately, if either $x \leq 0$ or $y \leq 0$, $F(x, y) = 0$. Now, for $0 < x < 2$ and $0 < y < 3$,

$$F(x, y) = P(X \leq x, Y \leq x)$$

$$= \int_0^y \int_0^x \frac{st + s^2}{17} ds \, dt$$

$$= \int_0^y \left[ \frac{s^2t}{34} + \frac{s^3y}{51} \right]_{s=x=0}^{s=x} dt$$

$$= \int_0^y \frac{x^2t^2}{34} + \frac{x^3}{51} dt$$

$$= \frac{x^2y^2}{68} + \frac{x^3y}{51}.$$

For $x \geq 2$ and $0 < y < 3$,

$$F(x, y) = \frac{y^2}{17} + \frac{8y}{51}.$$

For $0 < x < 2$ and $y \geq 3$,

$$F(x, y) = \frac{9x^2}{68} + \frac{x^3}{17}.$$

We can now write

$$F(x, y) = \begin{cases} 
0, & x \leq 0 \text{ or } y \leq 0; \\
\frac{x^2y^2}{68} + \frac{x^3y}{51}, & 0 < x < 2 \text{ and } 0 < y < 3; \\
\frac{y^2}{17} + \frac{8y}{51}, & x \geq 2 \text{ and } 0 < y < 3; \\
\frac{9x^2}{68} + \frac{x^3}{17}, & 0 < x < 2 \text{ and } y \geq 3; \\
1, & \text{else} 
\end{cases}$$

**Example 61.** Given the joint cumulative distribution

$$F(x, y) = \begin{cases} 
\frac{2\text{arctan}(xy)}{\pi}, & x > 0, \ y > 0; \\
0, & \text{else,} 
\end{cases}$$

find the corresponding joint probability density.

Let’s first take a derivative with respect to $y$ for $x > 0$ and $y > 0$:

$$\frac{\partial}{\partial y} F(x, y) = \frac{2}{\pi} \cdot \frac{x}{1 + x^2y^2}.$$
CHAPTER 3. PROBABILITY DISTRIBUTIONS AND DENSITIES

Now, a derivative with respect to \( x \) for \( x > 0 \) and \( y > 0 \):

\[
\frac{\partial^2}{\partial x \partial y} F(x, y) = \frac{\partial}{\partial x} \left[ \frac{2}{\pi} \cdot \frac{x}{1 + x^2 y^2} \right] = \frac{2}{\pi} \left[ \frac{1 - x^2 y^2}{(1 + x^2 y^2)^2} \right].
\]

Hence, the joint probability density is

\[
f(x, y) = \begin{cases} 
\frac{2}{\pi} \cdot \frac{1 - x^2 y^2}{(1 + x^2 y^2)^2}, & x > 0, \ y > 0; \\
0, & \text{else.}
\end{cases}
\]

3.4.3 Beyond the Bivariate

All of the notions above generalize to the multivariate case where we have random variables \( X_1, X_2, \ldots, X_n \).

**Definition 56.** If \( X_1, X_2, \ldots, X_n \) are discrete random variables over a sample space, the function given by

\[
f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]

for each tuple of values \((x_1, x_2, \ldots, x_n)\) in the range of \( X_1, X_2, \ldots, X_n \) is called the joint probability distribution of \( X_1, X_2, \ldots, X_n \).

**Example 62.** Given the joint probability distribution

\[
f(x, y, z) = \frac{x + y z}{90}, \quad x = 1, 2, 3; \ y = 1, 2, 3; \ z = 1, 2
\]

for discrete random variables \( X, Y, \) and \( Z \), find \( P(X = 3, Y + Z \leq 3) \).

Notice that the condition \( X = 3, Y + Z \leq 3 \) is satisfied only by the tuples \((3, 1, 1), (3, 1, 2), \) and \((3, 2, 1)\). So

\[
P(X = 3, Y + Z \leq 3) = f(3, 1, 1) + f(3, 1, 2) + f(3, 2, 1)
\]

\[= \frac{4}{90} + \frac{5}{90} + \frac{5}{90}
\]

\[= \frac{7}{45}.
\]

**Definition 57.** If \( X_1, X_2, \ldots, X_n \) are discrete random variables, the function

\[
F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)
\]

defined for all real values \( x_1, x_2, \ldots, x_n \) where \( f(t_1, t_2, \ldots, t_n) \) is the joint probability distribution of \( X_1, X_2, \ldots, X_n \) at \((t_1, t_2, \ldots, t_n)\) is called the joint cumulative distribution, or simply the joint distribution, of \( X_1, X_2, \ldots, X_n \).

**Definition 58.** A multivariate function \( f(x_1, x_2, \ldots, x_n) \) defined on \( \mathbb{R}^n \) is called a joint density function of the continuous random variables \( X_1, X_2, \ldots, X_n \) if

\[
P[(X_1, X_2, \ldots, X_n) \in A] = \int_A \cdots \int f(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \cdots \, dx_n
\]

for any region \( A \) of \( \mathbb{R}^n \).
3.5 Marginal Distributions and Densities

Definition 59. For continuous random variables $X_1, X_2, \ldots, X_n$, the function given by

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

$$= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, t_2, \ldots, t_n) \, dt_1 \, dt_2 \cdots dt_n$$

for all real values $x_1, x_2, \ldots, x_n$ where $f(t_1, t_2, \ldots, t_n)$ is the joint probability density of $X_1, X_2, \ldots, X_n$ is called the joint cumulative distribution of $X_1, X_2, \ldots, X_n$.

As before, whenever the derivative exists, we have

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x_1, x_2, \ldots, x_n).$$

3.5 Marginal Distributions and Densities

Given a multivariate distribution, we can isolate particular random variables using marginal distributions. For example, given a bivariate distribution between two random variables $X$ and $Y$, we can “collapse” the distribution down to $X$ by “accumulating” the extra information given by $Y$. In the discrete case, this will amount to addition and, in the continuous case, this will amount to integration.

Definition 60. Given a multivariate distribution with discrete random variables $X_1, X_2, \ldots, X_n$, their joint probability distribution function $f(x_1, x_2, \ldots, x_n)$, and indices

$$1 \leq j_1 < j_2 < \cdots < j_k \leq n,$$

the function given by

$$g(x_{j_1}, x_{j_2}, \ldots, x_{j_k}) = \sum_{\ell} \sum_{x_{\ell}} f(x_1, x_2, \ldots, x_n)$$

where the $\ell$ ranges over the indices between 1 and $n$ not included in $j_1, j_2, \ldots, j_k$ is called the marginal distribution of $X_{j_1}, X_{j_2}, \ldots, X_{j_k}$.

Note. In the particular case we have discrete random variables $X$ and $Y$ along with the joint probability distribution $f(x, y)$, then the marginal distribution of $X$ is given by

$$g(x) = \sum_y f(x, y).$$

Example 63. Refer to Example 54 to calculate the marginal distributions of $X$ and $Y$, separately.

Note that

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/44</td>
<td>9/44</td>
<td>9/44</td>
<td>1/44</td>
</tr>
<tr>
<td>1</td>
<td>3/44</td>
<td>9/44</td>
<td>9/44</td>
<td>1/44</td>
</tr>
<tr>
<td>2</td>
<td>1/44</td>
<td>27/44</td>
<td>3/7</td>
<td>1/7</td>
</tr>
<tr>
<td>3</td>
<td>6/7</td>
<td>27/44</td>
<td>3/7</td>
<td>1/7</td>
</tr>
</tbody>
</table>

Then the marginal distribution of $X$ is given by

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g(x)$</td>
<td>9/44</td>
<td>27/44</td>
<td>3/7</td>
<td>1/7</td>
</tr>
</tbody>
</table>
and the marginal distribution of \( Y \) is given by
\[
\begin{array}{c|cccc}
 y & 0 & 1 & 2 & 3 \\
h(y) & \frac{13}{44} & \frac{21}{44} & \frac{9}{44} & \frac{1}{44}
\end{array}
\]

For the sake of simplicity, we will define the marginal density in the continuous case for only one of the random variables.

**Definition 61.** Given continuous random variables \( X_1, X_2, \ldots, X_n \) and a joint density \( f(x_1, x_2, \ldots, x_n) \), the **marginal density** of \( X_j \) is given by
\[
g(x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \, \cdots \, dx_{j-1} \, dx_{j+1} \cdots dx_n.
\]

**Note.** In the particular case we have continuous random variables \( X \) and \( Y \) along with a joint probability density \( f(x, y) \), then the marginal density of \( X \) is given by
\[
g(x) = \int_{-\infty}^{\infty} f(x, y) \, dy.
\]

**Example 64.** Given the joint probability density
\[
f(x, y) = \begin{cases} 
10x^2y, & 0 < y < x < 1; \\
0, & \text{else}
\end{cases}
\]
find the marginal densities of \( X \) and \( Y \), separately.

Observe that the region for which \( f(x, y) \) takes on positive values is
\[
\begin{tikzpicture}
\draw[->] (-1,0) -- (1,0) node[right] {1};
\draw[->] (0,-1) -- (0,1) node[above] {1};
\draw (0,0) -- (1,1);
\end{tikzpicture}
\]
To find the marginal density of \( X \), we calculate
\[
g(x) = \int_0^x 10x^2y \, dy = \frac{10x^2y^2}{2} \bigg|_{y=0}^{y=x} = 5x^4
\]
for \( 0 < x < 1 \).

For the marginal density of \( Y \), we calculate
\[
h(y) = \int_y^1 10x^2y \, dx = \frac{10y^3}{3} - \frac{10y^4}{3} = \frac{10y}{3} (1 - y^3)
\]
for \( 0 < y < 1 \).
3.6 Conditional Distributions and Densities

As motivation to the upcoming definition, suppose we have a joint distribution \( f(x, y) \) with two discrete random variables \( X \) and \( Y \). Then, to calculate \( P(Y = y) \), we would calculate the marginal marginal distribution \( g(y) \) of \( Y \) at \( y \).

Now, in the way of conditional probability,

\[
P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{g(y)}
\]

provided that \( P(Y = y) = g(y) \neq 0 \).

**Definition 62.** If \( f(x, y) \) is the joint probability distribution (resp. density) of two discrete (resp. continuous) random variables \( X \) and \( Y \) and \( g(y) \) is the marginal distribution (resp. density) of \( Y \), then the conditional distribution (resp. conditional density) of \( X \) given \( Y = y \) is

\[
f(x | y) = \frac{f(x, y)}{g(y)}
\]
as long as \( g(y) \neq 0 \).

**Note.** To find the conditional distribution (resp. density) of \( Y \) given \( X = x \), we use the marginal distribution (resp. density) \( g(x) \) of \( X \) where \( g(x) \neq 0 \):

\[
f(y | x) = \frac{f(x, y)}{g(x)}.
\]

**Example 65.** Refer to Example 63 to calculate the conditional distribution of \( X \) given that \( Y = 1 \).

Observe that the marginal distribution of \( Y \) at \( y = 1 \) is \( h(1) = \frac{21}{44} \). Given that \( Y = 1 \), \( X \) can only take values 0, 1, and 2. So we compute the conditional distribution of \( X \) given \( Y = 1 \):

\[
\begin{align*}
f(0 | 1) &= \frac{f(0, 1)}{h(1)} = \frac{3}{220} \cdot \frac{44}{21} = \frac{1}{35}, \\
f(1 | 1) &= \frac{f(1, 1)}{h(1)} = \frac{9}{55} \cdot \frac{44}{21} = \frac{12}{35}, \\
f(2 | 1) &= \frac{f(2, 1)}{h(1)} = \frac{3}{10} \cdot \frac{44}{21} = \frac{22}{35}.
\end{align*}
\]

**Definition 63.** If \( f(x_1, x_2, \ldots, x_n) \) is the joint probability distribution (resp. density) of the discrete (resp. continuous) random variables \( X_1, X_2, \ldots, X_n \) and \( f_j(x_j) \) is the marginal distribution (resp. density) of \( X_j \) for each \( 1 \leq j \leq n \), then the \( n \) random variables \( X_1, X_2, \ldots, X_n \) are said to be independent if

\[
f(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} f_j(x_j)
\]

for all tuples \((x_1, x_2, \ldots, x_n)\). Otherwise, we say that the random variables \( X_1, X_2, \ldots, X_n \) are dependent.

**Example 66.** Refer to Example 63 to determine whether or not the random variables \( X \) and \( Y \) are independent.

Notice that

\[
f(0, 0) = \frac{1}{1540} \neq \frac{39}{1694} = \frac{6}{77} \cdot \frac{13}{44} = g(0) \cdot h(0).
\]
That is, \( X \) and \( Y \) are dependent.

**Example 67.** Refer to Example 64 and determine whether or not the two continuous random variables there are independent.

We found the marginal density of \( X \) to be \( 5x^4 \) and the marginal density of \( Y \) to be \( \frac{10y}{3} (1 - y^3) \). Now, since
\[
10x^2 y \neq 5x^4 \cdot \frac{10y}{3} (1 - y^3)
\]
for all \( 0 < y < x < 1 \), \( X \) and \( Y \) are dependent random variables.

**Example 68.** Given the joint probability density
\[
f(x, y) = \begin{cases} 
3xy^2, & 0 < x < 1, \ 0 < y < 2; \\
0, & \text{else} 
\end{cases}
\]
for continuous random variables \( X \) and \( Y \), determine whether or not \( X \) and \( Y \) are independent.

We first calculate the marginal density of \( X \) to be
\[
\int_0^1 \frac{3xy^2}{4} \, dy = \frac{xy^3}{4} \bigg|_{y=2}^{y=0} = 2x
\]
and the marginal density of \( Y \) to be
\[
\int_0^1 \frac{3xy^2}{4} \, dx = \frac{3x^2y^2}{8} \bigg|_{x=1}^{x=0} = \frac{3y^2}{8}.
\]
Now, observe that
\[
2x \cdot \frac{3y^2}{8} = \frac{3xy^2}{4}
\]
which establishes the independence of \( X \) and \( Y \).

In light of Examples 67 and 68, note that one of the key differences in the definitions of the probability densities is the following. In Example 67, the domain of the probability density is defined in terms of a relationship between \( x \) and \( y \). By contrast, in Example 68, the domain of the probability density imposes no relationship between the inputs \( x \) and \( y \). Hence, in the continuous case, independence of random variables is sometimes linked to what kinds of regions the probability density assumes positive values on.

**Example 69.** Consider the probability density
\[
f(x, y) = \begin{cases} 
x + y, & 0 < x < 1, \ 0 < y < 1; \\
0, & \text{otherwise} 
\end{cases}
\]
for two random variables \( X \) and \( Y \). Determine whether or not \( X \) and \( Y \) are independent.

Notice that the marginal density of \( X \) is given by
\[
g(x) = \int_0^1 x + y \, dy = xy + \frac{y^2}{2} \bigg|_{y=0}^{y=1} = x + \frac{1}{2}
\]
and the marginal density of $Y$ is given by
\[ h(y) = \int_0^1 x + y \, dx = \frac{x^2}{2} + xy \bigg|_0^1 = y + \frac{1}{2}. \]

Since
\[ f(x, y) \neq g(x) \cdot h(y), \]
$X$ and $Y$ are dependent.

**Example 70.** Suppose $X$ and $Y$ are independent continuous random variables where
\[ g_X(x) = \begin{cases} \frac{2 \exp(-x^2)}{\sqrt{\pi}}, & x > 0; \\ 0, & x \leq 0 \end{cases} \]
is the probability density of $X$ and
\[ g_Y(y) = \begin{cases} y, & 0 < y < 1; \\ 0, & \text{otherwise} \end{cases} \]
is the probability density of $Y$. Find the joint probability density and $P(Y^2 \leq X)$.

Since $X$ and $Y$ are independent, the joint probability density is
\[ f(x, y) = \begin{cases} \frac{2 \exp(-x^2)y}{\sqrt{\pi}}, & x > 0, \ 0 < y < 1; \\ 0, & \text{otherwise.} \end{cases} \]

To calculate $P(Y^2 \leq X)$, we compute, for $x > 0$,
\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\sqrt{x}} f(x, y) \, dy \, dx = \int_0^{\infty} \int_0^{\sqrt{x}} \frac{2 \exp(-x^2)y}{\sqrt{\pi}} \, dy \, dx = \int_0^{\infty} \frac{2 \exp(-x^2)}{\sqrt{\pi}} \cdot \frac{y^2}{2} \bigg|_{0}^{\sqrt{x}} \, dx = \int_0^{\infty} x \exp(-x^2) \, dx = \lim_{b \to \infty} \int_0^{b} \frac{x \exp(-x^2)}{\sqrt{\pi}} \, dx
\]
To complete the integral, we can use substitution: $u = x^2$ and $du = 2x \, dx$ which provides
\[
\lim_{b \to \infty} \int_0^{b} \frac{x \exp(-x^2)}{\sqrt{\pi}} \, dx = \lim_{b \to \infty} \int_{u=0}^{u=b^2} \frac{\exp(-u)}{2\sqrt{\pi}} \, du = \lim_{b \to \infty} \left(-\exp(-u)\right) \bigg|_{0}^{b^2} = \lim_{b \to \infty} \frac{1}{2\sqrt{\pi}} - \exp(-b^2) = \frac{1}{2\sqrt{\pi}} \approx 28.21\%.
\]
Chapter 4

Expected Value and Moments

4.1 Expected Value

We introduce the idea of expected value through an example.

**Example 71.** Consider a game where you pay $2 to roll a 6-sided die and the outcome of the roll determines an award.

<table>
<thead>
<tr>
<th>outcome</th>
<th>award</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$5</td>
</tr>
<tr>
<td>2</td>
<td>$1</td>
</tr>
<tr>
<td>3</td>
<td>$0</td>
</tr>
<tr>
<td>4</td>
<td>$1</td>
</tr>
<tr>
<td>5</td>
<td>$15</td>
</tr>
<tr>
<td>6</td>
<td>$1</td>
</tr>
</tbody>
</table>

Since each outcome is equally likely, we compute the average:

$$\frac{5 + 1 + 0 + 1 + 15 + 1}{6} = \frac{23}{6} \approx 3.83.$$

In the computation above, we should note that

$$5 \cdot \frac{1}{6} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{6} + 15 \cdot \frac{1}{6} = \frac{23}{6},$$

where the award amount is multiplied by the probability of its occurrence. On average, we expect to win $3.83 and, since we payed $2 to play, we expect to earn $1.83.

Another way to view this example is to play 60 times. At $2 per game, the total cost comes out to be $120. Based on probabilities, we should expect

- 10 of those games to award us with $5,
- 30 of those games to award us with $1,
- 10 of those games to award us with nothing, and
- 10 of those games to award us with $15.

So we can expect to be awarded

$$10 \cdot 5 + 30 \cdot 1 + 10 \cdot 15 = 230.$$

after adjusting by the $120 cost, we see that on average, per game, we made

$$\frac{110}{60} \approx 1.83.$$

We can also incorporate the cost beforehand:
CHAPTER 4. EXPECTED VALUE AND MOMENTS

<table>
<thead>
<tr>
<th>outcome</th>
<th>earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$3</td>
</tr>
<tr>
<td>2</td>
<td>$-1</td>
</tr>
<tr>
<td>3</td>
<td>$-2</td>
</tr>
<tr>
<td>4</td>
<td>$-1</td>
</tr>
<tr>
<td>5</td>
<td>$13</td>
</tr>
<tr>
<td>6</td>
<td>$-1</td>
</tr>
</tbody>
</table>

Then the expected earnings per game, on average, is

$$3 \cdot \frac{1}{6} + (-1) \cdot \frac{1}{2} + (-2) \cdot \frac{1}{6} + 13 \cdot \frac{1}{6} \approx 1.83.$$  

**Definition 64.** Given a

- discrete random variable $X$ and its probability distribution $f(x)$, the **expected value** of $X$ is defined to be

$$E(X) = \sum x \cdot f(x),$$

provided the sum is finite.

- continuous random variable $X$ and its probability density $f(x)$, the **expected value** of $X$ is defined to be

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) \, dx,$$

provided the integral is finite.

**Remark.** The expected value of a random variable is not to be confused with the most likely outcome. As we’ve already seen, in fact, the expected value of a random variable need not even be a possible outcome. The expected value should be thought of in the following light: Given a random variable $X$, suppose you run a sequence of trials and obtain outcomes $x_1, x_2, x_3, \ldots, x_n$. Then, for large enough $n$,

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \approx E(X).$$

**Example 72.** Suppose $X$ assumes all values $3^n$ for integers $n \geq 1$ and $f(3^n) = \frac{1}{2^n}$. Then $f(x)$ is a probability distribution. For this $X$, the expected value isn’t defined since the series

$$\sum_{n=1}^{\infty} \frac{3^n}{2^n}$$

diverges.

**Example 73.** Consider the probability density

$$f(x) = \frac{1}{\pi(1+x^2)}$$

for a continuous random variable $X$. Then the expected value of $X$ is not defined since the integral

$$\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} \, dx$$

diverges.

**Example 74.** Refer to Example 38 and determine the expected value of $X$. 
4.1. EXPECTED VALUE

In Example 44, we provided the table

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
<th>$x$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1/36</td>
<td>8</td>
<td>5/36</td>
</tr>
<tr>
<td>3</td>
<td>2/36</td>
<td>9</td>
<td>4/36</td>
</tr>
<tr>
<td>4</td>
<td>3/36</td>
<td>10</td>
<td>3/36</td>
</tr>
<tr>
<td>5</td>
<td>4/36</td>
<td>11</td>
<td>2/36</td>
</tr>
<tr>
<td>6</td>
<td>5/36</td>
<td>12</td>
<td>1/36</td>
</tr>
<tr>
<td>7</td>
<td>6/36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

which then allows us to compute the expected value

$$E(X) = 7.$$ 

Example 75. Refer to Example 42 and determine the expected value of $X$.

Using

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>220/969</td>
</tr>
<tr>
<td>1</td>
<td>462/969</td>
</tr>
<tr>
<td>2</td>
<td>252/969</td>
</tr>
<tr>
<td>3</td>
<td>35/969</td>
</tr>
</tbody>
</table>

we compute

$$E(X) = 0 \cdot \frac{220}{969} + 1 \cdot \frac{462}{969} + 2 \cdot \frac{252}{969} + 3 \cdot \frac{35}{969} = \frac{21}{19} \approx 1.105.$$ 

Example 76. Find the expected value of the probability density

$$f(x) = \begin{cases} 
\frac{2 \exp(-x^2)}{\sqrt{\pi}}, & x > 0; \\
0, & x \leq 0.
\end{cases}$$

We evaluate

$$\int_{0}^{\infty} \frac{2x \exp(-x^2)}{\sqrt{\pi}} \, dx = \lim_{b \to \infty} \int_{0}^{b} \frac{2x \exp(-x^2)}{\sqrt{\pi}} \, dx = \lim_{b \to \infty} \int_{u=0}^{u=b^2} \frac{\exp(-u)}{\sqrt{\pi}} \, du = \lim_{b \to \infty} - \frac{\exp(-u)}{\sqrt{\pi}} \bigg|_{u=0}^{u=b^2} = \lim_{b \to \infty} \frac{1}{\sqrt{\pi}} - \frac{\exp(-b^2)}{\sqrt{\pi}} = \frac{1}{\sqrt{\pi}}.$$ 

Remark. As we’ve seen before, we can create new random variables from old by applying algebraic operations to them. In fact, more generally, we can take any real-valued function $g$ where its domain contains the values of a random variable $X$ and apply it to $X$ to obtain a new random variable, denoted $g(X)$. 
**Example 77.** Suppose $X$ has density

$$f(x) = \begin{cases} 0.5, & -1 < x < 1; \\ 0, & \text{otherwise}. \end{cases}$$

Use this to find the density of $X^2$.

First, for $Y = X^2$, notice that $P(Y < 0) = 0$ and that $P(Y \leq 1) = 1$. For any real number $0 < t < 1$, notice that $Y < t$ if and only if $-\sqrt{t} < X < \sqrt{t}$. Hence,

$$P(Y < t) = P(-\sqrt{t} < X < \sqrt{t}) = \int_{-\sqrt{t}}^{\sqrt{t}} 0.5 \, dx = \sqrt{t}.$$

That is, the cumulative density function for $Y$ is given by

$$G(y) = \begin{cases} 0, & y < 0; \\ \sqrt{y}, & 0 \leq y < 1; \\ 1, & 1 \leq y. \end{cases}$$

Thus, the density for $Y$ is given by

$$g(y) = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 < y < 1; \\ 0, & \text{otherwise}. \end{cases}$$

**Theorem 65.** Given a discrete random variable $X$ and its probability distribution $f(x)$, the expected value of $g(X)$ is given by

$$E(g(X)) = \sum_x g(x)f(x).$$

In the case that $X$ is a continuous random variable and $f(x)$ is a probability density for $X$, the expected value of $g(X)$ is given by

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x) \, dx.$$

**Example 78.** A 6-sided die is rolled and $X$ is the value observed. For $g(x) = 2x^3 + 1$, find $E(g(X))$.

We calculate

$$E(g(X)) = \sum_{j=1}^{6} (2j^3 + 1) \cdot \frac{1}{6} = 148.$$

**Remark.** One thing that Example 78 illuminates is that the expected value $E(g(X))$ is not equal to $g(E(X))$. In particular, for a 6-sided dice roll where $X$ is the value of the roll and $g(x) = 2x^3 + 1$, $E(X) = 3.5$, $g(E(X)) = 86.75$, whereas $E(g(X)) = 148$. This should be reasonable since the outcomes of $2X^3 + 1$ are $3, 17, 55, 129, 251,$ and $433$.

**Example 79.** In Flatland,1 recent research has shown that the radius of circles has the following probability density:

$$f(r) = \begin{cases} \frac{2\sqrt{2}}{\pi(1+r^4)}, & r > 0; \\ 0, & \text{otherwise}. \end{cases}$$

Find the expected value of a circle’s area.

1see the novel by Edwin A. Abbott
Notice that the area is given by $A(r) = \pi r^2$. Then
\[ E(A(r)) = \int_0^\infty \frac{2\sqrt{2} \cdot \pi r^2}{\pi(1 + r^4)} \, dr = \pi. \]

Exercise 15. An ice machine produces ice cubes and the probability density of the side length $X$ of the cubes is given by
\[ f(x) = \begin{cases} 2, & 1.75 < x < 2.25 \\ 0, & \text{otherwise.} \end{cases} \]
Find the expected value of the volume of the cubes.

Theorem 66. If $a$ and $b$ are constants, then
\[ E(aX + b) = a \cdot E(X) + b \]
for any random variable $X$. If $b = 0$, then $E(aX) = a \cdot E(X)$ and, if $a = 0$, then $E(b) = b$.

Theorem 67. If $a_1, a_2, \ldots, a_n$ are constants and $g_1, g_2, \ldots, g_n$ are real-valued functions with domains that contain the values of a random variable $X$, then
\[ E \left( \sum_{j=1}^n a_j \cdot g_j(X) \right) = \sum_{j=1}^n a_j \cdot E(g_j(X)). \]

Example 80. Given the probability density
\[ f(x) = \begin{cases} \frac{3 - 3(x - 1)^2}{4}, & 0 < x < 2; \\ 0, & \text{otherwise}, \end{cases} \]
find $E(X^n)$ where $n$ is a positive integer and $E((3X + 2)^2)$.

First, notice that
\[
E(X^n) = \int_0^2 x^n \cdot \frac{3 - 3(x - 1)^2}{4} \, dx
= \int_0^2 x^n \cdot \frac{-3x^2 + 6x}{4} \, dx
= \int_0^2 \frac{3x^{n+1} - 3x^{n+2}}{2} \, dx
= \frac{3x^{n+2}}{2(n+2)} - \frac{3x^{n+3}}{4(n+3)} \Big|_0^2
= \frac{6 \cdot 2^n}{(n+2)(n+3)}.
\]

Then,
\[
E((3X + 2)^2) = E(9X^2 + 12X + 4)
= 9E(X^2) + 12E(X) + 4
= 9 \cdot \frac{6 \cdot 2^2}{(2+2)(2+3)} + 12 \cdot \frac{6 \cdot 2}{(1+2)(1+3)} + 4
= \frac{134}{5}.
\]
Theorem 68. Suppose $X$ and $Y$ are discrete random variables where $f(x,y)$ is their joint probability distribution. If $g(x,y)$ is a real-valued function, then the expected value of $g(X,Y)$ is

$$E(g(X,Y)) = \sum_x \sum_y g(x,y)f(x,y).$$

On the other hand, if $X$ and $Y$ are continuous random variables and $f(x,y)$ is a joint probability density, then

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y) \, dx \, dy.$$

Example 81. Two cosmic membranes vibrate with frequency $X$ and $Y$ where the joint density is given by

$$f(x,y) = \begin{cases} 
\frac{e^{-y}}{3}, & 1 < x < 4, \, y > 0; \\
0, & \text{otherwise} 
\end{cases}$$

The subliminal vibrations experienced by neighboring universes is the sum of the frequencies, $X + Y$. Find $E(X + Y)$.

We compute

$$E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f(x,y) \, dx \, dy$$

$$= \int_{0}^{\infty} \int_{1}^{4} (x + y) \frac{e^{-y}}{3} \, dx \, dy$$

$$= \int_{0}^{\infty} \int_{1}^{4} \frac{xe^{-y}}{3} + \frac{ye^{-y}}{3} \, dx \, dy$$

$$= \int_{0}^{\infty} x^{2}e^{-y/3} + \frac{xe^{-y}}{3} \bigg|_{x=4}^{x=1} \, dy$$

$$= \int_{0}^{\infty} 5e^{-y/2} + ye^{-y} \, dy$$

$$= \frac{7}{2}.$$

Exercise 16. A particle on a square plate has the following probability density:

$$f(x,y) = \begin{cases} 
\frac{4xy}{9}, & 1 < x < 2, \, 1 < y < 2; \\
0, & \text{otherwise} 
\end{cases}$$

A force acts on these particles and the force experienced is expressed by

$$\varphi(x,y) = \frac{1}{x^2 + y^2}.$$

Find the expected value $E(\varphi(X,Y))$ of this force on the particles.

The following is a generalization of Theorem 67.

Theorem 69. If $a_1, a_2, \ldots, a_n$ are constants, $g_1, g_2, \ldots, g_n$ are real-valued functions with domains that contain the values of random variables $X_1, X_2, \ldots, X_m$, then

$$E \left( \sum_{j=1}^{n} a_j \cdot g_j(X_1, X_2, \ldots, X_m) \right) = \sum_{j=1}^{n} a_j \cdot E(g_j(X_1, X_2, \ldots, X_m)).$$

Remark. The real power of this result is that it allows one to find the expected value of multivariate distributions without explicitly knowing the joint distribution or density. We will elaborate more on this in Section 4.5.
4.2 Moments, Variance, and Chebyshev’s Inequality

The expected value of a random variable $X$ gives us some information about it. Intuitively, $X$ is in some sense centered at its expected value. In this section we will determine other numerical tools to help classify random variables. To elaborate, we will first consider two random variables with the same expected value but that look quite different.

**Example 82.** Consider two random variables $X$ and $Y$ where the probability distributions are given below:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
<th>$y$</th>
<th>$g(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-2$</td>
<td>$1/24$</td>
<td>$-2$</td>
<td>$1/2$</td>
</tr>
<tr>
<td>$-1$</td>
<td>$1/12$</td>
<td>$2$</td>
<td>$1/2$</td>
</tr>
<tr>
<td>$0$</td>
<td>$3/4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1$</td>
<td>$1/12$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$2$</td>
<td>$1/24$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now examine the two probability histograms corresponding to two random variables $X$ and $Y$ keeping in mind that $E(X) = E(Y) = 0$.

**Definition 70.** For a non-negative integer $n$, the $n^{th}$ moment about the origin of a random variable $X$ is defined to be

$$
\mu_n' = E(X^n).
$$

**Remark.** Notice that $\mu_0' = 1$ for any random variable $X$.

**Definition 71.** Given a random variable $X$, the **mean** of $X$ is $\mu = \mu_1' = E(X)$.

**Definition 72.** For a non-negative integer $n$, the $n^{th}$ moment about the mean of a random variable $X$ is defined to be

$$
\mu_n = E((X - \mu)^n).
$$

**Remark.** Notice that $\mu_0 = 1$ and $\mu_1 = 0$ for any random variable $X$ with a finite expected value.

**Definition 73.** Given a random variable $X$, the **variance** of $X$ is defined to be $\text{Var}(X) = \mu_2 = E((X - \mu)^2)$.

**Definition 74.** Given a random variable $X$, the **standard deviation** of $X$ is defined to be $\sigma = \sqrt{\text{Var}(X)}$.

**Remark.** Sometimes the variance of $X$ is denoted by $\sigma_X^2$ or, more simply, $\sigma^2$.

**Example 83.** Refer to Example 82 and calculate the variance and standard deviation for both $X$ and $Y$.

Notice that

$$
\text{Var}(X) = 4 \cdot \frac{1}{24} + 1 \cdot \frac{1}{12} + 0 \cdot \frac{3}{4} + 1 \cdot \frac{1}{12} + 4 \cdot \frac{1}{24} = \frac{1}{2}
$$

and

$$
\text{Var}(Y) = 4 \cdot \frac{1}{2} + 4 \cdot \frac{1}{2} = 4.
$$

It follows that the standard deviation of $X$ is $\frac{1}{\sqrt{2}}$ and the standard deviation of $Y$ is 2.
**Theorem 75.** For a random variable $X$,

$$\text{Var}(X) = E(X^2) - E(X)^2.$$  

**Proof.** Note that, since $\mu = E(X)$,

$$\text{Var}(X) = E((X - \mu)^2)$$

$$= E(X^2 - 2\mu X + \mu^2)$$

$$= E(X^2) - 2\mu E(X) + \mu^2$$

$$= E(X^2) - E(X)^2,$$

the desired end. \hfill \square

**Example 84.** Recall

$$f(x) = \begin{cases} 
\frac{3 - 3(x - 1)^2}{4}, & 0 < x < 2; \\
0, & \text{otherwise},
\end{cases}$$

from Example 80. Find the variance of $X$.

Recall that we showed that

$$E(X^n) = \frac{6 \cdot 2^n}{(n + 2)(n + 3)}.$$  

It follows that $\mu = 1$ and

$$\mu'_2 = E(X^2) = \frac{6}{5}.$$  

By Theorem 75, we see that

$$\text{Var}(X) = \frac{6}{5} - 1 = \frac{1}{5}.$$  

**Example 85.** For the density function

$$f(x) = \begin{cases} 
1, & 0 < x < 1; \\
0, & \text{otherwise},
\end{cases}$$

find the variance of $X$.

Notice that

$$\mu = E(X) = \int_0^1 x \, dx = \frac{1}{2}.$$  

Also,

$$\mu'_2 = E(X^2) = \int_0^1 x^2 \, dx = \frac{1}{3}.$$  

By Theorem 75, we find that $\text{Var}(X) = \frac{1}{12}$.

**Example 86.** Find a constant $k > 0$ so that the probability density

$$f(x) = \begin{cases} 
\frac{1}{2k}, & -k < x < k; \\
0, & \text{otherwise}
\end{cases}$$

for random variable $X$ has variance 1.
4.2. MOMENTS, VARIANCE, AND CHEBYSHEV’S INEQUALITY

First, we compute the mean:

$$\mu = \int_{-\infty}^{\infty} \frac{x^2}{2k} \, dx = \frac{x^2 |_{-\infty}^{\infty}}{4k} = 0.$$ 

Now,

$$\mu' = \int_{-\infty}^{\infty} x \frac{x^2}{2k} \, dx = \frac{x^3 |_{-\infty}^{\infty}}{6k} = \frac{k^2}{3}.$$ 

Since

$$\text{Var}(X) = \frac{k^2}{3},$$

we see that $k = \sqrt{3}$.

Theorem 76. For constants $a$ and $b$ and a random variable $X$,

$$\text{Var}(aX + b) = a^2 \cdot \text{Var}(X).$$

Proof. First, recall that $E(aX + b) = a \cdot E(X) + b$. Then, notice that

$$\text{Var}(aX + b) = E((aX + b - E(aX + b))^2) = E((aX + b - a \cdot E(X) - b)^2)$$

$$= E(a^2(X - E(X))^2)$$

$$= a^2 \cdot E(X^2 - 2X \cdot E(X) + E(X)^2)$$

$$= a^2 \cdot (E(X^2) - 2E(X) \cdot E(X) + E(X)^2)$$

$$= a^2 \cdot (E(X^2) - E(X)^2)$$

$$= a^2 \cdot \text{Var}(X),$$

as promised. \qed

Chebyshev’s Inequality demonstrates that the standard deviation of any random variable does indeed have implications for how it is spread out and thus justifies the nomenclature.

Theorem 77 (Chebyshev’s Inequality). Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma \neq 0$. Then, for any constant $k \geq 1$,

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$ 

Proof. We will prove it for the case when $X$ is a continuous random variable. The proof for discrete $X$ is similar. Note that

$$\sigma^2 = \text{Var}(X) = E((X - \mu)^2)$$

$$= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx$$

$$= \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) \, dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) \, dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) \, dx$$

which provides

$$\int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) \, dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) \, dx \leq \sigma^2$$

since $(x - \mu)^2 f(x)$ is non-negative. Now, behold that
• for $x \leq \mu - k\sigma$, $0 < k\sigma \leq \mu - x$ which implies $k^2\sigma^2 \leq (x - \mu)^2$ and

• for $\mu + k\sigma \leq x$, $0 < k\sigma \leq x - \mu$ which implies $k^2\sigma^2 \leq (x - \mu)^2$.

Hence,

$$
\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x - \mu)^2 f(x) \, dx + \int_{\mu+k\sigma}^{\infty} (x - \mu)^2 f(x) \, dx
\geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x) \, dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x) \, dx
= k^2\sigma^2 \cdot \left[ \int_{-\infty}^{\mu-k\sigma} f(x) \, dx + \int_{\mu+k\sigma}^{\infty} f(x) \, dx \right]
= k^2\sigma^2 \cdot [P(X \leq \mu - k\sigma) + P(X \geq \mu + k\sigma)]
$$

which further provides

$$
P(X \leq \mu - k\sigma) + P(X \geq \mu + k\sigma) \leq \frac{1}{k^2}.
$$

By noting that $|X - \mu| \geq k\sigma$ is equivalent to the disjunction $X - \mu \leq -k\sigma$ or $X - \mu \geq k\sigma$, we see that

$$
P(|X - \mu| < k\sigma) = 1 - P(|X - \mu| \geq k\sigma) \geq 1 - \frac{1}{k^2},
$$

the desired result.

\[\square\]

**Example 87.** To see that Chebyshev’s Inequality can’t be improved upon, consider the discrete random variable $X$ with probability distribution $f(x)$ defined as follows, where $k \geq 1$:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>$\frac{1}{2k^2}$</td>
</tr>
<tr>
<td>0</td>
<td>$1 - \frac{1}{k^2}$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{1}{2k^2}$</td>
</tr>
</tbody>
</table>

Then $\mu = E(X) = 0$ and $\text{Var}(X) = \frac{1}{k^2}$ which provides $\sigma = \frac{1}{k}$. Hence,

$$
P(|X - \mu| < k\sigma) = P(|X| < 1) = 1 - \frac{1}{k^2}.
$$

**Remark.** Besides variance, skewness and kurtosis which are also defined in terms of moments are often used to give information about the shape of a distribution/density. Though we won’t study them in detail here, the definitions are offered below for the curious reader.

**Definition 78.** For a random variable $X$, **Pearson’s moment coefficient of skewness** of $X$ is defined to be

$$
E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right].
$$

**Definition 79.** For a random variable $X$, the **kurtosis** of $X$ is defined to be

$$
E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right].
$$
4.3 The Median

**Definition 80.** Generally speaking, for a random variable $X$, the median will be a real number $m = \text{Med}(X)$ so that

$$P(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad P(X \geq m) \geq \frac{1}{2}.$$ 

If there are two values $\alpha < \beta$ which satisfy the median property, we observe the convention that

$$\text{Med}(X) = \frac{\alpha + \beta}{2}.$$ 

In the following examples, we will not only compute medians, but compare them to means and observe that the difference between the mean and the median can capture ideas of skewness.

**Example 88.** Suppose $X$ has probability density

$$f(x) = \begin{cases} e^{-x}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Find the median of $X$ and compare it to the mean.

First, note that,

$$P(X \leq t) = \int_0^t e^{-x} \, dx = 1 - e^{-t}.$$ 

Then solve

$$1 - e^{-t} = \frac{1}{2} \quad \Rightarrow \quad \ln(1/2) = t$$

which establishes $\text{Med}(X) = \ln(2)$.

On the other hand,

$$E(X) = \int_0^\infty xe^{-x} \, dx = 1.$$ 

In particular, $\text{Med}(X) < E(X)$.

**Example 89.** Consider the following probability distribution:

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Find the mean and median.

Consider the table

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X \leq x)$</td>
<td>0.1</td>
<td>0.2</td>
<td>0.6</td>
<td>0.8</td>
<td>1</td>
</tr>
<tr>
<td>$P(X \geq x)$</td>
<td>1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Then $\text{Med}(X) = 3$.

For the mean, we compute

$$E(X) = (1)(0.1) + (2)(0.1) + (3)(0.4) + (4)(0.2) + (5)(0.2) = 3.3.$$
Example 90. Consider the following probability distribution:

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>10</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.35</td>
<td>0.1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Find the mean and median.

Consider the table

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>10</th>
<th>56</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X \leq x)$</td>
<td>0.1</td>
<td>0.2</td>
<td>0.5</td>
<td>0.85</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>$P(X \geq x)$</td>
<td>1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.5</td>
<td>0.15</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Since both 3 and 4 satisfy the conditions for the median, we see that

\[
\text{Med}(X) = \frac{3 + 4}{2} = 3.5.
\]

For the mean, we compute

\[
E(X) = (1)(0.1) + (2)(0.1) + (3)(0.3) + (4)(0.35) + (10)(0.1) + (56)(0.05) = 6.4.
\]

4.4 Moment-generating Functions

Vaguely speaking, a moment-generating function for a random variable $X$, when it exists, captures all of the $n^{\text{th}}$ moments about the origin of $X$ in some way, justifying the terminology. Theorem 82 will make this more explicit. Perhaps more importantly, under some general assumptions, a random variable’s distribution or density is determined by the sequence of its moments about the origin. In this way, random variables can be fully described by their moment-generating functions.

Definition 81. The moment-generating function for a random variable $X$ is

\[
M_X(t) = E(e^{tX})
\]

whenever it exists.

Example 91. Find the moment-generating function of $X$ where the probability density of $X$ is given by

\[
f(x) = \begin{cases} 
  e^{-x}, & x > 0; \\
  0, & \text{otherwise}. 
\end{cases}
\]

Notice that

\[
\int_{-\infty}^{\infty} e^{tx} f(x) \, dx = \int_{0}^{\infty} e^{tx} \cdot e^{-x} \, dx = \int_{0}^{\infty} e^{(t-1)x} \, dx = \lim_{b \to \infty} \frac{e^{(t-1)b}}{t-1} - \frac{1}{t-1}.
\]

If $t \geq 1$, the value is undefined. On the other hand, for $t < 1$, we see that

\[
M_X(t) = \frac{1}{1-t}.
\]

Exercise 17. Consider the probability density

\[
f(x) = \frac{\sqrt{2}}{\pi(1 + x^4)}
\]

and show that $\mu_0'$, $\mu_1'$, and $\mu_2'$ exist but $\mu_j'$ fails to exist for all $j \geq 3$. 
4.4. MOMENT-GENERATING FUNCTIONS

**Theorem 82.** For a random variable $X$ where all of its moments $\mu'_n$ about the origin and its moment-generating function $M_X(t)$ exist,

$$M_X(t) = \sum_{n=0}^{\infty} E(X^n) \cdot \frac{x^n}{n!}.$$

It follows that

$$\mu'_n = E(X^n) = \frac{d^n M_X(t)}{dt^n} \bigg|_{t=0}.$$

**Proof.** We will prove this for the case when $X$ is continuous. The case when $X$ is discrete is similar. Recall that

$$e^{tx} = \sum_{n=0}^{\infty} \frac{t^n x^n}{n!},$$

which provides

$$E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \cdot f(x) \, dx$$

$$= \int_{-\infty}^{\infty} \left( \sum_{n=0}^{\infty} \frac{t^n x^n}{n!} \right) \cdot f(x) \, dx$$

$$= \sum_{n=0}^{\infty} \frac{t^n}{n!} \cdot \int_{-\infty}^{\infty} x^n \cdot f(x) \, dx$$

$$= \sum_{n=0}^{\infty} \frac{t^n}{n!} \cdot E(X^n)$$

by properties of convergent series. It follows that

$$\frac{d^n M_X(t)}{dt^n} = \mu'_n + \sum_{j=n+1}^{\infty} \frac{n! \cdot t^{j-n}}{j!} \cdot \mu'_j$$

which yields

$$\mu'_n = \frac{d^n M_X(t)}{dt^n} \bigg|_{t=0}$$

concluding the proof.

**Example 92.** For the probability density

$$f(x) = \begin{cases} 1, & 0 < x < 1; \\ 0, & \text{otherwise}, \end{cases}$$

find the moment-generating function for $X$ and use it to find the expected value of $X$.

Note that

$$M_X(t) = \int_{0}^{1} e^{tx} \, dx = \frac{e^{tx}}{t} \bigg|_{x=0}^{x=1} = \frac{e^t - 1}{t}.$$ 

Now,

$$\frac{dM_X(t)}{dt} = \frac{te^t - e^t + 1}{t^2}.$$
so
\[
\left. \frac{dM_X(t)}{dt} \right|_{t=0} = \lim_{t \to 0} \frac{te^t - e^t + 1}{t^2} = \lim_{t \to 0} \frac{te^t + e^t - e^t}{2t} = \frac{e^0}{2} = \frac{1}{2}.
\]

**Theorem 83.** For constants \( a \) and \( b \),
\[
M_{X+b/a}(t) = \exp \left( \frac{bt}{a} \right) \cdot M_X \left( \frac{t}{a} \right).
\]

**Proof.** Behold that
\[
M_{X+b/a}(t) = E \left( \exp \left( \frac{X + b}{a} \cdot t \right) \right) = E \left( \exp \left( \frac{tX}{a} + \frac{bt}{a} \right) \right) = E \left( \exp \left( \frac{tX}{a} \right) \cdot \exp \left( \frac{bt}{a} \right) \right) = \exp \left( \frac{bt}{a} \right) \cdot E \left( \exp \left( \frac{t}{a} \cdot X \right) \right) = \exp \left( \frac{bt}{a} \right) \cdot M_X \left( \frac{t}{a} \right),
\]
the promised conclusion.

**Theorem 84.** Given a random variable \( X \) with finite mean \( \mu \) and finite standard deviation \( \sigma \), the random variable \( Y = \frac{X - \mu}{\sigma} \) has mean 0 and a standard deviation of 1.

Moreover, if the moment-generating function \( M_X(t) \) for \( X \) exists, then
\[
M_Y(t) = \exp \left( -\frac{\mu t}{\sigma} \right) \cdot M_X \left( \frac{t}{\sigma} \right).
\]

**Proof.** For the mean, notice that
\[
E(Y) = E \left( \frac{X - \mu}{\sigma} \right) = \frac{E(X) - \mu}{\sigma} = 0
\]
by Theorem 66.

Now, for the standard deviation, notice that
\[
\text{Var}(Y) = \text{Var} \left( \frac{X - \mu}{\sigma} \right) = \frac{\text{Var}(X)}{\sigma^2} = 1
\]
by Theorem 76.

The rest follows from Theorem 83.
4.5 Product Moments and Covariance

**Definition 85.** For two random variables $X$ and $Y$, the $n^{\text{th}}$ and $m^{\text{th}}$ product moment about the origin of $X$ and $Y$ is

$$\mu_{n,m}' = E(X^n Y^m).$$

**Definition 86.** For two random variables $X$ and $Y$, the $n^{\text{th}}$ and $m^{\text{th}}$ product moment about the mean of $X$ and $Y$ is

$$\mu_{n,m} = E((X - E(X))^n(Y - E(Y))^m).$$

**Definition 87.** For two random variables $X$ and $Y$, the covariance of $X$ and $Y$ is defined to be

$$\text{cov}(X,Y) = E((X - E(X))(Y - E(Y))).$$

The covariance of two random variables $X$ and $Y$ measures in some way how linearly $X$ and $Y$ are related. We’ll address this in more detail in Example 97.

**Theorem 88.** For a random variable $X$,

$$\text{cov}(X,X) = \text{Var}(X).$$

**Proof.** Notice that

$$\text{cov}(X,X) = E[(X - E(X))(X - E(X))]$$

$$= E[X^2 - 2X \cdot E(X) + E(X)^2]$$

$$= E(X^2) - 2E(X)^2 + E(X)^2$$

$$= E(X^2) - E(X)^2$$

$$= \text{Var}(X).$$

**Example 93.** Using the joint density provided in Example 64, find the covariance of $X$ and $Y$.

Recall that the joint density is

$$f(x,y) = \begin{cases} 
10x^2y, & 0 < y < x < 1; \\
0, & \text{else}
\end{cases}$$

the marginal density of $X$ is given by $g(x) = 5x^4$ for $0 < x < 1$ and the marginal density of $Y$ is given by $h(y) = \frac{10y}{3}(1 - y^3)$ for $0 < y < 1$. Then

$$E(X) = \int_0^1 5x^5 \, dx = \frac{5}{6}$$

and

$$E(Y) = \int_0^1 \frac{10y^2}{3} (1 - y^3) \, dy = \frac{5}{9}.$$ 

Lastly, we calculate

$$\text{cov}(X,Y) = \int_0^1 \int_0^x \left( x - \frac{5}{6} \right) \left( y - \frac{5}{9} \right) \cdot 10x^2y \, dy \, dx = \frac{5}{378}.$$ 

**Theorem 89.** For random variables $X$ and $Y$,

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y).$$
Proof. By properties of expected value,
\[
\text{cov}(X, Y) = E((X - E(X))(Y - E(Y)))
\]
\[
= E(XY - X \cdot E(Y) - Y \cdot E(X) + E(X) \cdot E(Y))
\]
\[
= E(XY) - E(X) \cdot E(Y) - E(Y) \cdot E(X) + E(X) \cdot E(Y)
\]
\[
= E(XY) - E(X) \cdot E(Y).
\]

Example 94. For the bivariate distribution given in Example 54, find the covariance.

Recall that the joint distribution and marginal distributions are given by

\[
\begin{array}{c|cccc}
\hline
x & 0 & 1 & 2 & 3 \\
\hline
y & \frac{1}{1540} & \frac{9}{385} & \frac{9}{70} & \frac{13}{77} \\
\hline
1 & \frac{3}{220} & \frac{9}{55} & \frac{3}{11} & \frac{21}{44} \\
2 & \frac{9}{220} & \frac{9}{55} & \frac{9}{44} & \frac{6}{77} \\
3 & \frac{1}{44} & \frac{27}{77} & \frac{3}{7} & \frac{1}{7} \\
\hline
\end{array}
\]

Then
\[
E(X) = \frac{27}{77} + 2 \cdot \frac{3}{7} + 3 \cdot \frac{1}{7} = \frac{18}{11}
\]

and
\[
E(Y) = \frac{21}{44} + 2 \cdot \frac{9}{44} + 3 \cdot \frac{1}{44} = \frac{21}{22}.
\]

Also,
\[
E(XY) = \frac{9}{55} + 2 \cdot \frac{3}{10} + 2 \cdot \frac{9}{55} = \frac{12}{11}.
\]

Finally,
\[
\text{cov}(X, Y) = \frac{12}{11} - \frac{18}{11} \cdot \frac{21}{22} = -\frac{57}{121}.
\]

Example 95. Rework Example 93 using Theorem 89.

Notice that
\[
E(XY) = \int_0^1 \int_0^x 10x^3y^2 \, dy \, dx = \frac{10}{21}.
\]

Then
\[
\text{cov}(X, Y) = \frac{10}{21} - \frac{5}{6} \cdot \frac{5}{9} = \frac{5}{378}.
\]

Theorem 90. If \( X \) and \( Y \) are independent, then \( E(XY) = E(X) \cdot E(Y) \) and \( \text{cov}(X, Y) = 0 \).
Proof. We will prove this for the continuous case. The discrete case is similar. Let \( f(x, y) \) be the joint density, \( g(x) \) be the marginal density of \( X \), and \( h(y) \) be the marginal density of \( Y \). By independence, \( f(x, y) = g(x) \cdot h(y) \). Then, notice that

\[
E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f(x, y) \, dx \, dy
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot g(x) \cdot h(y) \, dx \, dy
= \int_{-\infty}^{\infty} y \cdot h(y) \cdot \int_{-\infty}^{\infty} x \cdot g(x) \, dx \, dy
= E(X) \cdot \int_{-\infty}^{\infty} y \cdot h(y) \, dy
= E(X) \cdot E(Y).
\]

Lastly, as \( \text{cov}(X, Y) = E(XY) - E(X) \cdot E(Y) \), we see that \( \text{cov}(X, Y) = 0 \).

Though independent random variables have zero covariance, the converse is not generally true.

Example 96. Consider random variables \( X \) and \( Y \) with their joint density given by

\[
f(x, y) = \begin{cases} \frac{3}{4}, & -1 < x < 1, \ 0 < y < 1 - x^2; \\ 0, & \text{otherwise}. \end{cases}
\]

Show that \( X \) and \( Y \) are dependent and that \( \text{cov}(X, Y) = 0 \).

First, we note that the marginal density of \( X \) is

\[
g(x) = \int_{0}^{\sqrt{1-x^2}} \frac{3}{4} \, dy = \frac{3}{4} (1 - x^2)
\]

and that the marginal density of \( Y \) is

\[
h(y) = \int_{-\sqrt{1-y}}^{\sqrt{1-y}} \frac{3}{4} \, dx = \frac{3\sqrt{1-y}}{2}
\]

It is clear that \( g(x) \cdot h(y) \neq f(x, y) \) establishing that \( X \) and \( Y \) are dependent.

Now, we calculate

\[
E(XY) = \int_{-1}^{1} \int_{0}^{\sqrt{1-x^2}} \frac{3xy}{4} \, dy \, dx = 0.
\]

Also,

\[
E(X) = \int_{-1}^{1} \frac{3x}{4} (1 - x^2) \, dx = 0
\]

and

\[
E(Y) = \int_{0}^{1} \frac{3y \cdot \sqrt{1-y}}{2} \, dy = \frac{2}{5}.
\]

Hence,

\[
\text{cov}(X, Y) = E(XY) - E(X) \cdot E(Y) = 0.
\]

Proposition 91. For random variables \( X \) and \( Y \) and constants \( a, b, c, d \),

\[
\text{cov}(aX + b, cY + d) = ac \cdot \text{cov}(X, Y).
\]
Proof. Notice that

\[
\text{cov}(aX + b, cY + d) = E[(aX + b)(cY + d)] - E(aX + b) \cdot E(cY + d) = E[acXY + adX + bcY + bd] - [a \cdot E(X) + b] \cdot [c \cdot E(Y) + d] = ac \cdot E(XY) + ad \cdot E(X) + bc \cdot E(Y) + bd
\]

\[
- (ac \cdot E(X) \cdot E(Y) + ad \cdot E(X) + bc \cdot E(Y) + bd) = ac \cdot (E(XY) - E(X) \cdot E(Y)) = ac \cdot \text{cov}(X, Y),
\]

the promised conclusion. \qed

Example 97. Let \(X\) be a random variable and \(Y = aX + b\) for constants \(a\) and \(b\). Then note that, utilizing Proposition 91 and Theorem 88,

\[
\text{cov}(X, Y) = \text{cov}(X, aX + b) = a \cdot \text{cov}(X, X) = a \cdot \text{Var}(X).
\]

Also, recall that \(\text{Var}(Y) = a^2 \cdot \text{Var}(X)\). Then let \(\sigma_X = \sqrt{\text{Var}(X)}\) and

\[
\sigma_Y = \sqrt{\text{Var}(Y)} = a \cdot \sigma_X
\]

and notice that

\[
\frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{a \sigma_X^2} = \frac{\text{cov}(X, Y)}{a \cdot \text{Var}(X)} = 1.
\]

We will address some properties of this quantity

\[
\frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}
\]

in Theorem 92.

Remark. Behold that computing the covariance of \(X\) and \(Y\) in Example 97 did not require us to know the explicit joint distribution/density of \(X\) and \(Y\). In fact, finding an explicit joint distribution/density would be impossible since we weren’t given a distribution/density for \(X\). In Example 98 we will elaborate more on the task of finding a joint density of two random variables in a particular scenario.

Theorem 92. For two random variables \(X\) and \(Y\),

\[
E(XY)^2 \leq E(X^2) \cdot E(Y^2).
\]

Moreover,

\[
\text{cov}(X, Y)^2 \leq \text{Var}(X) \cdot \text{Var}(Y).
\]

In particular, if \(X\) and \(Y\) are random variables with finite and positive variance, we have

\[-1 \leq \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \leq 1.
\]

Proof. For any real number \(t\), notice that

\[
0 \leq E((tX - Y)^2) = E(t^2X^2 - 2tXY + Y^2) = t^2E(X^2) - 2t \cdot E(XY) + E(Y^2).
\]

Now, \(E(X^2)t^2 - 2E(XY)t + E(Y^2)\) is a quadratic in terms of \(t\) and has at most one zero by the inequality above. Hence,

\[
4E(XY)^2 - 4E(X^2) \cdot E(Y^2) \leq 0 \implies E(XY)^2 \leq E(X^2) \cdot E(Y^2).
\]
Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$. Using Proposition 91, we have
\[
\text{cov}(X, Y)^2 = \text{cov}(X - \mu_X, Y - \mu_Y)^2 = (E[(X - \mu_X)(Y - \mu_Y)] - E(X - \mu_X) \cdot E(Y - \mu_Y))^2 = E[(X - \mu_X)(Y - \mu_Y)]^2 \leq E((X - \mu_X)^2) \cdot E((Y - \mu_Y)^2) = \text{Var}(X) \cdot \text{Var}(Y).
\]

It follows that, if $\text{Var}(X)$ and $\text{Var}(Y)$ are positive and finite,
\[
\frac{\text{cov}(X, Y)^2}{\text{Var}(X) \cdot \text{Var}(Y)} \leq 1.
\]
The rest is straightforward.

**Example 98.** Suppose $X$ has distribution
\[
\begin{array}{c|c}
  x & f(x) \\
  \hline
  -1 & \frac{1}{3} \\
  0 & \frac{1}{3} \\
  1 & \frac{1}{3}
\end{array}
\]

Use this to find the density of $X^2$ and the joint density of $X$ and $X^2$. Then show that $\text{cov}(X, X^2) = 0$ and that $X$ and $X^2$ are dependent.

Immediately, the distribution of $Y = X^2$ is
\[
\begin{array}{c|c}
  y & g(y) \\
  \hline
  0 & \frac{1}{3} \\
  1 & \frac{2}{3}
\end{array}
\]
Notice that the joint distribution is given by
\[
\begin{array}{c|ccc}
  x & -1 & 0 & 1 \\
  \hline
  y & 0 & 1/3 & 0 & 1/3 \\
  1/3 & 1 & 1/3 & 2/3 & 1/3 \\
\end{array}
\]

Immediately, $X$ and $Y$ are dependent since
\[
P(X = -1, Y = 0) = 0 \neq \frac{1}{3} \cdot \frac{1}{3} = P(X = -1) \cdot P(Y = 0).
\]
Lastly, observe that
\[
\text{cov}(X, Y) = E(XY) - E(X) \cdot E(Y) = E(X^3) = 0.
\]

As a generalization of Theorem 90, we have

**Theorem 93.** If $X_1, X_2, \ldots, X_n$ are independent random variables,
\[
E(X_1 \cdot X_2 \cdots X_n) = E(X_1) \cdot E(X_2) \cdots E(X_n).
\]
Theorem 94. For constants $a_1, a_2, \ldots, a_n$ and random variables $X_1, X_2, \ldots, X_n$,
\[
E \left( \sum_{j=1}^{n} a_j X_j \right) = \sum_{j=1}^{n} a_j E(X_j)
\]
and
\[
\text{Var} \left( \sum_{j=1}^{n} a_j X_j \right) = \sum_{j=1}^{n} a_j^2 \cdot \text{Var}(X_j) + 2 \cdot \sum_{j=2}^{n-1} \sum_{k=1}^{j-1} a_j a_k \cdot \text{cov}(X_j, X_k).
\]

Corollary 95. For constants $a_1, a_2, \ldots, a_n$ and random variables $X_1, X_2, \ldots, X_n$ which are independent,
\[
\text{Var} \left( \sum_{j=1}^{n} a_j X_j \right) = \sum_{j=1}^{n} a_j^2 \cdot \text{Var}(X_j).
\]

The following generalizes Proposition 91.

Theorem 96. For constants $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ and random variables $X_1, X_2, \ldots, X_n$, if
\[
Y_1 = \sum_{j=1}^{n} a_j X_j \quad \text{and} \quad Y_2 = \sum_{j=1}^{n} b_j X_j,
\]
we have that
\[
\text{cov}(Y_1, Y_2) = \sum_{j=1}^{n} a_j b_j \cdot \text{Var}(X_j) + \sum_{j=2}^{n-1} \sum_{k=1}^{j-1} (a_j b_k + a_k b_j) \cdot \text{cov}(X_j, X_k).
\]

Corollary 97. For constants $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ and independent random variables $X_1, X_2, \ldots, X_n$, if
\[
Y_1 = \sum_{j=1}^{n} a_j X_j \quad \text{and} \quad Y_2 = \sum_{j=1}^{n} b_j X_j,
\]
we have that
\[
\text{cov}(Y_1, Y_2) = \sum_{j=1}^{n} a_j b_j \cdot \text{Var}(X_j).
\]

Definition 98. If $X$ is a discrete random variable and $f(x|y)$ is the conditional probability distribution of $X$ given $Y = y$, then, for a real-valued function $g(x)$ defined on the range of $X$, the conditional expectation of $g(X)$ given $Y = y$ is
\[
E(g(X)|y) = \sum_x g(x) \cdot f(x|y)
\]
where $x$ ranges over all possible values of $X$. If, on the other hand, $X$ is a continuous random variable and $f(x|y)$ is the conditional probability density of $X$ given $Y = y$, then the conditional expectation of $g(X)$ given $Y = y$ is
\[
E(g(X)|y) = \int_{-\infty}^{\infty} g(x) \cdot f(x|y) \, dx.
\]

Definition 99. Of particular interest would be the conditional mean of the random variable $X$ given $Y = y$ which is defined to be
\[
E(X|y)
\]
and the conditional variance of the random variable $X$ given $Y = y$ which is defined to be
\[
E((X - E(X|y))^2|y) = E(X^2|y) - E(X|y)^2.
\]
4.6 The Weak Law of Large Numbers

**Definition 100.** A finite collection of random variables \(X_1, X_2, \ldots, X_n\) are said to be **independent and identically distributed**, abbreviated IID, if they are all mutually independent and have the same distributions/densities. Moreover, the random variable defined to be

\[
\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

is called the **sample average**.

**Theorem 101.** For a sequence of IID random variables \(X_1, X_2, \ldots, X_n\) with \(\mu = E(X_j)\) and \(\sigma^2 = \text{Var}(X_j)\) for \(0 \leq j \leq n\),

\[
E(\bar{X}_n) = \mu
\]

and

\[
\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.
\]

**Proof.** This follows immediately from Theorem 94 and Corollary 95.

**Theorem 102 (Weak Law of Large Numbers).** For a sequence of IID random variables \(X_1, X_2, \ldots, X_n\) with \(\mu = E(X_j)\) and \(\sigma^2 = \text{Var}(X_j)\) for \(0 \leq j \leq n\),

\[
P(|\bar{X}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}
\]

for any \(\varepsilon > 0\).

**Proof.** By Theorem 101, we have that

\[
E(\bar{X}_n) = \mu
\]

and

\[
\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.
\]

Let \(k = \frac{\sqrt{n} \cdot \varepsilon}{\sigma}\) and notice that Chebyshev’s Inequality yields

\[
P\left(|\bar{X}_n - \mu| < k \cdot \frac{\sigma}{\sqrt{n}}\right) = P\left(|\bar{X}_n - \mu| < \varepsilon\right) \geq 1 - \frac{1}{k^2} = 1 - \frac{\sigma^2}{n \cdot \varepsilon^2}.
\]

\(\square\)
Chapter 5

Particular Probability Distributions and Densities

5.1 Discrete Distributions

5.1.1 Uniform Distributions

Definition 103. A discrete random variable $X$ with $n$ distinct values $x_1, x_2, \ldots, x_n$ has a discrete uniform distribution if its probability distribution is given by

$$f(x_j) = \frac{1}{n}.$$

Proposition 104. For a discrete uniform random variable $X$ with values $x_1, x_2, \ldots, x_n$,

$$E(X) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

and

$$\text{Var}(X) = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2.$$

Proposition 105. For a discrete uniform random variable $X$ with values $x_1, x_2, \ldots, x_n$, the moment-generating function is

$$M_X(t) = \frac{e^{tx_1} + e^{tx_2} + \cdots + e^{tx_n}}{n}.$$

Example 99. The rolling of a fair 6-sided die produces a uniformly distributed random variable.

5.1.2 Bernoulli Distributions

Definition 106. A random variable $X$ with two values coded by 0 and 1 where 1 is known as success and 0 is known as failure has a Bernoulli distribution if its probability distribution is given by

$$f(x) = \begin{cases} 1 - p, & x = 0; \\ p, & x = 1 \end{cases}$$

where $0 \leq p \leq 1$. We also denote Bernoulli distributions with

$$f(x; p) = p^x(1 - p)^{1-x}.$$

In this case, we will say that $X$ is a Bernoulli random variable with parameter $p$, denoted $X \sim \text{Bernoulli}(p)$. 

73
Proposition 107. For a Bernoulli random variable $X$ with parameter $p$, 

$$E(X) = p$$

and 

$$\text{Var}(X) = p(1 - p)$$

Proposition 108. For a Bernoulli random variable $X$ with parameter $p$, the corresponding moment-generating function is 

$$M_X(t) = pe^t - p + 1.$$ 

Example 100. The rolling of a fair 6-sided die where rolling a 6 is considered the success case is a Bernoulli random variable.

5.1.3 Binomial Distributions

A binomial experiment has the following components:

- a fixed number of trials, $n$
- each trial is independent
- each trial has a Bernoulli distribution with a shared probability of success, $p$

In the context of binomial experiments, we count the number of successes after the $n$ trials, a random variable. Then, the probability that there are $k$ successes out of the $n$ trials is 

$$\binom{n}{k} p^k (1 - p)^{n-k}$$

Note. We can think of the underlying sample space as $n$-length sequences of 0s and 1s; e.g.

$$(0, 0, 1, 1, 0, 1, 1, 1, \ldots, 0, 0, 0)$$

where each coordinate corresponds to the individual Bernoulli random variable. Then the binomial random variable $X$ is the sum of the entries in the sequence.

Definition 109. A binomial distribution is determined by a positive integer $n$ and $0 \leq p \leq 1$ and is given by 

$$b(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for $0 \leq k \leq n$. A random variable $X$ is called a binomial random variable with parameters $n$ and $p$, denoted $X \sim B(n, p)$ if its distribution is given by $b(k; n, p)$.

Remark. A random variable $X$ with a binomial distribution is the sum of a finite sequence of IID Bernoulli variables $Y_1, Y_2, \ldots, Y_n$; namely, 

$$X = Y_1 + Y_2 + \cdots + Y_n$$

Theorem 110. For a binomial distribution with $n$ and $p$ given, 

$$b(k; n, p) = b(n - k; n, 1 - p)$$

for any $0 \leq k \leq n$.

Proof. Note that 

$$b(n - k; n, 1 - p) = \binom{n}{n-k} (1-p)^{n-k} p^{n-(n-k)} = \binom{n}{k} p^k (1-p)^{n-k} = b(k; n, p).$$

\qed
Theorem 111. For a random variable \( X \sim B(n, p) \),
\[
E(X) = np
\]
and
\[
Var(X) = np(1 - p).
\]

Proof. Let \( Y \) be the underlying Bernoulli random variable and consider \( n \) independent copies \( Y_1, Y_2, \ldots, Y_n \) of \( Y \). Then \( X = Y_1 + Y_2 + \cdots + Y_n \). From this, we see that
\[
E(X) = \sum_{j=1}^{n} E(Y_j) = np.
\]
Similarly,
\[
Var(Y) = Var(Y_1 + Y_2 + \cdots + Y_n) = \sum_{j=1}^{n} Var(Y_j) = n(p - p^2) = np(1 - p).
\]

\( \Box \)

Theorem 112. The moment-generating function for a random variable \( X \sim B(n, p) \) is
\[
M_X(t) = (pe^t - p + 1)^n.
\]

Proof. Using the Binomial Theorem, notice that
\[
(pe^t - p + 1)^n = (pe^t + (1 - p))^n = \sum_{k=0}^{n} \binom{n}{k} p^k e^{tk} (1 - p)^{n-k} = e^{tx} \cdot \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = E(e^{tx}) = M_X(t).
\]

\( \Box \)

Example 101. Let rolling a fair 6-sided die where the success case is rolling a 6 be the underlying Bernoulli random variable. Then, if we roll the die 25 times and let \( X \) be the number of 6s appearing, \( X \) is a binomial random variable.

5.1.4 Negative Binomial and Geometric Distributions

Just as binomial distributions, a negative binomial distribution is concerned with a sequence of IID Bernoulli random variables. The difference here is that we are fixing the number \( k \) of successes and measuring the probability that the \( k \)th success happens on the \( n \)th trial. By noting the first place where the \( k \)th success is obtained, we have a random variable whose values form the set
\[
\{k, k + 1, k + 2, k + 3, \ldots\} = \{k + j : j \in \mathbb{N} \cup \{0\}\}.
\]

Let \( p \) be the probability of success in the underlying Bernoulli random variable. Then, for the \( k \)th success to occur on the \( n \)th trial, there must be \( k - 1 \) successes in the first \( n - 1 \) trials, which can be calculated using the binomial distribution for \( n \) and \( p \); namely, \( b(k - 1; n - 1, p) \). On the \( n \)th independent trial, we assume a success, which has probability \( p \). By independence, the probability that the \( k \)th success occurs on the \( n \)th trial is
\[
p \cdot b(k - 1; n - 1, p) = p \cdot \binom{n - 1}{k - 1} p^{k-1} (1 - p)^{n-1-(k-1)} = \binom{n - 1}{k - 1} p^k (1 - p)^{n-k}.
\]
Definition 113. A **negative binomial distribution** is determined by a positive integer \( k \) and \( 0 \leq p \leq 1 \) and is given by

\[
b^*(n; k, p) = \binom{n-1}{k-1} p^k (1-p)^{n-k}
\]

for \( n \geq k \). A random variable \( X \) is called a **negative binomial random variable** with parameters \( k \) and \( p \), denoted \( X \sim \text{NB}(k, p) \) if its distribution is given by \( b^*(n; k, p) \).

Theorem 114. Given \( k \) and \( p \),

\[
b^*(n; k, p) = \frac{k}{n} \cdot b(k; n, p).
\]

Proof. Notice that

\[
\frac{k}{n} \cdot b(k; n, p) = \frac{k}{n} \cdot \binom{n}{k} p^k (1-p)^{n-k}
\]

\[
= \frac{k}{n} \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}
\]

\[
= \frac{(n-1)!}{(k-1)!(n-k)!} p^k (1-p)^{n-k}
\]

\[
= \binom{n-1}{k-1} p^k (1-p)^{n-k}
\]

\[
= b^*(n; k, p).
\]

To justify the naming, let us recall a consequence of the Binomial Theorem:

\[
(1 + x)^n = \sum_{j=0}^{n} \binom{n}{j} x^j
\]

for positive integer \( n \). Note that this is the Maclaurin series (Taylor series centered at zero) for the function \((1 + x)^n\).

With this in mind, let’s also recall that

\[
(1 + x)^{-n} = \sum_{j=0}^{\infty} (-1)^j \binom{n+j-1}{j} x^j
\]

for positive integer \( n \). We define

\[
\binom{-n}{j} = (-1)^j \binom{n+j-1}{j}
\]

which allows us to write

\[
(1 + x)^{-n} = \sum_{j=0}^{\infty} \binom{-n}{j} x^j
\]

Now, we have a similar equality as in the Binomial Theorem:

\[
(x + y)^{-n} = y^{-n} \cdot \left(1 + \frac{x}{y}\right)^{-n}
\]

\[
= y^{-n} \cdot \sum_{j=0}^{\infty} \binom{-n}{j} \left(\frac{x}{y}\right)^j
\]

\[
= \sum_{j=0}^{\infty} \binom{-n}{j} x^j y^{-n-j}
\]
Theorem 115. For $k$ and $p$ given,

$$b^*(k + j; k, p) = \binom{k + j - 1}{j} p^k (1 - p)^j$$

where $j = 0, 1, 2, \ldots$

Proof. Using the symmetry of choosing,

$$b^*(k + j; k, p) = \binom{k + j - 1}{k - 1} p^k (1 - p)^{k + j - k}$$

$$= \binom{k + j - 1}{k + j - 1 - (k - 1)} p^k (1 - p)^j$$

$$= \binom{k + j - 1}{j} p^k (1 - p)^j$$

as promised. \qed

Now, we see that the coefficient in the representation of the negative binomial distribution offered by Theorem 115 is the combinatorial coefficient appearing in the Maclaurin series for $(1 + x)^{-n}$ earning the name of negative binomial distribution.

Theorem 116. For a random variable $X \sim \text{NB}(k, p)$,

$$E(X) = \frac{k}{p}$$

and

$$\text{Var}(X) = \frac{k}{p} \cdot \left( \frac{1}{p} - 1 \right).$$

Proof. We use the representation offered to us by Theorem 115 to compute

$$\frac{p}{k} \cdot E(X) = \frac{p}{k} \cdot \sum_{j=0}^{\infty} (k + j) \binom{k + j - 1}{j} p^k (1 - p)^j$$

$$= \frac{p}{k} \cdot \sum_{j=0}^{\infty} (k + j) \cdot \frac{(k + j - 1)!}{j!(k-1)!} p^k (1 - p)^j$$

$$= \sum_{j=0}^{\infty} \frac{(k + j)!}{j!k!} p^{k+1} (1 - p)^j$$

$$= \sum_{j=0}^{\infty} \binom{k + j}{k} p^{k+1} (1 - p)^j.$$
To get to the variance, let’s first compute \( E(X(X + 1)) \). Once we do so, we can use the fact that 
\[
E(X^2 + X) = E(X^2) + E(X)
\]

Obtain
\[
E(X^2) + E(X) - E(X) - E(X)^2 = \text{Var}(X).
\]

Notice that
\[
E(X^2) + E(X) = E(X(X + 1))
\]
\[
= \sum_{j=0}^{\infty} (k+j)(k+j+1) \binom{k+j}{j} p^k (1-p)^j
\]
\[
= \sum_{j=0}^{\infty} (k+j+1)(k+j) \cdot \frac{(k+j-1)!}{(k-1)!j!} \cdot p^k (1-p)^j
\]
\[
= \sum_{j=0}^{\infty} \frac{(k+1)k}{p^2} \cdot \frac{(k+j+1)!}{(k+1)!j!} \cdot p^{k+2} (1-p)^j
\]
\[
= \frac{k^2 + k}{p^2} \sum_{j=0}^{\infty} \binom{k+j+1}{k+1} p^{k+2} (1-p)^j
\]
and that
\[
b^*(k+2 + j; k+2, p) = \binom{k+j+1}{k+1} p^{k+2} (1-p)^{k+2+j-(k+2)} = \binom{k+j+1}{k+1} p^{k+2} (1-p)^j
\]

which provide
\[
E(X^2) + E(X) = \frac{k^2 + k}{p^2}.
\]

Lastly,
\[
\text{Var}(X) = E(X^2) + E(X) - E(X) - E(X)^2
\]
\[
= \frac{k^2}{p^2} + \frac{k}{p^2} - \frac{k}{p} - \frac{k^2}{p^2}
\]
\[
= \frac{k}{p} \left( 1 - \frac{1}{p} \right)
\]
as promised.

**Definition 117.** A geometric distribution is a negative binomial distribution where \( k = 1 \). In particular, the geometric distribution according to \( p \) is given by
\[
g(n; p) = p(1-p)^{n-1}.
\]

**Example 102.** Flip a fair coin until a heads appears. What is the probability that the first heads appears on the tenth toss?

This follows the geometric distribution:
\[
g(10; 1/2) = \left( \frac{1}{2} \right)^{10} = \frac{1}{1024}
\]
Example 103. Roll a fair 6-sided die until a 6 appears. What is the probability the first 6 appears on the fifth roll?

Here, we also use the geometric distribution:
\[ g(5, 1/6) = \frac{1}{6} \left( \frac{5}{6} \right)^4 = \frac{625}{7776} \]

5.1.5 Hypergeometric Distributions

Given \( N \) objects, let \( M \leq N \) be considered success items and the remaining \( N - M \) failure items. We are to select \( n \leq N \) objects. Then the corresponding random variable is how many of those \( n \) objects are success items.

Notice that if \( N - M < n \), then we necessarily select some failure items. In particular, any selection must contain at least \( n - N + M \) failure items. Hence, if we are to find the probability that there are \( k \) success items, the following inequality must be satisfied:

\[ \max\{0, n - N + M\} \leq k \leq \min\{n, M\}. \]

Definition 118. A hypergeometric distribution is determined by a number \( N \) of items, \( M \) of which are deemed as success items, and a number \( n \leq N \) to be selected. Then the probability that there are \( k \) success items contained in a selection is

\[ h(k; n, N, M) = \binom{M}{k} \binom{N-M}{n-k} \binom{N}{n} \]

for \( \max\{0, n - N + M\} \leq k \leq \min\{n, M\} \).

In particular, \( h(k; n, N, M) \) is the hypergeometric distribution.

Theorem 119. For a hypergeometric random variable \( X \) with \( N, M, \) and \( n \),

\[ E(X) = \frac{nM}{N} \]

and

\[ \text{Var}(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)}. \]

Example 104. A bag contains 580 marbles, 310 of which are green. A sample of 15 marbles is going to be picked. What is the probability that 7 of the 15 marbles are green?

Using the hypergeometric distribution, we compute

\[ h(7; 15, 580, 310) = \frac{\binom{310}{7} \binom{270}{8}}{\binom{580}{15}} \approx 0.178553 \]

Remark. For large enough \( N \) and small enough \( n \) (general rule of thumb is \( 20n < N \)), we have

\[ h(k; n, N, M) \approx \binom{M}{k} \left( \frac{N}{n} \right)^k. \]

Example 105. A bag contains 580 marbles, 310 of which are green. A sample of 15 marbles is going to be picked. Use a binomial distribution to approximate the probability that 7 of the 15 marbles are green and compare to the result of Example 104.
CHAPTER 5. PARTICULAR PROBABILITY DISTRIBUTIONS AND DENSITIES

Here, we use the binomial distribution:

\[ b\left(7; 15, \frac{310}{580}\right) = \binom{15}{7} \left(\frac{310}{580}\right)^7 \left(1 - \frac{310}{580}\right)^8 \approx 0.176836 \]

which isn’t too far from 0.178553 as we computed using the hypergeometric distribution.

5.1.6 Poisson Distributions

Just as binomial distributions can be used to approximate hypergeometric distributions, Poisson distributions can be used to approximate binomial distributions. Aside from their usefulness in approximations, Poisson distributions are also of interest in their own right.

Definition 120. A Poisson distribution depends on a parameter \( \lambda > 0 \) and is given by

\[ f(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \]

where \( k = 0, 1, 2, \ldots \). We will write \( X \sim \text{Pois}(\lambda) \) and/or say that \( X \) is a Poisson random variable with parameter \( \lambda \) to mean that \( X \) is a random variable with distribution given by \( f(k; \lambda) \).

To see how and why a Poisson distribution can approximate a binomial distribution, let \( n \) and \( p \) be given. Then let \( \lambda = np \) and notice that

\[
\begin{align*}
b(k; n, p) &= \binom{n}{k} p^k (1 - p)^{n-k} \\
&= \frac{n(n-1)(n-2) \cdots (n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \left(1 + \frac{\lambda}{n}\right)^n
\end{align*}
\]

Now, leaving \( k \) and \( \lambda \) fixed,

\[
\lim_{n \to \infty} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \left(1 + \frac{\lambda}{n}\right)^n = \frac{\lambda^k}{k!} e^{-\lambda}
\]

since

\[
\lim_{n \to \infty} \left(1 + \frac{-\lambda}{n}\right)^n = e^{-\lambda}
\]

It follows that, if we fix \( \lambda > 0 \) from the beginning, then

\[
\lim_{n \to \infty} b(k; n, \frac{\lambda}{n}) = f(k; \lambda)
\]

for any \( k = 0, 1, 2, \ldots \).

Example 106. The average number of callers to UCalc between 5pm and 6pm is known to be 13 (and the number of callers follows a Poisson distribution). What is the probability there will be at least 4 callers today between 5pm and 6pm?

The probability that there will be at least 4 callers is

\[
\sum_{k=4}^{\infty} e^{-13} \frac{13^k}{k!} = 1 - \sum_{k=0}^{3} e^{-13} \frac{13^k}{k!} \approx 0.99895
\]
Example 106 suggests another way to represent a Poisson distribution. Namely, suppose it is known that, on average, \( r \) events happen within 1 unit. Then the probability there are \( k \) events in \( t \) units is

\[
e^{-rt} \cdot \frac{(rt)^k}{k!}
\]

**Example 107.** A certain telescope on a distant planet receives, on average, \( 2.7 \times 10^{15} \) photons per second. What is the probability that this telescope receives \( 5.9 \times 10^{15} \) photons over the course of three seconds?

Notice that we use the Poisson distribution with \( \lambda = 3 \times (2.7 \times 10^{15}) = 8.1 \times 10^{15} \). Then the probability that this telescope receives \( 5.9 \times 10^{15} \) photons over the course of three seconds is

\[
e^{-8.1 \times 10^{15}} \cdot \frac{(8.1 \times 10^{15})^{5.9 \times 10^{15}}}{(5.9 \times 10^{15})!}
\]

which, to find a decimal approximation, requires a sophisticated calculator.

**Example 108.** It is known that 0.7% of mugs produced by UDrink have a deformed handle. Use a Poisson distribution to approximate the probability that 5 of 500 mugs have deformed handles.

Here, we use \( \lambda = (0.007)(500) = 3.5 \) to calculate

\[
e^{-3.5} \cdot \frac{(3.5)^5}{5!} \approx 0.132169
\]

Notice that, using the binomial distribution, we would compute the probability to be

\[
\binom{500}{5}(0.007)^5(1-0.007)^{495} \approx 0.132533
\]

**Theorem 121.** The moment-generating function for a random variable \( X \sim \text{Pois}(\lambda) \) is

\[M_X(t) = e^{\lambda(e^t-1)}\]

**Proof.** Notice that

\[
M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk} \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} \cdot e^{\lambda e^t} = e^{\lambda(e^t-1)}.
\]

**Corollary 122.** For a random variable \( X \sim \text{Pois}(\lambda) \),

\[E(X) = \lambda\]

and

\[\text{Var}(X) = \lambda\]
CHAPTER 5. PARTICULAR PROBABILITY DISTRIBUTIONS AND DENSITIES

Proof. By Theorem 121, we calculate
\[ \frac{d}{dt} M_X(t) = \lambda e^{t \lambda (e^t - 1)} \]
and
\[ \frac{d^2}{dt^2} M_X(t) = \lambda^2 e^{2t \lambda (e^t - 1)} + \lambda e^{t \lambda (e^t - 1)}. \]
Then, notice that
\[ E(X) = \left. \frac{d}{dt} M_X(t) \right|_{t=0} = \lambda \]
and
\[ E(X^2) = \left. \frac{d^2}{dt^2} M_X(t) \right|_{t=0} = \lambda^2 + \lambda \]
which provide
\[ \text{Var}(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \]

Theorem 123. Suppose \( X_1 \sim \text{Pois}(\lambda_1) \) and \( X_2 \sim \text{Pois}(\lambda_2) \) where \( X_1 \) and \( X_2 \) are independent. Then \( X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2) \). More generally, if \( X_j \sim \text{Pois}(\lambda_j) \) for \( j = 1, 2, \ldots, n \) and \( X_1, X_2, \ldots, X_n \) are independent,
\[ X_1 + X_2 + \cdots + X_n \sim \text{Pois}(\lambda_1 + \lambda_2 + \cdots + \lambda_n). \]

Proof. First, notice that \( X_1 + X_2 = k \) if and only if, for some \( j \leq k \), \( X_1 = j \) and \( X_2 = k - j \). It follows that
\[
\begin{align*}
P(X_1 + X_2 = k) &= \sum_{j=0}^{k} P(X_1 = j \text{ and } X_2 = k - j) \\
&= \sum_{j=0}^{k} P(X_1 = j) \cdot P(X_2 = k - j) \\
&= \sum_{j=0}^{k} e^{-\lambda_1} \frac{\lambda_1^j}{j!} \cdot e^{-\lambda_2} \frac{\lambda_2^{k-j}}{(k-j)!} \\
&= e^{-(\lambda_1 + \lambda_2)} \cdot \sum_{j=0}^{k} \frac{\lambda_1^j \lambda_2^{k-j}}{j!(k-j)!}.
\end{align*}
\]
With sights forward, observe that
\[
(\lambda_1 + \lambda_2)^k = \sum_{j=0}^{k} \binom{k}{j} \lambda_1^j \lambda_2^{k-j} = \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \cdot \lambda_1^j \lambda_2^{k-j}
\]
by the Binomial Theorem. Thus,
\[
\begin{align*}
e^{-(\lambda_1 + \lambda_2)} \cdot \frac{(\lambda_1 + \lambda_2)^k}{k!} &= e^{-(\lambda_1 + \lambda_2)} \frac{1}{k!} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \cdot \lambda_1^j \lambda_2^{k-j} \\
&= e^{-(\lambda_1 + \lambda_2)} \cdot \sum_{j=0}^{k} \frac{\lambda_1^j \lambda_2^{k-j}}{j!(k-j)!} \\
&= P(X_1 + X_2 = k).
\end{align*}
\]
Therefore, \( X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2) \).
The more general statement about longer sums follows by induction. \( \square \)
5.1.7 Multinomial Distributions

A multinomial distribution can be seen to be a multivariate generalization of binomial distributions. Let \( X_1, X_2, \ldots, X_n \) be IID random variables where each random variable has \( k \) distinct possible outcomes \( a_1, a_2, \ldots, a_k \) and the \( j \)th outcome \( a_j \) has probability \( p_j \); i.e., \( P(X_\ell = a_j) = p_j \) for each \( 1 \leq \ell \leq n \) and \( 1 \leq j \leq k \). Then, we calculate the probability that, for a given sequence \( x_1, x_2, \ldots, x_k \) where \( x_1 + x_2 + \cdots + x_k = n \), there are \( x_j \) occurrences of outcome \( a_j \) to be

\[
\binom{n}{x_1, x_2, \ldots, x_n} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}
\]

The factor of

\[
\binom{n}{x_1, x_2, \ldots, x_n}
\]

comes from considering the number of ways in which one can build an \( n \)-length sequence of the \( a_j \) so that there are \( x_j \) occurrences of \( a_j \).

**Definition 124.** Given a sequence \( X_1, X_2, \ldots, X_n \) of IID random variables where each random variable has \( k \) distinct possible outcomes \( a_1, a_2, \ldots, a_k \) where the \( j \)th outcome \( a_j \) has probability \( p_j \), the corresponding **multinomial distribution** is

\[
f(x_1, x_2, \ldots, x_k; n, p_1, p_2, \ldots, p_k) = \binom{n}{x_1, x_2, \ldots, x_n} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}
\]

where \( 0 \leq x_j \leq n \) and \( x_1 + x_2 + \cdots + x_k = n \).

**Remark.** The name multinomial coefficient is inspired by Corollary 15 involving multinomial coefficients.

**Example 109.** A weighted 4-sided die has the following probability distribution:

<table>
<thead>
<tr>
<th>outcome</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

If we roll the die 10 times, what’s the probability we get fives 1s, three 2s, and two 3s?

Here, we compute

\[
f(5, 3, 2, 0; 4, 0.1, 0.3, 0.4, 0.2) = \binom{10}{5, 3, 2, 0} (0.1)^5(0.3)^3(0.4)^2(0.2)^0 = 0.000108864
\]

5.1.8 Multivariate Hypergeometric Distributions

Just as the multinomial distribution is a multivariate extension of binomial distributions, **multivariate hypergeometric distributions** extend hypergeometric distributions. Suppose we have a partition of \( N \) items into collections of sizes \( M_1, M_2, \ldots, M_k \) where \( 0 < M_j \) and \( M_1 + M_2 + \cdots + M_k = N \). We are to randomly select \( n \leq N \) items and want to find the probability that there are \( x_j \leq M_j \) success items in our sample for a sequence \( x_1, x_2, \ldots, x_k \) with \( x_j \geq 0 \) and \( x_1 + x_2 + \cdots + x_k = n \). That probability is

\[
\binom{M_1}{x_1} \binom{M_2}{x_2} \cdots \binom{M_k}{x_k} \binom{N}{n}
\]
Definition 125. A **multivariate hypergeometric distribution** is determined by a number \(N\), a sequence \(M_1, M_2, \ldots, M_k\) of positive integers with \(M_1 + M_2 + \cdots + M_k = N\), and a number \(n \leq N\). The multivariate hypergeometric distribution is then defined to be
\[
f(x_1, x_2, \ldots, x_k; n, N, M_1, M_2, \ldots, M_k) = \frac{\binom{M_1}{x_1} \binom{M_2}{x_2} \cdots \binom{M_k}{x_k}}{\binom{N}{n}}
\]
for \(0 \leq x_j\) and \(\sum_{j=1}^{k} x_j = n\).

**Example 110.** A bag of jellybeans has 52 red, 23 green, 17 yellow, and 10 blue jellybeans. A sample of 7 jellybeans is to be randomly selected. What is the probability that the sample consists of 4 red, 2 green, and 1 yellow?

We compute
\[
f(4, 2, 1, 0; 7, 102, , 52, 23, 17, 10) = \frac{\binom{52}{4} \binom{23}{2} \binom{17}{1} \binom{10}{0}}{\binom{102}{7}} \approx 0.0630525
\]

### 5.2 Continuous Densities

#### 5.2.1 Uniform Distributions

**Definition 126.** A continuous random variable \(X\) is said to have a **uniform distribution** if its probability density is given by
\[
f(x) = \begin{cases} 
\frac{1}{b - a}, & a < x < b; \\
0, & \text{otherwise}.
\end{cases}
\]

For a continuous random variable \(X\) with a uniform distribution, we will say \(X\) is **uniformly distributed** on the interval \([a, b]\).

**Theorem 127.** For a uniformly distributed continuous random variable \(X\) on the interval \([a, b]\),
\[
E(X) = \frac{a + b}{2}, \\
\text{Var}(X) = \frac{(b - a)^2}{12},
\]
and the moment-generating function is given by
\[
M_X(t) = \frac{e^{tb} - e^{ta}}{t(b - a)}.
\]

**Proof.** Notice that
\[
E(X) = \int_{a}^{b} x \frac{1}{b - a} \, dx = \left. \frac{x^2}{2(b - a)} \right|_{a}^{b} = \frac{b^2 - a^2}{2(b - a)} = \frac{(a + b)(b - a)}{2(b - a)} = \frac{a + b}{2}.
\]

Then, to find \(\text{Var}(X)\), we first compute
\[
E(X^2) = \int_{a}^{b} \frac{x^2}{b - a} \, dx = \left. \frac{x^3}{3(b - a)} \right|_{a}^{b} = \frac{b^3 - a^3}{3(b - a)} = \frac{(b - a)(a^2 + ab + b^2)}{3(b - a)} = \frac{a^2 + ab + b^2}{3}.
\]
Then
\[ \text{Var}(X) = a^2 + ab + b^2 - \frac{(a + b)^2}{4} = \frac{4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2}{12} = \frac{a^2 - 2ab + b^2}{12} = \frac{(a - b)^2}{12}. \]

Lastly,
\[ M_X(t) = E(e^{tX}) = \int_a^b \frac{e^{tx}}{b - a} \, dx = \frac{e^{tx}}{t(b - a)} \bigg|_a^b = \frac{e^{tb} - e^{ta}}{t(b - a)}. \]

### 5.2.2 Gamma, Exponential, and Chi-Square Distributions

Before we discover gamma distributions, let’s review the gamma function, a necessary component in gamma distributions.

**Definition 128.** Recall that the **gamma function** is given by
\[ \Gamma(x) = \int_0^\infty t^{x-1}e^{-t} \, dt \]
for \( x > 0. \)

**Exercise 18.** Show that
\[ \int_0^\infty t^{x-1}e^{-t} \, dt \]
is a convergent integral for each \( x > 0. \) That is, show that \( \Gamma(x) \) is defined for all \( x > 0. \)

**Note.** Observe that \( \Gamma(x) > 0 \) for all \( x > 0 \) since \( t^{x-1}e^{-t} > 0 \) for \( t > 0. \)

**Exercise 19.** For \( x > 0, \) show that
\[ \Gamma(x + 1) = x \cdot \Gamma(x) \]
by using integration by parts. Use this to conclude that, for positive integer \( n, \) \( \Gamma(n) = (n - 1)! \).

Suppose \( Y \sim \text{Pois}(rt) \) where \( r \) is the rate of success events per unit time and that success events for \( Y \) cannot happen simultaneously. Let \( X \) be the amount of time it takes for the \( k^{\text{th}} \) success event to occur. Notice that \( P(X > t) \) is the probability that it takes more than \( t \) time units for \( k \) success events to occur. In other words, \( P(X > t) \) is the probability that at most \( k - 1 \) success events occur between time 0 and time \( t; \) i.e.,
\[ P(X > t) = \sum_{j=0}^{k-1} e^{-rt} \cdot \frac{(rt)^j}{j!} = e^{-rt} \cdot \sum_{j=0}^{k-1} \frac{(rt)^j}{j!}. \]

It follows that
\[ P(X \leq t) = 1 - P(X > t) = 1 - e^{-rt} \cdot \sum_{j=0}^{k-1} \frac{(rt)^j}{j!}. \]

Since \( F(t) := P(X \leq t) \) is the cumulative density function for \( X, \) to find the corresponding probability density, we need only differentiate:
\[ \frac{d}{dt} F(t) = \frac{d}{dt} \left[ 1 - e^{-rt} \sum_{j=0}^{k-1} \frac{(rt)^j}{j!} \right]. \]
\[
\frac{d}{dt} \left[ 1 - e^{-rt} \cdot \left( 1 + rt + \frac{r^2 t^2}{2} + \cdots + \frac{r^{k-1} t^{k-1}}{(k-1)!} \right) \right]
\]
\[
= (-1) \cdot \left[ e^{-rt} \cdot \left( 1 + rt + \frac{r^2 t^2}{2} + \cdots + \frac{r^{k-1} t^{k-1}}{(k-2)!} \right) - re^{-rt} \cdot \sum_{j=0}^{k-1} \frac{(rt)^j}{j!} \right]
\]
\[
= re^{-rt} \cdot \sum_{j=0}^{k-1} \frac{r^j t^j}{j!} - e^{-rt} \cdot \sum_{j=1}^{k-1} \frac{r^j t^j}{(j-1)!}
\]
\[
= re^{-rt} \cdot \sum_{j=0}^{k-1} \frac{r^j t^j}{j!} - re^{-rt} \cdot \sum_{j=1}^{k-1} \frac{r^{j-1} t^{j-1}}{(j-1)!}
\]
\[
= re^{-rt} \cdot \sum_{j=0}^{k-1} \frac{r^j t^j}{j!} - re^{-rt} \cdot \sum_{j=0}^{k-2} \frac{r^j t^j}{j!}
\]
\[
= re^{-rt} \cdot \frac{r^{k-1} t^{k-1}}{(k-1)!}
\]
\[
= e^{-rt} \cdot \frac{r^k t^{k-1}}{\Gamma(k)}.
\]

Hence, we have
\[
f(t) = \begin{cases} 
  e^{-rt} \cdot \frac{r^k t^{k-1}}{\Gamma(k)}, & t > 0; \\
  0, & t \leq 0.
\end{cases}
\]

For this to be a probability density, we need
\[
1 = \int_{-\infty}^{\infty} f(t) \, dt 
= \int_{0}^{\infty} e^{-rt} \cdot \frac{r^k t^{k-1}}{\Gamma(k)} \, dt 
= \frac{1}{\Gamma(k)} \cdot \int_{0}^{\infty} r^k e^{-rt} t^{k-1} \, dt 
\]

which provides
\[
\int_{0}^{\infty} t^{k-1} e^{-t} \, dt = \int_{0}^{\infty} r^k t^{k-1} e^{-rt} \, dt.
\]

Notice that, \( \lambda = rt \) gives \( d\lambda = r \, dt \) which means
\[
\int_{0}^{\infty} r^k t^{k-1} e^{-rt} \, dt = \int_{0}^{\infty} r^k \left( \frac{\lambda}{r} \right)^{k-1} \cdot e^{-\lambda} \cdot \frac{1}{r} \, d\lambda 
= \int_{0}^{\infty} \lambda^{k-1} \cdot e^{-\lambda} \, d\lambda 
= \Gamma(k).
\]

Therefore, \( f(t) \) is a probability density which, using \( x = t, \alpha = k, \) and \( \beta = r, \) we can rewrite as
\[
f(x) = \begin{cases} 
  \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, & x > 0; \\
  0, & x \leq 0.
\end{cases}
\]

**Definition 129.** A continuous random variable \( X \) has a **gamma distribution** with parameters \( \alpha > 0 \)
and \( \beta > 0 \), denoted \( X \sim \Gamma(\alpha, \beta) \), if the probability density for \( X \) is given by

\[
f(x) = \begin{cases} 
\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, & x > 0; \\
0, & x \leq 0.
\end{cases}
\]

Whenever \( X \sim \Gamma(\alpha, \beta) \), we also say that \( X \) is a gamma random variable.

**Remark.** Using \( \theta = \frac{1}{\beta} \implies \beta = \frac{1}{\theta} \), we can also write a gamma distribution as

\[
f(x) = \begin{cases} 
\frac{x^{\alpha-1} e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)}, & x > 0; \\
0, & x \leq 0,
\end{cases}
\]

**Definition 130.** For \( X \sim \Gamma(\alpha, \beta) \), \( \alpha \) is called the shape parameter and, under the assumption that the corresponding probability density is

\[
f(x) = \begin{cases} 
\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, & x > 0; \\
0, & x \leq 0,
\end{cases}
\]

\( \beta \) is called the rate parameter. If we use \( \theta = \frac{1}{\beta} \) as the parameter, that is, that the corresponding probability density is

\[
f(x) = \begin{cases} 
\frac{x^{\alpha-1} e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)}, & x > 0; \\
0, & x \leq 0,
\end{cases}
\]

then \( \theta \) is called the scale parameter.

**Theorem 131.** If \( X \sim \Gamma(\alpha, \beta) \), then the \( n^{\text{th}} \) moment about the origin is given by

\[
\mu'_n = E(X^n) = \frac{\Gamma(\alpha + n)}{\beta^n \cdot \Gamma(\alpha)}
\]

where \( \beta \) is the rate parameter and \( \theta \) is the scale parameter.

**Proof.** Notice that

\[
E(X^n) = \int_0^\infty x^n \cdot \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \, dx
\]

\[
= \frac{1}{\Gamma(\alpha)} \cdot \int_0^{\infty} \frac{\beta^{\alpha+n-1}}{\beta^{n-1}} \cdot x^{\alpha+n-1} e^{-\beta x} \, dx
\]

\[
= \frac{1}{\beta^{n-1} \cdot \Gamma(\alpha)} \cdot \int_0^{\infty} (\beta x)^{\alpha+n-1} e^{-\beta x} \, dx
\]

which, by \( y = \beta x \) and \( dy = \beta \, dx \), gives

\[
E(X^n) = \frac{1}{\beta^{n-1} \cdot \Gamma(\alpha)} \cdot \int_0^{\infty} y^{(\alpha+n)-1} e^{-y} \cdot \frac{1}{\beta} \, dy
\]

\[
= \frac{\Gamma(\alpha + n)}{\beta^n \cdot \Gamma(\alpha)}.
\]

**Corollary 132.** For \( X \sim \Gamma(\alpha, \beta) \) where \( \beta \) is the rate parameter and \( \theta \) is the scale parameter,

\[
E(X) = \frac{\alpha}{\beta} = \alpha \theta
\]

and

\[
\text{Var}(X) = \frac{\alpha}{\beta^2} = \alpha \theta^2.
\]
CHAPTER 5. PARTICULAR PROBABILITY DISTRIBUTIONS AND DENSITIES

Proof. By Theorem 131,
\[ E(X) = \mu_1 = \frac{\Gamma(\alpha + 1)}{\beta \cdot \Gamma(\alpha)} = \frac{\alpha \cdot \Gamma(\alpha)}{\beta \cdot \Gamma(\alpha)} = \frac{\alpha}{\beta} \]
and
\[ E(X^2) = \mu_2 = \frac{\Gamma(\alpha + 2)}{\beta^2 \cdot \Gamma(\alpha)} = \frac{(\alpha + 1)\alpha \cdot \Gamma(\alpha)}{\beta^2 \cdot \Gamma(\alpha)} = \frac{\alpha^2 + \alpha}{\beta^2}. \]
Hence,
\[ \text{Var}(X) = \frac{\alpha^2 + \alpha}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}. \]

Theorem 133. For \( X \sim \Gamma(\alpha, \beta) \), \( \theta = 1/\beta \), the moment-generating function is given by
\[ M_X(t) = \frac{\beta^\alpha}{(\beta - t)^\alpha} = (1 - \theta t)^{-\alpha} \]
for \(|t| < \beta\).

Proof. Observe that
\[ E(e^{tX}) = \int_0^\infty e^{tx} \cdot \frac{\beta^\alpha \cdot x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \, dx \]
which, by using \( y = (\beta - t)x \) and \( dy = (\beta - t) \, dx \),
\[ E(e^{tX}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \int_0^\infty \left(\frac{y}{\beta - t}\right)^{\alpha-1}e^{-y} \cdot \frac{1}{\beta - t} \, dy \]
\[ = \frac{\beta^\alpha}{(\beta - t)^\alpha \cdot \Gamma(\alpha)} \cdot \int_0^\infty y^{\alpha-1}e^{-y} \, dy \]
\[ = \frac{\beta^\alpha}{(\beta - t)^\alpha \cdot \Gamma(\alpha)} \cdot \Gamma(\alpha) \]
\[ = \frac{\beta^\alpha}{(\beta - t)^\alpha}. \]
For \( \theta = 1/\beta \),
\[ \frac{\beta^\alpha}{(\beta - t)^\alpha} = \left(\frac{1}{\theta}\right)^\alpha \frac{1}{\theta^\alpha} \left(\frac{1 - \theta t}{\theta}\right)^{-\alpha} = (1 - \theta t)^{-\alpha}. \]

Example 111. Suppose \( X \) is a continuous random variable with a gamma distribution so that \( E(X) = 6 \) and \( \text{Var}(X) = 18 \). Find the probability density for \( X \).

We need to find \( \alpha \) and \( \beta \) so that
\[ \frac{\alpha}{\beta} = 6 \]
and
\[ \frac{\alpha}{\beta^2} = 18. \]
From these two, we find that \( \alpha = 6\beta = 18\beta^2 \) which provides
\[ 6\beta(1 - 3\beta) = 0. \]
Since $\beta > 0$, we see that $\beta = \frac{1}{3}$ and $\alpha = 2$. Therefore, the probability density for $X$ is given by

$$f(x) = \begin{cases} \frac{x e^{-x/3}}{9}, & x > 0; \\ 0, & x \leq 0. \end{cases}$$

**Definition 134.** A continuous random variable $X$ is said to have an exponential distribution if $X \sim \Gamma(1, \lambda)$. That is, the probability density for $X$ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0; \\ 0, & x \leq 0. \end{cases}$$

Since exponential distributions only rely on one parameter, we will write $X \sim \text{Exp}(\lambda)$ to denote that $X \sim \Gamma(1, \lambda)$ and we will say that $X$ is an exponential random variable with parameter $\lambda$.

As in the discussion motivating the gamma distribution, the exponential distribution can be used to model the waiting time for the first Poisson event to occur given a discrete random variable $Y \sim \text{Pois}(\lambda)$. In fact, let $X$ be the amount of time it takes for the first success event for $Y$. Then

$$P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t}.$$  

Then we see that the probability density for $X$ is given by

$$\frac{d}{dt} \left[1 - e^{-\lambda t}\right] = \lambda e^{-\lambda t}$$

for $t > 0$, an exponential distribution.

Assuming there was a Poisson event at time $t_0 \geq 0$, then the same reasoning shows that the probability the next Poisson event occurs $t$ time units later is still $1 - e^{-\lambda t}$. Hence, exponential distributions are not only useful for finding the waiting time for the first Poisson event, but also for finding the waiting time between Poisson events.

**Example 112.** A jaywalker is waiting to cross a street. The number of cars passing directly in front of the jaywalker per hour is a random variable with a Poisson distribution of parameter $\lambda = 7.6$. What is the probability that the jaywalker has less than 2 minutes to cross between passing cars?

Since the Poisson distribution is measured in hours and we wish to compute a probability for 2 minutes, we use

$$\frac{2}{60} = \frac{1}{30}$$

hours. That is, the desired probability is

$$\int_{0}^{1/30} 7.6e^{-7.6x} \, dx \approx 0.223791$$

**Corollary 135.** For $X \sim \text{Exp}(\lambda)$,

$$E(X) = \frac{1}{\lambda},$$

$$\text{Var}(X) = \frac{1}{\lambda^2},$$

and

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

for $|t| < \lambda$.‌
Proof. Apply Corollary 132 and Theorem 133.

Another special case of a gamma distribution is a *chi-square* distribution which we will revisit more in Section 5.2.4.

**Definition 136.** A continuous random variable $X$ has a **chi-square distribution** with $k$ degrees of freedom if $X \sim \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$ which we denote as $X \sim \chi^2(k)$. If $X \sim \chi^2(k)$, then the probability density for $X$ is given by

$$f(x) = \begin{cases} \frac{x^{(k-2)/2} \cdot e^{-x/2}}{2^{k/2} \Gamma(k/2)}, & x > 0; \\ 0, & \text{otherwise}. \end{cases}$$

**Corollary 137.** If $X \sim \chi^2(k)$, then

$$E(X) = k,$$

$$\text{Var}(X) = 2k,$$

and

$$M_X(t) = (1 - 2t)^{-k/2}$$

for $|t| < \frac{1}{2}$.

Proof. Apply Corollary 132 and Theorem 133.

### 5.2.3 Beta Distributions

Before we discuss beta distributions, we introduce the beta function.

**Definition 138.** The **beta function** is defined to be

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

for $\alpha, \beta > 0$.

**Theorem 139.** For $\alpha, \beta > 0$,

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \, dx.$$  

Proof. First, notice that

$$\Gamma(\alpha) \Gamma(\beta) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx \cdot \int_0^\infty y^{\beta-1} e^{-y} \, dy = \int_0^\infty \int_0^\infty x^{\alpha-1} y^{\beta-1} e^{-(x+y)} \, dx \, dy.$$  

Let $x = u^2$ and $y = v^2$ for $u, v > 0$ and notice that $dx = 2u \, du$ and $dy = 2v \, dv$. Then

$$\Gamma(\alpha) \Gamma(\beta) = 4 \int_0^\infty \int_0^\infty u^{2\alpha-1} v^{2\beta-1} e^{-(u^2+v^2)} \, du \, dv.$$  

Now, using polar coordinates, we have the change of variables $u = r \cos(\theta)$ and $v = r \sin(\theta)$ for $r > 0$ and $0 < \theta < \pi/2$. This leads us to

$$\Gamma(\alpha) \Gamma(\beta) = 4 \int_0^{\pi/2} \int_0^\infty (r \cos(\theta))^{2\alpha-1} (r \sin(\theta))^{2\beta-1} e^{-r^2} \cdot r \, dr \, d\theta$$

$$= 4 \int_0^{\pi/2} \int_0^\infty r^{2\alpha+2\beta-1} e^{-r^2} \cos(\theta)^{2\alpha-1} \sin(\theta)^{2\beta-1} \, dr \, d\theta$$

$$= 4 \cdot \int_0^{\pi/2} \cos(\theta)^{2\alpha-1} \sin(\theta)^{2\beta-1} \, d\theta \cdot \int_0^\infty r^{2\alpha+\beta-1} e^{-r^2} \, dr.$$
Similarly,

\[
\Gamma(\alpha + \beta) = \int_0^\infty x^{\alpha+\beta-1} e^{-x} \, dx = \int_0^\infty (y^2)^{\alpha+\beta-1} e^{-y^2} \cdot 2y \, dy = 2 \int_0^\infty y^{2\alpha+2\beta-1} e^{-y^2} \, dy.
\]

It follows that

\[
B(\alpha, \beta) = 2 \int_0^{\pi/2} \cos(\theta)^{2\alpha-1} \sin(\theta)^{2\beta-1} \, d\theta.
\]

Lasly, we consider

\[
\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx
\]

using the change of variables \(x = \cos^2(\theta)\) for \(0 < \theta < \pi/2:\)

\[
\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \, dx = \int_{\pi/2}^0 (\cos^2(\theta))^{\alpha-1}(1 - \cos^2(\theta))^{\beta-1} \cdot (-2\sin(\theta)\cos(\theta)) \, d\theta = 2 \int_{\pi/2}^0 \cos(\theta)^{2\alpha-1} \sin(\theta)^{2\beta-1} \, d\theta = B(\alpha, \beta).
\]

Suppose \(Y\) is a Bernoulli variable so that each trial is independent of the others but the probability of success \(p\) is unknown. Let \(X_0\) be uniformly distributed on the interval \([0, 1]\) which we use as our guess for \(p\). To approximate \(p\), we could conduct a large number \(n\) of trials of \(Y\) and count the number of successes \(k\). With that said, let \(Y_n = \sum_{j=1}^n Y_j\), a binomial variable. For reasons exceeding the scope of this course, we can form a joint probability density \(f(k, x)\) for \(Y_n\) and \(X_0\). Note that the marginal density of \(X_0\) is given by

\[
g(x) = \begin{cases} 
1, & 0 < x < 1; \\
0, & \text{otherwise}.
\end{cases}
\]

which provides

\[
\frac{f(k, x)}{g(x)} = f(k|x) = P(Y_n = k|X_0 = x) = \binom{n}{k} x^k (1-x)^{n-k}.
\]

The marginal distribution for \(Y_n\) is given by

\[
h(k) = \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} \, dx = \binom{n}{k} \cdot B(k+1, n-k+1)
\]

by Theorem 139. Moreover,

\[
h(k) = \binom{n}{k} \cdot B(k+1, n-k+1)
\]

\[
= \frac{n!}{(n-k)!k!} \cdot \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)}
\]

\[
= \frac{n!}{(n-k)!k!(n-k)!} \cdot \frac{k!(n-k)!}{(n+1)!}
\]

\[
= \frac{1}{n+1}.
\]
Thus,
\[ P(X_0 = x | Y_n = k) = \binom{n}{k} x^k (1-x)^{n-k} \cdot (n+1) \]
\[ = \frac{(n+1)!}{(n-k)!k!} \cdot x^k (1-x)^{n-k} \]
\[ = \frac{\Gamma(n+2)}{\Gamma(n-k+1)\Gamma(k+1)} \cdot x^k (1-x)^{n-k} \]
\[ = \frac{x^k (1-x)^{n-k}}{B(k+1, n-k+1)}. \]

That is, after the \( n \) trials, we can consider the conditional random variable \( X \) where the probability density of \( X \) is given by
\[ \varphi(x) = \frac{x^k (1-x)^{n-k}}{B(k+1, n-k+1)} = P(X_0 = x | Y_n = k) \]
for \( 0 < x < 1. \)

**Definition 140.** A continuous random variable \( X \) has a beta distribution with parameters \( \alpha, \beta > 0 \) if the probability density for \( X \) is given by
\[ f(x) = \begin{cases} 
  x^{\alpha-1}(1-x)^{\beta-1} \frac{1}{B(\alpha, \beta)}, & 0 < x < 1; \\
  0, & \text{otherwise}. 
\end{cases} \]

In such a case, we say that \( X \) is a beta random variable with parameters \( \alpha, \beta \), denoted \( X \sim \text{Beta}(\alpha, \beta) \).

**Example 113.** Suppose \( Y \sim \text{Bernoulli}(p) \) and that each trial of \( Y \) is independent from others. We have conducted 47 experiments, 14 of which were success events. Use a beta distribution to find the probability that \( 0.25 < p < 0.35 \).

As discussed above, we can use \( n = 47 \) and \( k = 14 \) to obtain the parameters \( \alpha = 15 \) and \( \beta = 34 \). Then, notice that
\[ P(0.25 < p < 0.35) = \int_{0.25}^{0.35} \frac{x^{14}(1-x)^{33}}{B(15, 34)} \, dx \approx 0.554201 \]

**Example 114.** Suppose \( Y \sim \text{Bernoulli}(p) \) and that each trial of \( Y \) is independent from others. We have conducted 157 experiments, 47 of which were success events. Use a beta distribution to find the probability that \( 0.25 < p < 0.35 \).

As discussed above, we can use \( n = 157 \) and \( k = 47 \) to obtain the parameters \( \alpha = 48 \) and \( \beta = 111 \). Then, notice that
\[ P(0.25 < p < 0.35) = \int_{0.25}^{0.35} \frac{x^{47}(1-x)^{110}}{B(48, 111)} \, dx \approx 0.831551 \]

Notice that \( \frac{14}{47} \approx 0.29787 \) and \( \frac{47}{157} \approx 0.29936 \) which would suggest that the Bernoulli variable \( Y \) in Examples 113 and 114 has a probability of success \( p \approx 0.3 \). What the beta distribution computations tell us in both cases is that, with more trials, we have even more reason to believe that \( p \approx 0.3 \), which adheres to naive intuition. The benefit here is that the beta distributions offer a quantitative foundation which can be used to evaluate the validity of our assertions.

**Theorem 141.** Let \( X \sim \text{Beta}(\alpha, \beta) \). Then the \( n^{th} \) moment about the origin is
\[ E(X^n) = \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)} = \prod_{j=0}^{n-1} \frac{\alpha + j}{\alpha + \beta + j}. \]
Proof. Note that
\[
E(X^n) = \int_0^1 x^n \cdot \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \, dx
\]
\[
= \frac{1}{B(\alpha, \beta)} \cdot \int_0^1 x^{\alpha+n-1}(1-x)^{\beta-1} \, dx
\]
\[
= \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)}
\]
by Theorem 139. Now, by the definition of the beta function,
\[
E(X^n) = \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + n)\Gamma(\beta)}{\Gamma(\alpha + \beta + n)} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}
\]
\[
= \frac{(\alpha + n - 1)(\alpha + n - 2) \cdots \alpha \cdot \Gamma(\alpha)\Gamma(\beta)}{(\alpha + \beta + n - 1)(\alpha + \beta + n - 2) \cdots (\alpha + \beta)} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}
\]
\[
= \prod_{j=0}^{n-1} \frac{\alpha + j}{\alpha + \beta + j}
\]
establishing the asserted equality.

Corollary 142. Let \(X \sim \text{Beta} (\alpha, \beta)\). Then
\[
E(X) = \frac{\alpha}{\alpha + \beta}
\]
and
\[
\text{Var}(X) = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\]
Proof. By Theorem 141, we see that
\[
E(X) = \frac{\alpha}{\alpha + \beta}
\]
and that
\[
E(X^2) = \frac{\alpha}{\alpha + \beta} \cdot \frac{\alpha + 1}{\alpha + \beta + 1}.
\]
Then
\[
\text{Var}(X) = \frac{\alpha}{\alpha + \beta} \cdot \frac{\alpha + 1}{\alpha + \beta + 1} - \left(\frac{\alpha}{\alpha + \beta}\right)^2
\]
\[
= \frac{\alpha + \beta}{\alpha + \beta + 1} \cdot \frac{\alpha}{\alpha + \beta} - \frac{\alpha}{\alpha + \beta} \cdot \frac{\alpha + 1}{\alpha + \beta + 1} - \left(\frac{\alpha}{\alpha + \beta}\right)^2 \cdot \frac{\alpha + \beta + 1}{\alpha + \beta + 1}
\]
\[
= \frac{\alpha^2(\alpha + \beta) + \alpha(\alpha + \beta) - \alpha^2(\alpha + \beta) - \alpha^2}{(\alpha + \beta)^2(\alpha + \beta + 1)}
\]
\[
= \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\]

Corollary 143. If \(X \sim \text{Beta} (\alpha, \beta)\), the moment-generating function is given by
\[
M_X(t) = \sum_{n=0}^{\infty} \frac{B(\alpha + n, \beta)}{B(\alpha, \beta)} \cdot \frac{x^n}{n!}.
\]
Proof. This follows from Theorems 141 and 82.

Before we move on to the next section, we provide a result concerning gamma distributions utilizing the beta function.

Theorem 144. Suppose $X_j \sim \Gamma(\alpha_j, \beta)$ for $j = 1, 2, \ldots, n$ and that $X_1, X_2, \ldots, X_n$ are independent. Then $X_1 + X_2 + \cdots + X_n \sim \Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_n, \beta)$.

Proof. We will prove the statement for $X \sim \Gamma(\alpha_1, \beta)$ and $Y \sim \Gamma(\alpha_2, \beta)$ where $X$ and $Y$ are independent. The more general statement follows by induction.

Let’s compute $P(X + Y \leq z)$ by first examining the region of integration. Since $X$ and $Y$ are gamma variables, $P(X < 0) = P(Y < 0) = 0$. Hence, we integrate over

\[
P(X + Y \leq z) = \int_0^z \int_0^{z-x} \frac{\beta^{\alpha_1} x^{\alpha_1-1} e^{-\beta x}}{\Gamma(\alpha_1)} \cdot \frac{\beta^{\alpha_2} y^{\alpha_2-1} e^{-\beta y}}{\Gamma(\alpha_2)} \, dy \, dx
\]

Let $x = st$ and $y = t - st$. The corresponding Jacobian is

\[
\begin{vmatrix}
  t & s \\
  -t & 1-s
\end{vmatrix} = t - st + st = t.
\]

To see how this change of variables affects the region of integration, observe that

- if $y = 0$, $t = st$ which means that $s = 1$ as long as $t \neq 0$,
- if $y = z - x$, $t - st = z - st$ which implies that $z = t$, and
- if $x = 0$, $s = 0$ or $t = 0$.

Also, notice that $0 \leq x = st$ and $0 \leq y = t - st$ imply that $0 \leq st \leq t$ so $0 \leq s \leq 1$ and $t \geq 0$. Moreover, since $y \leq z - x$, $t - st \leq z - st$ which implies that $t \leq z$. Thus, the new region of integration is
5.2. CONTINUOUS DENSITIES

This provides

\[
P(X + Y \leq z) = \frac{\beta^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z \int_0^{x+y} e^{-\beta(x+y)} x^{\alpha_1-1} y^{\alpha_2-1} \, dy \, dx
\]

\[
= \frac{\beta^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z \int_0^1 e^{-\beta t} t^{\alpha_1-1} (t-st)^{\alpha_2-1} \cdot t \, ds \, dt
\]

\[
= \frac{\beta^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z \int_0^1 e^{-\beta t} t^{\alpha_1+\alpha_2-1} (1-s)^{\alpha_2-1} \, ds \, dt
\]

\[
= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot B(\alpha_1, \alpha_2) \cdot \int_0^z \beta^{\alpha_1+\alpha_2} t^{\alpha_1+\alpha_2-1} e^{-\beta t} \, dt
\]

\[
= \frac{B(\alpha_1, \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot \int_0^z \beta^{\alpha_1+\alpha_2} t^{\alpha_1+\alpha_2-1} e^{-\beta t} \, dt
\]

\[
= \int_0^z \beta^{\alpha_1+\alpha_2} t^{\alpha_1+\alpha_2-1} e^{-\beta t} \, dt.
\]

It follows that the probability density for \(X + Y\) is given by

\[
\frac{\beta^{\alpha_1+\alpha_2} t^{\alpha_1+\alpha_2-1} e^{-\beta t}}{\Gamma(\alpha_1+\alpha_2)}
\]

for \(t > 0\), which is to say that \(X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)\).

5.2.4 Normal Distributions

Recall Exercise 14 where it was asked to show that

\[
\int_{-\infty}^{\infty} \exp \left( -\frac{x^2}{2\sigma^2} \right) \frac{dx}{\sqrt{\pi}} = 1.
\]

In this section, this fact will be used.

**Definition 145.** A continuous random variable \(X\) is said to have a **normal distribution** if its probability density is given by

\[
f(x) = \frac{\exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)}{\sqrt{2\pi}\sigma^2}
\]

where \(\mu \in \mathbb{R}\) and \(\sigma > 0\). If \(X\) has a normal distribution, we will say that \(X\) is a **normal random variable** with parameters \(\mu\) and \(\sigma^2\), denoted by \(X \sim \mathcal{N}(\mu, \sigma^2)\).

**Remark.** A normal distribution is commonly referred to as a **bell curve** due to the shape of the graph of the probability density function.

**Exercise 20.** Verify that

\[
\int_{-\infty}^{\infty} \frac{\exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)}{\sqrt{2\pi}\sigma^2} \, dx = 1.
\]

**Definition 146.** The **standard normal distribution** is \(\mathcal{N}(0,1)\); namely, it is the probability density

\[
f(x) = \frac{\exp \left( -\frac{x^2}{2} \right)}{\sqrt{2\pi}}.
\]

**Proposition 147.** For

\[
f(x) = \frac{\exp \left( -\frac{x^2}{2} \right)}{\sqrt{2\pi}},
\]
\( f(-x) = f(x) \) which implies that
\[
\int_{-a}^{a} f(x) \, dx = 2 \cdot \int_{0}^{a} f(x) \, dx.
\]

**Theorem 148.** For \( X \sim N(\mu, \sigma^2) \), the moment-generating function is given by
\[
M_X(t) = \exp \left( \mu t + \frac{\sigma^2 t^2}{2} \right).
\]

From this, we obtain
\[
E(X) = \mu
\]
and
\[
\text{Var}(X) = \sigma^2.
\]

**Proof.** Observe that
\[
M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \frac{\exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)}{\sqrt{2\pi\sigma^2}} \, dx
\]
\[
= \int_{-\infty}^{\infty} \exp \left( tx - \frac{(x-\mu)^2}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi\sigma^2}} \, dx
\]
\[
= \int_{-\infty}^{\infty} \exp \left( \frac{-[(x-\mu)^2 - 2\sigma^2 tx]}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi\sigma^2}} \, dx.
\]

Now,
\[
(x - \mu)^2 - 2\sigma^2 tx = x^2 - 2(\mu + \sigma^2 t)x + \mu^2
\]
\[
= [x - (\mu + \sigma^2 t)]^2 + \mu^2 - (\mu + \sigma^2 t)^2
\]
\[
= [x - (\mu + \sigma^2 t)]^2 - \sigma^2(2\mu t + \sigma^2 t^2)
\]
which provides
\[
\frac{-(x - \mu)^2 - 2\sigma^2 tx}{2\sigma^2} = \frac{-(x - (\mu + \sigma^2 t))^2}{2\sigma^2} + \left( \mu t + \frac{\sigma^2 t^2}{2} \right).
\]

It follows that
\[
M_X(t) = \int_{-\infty}^{\infty} \exp \left( \frac{-[(x-\mu)^2 - 2\sigma^2 tx]}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi\sigma^2}} \, dx
\]
\[
= \exp \left( \mu t + \frac{\sigma^2 t^2}{2} \right) \cdot \int_{-\infty}^{\infty} \exp \left( \frac{-[(x-(\mu+\sigma^2 t))^2]}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi\sigma^2}} \, dx
\]
\[
= \exp \left( \mu t + \frac{\sigma^2 t^2}{2} \right)
\]
as
\[
\int_{-\infty}^{\infty} \exp \left( \frac{-[(x-(\mu+\sigma^2 t))^2]}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi\sigma^2}} \, dx = 1.
\]

Note that
\[
M'_X(t) = (\mu + \sigma^2 t) \cdot M_X(t)
\]
and
\[
M''_X(t) = (\mu + \sigma^2 t)^2 \cdot M_X(t) + \sigma^2 \cdot M_X(t)
\]
which provide

\[ E(X) = M'_X(0) = \mu \]

and

\[ E(X^2) = M''_X(0) = \mu^2 + \sigma^2. \]

Lastly,

\[ \text{Var}(X) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2, \]

finishing the proof.

**Theorem 149.** If \( X \sim N(\mu, \sigma^2) \), then, for \( Y = \frac{X - \mu}{\sigma} \), \( Y \sim N(0, 1) \).

**Proof.** Observe that

\[
Y \leq y \iff \frac{X - \mu}{\sigma} \leq y \iff X \leq \sigma y + \mu.
\]

That is,

\[
P(Y \leq y) = P(X \leq \sigma y + \mu) = \int_{-\infty}^{\sigma y + \mu} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \, dx.
\]

Let \( y = \frac{x - \mu}{\sigma} \). Then \( dy = \frac{dx}{\sigma} \) and

\[
P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \sigma \, dy = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \, dy
\]

which is the cumulative density for \( Y \). Finally, differentiating with respect to \( y \), we see that the probability density for \( Y \) is given by

\[ g(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right), \]

the standard normal distribution.

Since integrals involving normal distributions can’t be calculated by hand, Theorem 149 allows us to appeal only to the standard normal distribution when finding probabilities.

**Example 115.** Suppose the length of a metal rod is normally distributed with a mean of 5 cm and a variance of 0.04. Find the probability that a randomly selected rod is between 4.9 and 5.3 cm long.

Let \( Y \) be the length of the metal rod and

\[ X = \frac{Y - 5}{\sqrt{0.04}} = \frac{Y - 5}{0.2}. \]

Notice that

\[ Y = 4.9 \implies X = -0.5 \]

and

\[ Y = 5.3 \implies X = 1.5. \]
Then, we can use a computational engine to calculate
\[
\int_{-0.5}^{1.5} \exp\left(\frac{-x^2}{2}\right) \frac{dx}{\sqrt{2\pi}} \approx 0.624655
\]

**Theorem 150.** If \( X \sim \mathcal{N}(0, 1) \), then \( X^2 \sim \chi^2(1) \). Moreover, if \( X_j \sim \mathcal{N}(0, 1) \) for \( j = 1, 2, \ldots, n \) where \( X_1, X_2, \ldots, X_n \) are independent, then
\[
X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n).
\]

**Proof.** Notice that, for \( x > 0 \),
\[
P(X^2 \leq x) = P(-\sqrt{x} \leq X \leq \sqrt{x})
= 2 \cdot P(0 \leq X \leq \sqrt{x})
= 2 \cdot \int_{0}^{\sqrt{x}} \frac{\exp\left(\frac{-t^2}{2}\right)}{\sqrt{2\pi}} dt
\]
Let \( u = t^2 \) and notice that \( du = 2t \, dt \), \( u = 0 \) when \( t = 0 \), and \( u = x \) when \( t = \sqrt{x} \). Also, \( \sqrt{u} = t \) so \( 2t = 2\sqrt{u} \). Hence,
\[
P(X^2 \leq x) = 2 \cdot \int_{0}^{\sqrt{x}} \frac{\exp\left(\frac{-u}{2}\right)}{\sqrt{2\pi}} \cdot \frac{1}{2\sqrt{u}} \, du
= \int_{0}^{x} \frac{u^{-1/2}e^{-u/2}}{\sqrt{2\pi}} \, du.
\]
By differentiating with respect to \( x \), we obtain the probability density for \( X \):
\[
f(x) = \begin{cases} 
\frac{x^{-1/2}e^{-x/2}}{\sqrt{2\pi}}, & x > 0; \\
0, & x \leq 0.
\end{cases}
\]
Recall that the probability density corresponding to \( \chi^2(1) \) is
\[
g(x) = \begin{cases} 
\frac{x^{-1/2}e^{-x/2}}{\sqrt{2} \cdot \Gamma(1/2)}, & x > 0; \\
0, & x \leq 0.
\end{cases}
\]
Thus,
\[
\int_{0}^{\infty} \frac{x^{-1/2}e^{-x/2}}{\sqrt{2\pi}} \, dx = 1 = \int_{0}^{\infty} \frac{x^{-1/2}e^{-x/2}}{\sqrt{2} \cdot \Gamma(1/2)} \, dx
\]
which means that \( \Gamma(1/2) = \sqrt{\pi} \). In particular, \( X^2 \sim \chi^2(1) \).

For the more general statement, suppose \( X_1, X_2, \ldots, X_n \) are as in the statement of the theorem. Then \( X_j^2 \sim \chi^2(1) \). Since \( \chi^2(1) \equiv \Gamma(1/2, 1/2) \), we apply Theorem 144 to obtain that
\[
X_1^2 + X_2^2 + \cdots + X_n^2 \sim \Gamma(n/2, 1/2) \equiv \chi^2(n).
\]
5.2.5 Bivariate Normal Distributions

The normal distribution can be generalized to higher dimensions. For the sake of sanity, we will only mention the bivariate case.

**Definition 151.** Two continuous random variables $X$ and $Y$ are said to have a bivariate normal distribution if their joint probability density is given by

$$f(x, y) = \frac{\exp \left( \frac{-1}{2(1-\rho^2)} \cdot \left[ \left( \frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x-\mu_1}{\sigma_1} \right) \left( \frac{y-\mu_2}{\sigma_2} \right) + \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right)}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}}$$

for $x, y \in \mathbb{R}$, $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, and $-1 < \rho < 1$. The quantity $\rho$ is known as the **correlation coefficient**.

**Remark.** In fact, the quantity $\rho$ for a bivariate normal distribution is related to the quantity discussed in Theorem 92 as will be directly addressed in Definition 155.

**Note.** It will be helpful to identify particular cases of bivariate normal distributions like the one with parameters $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$:

$$f(x, y) = \frac{\exp \left( \frac{-(x^2 + y^2)}{2(1-\rho^2)} \right)}{2\pi \sqrt{1-\rho^2}}.$$

**Theorem 152.** If $X$ and $Y$ have a bivariate normal distribution, then the marginal density of $X$ is given by

$$g(x) = \frac{\exp \left( \frac{-(x-\mu_1)^2}{2\sigma_1^2} \right)}{\sqrt{2\pi \sigma_1^2}}$$

and the marginal density of $Y$ is given by

$$h(y) = \frac{\exp \left( \frac{-(y-\mu_2)^2}{2\sigma_2^2} \right)}{\sqrt{2\pi \sigma_2^2}}$$

Particularly, $E(X) = \mu_1$, $\text{Var}(X) = \sigma_1^2$, $E(Y) = \mu_2$, and $\text{Var}(Y) = \sigma_2^2$.

**Proof.** By symmetry, we will only prove the statement for the marginal density of $X$. Let

$$u = \frac{x - \mu_1}{\sigma_1} \text{ and } v = \frac{y - \mu_2}{\sigma_2}$$

and notice that

$$g(x) = \int_{-\infty}^{\infty} \exp \left( \frac{-1}{2(1-\rho^2)} \cdot \left[ \left( \frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x-\mu_1}{\sigma_1} \right) \left( \frac{y-\mu_2}{\sigma_2} \right) + \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right) \cdot \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} dy$$

$$= \exp \left( \frac{-u^2}{2(1-\rho^2)} \right) \cdot \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} \exp \left( \frac{2\rho uv - v^2}{2(1-\rho^2)} \right) \cdot \sigma_2 \cdot dv$$

$$= \exp \left( \frac{-u^2}{2(1-\rho^2)} \right) \cdot \frac{1}{2\pi \sigma_1 \sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} \exp \left( \frac{2\rho uv - v^2}{2(1-\rho^2)} \right) \cdot dv$$

Observe that

$$v^2 - 2\rho uv = v^2 - 2\rho uv + \rho^2 u^2 - \rho^2 u^2 = (v - \rho u)^2 - \rho^2 u^2.$$
Then

\[ g(x) = \frac{\exp \left( \frac{-u^2}{2(1-\rho^2)} \right)}{2\pi\sigma_1 \sqrt{1-\rho^2}} \cdot \int_{-\infty}^{\infty} \exp \left( \frac{\rho^2 u^2 - (v - \rho u)^2}{2(1-\rho^2)} \right) dv \]

\[ = \frac{\exp \left( \frac{-u^2}{2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_1^2}} \cdot \exp \left( \frac{\rho^2 u^2}{2(1-\rho^2)} \right) \int_{-\infty}^{\infty} \exp \left( \frac{(v - \rho u)^2}{2(1-\rho^2)} \right) dv \]

\[ = \frac{\exp \left( \frac{-u^2}{2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_1^2}} \cdot \exp \left( \frac{\rho^2 u^2}{2(1-\rho^2)} \right) \exp \left( \frac{\rho^2 v^2}{2(1-\rho^2)} \right) \]

\[ = \frac{\exp \left( -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right)}{\sqrt{2\pi\sigma_1^2}} \]

\[ \square \]

**Theorem 153.** If \( X \) and \( Y \) have a bivariate normal distribution, then the conditional density of \( X \) given \( Y \) is

\[ f(x|y) = \frac{\exp \left( -\frac{(x-(\mu_1+\rho \frac{\sigma_1}{\sigma_2} (y-\mu_2))^2}{2\sigma_1^2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \]

and the conditional density of \( Y \) given \( X \) is

\[ f(y|x) = \frac{\exp \left( -\frac{(y-(\mu_2+\rho \frac{\sigma_2}{\sigma_1} (x-\mu_1))^2}{2\sigma_2^2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \]

**Proof.** By symmetry, we will only provide the statement for \( f(x|y) \). Let

\[ u = \frac{x - \mu_1}{\sigma_1} \quad \text{and} \quad v = \frac{y - \mu_2}{\sigma_2} \]

and notice that the joint density is

\[ \exp \left( -\frac{(u^2 - 2\rho uv + v^2)}{2(1-\rho^2)} \right) \frac{2\pi\sigma_1\sigma_2}{\sqrt{1-\rho^2}}. \]

By Theorem 152, we see that the marginal density for \( Y \) is given by

\[ \frac{\exp \left( -\frac{1}{2} \cdot \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right)}{\sqrt{2\pi\sigma_2^2}} = \exp \left( -\frac{v^2}{2} \right) \]

Then, the conditional density is computed by

\[ \frac{\exp \left( -\frac{(u^2 - 2\rho uv + v^2)}{2(1-\rho^2)} \right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \sqrt{2\pi\sigma_2^2} = \frac{\exp \left( \frac{v^2(1-\rho^2) - (u^2 - 2\rho uv + v^2)}{2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \]

\[ = \frac{\exp \left( \frac{v^2 - \rho^2 v^2 - u^2 + 2\rho uv - v^2}{2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \]

\[ = \frac{\exp \left( -\frac{(u^2 - 2\rho uv + v^2)}{2(1-\rho^2)} \right)}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \]
5.2. CONTINUOUS DENSITIES

\[
= \exp \left( \frac{-(u - \mu v)^2}{2(1 - \rho^2)} \right)
\]

To finish up, we need only manipulate the exponent:

\[
\frac{-(u - \mu v)^2}{2(1 - \rho^2)} = -\frac{1}{2(1 - \rho^2)} \cdot \left[ \frac{x - \mu_1 - \rho \cdot \sigma_1}{\sigma_2} \cdot \frac{y - \mu_2}{\sigma_2} \right]^2
\]

\[
= -\frac{1}{2(1 - \rho^2)} \cdot \left[ \frac{x - \mu_1 - \rho \cdot \sigma_1}{\sigma_1} \cdot \frac{y - \mu_2}{\sigma_2} \right]^2
\]

\[
= \frac{1}{2\sigma_1^2(1 - \rho^2)} \cdot \left[ x - \mu_1 + \rho \cdot \frac{\sigma_1}{\sigma_2} \cdot (y - \mu_2) \right]^2
\]

\[
= \rho \cdot \frac{\sigma_1}{\sigma_2} \cdot (y - \mu_2).
\]

**Theorem 154.** If \(X\) and \(Y\) have a bivariate normal distribution, then \(\text{cov}(X, Y) = \rho\sigma_1\sigma_2\).

**Proof.** Note that, by Theorem 153,

\[
E[(X - \mu_1)|Y = y] = \int_{-\infty}^{\infty} \frac{(x - \mu_1) f(x|y)}{h(y)} \, dx
\]

Recall that

\[
f(x, y) = f(x|y) \cdot h(y)
\]

where \(h(y)\) is the marginal density of \(Y\). Then, appealing to Theorem 152 when necessary,

\[
\text{cov}(X, Y) = E((X - \mu_1)(Y - \mu_2))
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_1)(y - \mu_2) f(x, y) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_1)(y - \mu_2) f(x|y) h(y) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} (y - \mu_2) h(y) \cdot \int_{-\infty}^{\infty} (x - \mu_1) f(x|y) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} (y - \mu_2) h(y) \cdot \rho \cdot \frac{\sigma_1}{\sigma_2} (y - \mu_2) \, dy
\]

\[
= \rho \cdot \frac{\sigma_1}{\sigma_2} \cdot \text{Var}(Y)
\]

\[
= \rho \cdot \frac{\sigma_1}{\sigma_2} \cdot \frac{\sigma_2^2}{\sigma_2^2}
\]

\[
= \rho \sigma_1 \sigma_2.
\]
Definition 155. For two random variables $X$ and $Y$, the correlation coefficient is defined to be
\[
\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}
\]
where $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$, respectively.

Theorem 156. Two random variables $X$ and $Y$ with a bivariate normal distribution are independent if and only if their correlation coefficient $\rho = 0$.

Proof. First, suppose $X$ and $Y$ are independent and let
\[
f(x, y) = \exp \left( \frac{-1}{2(1-\rho^2)^2} \left[ \left( \frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x-\mu_1}{\sigma_1} \right) \left( \frac{y-\mu_2}{\sigma_2} \right) + \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right) \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-\rho^2}},
\]
the joint probability density of $X$ and $Y$. By Theorem 152 the marginal density of $X$ is given by
\[
g(x) = \exp \left( \frac{-(x-\mu_1)^2}{2\sigma_1^2} \right) \sqrt{\frac{1}{2\pi\sigma_1^2}}
\]
and the marginal density of $Y$ is given by
\[
h(y) = \exp \left( \frac{-(y-\mu_2)^2}{2\sigma_2^2} \right) \sqrt{\frac{1}{2\pi\sigma_2^2}}
\]
Since $X$ and $Y$ are independent, $g(x)h(y) = f(x, y)$.

\[
\frac{\exp \left( - \left( \frac{(x-\mu_1)^2 + (y-\mu_2)^2)}{2\sigma_1^2 \sigma_2^2} \right) \right)}{2\pi\sigma_1\sigma_2} = \frac{\exp \left( - \frac{1}{2(1-\rho^2)^2} \left[ \left( \frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x-\mu_1}{\sigma_1} \right) \left( \frac{y-\mu_2}{\sigma_2} \right) + \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right)}{2\pi\sigma_1\sigma_2 \sqrt{1-\rho^2}}.
\]
In particular, at $(\mu_1, \mu_2)$, this provides
\[
1 = \frac{1}{\sqrt{1-\rho^2}} \implies 1 = \sqrt{1-\rho^2} \implies \rho^2 = 0 \implies \rho = 0.
\]

On the other hand, if $\rho = 0$, then the joint probability density for $X$ and $Y$ is
\[
f(x, y) = \exp \left( \frac{-1}{2\sigma_1^2 \sigma_2^2} \left[ \left( \frac{x-\mu_1}{\sigma_1} \right)^2 + \left( \frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right) = \exp \left( - \left( \frac{(x-\mu_1)^2 + (y-\mu_2)^2)}{2\sigma_1^2 \sigma_2^2} \right) \right)
\]
By Theorem 152 the marginal density of $X$ is given by
\[
g(x) = \exp \left( \frac{-(x-\mu_1)^2}{2\sigma_1^2} \right) \sqrt{\frac{1}{2\pi\sigma_1^2}}
\]
and the marginal density of $Y$ is given by
\[
h(y) = \exp \left( \frac{-(y-\mu_2)^2}{2\sigma_2^2} \right) \sqrt{\frac{1}{2\pi\sigma_2^2}}
\]
Hence, as $g(x)h(y) = f(x, y)$, we see that $X$ and $Y$ are independent. \qed


Exercise 21. Let
\[ f(x, y) = \frac{\exp\left(\frac{-\left(x^2+y^2\right)}{2}\right)}{2\pi}. \]

Also, let
\[ R_1 = \{(x, y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < 1\} \cup \{(x, y) \in \mathbb{R}^2 : -1 < x < 0, -1 < y < 0\} \]

and
\[ R_2 = \{(x, y) \in \mathbb{R}^2 : -1 < x < 0, 0 < y < 1\} \cup \{(x, y) \in \mathbb{R}^2 : 0 < x < 1, -1 < y < 0\}, \]

These regions are visualized as follows:

Then let \( X \) and \( Y \) have joint probability density
\[ g(x, y) = \begin{cases} 
 0, & (x, y) \in R_1; \\
 2 \cdot f(x, y), & (x, y) \in R_2; \\
 f(x, y), & \text{otherwise}. 
\end{cases} \]

Verify that the marginal densities for \( X \) and \( Y \) are normal distributions.
Chapter 6

Using the Standard Normal Distribution

6.1 Producing and Using $z$-tables

6.1.1 Approximating Standard Normal Probability

Notice that the standard normal distribution is the function

$$e^{-x^2/2} = \frac{1}{\sqrt{2\pi}}, \sum_{k=0}^{\infty} \frac{(-x^2/2)^k}{k!} = \sum_{k=0}^{\infty} \frac{(-1)^k \cdot x^{2k}}{2^k k! \sqrt{2\pi}}$$

So we can approximate the standard normal distribution with these Taylor polynomials

$$\sum_{k=0}^{n} \frac{(-1)^k \cdot x^{2k}}{2^k k! \sqrt{2\pi}}$$

for large enough values of $n$. Please visit Desmos.com for an interactive visualization of these approximations.

Hence, for values of $x$ close enough to 0 and large enough values of $n$,

$$\int_{-\infty}^{x} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} \approx C + \sum_{k=0}^{n} \frac{(-1)^k \cdot x^{2k+1}}{2k + 1)! 2^k \sqrt{2\pi}}.$$

To find the constant $C$, notice that

$$\frac{1}{2} = \int_{-\infty}^{0} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}$$

and that

$$C + \sum_{k=0}^{n} \frac{(-1)^k \cdot 0^{2k+1}}{(2k + 1)! 2^k \sqrt{2\pi}} = C.$$

So we can approximate

$$\int_{-\infty}^{x} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}$$

with

$$\frac{1}{2} + \sum_{k=0}^{n} \frac{(-1)^k \cdot x^{2k+1}}{(2k + 1)! 2^k \sqrt{2\pi}}.$$

In particular, if $Z \sim \mathcal{N}(0,1)$,

$$P(Z \leq z) \approx \frac{1}{2} + \sum_{k=0}^{n} \frac{(-1)^k \cdot z^{2k+1}}{(2k + 1)! 2^k \sqrt{2\pi}}.$$
6.1.2 Approximating $z$ Given a Probability

Sometimes, we wish to find a value $z$ so that

$$
\int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx = p
$$

for some $0 \leq p \leq 1$. The most immediate way is to use a $z$-table and search for values of $z$ that satisfy the desired equality.

Another way is to use the Newton-Raphson method, a numerical method for approximating zeros of any differentiable function $f(z)$. Notice that our task is finding $z$ so that

$$
\int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx - p = 0.
$$

Using our approximations, we can try to solve

$$
\frac{1}{2} - p + \sum_{k=0}^{n} \frac{(-1)^k \cdot z^{2k+1}}{(2k+1)2^k k!\sqrt{2\pi}} = 0.
$$

Recall the Newton-Raphson method tells us to start with some initial guess $z_0$ and then recursively define

$$
z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}
$$

as long as $f'(z_n) \neq 0$. Then, under the right conditions, $z_n \to z$ where $f(z) = 0$.

In this particular context, we have that

$$f(z) = \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx - p
$$

and

$$f'(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}.$$

**Example 116.** Suppose the spiritual energy level $X$ caused by a cosmic whisper is a normally distributed random variable with mean 133 consciousness units and standard deviation 6 consciousness units. Xenologists experience Acute Celestial Awareness if the spiritual energy level registered from a cosmic whisper is in the top 1%. Find how many consciousness units invoke Acute Celestial Awareness.

Using the standard normal distribution as a point of reference, we need to find $z$ so that

$$0.01 = \int_{z}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx = \int_{-\infty}^{-z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx.$$

Using a $z$-table or the Newton-Raphson method, $z \approx 2.326$. Now, let $A$ be the consciousness units which invoke Acute Celestial Awareness and notice that

$$\frac{A - 133}{6} \approx 2.326 \implies A \approx 146.956$$

Therefore, it takes about 147 consciousness units to invoke Acute Celestial Awareness.
6.2 Approximating the Binomial Distribution

Theorem 157. Suppose $X$ and $Y$ are random variables so that $M_X(t)$ and $M_Y(t)$ exist on some non-degenerate interval $I$ about 0 and that $M_X(t) = M_Y(t)$ for all $t \in I$. Then $X$ and $Y$ have the same probability distribution/density.

Although a proof for the general statement exceeds the scope of this course, we will prove it for random variables with finite domains consisting of positive integers as preliminary justification.

Proof. Suppose $X$ and $Y$ are random variables with values contained in $A = \{1, 2, \ldots, n\}$ where $f : A \to [0, 1]$ is the probability distribution for $X$ and $g : A \to [0, 1]$ is the probability distribution for $Y$. Notice that, if $y = e^t$ for $t \in I$,

$$0 = M_X(t) - M_Y(t) = E(e^{tX}) - E(e^{tY}) = \left[ \sum_{k=1}^{n} e^{tk}f(k) \right] - \left[ \sum_{k=1}^{n} e^{tk}g(k) \right] = \sum_{k=1}^{n} (f(k) - g(k))y^k,$$

a polynomial in $y$ where $y$ ranges over some non-degenerate interval. Since the only polynomial which assumes the value of zero on a non-degenerate interval is the constant zero polynomial, we see that $f(k) = g(k)$ for each $k \in A$, establishing that $X$ and $Y$ have the same probability density. \qed

Before we continue let’s revisit the discussion alluded to in Section 3.3 concerning discrete random variables and continuous random variables. We can represent a binomial distribution $B(n, p)$ with

$$f(x) = \begin{cases} 0, & x < -0.5; \\ \left(\frac{n}{x+0.5}\right)p^{|x+0.5|}(1-p)^{n-|x+0.5|}, & -0.5 \leq x < n + 0.5; \\ 0, & x \geq n + 0.5. \end{cases}$$

This a piece-wise continuous function for $x \in \mathbb{R}$ and, for $X \sim B(n, p)$,

$$P(X = k) = \int_{k-0.5}^{k+0.5} f(x) \, dx$$

for $k = 0, 1, \ldots, n$.

Theorem 158. If $X \sim B(n, p)$, then the moment-generating function for

$$Y = \frac{X - np}{\sqrt{np(1-p)}}$$

for large enough $n$ is approximately the moment-generating function for a standard normal distribution. Formally, if $X_n \sim B(n, p)$ and

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}},$$

then, for any $t$,

$$\lim_{n \to \infty} M_{Y_n}(t) = e^{t^2/2}.$$

Proof. By Theorem 112, we know that

$$M_{X_n}(t) = [pe^t - p + 1]^n$$
and Theorem 84 yields

\[ M_{Y_n}(t) = \exp\left(\frac{-npt}{\sigma}\right) \left[ p \cdot \exp\left(\frac{t}{\sigma}\right) - p + 1 \right]^n \]

where \( \sigma = \sqrt{np(1-p)} \). Now, observe that

\[
\ln \left[ M_{Y_n}(t) \right] = \frac{-npt}{\sigma} + n \cdot \ln \left[ p \cdot \exp\left(\frac{t}{\sigma}\right) - p + 1 \right]
= \frac{-npt}{\sigma} + n \cdot \ln \left( 1 + p \cdot \left[ \exp\left(\frac{t}{\sigma}\right) - 1 \right] \right)
\]

Recall that, using the Taylor series expansion of \( \ln(1 + x) \),

\[
\ln \left( 1 + p \cdot \left[ \exp\left(\frac{t}{\sigma}\right) - 1 \right] \right) = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{p^k}{k} \left[ \exp\left(\frac{t}{\sigma}\right) - 1 \right]^k
\]

Also using series,

\[
\exp\left(\frac{t}{\sigma}\right) - 1 = \left[ \sum_{k=0}^{\infty} \frac{1}{k!} \cdot \left(\frac{t}{\sigma}\right)^k \right] - 1
= \sum_{k=1}^{\infty} \frac{1}{k!} \cdot \left(\frac{t}{\sigma}\right)^k
\]

Putting these together,

\[
\ln \left( 1 + p \cdot \left[ \exp\left(\frac{t}{\sigma}\right) - 1 \right] \right) = \sum_{k=1}^{\infty} (-1)^{k+1} \cdot \frac{p^k}{k} \left[ \sum_{j=1}^{\infty} \frac{1}{j!} \cdot \left(\frac{t}{\sigma}\right)^j \right]^k
= p \cdot \left[ \sum_{j=1}^{\infty} \frac{1}{j!} \cdot \left(\frac{t}{\sigma}\right)^j \right] - \frac{p^2}{2} \left[ \frac{t}{\sigma} + \left(\frac{t}{\sigma}\right)^2 + O\left(\frac{t^3}{\sigma^3}\right) \right]^2
+ O\left(\frac{t^3}{\sigma^3}\right)
= \frac{pt}{\sigma} + \frac{pt^2}{\sigma^2} + O\left(\frac{t^3}{\sigma^3}\right)
- \frac{p^2}{2} \left[ \frac{t^2}{\sigma^2} + O\left(\frac{t^3}{\sigma^3}\right) \right]
+ O\left(\frac{t^3}{\sigma^3}\right)
= \frac{pt}{\sqrt{np(1-p)}} + \frac{pt^2(1-p)}{2np(1-p)} + O\left(\frac{t^3}{\sigma^3}\right)
= \frac{pt}{\sqrt{np(1-p)}} + \frac{t^2}{2n} + O\left(\frac{t^3}{\sigma^3}\right)
\]
Hence,
\[
\ln [M_{Y_n}(t)] = -\frac{npt}{\sigma} + n \cdot \ln \left(1 + p \cdot \left[\exp \left(\frac{t}{\sigma}\right) - 1\right]\right)
\]
\[
= \frac{-npt}{\sqrt{np(1-p)}} + n \cdot \frac{pt}{\sqrt{np(1-p)}} + n \cdot \frac{t^2}{2n} + n \cdot O\left(\frac{t^3}{\sigma^3}\right)
\]
\[
= \frac{t^2}{2} + n \cdot O\left(\frac{t^3}{\sigma^3}\right).
\]

Now,
\[
n \cdot O\left(\frac{t^3}{\sigma^3}\right)
\]
is a series consisting of terms of the form
\[
\frac{(t^3)^k}{n^{k/2}p^{3k/2}(1 - p)^{3k/2}}
\]
That is,
\[
\lim_{n \to \infty} n \cdot O\left(\frac{t^3}{\sigma^3}\right) = 0.
\]

Finally, as
\[
\lim_{n \to \infty} \ln [M_{Y_n}(t)] = \frac{t^2}{2},
\]
we see that
\[
\lim_{n \to \infty} M_{Y_n}(t) = e^{t^2/2}.
\]

As we will see later, Theorem 158 implies that the \textit{standardized} binomial distributions $Y_n$ actually converge to the standard normal distribution. For an interactive graph relating standardized binomial distributions and the standard normal distribution, please visit Desmos.com. For now, let’s elaborate on this via example.

\textbf{Example 117.} Suppose a coin is weighted so that it lands on heads with a probability of 62\%. Use a normal distribution to approximate the probability that, of 100 tosses, exactly 38 tails appear.

Recall that we can model probability distributions with probability densities by extending the probability distribution to intervals of base length one. As getting 38 tails is equivalent to getting 62 heads, we are thus interested in the interval 61.5 to 62.5. Since the mean of the given random variable is $100 \cdot 0.62 = 62$ and the standard deviation is
\[
\sqrt{100 \cdot 0.62 \cdot 0.38} \approx 4.85386,
\]
we convert
\[
\frac{61.5 - 62}{4.85386} \approx -0.103 \quad \text{and} \quad \frac{62.5 - 62}{4.85386} \approx 0.103
\]
Now, using the standard normal distribution,
\[
\int_{-0.103}^{0.103} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \approx 0.082037
\]

\textbf{Remark.} Had we used the binomial distribution in Example 117 instead, notice that
\[
\binom{100}{62} (0.62)^{62} (0.38)^{38} \approx 0.0819687
\]
Remark. Recall that binomial distributions can be used to approximate hypergeometric distributions and that Poisson distributions can be used to approximate binomial distributions. In fact, normal distributions can be used to approximate all of these. As we will see later, this alludes to an important and central result of statistics.

6.3 Normal Score Plots

Suppose you have \( n \) numerical observations for some random variable \( X \) and you want to determine if \( X \) is approximately normally distributed. So let \( Z \sim N(0, 1) \) and order your observations from least to greatest:

\[
x_1 \leq x_2 \leq \cdots \leq x_n.
\]

Now, with these data, we stipulate

\[
P(X \leq x_j) = \frac{j}{n+1}.
\]

For each \( j = 1, 2, \ldots, n \), let \( z_j \) be so that \( P(Z \leq z_j) = \frac{j}{n+1} \). Then the normal scores plot is

\[
\{(z_j, x_j) : j = 1, 2, \ldots, n\}.
\]

If the points of this plot exhibit a linear relationship (with non-zero slope), then \( X \) appears to have a distribution similar to \( aY + b \) for constants \( a \) and \( b \) where \( a > 0 \). In other words, \( X \) appears to be normally distributed.

Example 118. Suppose we’ve made four measurements: 6, 3, 7, 9. If we view them on a number line,

we see that the number line is split up into five pieces. And hence, we weigh each piece equally:

Organizing this into a table:

\[
\begin{array}{c|c}
  x & P(X \leq x) \\
  \hline
  3 & 1/5 \\
  6 & 2/5 \\
  7 & 3/5 \\
  9 & 4/5 \\
\end{array}
\]

The corresponding \( z \)-values would be

\[
\begin{array}{c|c}
  z & P(Z \leq z) \\
  \hline
  -0.8416 & 1/5 \\
  -0.2533 & 2/5 \\
  0.2533 & 3/5 \\
  0.8416 & 4/5 \\
\end{array}
\]

Now, the corresponding normal score plot is
These four points are somewhat co-linear so, based on these data alone, we suspect that $X$ is approximately normally distributed.
Part II

Statistics
Chapter 7

Functions of Random Variables

As we have seen in the Introduction to Probability Theory, given a collection of random variables \(X_1, X_2, \ldots, X_n\) along with their probability distributions or densities, we may be interested in finding the probability distribution or density of \(T(X_1, X_2, \ldots, X_n)\) where \(T\) is a real-valued function. For example, we have found the probability density of \(X^2\) where \(X\) has the standard normal distribution. Here we will cover three different techniques to accomplish such tasks, in general.

7.1 Distribution Function Technique

Given random variables \(X_1, X_2, \ldots, X_n\) and a real-valued function \(Y = T(X_1, X_2, \ldots, X_n)\), if we can find a differentiable expression for

\[
F(y) = P(Y \leq y) = P[T(X_1, X_2, \ldots, X_n) \leq y],
\]

then the probability density for \(Y\) is given by

\[
\frac{d}{dy} F(y).
\]

Exercise 22. Suppose \(X\) is uniformly distributed on the interval \([0, 1]\). Use the Distribution Function Technique to find the probability density of \(X^2\).

Example 119. Suppose the joint probability density for \(X\) and \(Y\) is given by

\[
f(x, y) = \begin{cases} 
6e^{-2x-3y}, & x, y > 0; \\
0, & \text{otherwise}.
\end{cases}
\]

Find the probability density for \(Z = X + Y\).

Consider the \(xy\)-plane and the inequality

\[
Z = X + Y \leq z.
\]

Graphically, we have

115
To find $P(Z \leq z)$, we integrate

$$
\int_0^z \int_0^{-x+z} 6e^{-2x-3y} \, dy \, dx = 6 \cdot \int_0^z e^{-2x} \cdot \int_0^{-x+z} e^{-3y} \, dy \, dx
$$

$$
= 6 \cdot \int_0^z e^{-2x} \cdot \left[ e^{-3y} \right]_{-x+z}^0 \, dx
$$

$$
= -2 \cdot \int_0^z e^{-2x} \cdot e^{3x-3z} - e^{-2x} \, dx
$$

$$
= -2 \cdot \left[ e^{-3z} \cdot e^x + e^{-2x} \right]_{0}^{z}
$$

$$
= -2 \cdot \left[ e^{-2z} - e^{-3z} + e^{-2z} - \frac{1}{2} \right]
$$

$$
= -3e^{-2z} + 2e^{-3z} + 1.
$$

Differentiating, we find the probability density for $Z = X + Y$ to be

$$
f(z) = \begin{cases} 
6e^{-2z} - 6e^{-3z}, & z \geq 0; \\
0, & \text{otherwise.}
\end{cases}
$$

### 7.2 Transformation Technique

#### 7.2.1 Single Variable

**Example 120.** Suppose $X$ is the number of green balls in a random sample of four balls taken from a bag which contains twelve red balls and eight green balls. Find the probability distribution of $(X - 2)^2$.

Consider

<table>
<thead>
<tr>
<th>$X$</th>
<th>$(X - 2)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

and that the probability distribution for $X$ is given by

$$
f(x) = \frac{\binom{8}{x} \binom{12}{4-x}}{\binom{20}{4}}, \quad x = 0, 1, 2, 3, 4.
$$
Finally, to find the probability distribution \( g(y) \) for \( Y = (X - 2)^2 \), observe that
\[
\begin{align*}
g(0) &= f(2), \\
g(1) &= f(1) + f(3), \\
g(4) &= f(0) + f(4).
\end{align*}
\]

**Theorem 159.** Suppose \( f(x) \) is the probability density for a continuous random variable \( X \). Also, suppose that \( y = u(x) \) is differentiable and either strictly increasing or decreasing on the set \( D = \{ x : f(x) \neq 0 \} \). Let \( v(x) \) be the inverse of \( u(x) \) on \( D \) which provides \( x = v(y) \) for all \( x \in D \). Then the probability density for \( Y = u(X) \) is given by
\[
g(y) = \begin{cases} f(v(y)) \cdot |v'(y)|, & x = v(y) \in D, \ u'(x) \neq 0; \\ 0, & \text{otherwise} \end{cases}
\]

**Proof.** We first consider the case when \( u(x) \) is strictly increasing. Notice that, for \( x = v(y) \), we have
\[
dx = v'(y) \ dy
\]
which provides
\[
P(a < Y < b) = P(v(a) < X < v(b))
= \int_{v(a)}^{v(b)} f(x) \ dx
= \int_{a}^{b} f(v(y)) \cdot v'(y) \ dy.
\]
By the facts that \( y = u(x) \implies \frac{dy}{dx} = u'(x) \geq 0 \) and
\[
x = v(y) \implies 1 = v'(y) \cdot \frac{dy}{dx} \implies v'(y) = \frac{1}{u'(x)}
\]
whenever \( u'(x) \neq 0 \), we see that \( v'(y) \geq 0 \). Hence, the probability density for \( Y \) is as proposed.

Now we consider the case when \( u(x) \) is strictly decreasing. In a similar fashion to the increasing case, we compute
\[
P(a < Y < b) = P(v(b) < X < v(a))
= \int_{v(b)}^{v(a)} f(x) \ dx
= \int_{a}^{b} f(v(y)) \cdot v'(y) \ dy
= -\int_{a}^{b} f(v(y)) \cdot v'(y) \ dy.
\]
As above, we see that \( v'(y) \leq 0 \) for \( x \in D \) so that \( u'(x) \neq 0 \). Hence, \( |v'(y)| = -v'(y) \) and we obtain
\[
P(a < Y < b) = \int_{a}^{b} f(v(y)) \cdot |v'(y)| \ dy.
\]
This finishes the proof.

A trivial consequence of this is the following.

**Proposition 160.** If the probability density function for a continuous random variable \( X \) is \( f(x) \), then the probability density for \( Y = aX + b \) where \( a \) and \( b \) are constants with \( a \neq 0 \) is
\[
g(y) = \frac{f((y - b)/a)}{|a|}.
\]
Proof. Notice that

\[ y = ax + b \implies x = \frac{y - b}{a} \implies \frac{dx}{dy} = \frac{1}{a}. \]

The conclusion is then a direct application of Theorem 159.

**Example 121.** If \( X \) has the exponential distribution given by

\[ f(x) = \begin{cases} e^{-x}, & x > 0; \\ 0, & \text{otherwise}. \end{cases} \]

find the probability density of \( Y = \sqrt{X} \).

Since \( y = \sqrt{x} \) is a strictly increasing function when \( x > 0 \), we can apply Theorem 159. The inverse is given by \( x = v(y) = y^2 \) which provides

\[ v'(y) = \frac{dx}{dy} = 2y. \]

Therefore,

\[ g(y) = \begin{cases} 2ye^{-y^2}, & y > 0; \\ 0, & \text{otherwise}. \end{cases} \]

**Example 122.** Suppose that \( \Theta \) is a random variable which is uniformly distributed on the interval \([-\pi/2, \pi/2]\); i.e., its probability density is given by

\[ f(\theta) = \begin{cases} 1/\pi, & -\pi/2 < \theta < \pi/2; \\ 0, & \text{otherwise}. \end{cases} \]

Given the resulting angle \( \theta \), a particular point \((1, y)\) is determined on the line \( x = 1 \) visualized as follows.

From the diagram, we see that

\[ y = \tan(\theta) \]

for \(-\pi/2 < \theta < \pi/2\). Tangent is differentiable and strictly increasing on this interval so we can apply Theorem 159. In particular, we see that

\[ \theta = v(y) = \arctan(y) \implies \frac{d\theta}{dy} = \frac{1}{1+y^2}. \]
Since \( f(v(y)) = 1/\pi \), we see that
\[
g(y) = \frac{1}{\pi + \pi y^2}
\]
for \( y \in \mathbb{R} \).

**Example 123.** Suppose the random variable \( \Theta \) is uniformly distributed on the interval \([0, \pi]\). Find the probability density for \( X = \cos(\Theta) \).

Notice that \( \cos(\theta) \) is strictly decreasing when \( 0 \leq \theta \leq \pi \). Also,
\[
x = \cos(\theta) \implies \theta = v(x) = \arccos(x)
\]
which gives
\[
v'(x) = \frac{d\theta}{dx} = \frac{-1}{\sqrt{1 - x^2}}
\]
for \(-1 < x < 1\). Therefore, by Theorem 159, we see that the probability density for \( X \) is
\[
f(x) = \begin{cases} \frac{1}{\pi \sqrt{1 - x^2}}, & -1 < x < 1; \\ 0, & \text{otherwise}. \end{cases}
\]

### 7.2.2 Several Variables

Recall Theorem 122 from first semester’s notes. We phrase it here slightly differently.

**Example 124.** Suppose \( X_1 \sim \text{Pois}(\lambda_1) \), \( X_2 \sim \text{Pois}(\lambda_2) \), and that \( X_1 \) and \( X_2 \) are independent. Find the probability distribution for \( Y = X_1 + X_2 \).

By independence of \( X_1 \) and \( X_2 \), we see that the joint distribution is given by
\[
e^{-(\lambda_1 + \lambda_2)} \frac{\lambda_1^{x_1} \lambda_2^{x_2}}{x_1! x_2!}
\]
for \( x_1 = 0, 1, 2, \ldots \) and \( x_2 = 0, 1, 2, \ldots \). Notice that \( y = x_1 + x_2 \geq 0 \). Hence, using the substitution \( x_1 = y - x_2 \), we see that
\[
0 \leq x_1 = y - x_2 \implies x_2 \leq y.
\]
So we write the joint distribution of \( Y \) and \( X_2 \):
\[
g(y, x_2) = e^{-(\lambda_1 + \lambda_2)} \frac{\lambda_1^{y-x_2} \lambda_2^{x_2}}{(y-x_2)! x_2!}
\]
for \( y = 0, 1, 2, \ldots \) and \( x_2 = 0, 1, 2, \ldots, y \).

To get the marginal density of \( Y \), we compute
\[
h(y) = \sum_{x_2=0}^{y} \frac{e^{-(\lambda_1 + \lambda_2)} \lambda_1^{y-x_2} \lambda_2^{x_2}}{(y-x_2)! x_2!}
\]

\[
= \frac{e^{-(\lambda_1 + \lambda_2)}}{y!} \sum_{x_2=0}^{y} \frac{y!}{(y-x_2)! x_2!} \lambda_1^{y-x_2} \lambda_2^{x_2}
\]

\[
= \frac{e^{-(\lambda_1 + \lambda_2)}}{y!} \sum_{x_2=0}^{y} \binom{y}{x_2} \lambda_1^{y-x_2} \lambda_2^{x_2}
\]


\[ = e^{-(\lambda_1 + \lambda_2)} \cdot y! \cdot (\lambda_1 + \lambda_2)^y. \]

Hence, \( Y \sim \text{Pois}(\lambda_1 + \lambda_2). \)

---

**Example 125.** Suppose the joint probability density for \( X \) and \( Y \) is given by

\[
f(x, y) = \begin{cases} 
  e^{-(x+y)}, & x, y > 0; \\
  0, & \text{otherwise}.
\end{cases}
\]

Find the probability density for \( Z = \frac{X}{X + Y}. \)

If we hold \( X \) constant, notice that \( Z \) is strictly decreasing as \( Y \) increases and that, as a function of \( y \), \( z = \frac{x}{x + y} \) is differentiable. Hence, we can appeal to Theorem 159. Observe that

\[
z = \frac{x}{x + y} \implies yz = x - xz \implies y = x \cdot \frac{1 - z}{z}.
\]

Then

\[
\frac{\partial y}{\partial z} = -\frac{x}{z^2}.
\]

Note that

\[
x + y = x + x \cdot \frac{1 - z}{z} = x \cdot \left[ 1 + \frac{1 - z}{z} \right] = \frac{x}{z}
\]

and that \( 0 < z < 1 \). It follows that the joint probability density for \( X \) and \( Z \) is

\[
e^{-x/z} \cdot \left| \frac{\partial y}{\partial z} \right| = e^{-x/z} \cdot \frac{x}{z^2}
\]

for \( x > 0 \) and \( 0 < z < 1 \).

Lastly, to get the marginal density for \( Z \), we integrate out the \( x \) using the substitution \( u = x/z \):

\[
h(z) = \int_0^\infty e^{-x/z} \cdot \frac{x}{z^2} \, dx
\]

\[
= \int_0^\infty e^{-u} \cdot \frac{u}{z} \cdot z \, du
\]

\[
= \int_0^\infty u \cdot e^{-u} \, du
\]

\[
= 1
\]

for \( 0 < z < 1 \). That is, \( Z \) is uniformly distributed on the interval \([0, 1]\).

---

Recall the Jacobian and how it’s used for integration where we do a substitution involving several variables. In particular, say we have a change of variables \( x = g(u, v) \) and \( y = h(u, v) \). Then the corresponding Jacobian is the determinant

\[
\begin{vmatrix} 
  \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\
  \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v}
\end{vmatrix}
= \frac{\partial x}{\partial u} \cdot \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \cdot \frac{\partial y}{\partial u}.
\]

**Example 126.** Suppose the joint probability of \( X \) and \( Y \) is given by

\[
f(x, y) = \begin{cases} 
  e^{-(x+y)}, & x, y > 0; \\
  0, & \text{otherwise}.
\end{cases}
\]

Find the joint probability density of \( U = X + Y \) and \( V = \frac{X}{X + Y} \).
First, notice that \( u(x, y) = x + y \) and \( v(x, y) = \frac{x}{x+y} \) are partially differentiable. Also, we check that
\[
(x, y) \mapsto (u, v)
\]
is a one-to-one transformation of the plane.

To check this, suppose \((x_1, y_1)\) and \((x_2, y_2)\) are so that
\[
x_1 + y_1 = x_2 + y_2 \quad \text{and} \quad \frac{x_1}{x_1+y_1} = \frac{x_2}{x_2+y_2}.
\]
Then we see that
\[
\frac{x_1}{x_1+y_1} = \frac{x_1}{x_2+y_2} = \frac{x_2}{x_2+y_2} \implies x_1 = x_2.
\]
Moreover,
\[
x_1 + y_1 = x_2 + y_1 = x_2 + y_2 \implies y_1 = y_2.
\]
That is, this transformation of the plane is one-to-one.

Now, we solve for the inverse representations:
\[
y = u - x \implies v = \frac{x}{x+u-x} \implies v = \frac{x}{u} \implies x = uv \implies y = u - uv = u(1-v).
\]
The corresponding Jacobian is
\[
\begin{vmatrix}
  v & u \\
  1-v & -u \\
\end{vmatrix} = -uv - u(1-v) = -u.
\]
Hence, the joint probability density for \( U \) and \( V \) is
\[
g(u, v) = f(uv, u(1-v)) \cdot | -u | = e^{-(uv+u-u)} \cdot u = u e^{-u}
\]
whenever \( u > 0 \) and \( 0 < v < 1 \).

### 7.3 Moment-generating Function Technique

In this section we will again be using the fact that probability distributions/densities are determined by their moment-generating function, supposing it exists. Recall that

**Theorem 161.** If \( X_1, X_2, \ldots, X_n \) are independent random variables and
\[
Y = X_1 + X_2 + \cdots + X_n,
\]
then the moment-generating function for \( Y \) is
\[
M_Y(t) = \prod_{j=1}^{n} M_{X_j}(t)
\]
where \( M_{X_j}(t) \) is the moment-generating function for \( X_j \), \( j = 1, 2, \ldots, n \).

**Example 127.** Suppose \( X_1 \sim \text{Pois}(\lambda_1) \) and \( X_2 \sim \text{Pois}(\lambda_2) \) are independent. Using the moment-generating function technique, show that \( X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2) \).
Recall that
\[ M_{X_j}(t) = \exp \left( \lambda_j (e^t - 1) \right) \]
so
\[ M_Y(t) = \exp \left( \lambda_1 (e^t - 1) \right) \cdot \exp \left( \lambda_2 (e^t - 1) \right) = \exp \left( (\lambda_1 + \lambda_2)(e^t - 1) \right) \]
which shows that \( Y \sim \text{Pois}(\lambda_1 + \lambda_2) \).

**Example 128.** Suppose \( X_1 \sim \Gamma(\alpha_1, \beta) \) and \( X_2 \sim \Gamma(\alpha_2, \beta) \) are independent. Using the moment-generating function technique, show that \( X_1 + X_2 \sim \Gamma(\alpha + \alpha_2, \beta) \).

Recall that
\[ M_{X_j}(t) = (1 - \theta t)^{-\alpha_j} \]
where \( \theta = 1/\beta \) so
\[ M_Y(t) = (1 - \theta t)^{-\alpha_1} \cdot (1 - \theta t)^{-\alpha_2} = (1 - \theta t)^{-(\alpha_1 + \alpha_2)} \]
which shows that \( Y \sim \Gamma(\alpha_2 + \alpha_2, \beta) \).
Chapter 8

Sampling

Statistics is primarily concerned with drawing conclusions and making predictions based on experiments and investigations. In essence, statistical inference is based on collecting data from a population and inferring something about the population based on the data.

**Definition 162.** The population consists of all subjects being studied. Generally speaking, a sample is a collection of subjects from the population.

**Definition 163.** Formally, a random sample (from an infinite population) is a finite collection of IID random variables where their shared distribution is the population.

**Definition 164.** For a random sample \(X_1, X_2, \ldots, X_n\), the sample mean is

\[
\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j
\]

and the sample variance is

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

We will discover why the sample variance is divided by \(n - 1\) instead of \(n\) a bit later in Section 8.2.

These are examples of what we colloquially know as statistics.

**Definition 165.** A statistic is a function involving random samples and constants.

In Section 7 we studied three ways which can be used to find the probability distribution/density of transformations of random variables. Since statistics are transformations of random variables, those techniques may help us understand the distributions of statistics.

8.1 The Sample Mean

Statistics, generally speaking, are also random variables in their own right. Here, we’ll investigate the statistic of sample mean.

**Theorem 166.** If \(X_1, X_2, \ldots, X_n\) is a random sample from an infinite population with mean \(\mu\) and variance \(\sigma^2\), then

\[
E(\bar{X}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.
\]

**Proof.** The fact that \(E(\bar{X}) = \mu\) is immediate. For the variance, notice that

\[
\text{Var}(\bar{X}) = \text{Var} \left( \frac{1}{n} \sum_{j=1}^{n} X_j \right)
\]
\[ \begin{align*}
&\sum_{j=1}^{n} \text{Var} \left( \frac{X_j}{n} \right) \\
&= \sum_{j=1}^{n} \frac{1}{n^2} \cdot \text{Var} (X_j) \\
&= \frac{1}{n^2} \sum_{j+1}^{n} \sigma^2 \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}.
\end{align*} \]

The quantity \( \sqrt{\text{Var}(\bar{X})} \) is usually referred to as the **standard error of the mean**. Notice that, as \( n \) grows, \( \sqrt{\text{Var}(\bar{X})} \) approaches zero. Intuitively, this means that increasing the sample size reduces the error; i.e., our approximations of the population mean improve with larger samples.

**Theorem 167.** For any positive constant \( c \),
\[ P(\mu - c \leq \bar{X} \leq \mu + c) \geq 1 - \frac{\sigma^2}{nc^2}. \]

**Remark.** Recall that we proved Theorem 167 last semester. It is known as the Weak Law of Large Numbers.

**Theorem 168** (Central Limit Theorem). If \( X_1, X_2, \ldots, X_n \) forms a random sample for an infinite population with mean \( \mu \) and variance \( \sigma^2 \), then, as long as the moment-generating function function exists for the population, the limiting distribution of
\[ Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \]
is the standard normal distribution.

**Proof.** First, observe that
\[ Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} \cdot n}{\sigma/\sqrt{n}} - \frac{\mu \sqrt{n}}{\sigma} = \frac{n\bar{X}}{\sigma/\sqrt{n}} - \frac{\mu \sqrt{n}}{\sigma}. \]

Hence,
\[ M_Z(t) = E(e^{zt}) = E \left( \exp \left( n\bar{X} \cdot \frac{t}{\sigma/\sqrt{n}} - \frac{\mu \sqrt{n}}{\sigma} \cdot t \right) \right) = \exp \left( -\frac{\sqrt{n} \cdot \mu t}{\sigma} \right) \cdot M_{n\bar{X}} \left( \frac{t}{\sigma/\sqrt{n}} \right). \]

As \( n\bar{X} = X_1 + X_2 + \cdots + X_n \) and the \( X_j \) are IID,
\[ M_Z(t) = \exp \left( -\frac{\sqrt{n} \cdot \mu t}{\sigma} \right) \cdot M_X \left( \frac{t}{\sigma/\sqrt{n}} \right)^n. \]

Then,
\[ \ln (M_Z(t)) = -\frac{\sqrt{n} \cdot \mu t}{\sigma} + n \cdot \ln \left[ M_X \left( \frac{t}{\sigma/\sqrt{n}} \right) \right]. \]
Now, we’ll isolate \( M_X \left( \frac{t}{\sigma \sqrt{n}} \right) \). Observe that

\[
M_X \left( \frac{t}{\sigma \sqrt{n}} \right) = E \left( \exp \left( X \cdot \frac{t}{\sigma \sqrt{n}} \right) \right)
\]

\[
= E \left( 1 + \sum_{j=1}^{\infty} \frac{X^j}{j! \sigma^j n^{j/2}} \right)
\]

\[
= 1 + \sum_{j=1}^{\infty} E(X^j) \frac{t^j}{j! \sigma^j n^{j/2}}.
\]

Then, using the fact that

\[
\ln(1 + x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k},
\]

we see that

\[
\ln \left[ M_X \left( \frac{t}{\sigma \sqrt{n}} \right) \right] = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k} \left[ \sum_{j=1}^{\infty} \frac{E(X^j) t^j}{j! \sigma^j n^{j/2}} \right]^k
\]

\[
= \left[ \frac{E(X)t}{\sigma n^{1/2}} + \frac{E(X^2)t^2}{2\sigma^2 n} + \frac{E(X^3)t^3}{6\sigma^3 n^{3/2}} + \cdots \right]
\]

\[
- \frac{1}{2} \cdot \left[ \frac{E(X)t}{\sigma n^{1/2}} + \frac{E(X^2)t^2}{2\sigma^2 n} + \frac{E(X^3)t^3}{6\sigma^3 n^{3/2}} + \cdots \right]^2
\]

\[
+ \frac{1}{3} \cdot \left[ \frac{E(X)t}{\sigma n^{1/2}} + \frac{E(X^2)t^2}{2\sigma^2 n} + \frac{E(X^3)t^3}{6\sigma^3 n^{3/2}} + \cdots \right]^3
\]

\[
= \frac{E(X)}{\sigma n^{1/2}} \cdot t + \frac{E(X^2) - E(X)^2}{2\sigma^2 n} \cdot t^2 + \frac{E(X^3) - 2E(X)E(X^2)}{6\sigma^3 n^{3/2}} + \frac{E(X)^3}{\sigma^3 n^{3/2}} \cdot t^3 + \cdots
\]

\[
= \frac{\mu t}{\sigma n^{1/2}} + \frac{t^2}{2n} + \cdots
\]

Hence,

\[
\ln (M_Z(t)) = -\sqrt{n} \cdot \frac{\mu t}{\sigma} + n \cdot \left[ \frac{\mu t}{\sigma n^{1/2}} + \frac{t^2}{2n} + \cdots \right]
\]

\[
= -\sqrt{n} \cdot \frac{\mu t}{\sigma} + \frac{n\mu t}{\sigma n^{1/2}} + \frac{t^2}{2} + \cdots
\]

\[
= -\sqrt{n} \cdot \frac{\mu t}{\sigma} + \frac{\sqrt{n} \cdot \mu t}{\sigma} + \frac{t^2}{2} + \cdots
\]

\[
= \frac{t^2}{2} + \cdots
\]

The key here is that the terms summarized by the “...” go to zero as \( n \to \infty \). That is,

\[
\lim_{n \to \infty} \ln (M_Z(t)) = \frac{t^2}{2} \implies \lim_{n \to \infty} M_Z(t) = \exp \left( \frac{t^2}{2} \right)
\]
which is the moment-generating function for the standard normal distribution.

**Lemma 169.** If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ where $a \neq 0$, then

$$Y \sim \mathcal{N}(a\mu + b, (a\sigma)^2).$$

In light of Lemma 169, the Central Limit Theorem is asserting that, for large sample size $n$, the sample mean $\bar{X}$ is approximately normally distributed with mean $\mu$ and variance $\sigma^2/n$. In fact, if $X$ is normally distributed, we get a stronger claim.

**Theorem 170.** If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_1, X_2, \ldots, X_n$ forms a random sample for $X$, then $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

**Proof.** Notice that $X_i/n = X/n \sim \mathcal{N}(\mu/n, (\sigma/n)^2)$ by Lemma 169. Hence,

$$M_{X_i/n}(t) = M_{X/n}(t) = \exp \left( \frac{\mu}{n} \cdot t + \frac{\sigma^2}{n^2} \cdot \frac{t^2}{2} \right).$$

By the moment-generating functions,

$$M_{\bar{X}}(t) = \left[M_{X_i/n}(t)\right]^n = \left[\exp \left( \frac{\mu}{n} \cdot t + \frac{\sigma^2}{n^2} \cdot \frac{t^2}{2} \right)\right]^n = \exp \left( \mu t + \frac{\sigma^2}{n} \cdot \frac{t^2}{2} \right),$$

which establishes that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

**Note.** Theorem 170 and Lemma 169 show that, for a random sample $X_1, X_2, \ldots, X_n$ from a normal population with mean $\mu$ and variance $\sigma^2$, $X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$.

**Example 129.** The length of a widget is a continuous random variable with a mean of 512mm and a standard deviation of 3mm.

(a) Use a normal distribution to find the probability that the average length of a random sample consisting of 49 widgets is greater than 513mm.

(b) How big of a sample should we take to have a 95% probability that the sample mean $\bar{X}$ is within 0.5mm of the population mean?

Using the Central Limit Theorem, we suppose the sample mean $\bar{X}$ is approximately normally distributed with mean $\mu = 512$ and variance $\sigma^2 = 9/49$. It follows that the standard deviation of $\bar{X}$ is $\sigma = 3/7$.

Observe that $Z = \frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ and that

$$\frac{513 - 512}{(3/7)} = \frac{7}{3}.$$

Therefore,

$$P(\bar{X} > 512) = P(Z > 7/3) \approx 0.0098153286286.$$

Now, we wish to find $n$ so that

$$P(|\bar{X} - \mu| < 0.5) = 95\%.$$

Notice that

$$P(|\bar{X} - \mu| < 0.5) = P(-0.5 < \bar{X} - \mu < 0.5) = P \left( \frac{-0.5}{3/\sqrt{n}} < \frac{\bar{X} - \mu}{3/\sqrt{n}} < \frac{0.5}{3/\sqrt{n}} \right)$$
and since \( Z = \frac{\bar{X} - \mu}{3/\sqrt{n}} \) can be assumed to have approximately the standard normal distribution, we can use a \( z \)-table to see that
\[
\frac{0.5}{3/\sqrt{n}} \approx 1.96 \implies n \approx 138.3.
\]
Hence, if we take a sample of 139 widgets, there is a 95% probability that the sample mean will be within 0.5mm of the population mean.

Although the Central Limit Theorem is quite useful when taking large samples from distributions which may not be normal, we don’t always need to use it.

**Example 130.** Suppose \( X \) has an exponential distribution given by
\[
f(x) = \frac{e^{-x/5}}{5} \quad \text{for } x > 0.
\]
Take a random sample of size \( n = 121 \).

(a) Use the Central Limit Theorem to approximate \( P(\bar{X} < 4.3) \).

(b) Use the moment-generating function technique to find the distribution of \( \bar{X} \). Then calculate \( P(\bar{X} < 4.3) \).

(a) Using the Central Limit Theorem, we assume that
\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 5}{5/\sqrt{121}}
\]
has the standard normal distribution. Then
\[
P(\bar{X} < 4.3) \approx P(Z < -1.54) \approx 0.061780177.
\]

(b) Recall that the moment-generating function for this exponential distribution is given by
\[
M_X(t) = (1 - 5t)^{-1}.
\]
It follows that
\[
M_{n \cdot \bar{X}}(t) = (1 - 5t)^{-n}
\]
which is the moment-generating function for a gamma distribution. That is, the probability density for \( Y = n \cdot \bar{X} \) is given by
\[
g(x) = \frac{x^{120} \cdot e^{-x/5}}{5^{121} \cdot \Gamma(121)} \quad \text{for } x > 0.
\]
Hence,
\[
P(\bar{X} < 4.3) = P(Y < 520.3) = \int_{0}^{520.3} \frac{x^{120} \cdot e^{-x/5}}{5^{121} \cdot \Gamma(121)} \, dx \approx 0.0561775.
\]
CHAPTER 8. SAMPLING

8.1.1 Finite Populations

Now that we’ve covered the sample mean in the context of infinite populations, let’s turn our attention to finite populations. We consider taking samples without replacement from a finite population \( \{c_1, c_2, \ldots, c_N\} \) of numerical values. To more easily allow for repeated values, we can represent the population through a function \( \gamma : \{1, 2, \ldots, N\} \rightarrow \mathbb{R} \) where \( c_k = \gamma(k) \). Observe that

\[
\mu = \frac{1}{N} \cdot \sum_{j=1}^{N} \gamma(j)
\]

and

\[
\sigma^2 = \frac{1}{N} \cdot \sum_{j=1}^{N} (\gamma(j) - \mu)^2.
\]

The probability of selecting the ordered random sample \((x_1 = \gamma(j_1), x_2 = \gamma(j_2), \ldots, x_n = \gamma(j_n))\) is

\[
f(x_1, x_2, \ldots, x_n) = \frac{1}{P(N, n)}.
\]

Since there are \( n! \) to permute the selection \((\gamma(j_1), \gamma(j_2), \ldots, \gamma(j_n))\), we see that the probability of our random sample consisting of the values \(\{\gamma(j_1), \gamma(j_2), \ldots, \gamma(j_n)\}\) is

\[
\frac{n!}{P(N, n)} = \frac{1}{\binom{N}{n}}.
\]

Another way to see this is that there are \( \binom{N}{n} \) different collections of size \( n \) from the population of size \( N \) and we impose that the selections are equally likely.

To find the marginal probability distribution function for \(X_k\), notice that

\[
f(x_k) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{k-1}} \sum_{x_{k+1}} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n)
\]

\[
= \frac{P(N-1, n-1)}{P(N, n)}
\]

\[
= \frac{(N-1)! (N-n)!}{(N-1-(n-1))! N!}
\]

\[
= \frac{1}{N}.
\]

Similarly, the marginal probability distribution for \(X_k\) and \(X_\ell\) is given by

\[
f(x_k, x_\ell) = \frac{1}{N(N-1)}.
\]

**Theorem 171.** If \(X_k\) and \(X_\ell\) are the \(k\)th and \(\ell\)th random variables of a random sample of size \( n \) from a population described by \( \gamma : \{1, 2, \ldots, N\} \rightarrow \mathbb{R} \), then

\[
\text{cov}(X_k, X_\ell) = -\frac{\sigma^2}{N-1}.
\]

**Proof.** First, note that

\[
\sum_{j=1}^{N} (\gamma(j) - \mu) = \left[ \sum_{j=1}^{N} \gamma(j) \right] - N \cdot \mu = N \cdot \mu - N \cdot \mu = 0.
\]
Now, recalling the definition of covariance, observe that
\[
\text{cov}(X_k, X_\ell) = E((X_k - \mu)(X_\ell - \mu))
\]
\[
= \frac{1}{N(N-1)} \sum_{i=1}^{N} \left( (\gamma(i) - \mu) \cdot \sum_{j \neq i} (\gamma(j) - \mu) \right)
\]
\[
= \frac{1}{N(N-1)} \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} (\gamma(i) - \mu)^2
\]
\[
= \frac{-\sigma^2}{N-1}.
\]

**Theorem 172.** If \( \bar{X} \) is the sample mean of a random sample of size \( n \) from a finite population with mean \( \mu \) and variance \( \sigma^2 \), then \( E(\bar{X}) = \mu \) and

\[
\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.
\]

**Proof.** First, note that

\[
E(X_k) = \sum_{j=1}^{N} \gamma(j) \cdot f(\gamma(j)) = \sum_{j=1}^{N} \frac{\gamma(j)}{N} = \mu.
\]

Then

\[
E(\bar{X}) = \frac{1}{N} \cdot \sum_{k=1}^{N} E(X_k) = \mu.
\]

For the variance, recall that

\[
\text{Var}(\bar{X}) = \text{Var}\left( \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n} \right)
\]
\[
= \left[ \sum_{j=1}^{n} \text{Var}(X_j) \cdot \frac{1}{n^2} \right] + 2 \cdot \left[ \sum_{j=2}^{n} \sum_{k=1}^{j-1} \frac{\text{cov}(X_j, X_k)}{n^2} \right]
\]
\[
= \frac{\sigma^2}{n} + 2 \cdot \frac{-\sigma^2}{n^2(N-1)} \cdot \sum_{j=2}^{n} \sum_{k=1}^{j-1} 1
\]
\[
= \frac{\sigma^2}{n} - \frac{2\sigma^2}{n^2(N-1)} \cdot \sum_{j=2}^{n} (j-1)
\]
\[
= \frac{\sigma^2}{n} - \frac{2\sigma^2}{n^2(N-1)} \cdot \left[ \frac{n(n+1)}{2} - (n-1) \right]
\]
\[
= \frac{n^2 \sigma^2}{n^2(N-1)} - \frac{2\sigma^2}{n^2(N-1)} \cdot \left[ \frac{n(n-1)}{2} \right]
\]
\[ \frac{n \sigma^2 (N - 1) - n(n - 1) \sigma^2}{n^2 (N - 1)} = \frac{\sigma^2}{n} \cdot \frac{(N - 1) - (n - 1)}{N - 1} = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}. \]

### 8.2 A Note on Sample Variance

**Theorem 173.** For a continuous random variable \( X \) with variance \( \sigma^2 \),

\[ E(S^2) = \sigma^2. \]

**Proof.** First, for a random sample \( X_1, X_2, \ldots, X_n \) of \( X \), note that

\[ (X_j - \bar{X})^2 = (X_j - \mu + \mu - \bar{X})^2 = (X_j - \mu)^2 + 2(X_j - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2. \]

By Theorem 166, we know that \( E(\bar{X}) = \mu \) which implies that

\[ E((\bar{X} - \mu)^2) = E((\bar{X} - E(\bar{X}))^2) = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}. \]

Similarly,

\[ E((X_j - \mu)^2) = \text{Var}(X) = \sigma^2. \]

Since

\[ (X_j - \mu)(\mu - \bar{X}) = \mu X_j - X_j \cdot \bar{X} - \mu^2 + \mu \cdot \bar{X}, \]

we see that

\[ E((X_j - \mu)(\mu - \bar{X})) = \frac{1}{n} \cdot \frac{\sum_{k=1}^{n} X_k}{n} \cdot \frac{1}{n} \cdot \sum_{k=1}^{n} X_k \cdot X_j \]

which implies that

\[ E(X_j \cdot \bar{X}) = \frac{1}{n} \cdot \sum_{k=1}^{n} E(X_k \cdot X_j). \]

Since

\[ \sigma^2 = \text{Var}(X_j) = E(X_j^2) - (E(X_j))^2 = E(X_j^2) - \mu^2, \]

we see that

\[ E(X_j^2) = \sigma^2 + \mu^2. \]
For $k \neq j$, $E(X_k \cdot X_j) = E(X_k) \cdot E(X_j) = \mu^2$ by independence. Hence,

$$E(X_j \cdot \bar{X}) = \frac{1}{n} \cdot \sum_{k=1}^{n} E(X_k \cdot X_j)$$

$$= \frac{1}{n} \cdot [(n-1)\mu^2 + \sigma^2 + \mu^2]$$

$$= \mu^2 + \frac{\sigma^2}{n}.$$  

Thus,

$$E[(X_j - \bar{X})^2] = \sigma^2 + 2\mu^2 - 2 \cdot E(X_j \cdot \bar{X}) + \frac{\sigma^2}{n}$$

$$= \sigma^2 + 2\mu^2 - 2\mu^2 - 2 \cdot \frac{\sigma^2}{n} + \frac{\sigma^2}{n}$$

$$= \frac{n-1}{n} \cdot \sigma^2.$$  

Finally,

$$E(S^2) = E \left[ \frac{1}{n-1} \cdot \sum_{j=1}^{n} (X_j - \bar{X})^2 \right]$$

$$= \frac{1}{n-1} \cdot \sum_{j=1}^{n} E[(X_j - \bar{X})^2]$$

$$= \frac{1}{n-1} \cdot \sum_{j=1}^{n} \frac{n-1}{n} \cdot \sigma^2$$

$$= \sigma^2.$$  

8.3 The Chi-Square Distribution

Recall that a random variable $X$ has a chi-square distribution with $\nu$ degrees of freedom, denoted $X \sim \chi(\nu)$, if the probability density function for $X$ is given by

$$f(x) = \begin{cases} 
\frac{x^{(\nu-2)/2} \cdot e^{-x/2}}{2^{\nu/2} \cdot \Gamma(\nu/2)}, & x > 0; \\
0, & \text{otherwise}.
\end{cases}$$

For a graph of this distribution, please visit Desmos.com. Also recall that the mean and variance are $\nu$ and $2\nu$, respectively. Moreover, the moment-generating function is given by

$$M_X(t) = (1 - 2t)^{-\nu/2}.$$  

**Theorem 174.** If $X_1, X_2, \ldots, X_n$ form a random sample of the standard normal distribution, then

$$Y = \sum_{j=1}^{n} X_j^2$$

has a chi-square distribution with $n$ degrees of freedom.
Proof. First, suppose that \(X \sim N(0, 1)\) and note that
\[
P(0 \leq X^2 < t) = P(-\sqrt{t} < X < \sqrt{t})
\]
\[
= 2 \cdot P(0 \leq X < \sqrt{t})
\]
\[
= 2 \cdot \int_0^{\sqrt{t}} e^{-x^2/2} \, dx.
\]

Let \(u = x^2\) for \(x \geq 0\) and note that \(du = 2x \, dx\). Moreover, \(x = \sqrt{u}\) which provides
\[
dx = \frac{1}{2\sqrt{u}} \, du.
\]
When \(x = 0\), \(u = 0\) and when \(x = \sqrt{t}\), \(u = t\). Hence,
\[
P(0 \leq X^2 < t) = 2 \cdot \int_0^{\sqrt{t}} e^{-x^2/2} \, dx
\]
\[
= 2 \cdot \int_0^{t} e^{-u/2} \, \frac{1}{2\sqrt{u}} \, du
\]
\[
= \frac{1}{\sqrt{2\pi}} \cdot \int_0^{t} e^{-u/2} \, du.
\]
Differentiating with respect to \(t\) gives us the probability density function for \(Y = X^2\):
\[
g(y) = \frac{y^{-1/2} \cdot e^{-y/2}}{\sqrt{2\pi}} = \frac{y^{-1/2} \cdot e^{-y/2}}{2^{1/2} \cdot \Gamma(1/2)},
\]
for \(y > 0\). Hence, we see that \(X^2\) has a chi-square distribution with one degree of freedom.

For the rest, we use the moment-generating function technique. Notice that \(M_{X_j}(t) = (1 - 2t)^{-1/2}\) which means that the moment-generating function for
\[
Y = \sum_{j=1}^{n} X_j^2
\]
is
\[
M_Y(t) = [(1 - 2t)^{-1/2}]^n = (1 - 2t)^{-n/2}
\]
which establishes that \(Y\) has a chi-square distribution with \(n\) degrees of freedom. \(\square\)

In a similar fashion,

Theorem 175. If \(X_1, X_2, \ldots, X_n\) are independent random variables having chi-square distributions with \(\nu_1, \nu_2, \ldots, \nu_n\) degrees of freedom, respectively, then
\[
Y = \sum_{j=1}^{n} X_j
\]
has a chi-square distribution with \(\nu_1 + \nu_2 + \cdots + \nu_n\) degrees of freedom.

Moreover, this reverses.

Theorem 176. If \(X\) and \(Y\) are independent random variables where \(X\) and \(X + Y\) have chi-square distributions with \(\nu_1\) and \(\nu_2\) degrees of freedom, respectively, then \(Y\) has a chi-square distribution with \(\nu_2 - \nu_1\) degrees of freedom.
8.3. THE CHI-SQUARE DISTRIBUTION

Proof. First, note that
\[ M_X(t) = (1 - 2t)^{-\nu_1/2} \]
and
\[ M_{X+Y}(t) = (1 - 2t)^{-\nu_2/2}. \]
By independence,
\[ (1 - 2t)^{-\nu_2/2} = M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = (1 - 2t)^{-\nu_1/2} \cdot M_Y(t) \]
which implies that
\[ M_Y(t) = \frac{(1 - 2t)^{-\nu_2/2}}{(1 - 2t)^{-\nu_1/2}} = (1 - 2t)^{-\left(\nu_2 - \nu_1\right)/2}. \]

\[ \square \]

Theorem 177. If \( X_1, X_2, \ldots, X_n \) is a random sample of a normally distributed random variable, then \( \bar{X} \) and \( S^2 \) are independent. Consequently, given the population mean \( \mu \) and variance \( \sigma^2 \), the random variable \( \frac{n-1}{\sigma^2} \cdot S^2 \) has a chi-square distribution with \( n - 1 \) degrees of freedom.

We will not provide the full proof here but we will prove the special case when \( n = 2 \) and the variables have the standard normal distribution.

Proof. Let
\[ f(x_1, x_2) = \frac{\exp\left(-x_1^2/2\right) \cdot \exp\left(-x_2^2/2\right)}{2\pi} \]
be the joint probability distribution of \( X_1 \) and \( X_2 \).

First we will show that \( \bar{X} \) and \( X_j - \bar{X} \) are independent where \( j = 1, 2 \). Let \( Y = \bar{X} \) and \( Y_j = X_j - \bar{X} \).

We will see that the transformation \((X_1, X_2) \mapsto (Y, Y_j)\) is one-to-one. Suppose \((X_1, X_2)\) and \((U_1, U_2)\) are so that
\[ \frac{X_1 + X_2}{2} = \frac{U_1 + U_2}{2} \]
and
\[ \frac{X_j - X_1 + X_2}{2} = \frac{U_j - U_1 + U_2}{2}. \]
It follows immediately that \( X_j = U_j \). Hence, the transformation \((X_1, X_2) \mapsto (Y, Y_j)\) is one-to-one.

Check that
\[
\begin{align*}
Y &= (X_1 + X_2)/2 \\
Y_j &= X_j - (X_1 + X_2)/2
\end{align*}
\Rightarrow
\begin{align*}
2Y &= X_1 + X_2 \\
X_j &= Y + Y_j
\end{align*}
\]
which leads to two cases:
\[
\begin{align*}
X_1 &= Y + Y_1 \\
X_2 &= Y - Y_1
\end{align*}\quad \text{or} \quad 
\begin{align*}
X_1 &= Y - Y_2 \\
X_2 &= Y + Y_2
\end{align*}
\]
The corresponding Jacobian is
\[
\begin{vmatrix}
1 & \pm 1 \\
1 & \mp 1
\end{vmatrix} = \mp 2.
\]
Hence, the joint probability density of \( Y = \bar{X} \) and \( Y_j = X_j - \bar{X} \) is
\[
g(y, y_j) = 2 \cdot f(y \pm y_j, y \mp y_j) \\
= 2 \cdot \exp\left[-\frac{1}{2} \cdot ((y \pm y_j)^2 + (y \mp y_j)^2)\right] \\
= \frac{1}{\pi} \cdot \exp\left(-\frac{1}{2} \cdot (2y^2 + 2y_j^2)\right) \\
= \frac{1}{\pi} \cdot \exp (-y^2) \cdot \exp (-y_j^2).
\]
By Theorem 170, we know that
\[ \bar{X} \sim \mathcal{N}(0, 1/2) \]
which means the probability density for \( Y = \bar{X} \) is
\[ h(y) = \frac{1}{\sqrt{2\pi \cdot (1/2)}} \cdot \exp \left[ -\frac{y^2}{2(1/2)} \right] = \frac{\exp(-y^2)}{\sqrt{\pi}}. \]
Hence,
\[ g(y, y_j) = h(y) \cdot \frac{\exp(-y_j^2)}{\sqrt{\pi}}. \]
Hence, \( \bar{X} \) and \( X_j - \bar{X} \) are independent. It follows that \( \bar{X} \) and \( S^2 \) are independent.

Now, we assume that \( \bar{X} \) and \( S^2 \) are independent. Observe that
\[ (X_j - \mu)^2 = (X_j - \bar{X} + \bar{X} - \mu)^2 = (X_j - \bar{X})^2 + 2(X_j - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2. \]
It follows that
\[ \sum_{j=1}^{n} (X_j - \mu)^2 = \sum_{j=1}^{n} [(X_j - \bar{X})^2 + 2(X_j - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2] = \sum_{j=1}^{n} (X_j - \bar{X})^2 + 2(\bar{X} - \mu) \cdot \left[ \sum_{j=1}^{n} (X_j - \bar{X}) \right] + n \cdot (\bar{X} - \mu)^2 \]
Moreover,
\[ \sum_{j=1}^{n} \left( \frac{X_j - \mu}{\sigma} \right)^2 = \frac{n - 1}{\sigma^2} \cdot S^2 + \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right]^2 \]
By Theorem 174, we know that
\[ \sum_{j=1}^{n} \left( \frac{X_j - \mu}{\sigma} \right)^2 \]
has a chi-square distribution with \( n \) degrees of freedom. Similarly,
\[ \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right]^2 \]
has a chi-square distribution with one degree of freedom. Since
\[ \frac{n - 1}{\sigma^2} \cdot S^2 \quad \text{and} \quad \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right]^2 \]
are independent, Theorem 176 guarantees that \( \frac{n - 1}{\sigma^2} \cdot S^2 \) has a chi-square distribution with \( n - 1 \) degrees of freedom.

**Example 131.** A random sample of size 25 is taken from a normally distributed population for which it is claimed that \( \sigma^2 = 4 \). What is the probability that the sample variance is greater than 5?

By Theorem 177, we know that
\[ \frac{n - 1}{\sigma^2} \cdot S^2 = \frac{24}{4} \cdot S^2 = 6S^2 \sim \chi(24). \]
Thus,
\[ P(S^2 > 5) = P(6S^2 > 30) = \int_{30}^{\infty} \frac{x^{11} e^{-x/2}}{2^{12} \cdot \Gamma(12)} \, dx \approx 0.1847518. \]
8.4 Student’s $t$-distribution

As we’ve seen, the sample mean of a random sample from a normal population is itself normally distributed. Namely, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then
\[
\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)
\]
where $n$ is the size of the random sample. In most practical applications, both $\mu$ and $\sigma$ are unknown, though we can approximate both. With this in mind, we study
\[
\frac{\bar{X} - \mu}{S/\sqrt{n}}
\]
where $S$ is the sample standard deviation for a random sample of size $n$.

**Definition 178.** A random variable $T$ has a $t$-distribution with $\nu$ degrees of freedom if its probability density function is given by
\[
f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \pi} \cdot \Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}
\]
for $t \in \mathbb{R}$.

**Remark.** Observe that the probability density function for a random variable with a $t$-distribution is symmetric about the origin.

**Theorem 179.** If $Y \sim \chi^2(\nu)$ and $Z \sim \mathcal{N}(0, 1)$ are independent, then
\[
T = \frac{Z}{\sqrt{Y/\nu}}
\]
has a $t$-distribution with $\nu$ degrees of freedom.

**Proof.** By independence of $Y$ and $Z$, their joint probability density function is given by
\[
g(y, z) = \frac{y^{\nu/2} \cdot \exp(-y/2) \cdot \exp(-z^2/2)}{2^{\nu/2} \cdot \Gamma(\nu/2) \cdot \sqrt{2\pi}}
\]
for $y > 0$ and $z \in \mathbb{R}$. Fix $y$ and let
\[
t = \frac{z}{\sqrt{y/\nu}}.
\]
Since $y > 0$, we see that $t$ is strictly increasing as a function of $z$. Then
\[
z = t \cdot \left(\frac{y}{\nu}\right)^{1/2}
\]
which yields
\[
\frac{\partial z}{\partial t} = \left(\frac{y}{\nu}\right)^{1/2}.
\]
Hence, the joint probability density function for $Y$ and $T$ is given by
\[
h(y, t) = \frac{y^{\nu/2} \cdot \exp(-y/2) \cdot \exp\left(-\frac{t^2 y}{2\nu}\right)}{2^{\nu/2} \cdot \Gamma(\nu/2) \cdot \sqrt{2\pi}} \cdot \left(\frac{y}{\nu}\right)^{1/2} = \frac{y^{\nu/2} \cdot \exp\left(-\frac{y}{2} \cdot \left(1 + \frac{t^2}{\nu}\right)\right)}{2^{\nu/2} \cdot \sqrt{2\pi} \nu^{\nu/2} \cdot \Gamma(\nu/2)}
\]
for $y > 0$ and $t \in \mathbb{R}$.

Now, to find the marginal probability density function for $T$, we must integrate out the $y$. Toward this end, let
\[
u = \frac{y}{2} \cdot \left(1 + \frac{t^2}{\nu}\right) = \frac{y}{2} \cdot \left(\frac{\nu + t^2}{\nu}\right) \implies y = \frac{2\nu}{\nu + t^2} \cdot u \implies dy = \frac{2\nu}{\nu + t^2} \, du.
Then observe that
\[
\int_0^\infty h(y,t) \, dy = \int_0^\infty \left( \frac{2^\nu \cdot \nu \cdot t^t \cdot \exp(-u)}{\nu + t^2} \right) \cdot \frac{2\nu}{\nu + t^2} \, du
\]
\[
= \int_0^\infty 2^{\nu - 1} \cdot \left( \frac{1 + t^2}{\nu} \right)^{\nu - 1} \cdot \frac{\nu^\nu \cdot \exp(-u)}{\nu \cdot \Gamma(\nu/2)} \cdot 2 \cdot \left( 1 + \frac{t^2}{\nu} \right)^{-1} \, du
\]
\[
= \left( 1 + \frac{t^2}{\nu} \right)^{-\nu - 1} \cdot \frac{\Gamma(\nu/2)}{\nu \cdot \Gamma(\nu/2)} \cdot \int_0^\infty u^{\nu - 1} \cdot \exp(-u) \, du
\]
This finishes the proof.

**Corollary 180.** If \( \bar{X} \) and \( S^2 \) are the sample mean and sample variance, respectively, of a random sample of size \( n \) from a normally distributed random variable with mean \( \mu \) and variance \( \sigma^2 \), then
\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}}
\]
has a \( t \)-distribution with \( n - 1 \) degrees of freedom.

**Proof.** From Theorem 170, we know that
\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).
\]
Theorem 177 informs us that
\[
Y = \frac{n-1}{\sigma^2} \cdot S^2 \sim \chi(n-1)
\]
and that \( Y \) and \( Z \) are independent. Hence, Theorem 179 guarantees that
\[
T = \frac{Z}{\sqrt{Y/(n-1)}}
\]
\[
= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sqrt{n-1}}{\sqrt{n-1} \cdot S/\sqrt{n}}
\]
\[
= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S}
\]
\[
= \frac{\bar{X} - \mu}{S/\sqrt{n}}
\]
has the \( t \)-distribution with \( n - 1 \) degrees of freedom.

**Example 132.** A random sample of size 23 is taken from a normally distributed random variable. The sample mean is 17.3 and the sample standard deviation is 1.13. Test the claim that the population mean is \( \mu = 15 \).

We know that
\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}}
\]
has the \( t \)-distribution with \( n - 1 \) degrees of freedom. In particular
\[
\frac{\bar{X} - \mu}{S/\sqrt{23}}
\]
has the t-distribution with 22 degrees of freedom. Assuming that $\mu = 15$, we calculate

$$t = \frac{17.3 - 15}{1.13/\sqrt{23}} \approx 9.76143.$$  

Referring to a t-table, we note that

$$P(T > 3.792) \approx 0.0005$$

which means that $t = 9.76143$ is highly improbable. Therefore, it is quite unlikely that $\mu = 15$.

Theorem 181. The t-distribution with $\nu$ degrees of freedom approaches the standard normal distribution as $\nu \to \infty$. That is,

$$\lim_{\nu \to \infty} \frac{\Gamma \left( \frac{\nu+1}{2} \right)}{\sqrt{\nu \pi} \cdot \Gamma(\nu/2)} \cdot \left( 1 + \frac{t^2}{\nu} \right)^{-\frac{\nu+1}{2}} = \frac{e^{-t^2/2}}{\sqrt{2\pi}}$$

for $t \in \mathbb{R}$.

For an interactive graph of the t-distribution converging to the standard normal distribution, please visit Desmos.com.

8.5 The Fisher-Snedecor Distribution

Definition 182. A random variable $X$ has an $F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom, denoted $X \sim F(\nu_1, \nu_2)$, provided that its probability density function is given by

$$f(x) = \frac{\Gamma \left( \frac{\nu_1 + \nu_2}{2} \right)}{\Gamma(\nu_1/2) \cdot \Gamma(\nu_2/2)} \cdot \left( \frac{\nu_1}{\nu_2} \right)^{\nu_1/2} \cdot x^{\frac{\nu_1}{2} - 1} \cdot \left( 1 + \frac{\nu_1}{\nu_2} \cdot x \right)^{-\frac{\nu_1 + \nu_2}{2}}$$

for $x > 0$.

For an interactive graph of the F-distribution, please visit Desmos.com.

Theorem 183. If $U \sim \chi(\nu_1)$ and $V \sim \chi(\nu_2)$ are independent, then

$$\frac{U}{\nu_1} / \frac{V}{\nu_2} \sim F(\nu_1, \nu_2).$$

Primarily, F-distributions are used to compare variances of two normal populations. Particularly, we can study

$$\frac{\sigma_1^2}{\sigma_2^2}$$

which can give some indication whether $\sigma_1^2 = \sigma_2^2$.

Take a random sample of size $n_1$ from a normal population with variance $\sigma_1^2$ and a random sample of size $n_2$ from another normal population with variance $\sigma_2^2$. Let $S_1^2$ and $S_2^2$ be the sample variances for the two populations. Assuming that the collection of $n_1 + n_2$ random variables are independent, we can say that

$$\frac{n_1 - 1}{\sigma_1^2} \cdot S_1^2 \sim \chi(n_1 - 1) \quad \text{and} \quad \frac{n_2 - 1}{\sigma_2^2} \cdot S_2^2 \sim \chi(n_2 - 1)$$

are independent. Therefore, according to Theorem 183,

$$\frac{n_1 - 1}{\sigma_1^2} \cdot \frac{S_1^2}{n_1 - 1} \cdot \frac{n_2 - 1}{\sigma_2^2} \cdot \frac{S_2^2}{n_2 - 1} \sim F(n_1 - 1, n_2 - 1).$$
Example 133. Take independent random samples of sizes \( n_1 = 9 \) and \( n_2 = 11 \) from two normal variables with equal variances. Find a value \( x \) so that
\[
P\left(\frac{S_1^2}{S_2^2} < x\right) = 0.95.
\]
First, notice that
\[
\frac{S_1^2}{S_2^2} \sim F(8, 10).
\]
Also,
\[
0.95 = P\left(\frac{S_1^2}{S_2^2} < x\right) = 1 - P\left(\frac{S_1^2}{S_2^2} \geq x\right) \iff P\left(\frac{S_1^2}{S_2^2} \geq x\right) = 0.05.
\]
Referring to an \( F \)-table produces \( x \approx 3.0717 \). In words, there is a 95% likelihood that
\[
S_1^2 \geq (3.0717) \cdot S_2^2.
\]

8.6 Order Statistics

We’ve concerned ourselves mostly with normally distributed populations. In this section, we will consider populations which may not be normally distributed. Statistics without any assumption on the population is known as non-parametric statistics. Order statistics are a particular subfield of non-parametric statistics.

Definition 184. Consider a random sample \( X_1, X_2, \ldots, X_n \) of size \( n \) from a population modeled by a continuous random variable. Then arrange the sample data in ascending order to obtain the order statistics \( Y_1, Y_2, \ldots, Y_n \). Note that
\[
Y_1 = \min\{X_1, X_2, \ldots, X_n\}
\]
and
\[
Y_n = \max\{X_1, X_2, \ldots, X_n\}.
\]
\( Y_k \) is referred to as the \( k^{th} \) order statistic.

Theorem 185. Suppose \( X \) is a random variable with probability density function \( f(x) \). For a random sample \( X_1, X_2, \ldots, X_n \) of \( X \), the \( k^{th} \) order statistic \( Y_k \) has probability density function
\[
g(y_k) = \frac{n!}{(k-1)!(n-k)!} \cdot \left[ \int_{-\infty}^{y_k} f(x) \, dx \right]^{k-1} \cdot f(y_k) \cdot \left[ \int_{y_k}^{\infty} f(x) \, dx \right]^{n-k}.
\]

Example 134. Take a random sample \( X_1, X_2, \ldots, X_5 \) of size 5 from a random variable with an exponential distribution where the corresponding probability density function is
\[
f(x) = \frac{e^{-x/3}}{3}
\]
for \( x > 0 \). Find the probability that \( Y_1 < 2 \).

By Theorem 185, we know that the probability density function for \( Y_1 \) is given by
\[
f(y_1) = \frac{5!}{0! \cdot 4!} \cdot \frac{e^{-y_1/3}}{3} \cdot \left[ \int_{y_1}^{\infty} \frac{e^{-x/3}}{3} \, dx \right]^4
\]
\[
= 5 \cdot \frac{e^{-y_1/3}}{3} \cdot \left[ \lim_{k \to \infty} -e^{-x/3}\bigg|_{y_1}^{k} \right]^4
\]
\[
= 5 \cdot \frac{e^{-y_1/3}}{3} \cdot \left[ e^{-y_1/3} \right]^4
\]
\[
= 5 \cdot \frac{e^{-5y_1/3}}{3} \cdot e^{-4y_1/3}
\]
\[
= 5 \cdot \frac{e^{-y_1/3}}{3}.
\]
Hence,
\[ P(Y_1 < 2) = \int_0^2 \frac{5e^{-5y_1/3}}{3} \, dy_1 = 1 - \frac{1}{e^{10/3}} \approx 0.9643260066527. \]

**Example 135.** A random sample \( X_1, X_2, \ldots, X_7 \) of size 7 is taken from a random variable with an exponential distribution with an expected value of 5. Find values \( \alpha \) and \( \beta \) so that \( P(Y_1 < \alpha) = P(Y_7 > \beta) = 0.95 \).

First, we note that the exponential distribution has the probability density function
\[ f(x) = \frac{e^{-x/5}}{5} \]
for \( x > 0 \). Next, we appeal to Theorem 185 to obtain the probability density function for \( Y_1 \) which is given by
\[ g(y_1) = \frac{7!}{6!0!} \cdot \frac{e^{-y_1/5}}{5} \cdot \left[ \int_{y_1}^{\infty} e^{-x/5} \, dx \right]^6 \]
\[ = \frac{7}{5} \cdot \frac{e^{-y_1/5}}{5} \cdot e^{-6y_1/5} \]
\[ = \frac{7}{5} \cdot e^{-7y_1/5} \]
and the probability density function for \( Y_7 \) which is given by
\[ h(y_7) = \frac{7!}{6!0!} \left[ \int_{0}^{y_7} \frac{e^{-x/5}}{5} \, dx \right]^6 \cdot \frac{e^{-y_7/5}}{5} \]
\[ = \frac{7}{5} \cdot \left[ 1 - e^{-y_7/5} \right]^6 \cdot e^{-y_7/5}. \]

Now,
\[ P(Y_1 < \alpha) = \int_0^\alpha \frac{7}{5} \cdot e^{-7y_1/5} \, dy_1 = 1 - e^{-7\alpha/5}. \]
Solving,
\[ P(Y_1 < \alpha) = 0.95 \implies 1 - e^{-7\alpha/5} = 0.95 \]
\[ \implies 0.05 = e^{-7\alpha/5} \]
\[ \implies \alpha = \frac{5 \cdot \ln(0.5)}{7} \approx 0.495105. \]

For \( Y_7 \), let
\[ u = 1 - e^{-y_7/5} \implies du = \frac{e^{-y_7/5}}{5} \, dy_7. \]
It follows that
\[ P(Y_7 > \beta) = \int_\beta^\infty \frac{7}{5} \cdot \left[ 1 - e^{-y_7/5} \right]^6 \cdot e^{-y_7/5} \, dy_7 \]
\[ = \int_{1-e^{-\beta/5}}^1 7u^6 \, du \]
\[ = u^7 \bigg|_{1-e^{-\beta/5}}^1 \]
\[ = 1 - \left[ 1 - e^{-\beta/5} \right]^7. \]
Solving
\[ P(Y_7 > \beta) = 0.95 \implies 1 - \left[ 1 - e^{-\beta/5} \right]^7 = 0.95 \]
\[ \implies 0.05 = \left[ 1 - e^{-\beta/5} \right]^7 \]
\[ \implies (0.05)^{1/7} = 1 - e^{-\beta/5} \]
\[ \implies e^{-\beta/5} = 1 - (0.05)^{1/7} \]
\[ \implies \beta = -5 \cdot \ln(1 - (0.05)^{1/7}) \approx 5.275413. \]

**Example 136.** Suppose \( X_1 \) and \( X_2 \) forms a random sample of a random variable with an exponential distribution given by \( f(x) = e^{-x} \) for \( x > 0 \). Find the conditional probability density for the second order statistic \( Y_2 \) given that \( Y_1 = \alpha > 0 \).

If \( Y_1 = \alpha \), note that either
- \( X_1 = \alpha \) and \( X_2 = \beta > \alpha \)
- \( X_1 = \beta > \alpha \) and \( X_2 = \alpha \).

By independence, the joint probability density for \( X_1 \) and \( X_2 \) is given by
\[ g(x_1, x_2) = e^{-(x_1 + x_2)} \]
where \( x_1, x_2 > 0 \). Now, observe that the joint probability density for \( Y_1 \) and \( Y_2 \) is given by
\[ h(y_1, y_2) = g(x_1, x_2) + g(x_2, x_1) = 2e^{-(y_1 + y_2)} \]
for \( 0 < y_1 \leq y_2 \). The marginal density for \( Y_1 \) is thus
\[ h_1(y_1) = \int_{y_1}^{\infty} 2e^{-(y_1 + y_2)} \, dy_2 = 2e^{-2y_1} \]
for \( y_1 > 0 \).

Hence, the conditional probability density of \( Y_2 \) given that \( Y_1 = \alpha \) is given by
\[ \varphi(y_2 | y_1 = \alpha) = \frac{h(\alpha, y_2)}{h_1(\alpha)} = \frac{2e^{-(\alpha + y_2)}}{2e^{-2\alpha}} = e^{\alpha - y_2} \]
for \( y_2 > \alpha \).

**Example 137.** For a random sample \( X_1, X_2, \ldots, X_n \) from a uniformly distributed random variable on \([0, \beta]\), find \( E(Y_n) \).

By Theorem 185, we know that
\[ g(y_n) = \frac{n!}{(n-1)! (n-1)!} \cdot \left[ \int_0^{y_n} \frac{1}{\beta} \right]^{n-1} \cdot \frac{1}{\beta} \]
\[ = n \cdot \left( \frac{y_n}{\beta} \right)^{n-1} \cdot \frac{1}{\beta} \]
\[ = ny_n^{n-1} \frac{1}{\beta^n} \]
8.6. ORDER STATISTICS

for $0 < y_n < \beta$. Therefore,

$$E(Y_n) = \int_0^\beta y_n \cdot \frac{ny_n^{n-1}}{\beta^n} \, dy_n = \int_0^\beta \frac{ny_n^n}{\beta^n} \, dy_n = \frac{n}{\beta^n} \cdot \frac{y_n^{n+1}}{n+1} \bigg|_0^\beta = \frac{n\beta}{n+1}.$$ 

8.6.1 The Sample Median

**Definition 186.** For a random sample from an infinite population, the sample median is defined,

- for a sample size of $2n + 1$, to be the $(n + 1)^{th}$ order statistic $Y_{n+1}$.
- for a sample size of $2n$, to be

$$\frac{Y_n + Y_{n+1}}{2}.$$
Chapter 9

Point Estimation

Definition 187. Using the value of a sample statistic to estimate the value of a population parameter is called point estimation. The value of the statistic is known as a point estimate. The statistic used for point estimation will be referred to as an estimator of the parameter $\theta$.

Example 138. The sample mean $\bar{X}$ is an estimator for the population mean $\mu$.

Example 139. A random sample from a random variable yields a sample mean of 12.5. Then 12.5 is a point estimate of the population mean.

9.1 Properties of Estimators

Definition 188. A statistic $\hat{\Theta}$ is said to be an unbiased estimator of a parameter $\theta$ if $E(\hat{\Theta}) = \theta$ for all possible values of $\theta$. Otherwise, $\hat{\Theta}$ is said to be a biased estimator.

Example 140. The sample mean $\bar{X}$ of a random sample from an infinite population is an unbiased estimator of the population mean $\mu$ by Theorem 166.

Example 141. The sample variance $S^2$ of a random sample from an infinite population is an unbiased estimator of the population variance $\sigma^2$ by Theorem 173.

Example 142. Suppose $Y_n$ is the $n^{th}$ order statistic from a random sample of size $n$ from a random variable which is uniformly distributed on the interval $[0, \beta]$. By Example 137, $\hat{\Theta} = \frac{n+1}{n} \cdot Y_n$ is an unbiased estimator of $\beta$.

Example 143. Suppose $X \sim B(n, p)$. Then

$$\hat{\Theta} = \frac{X + \frac{1}{2} \cdot \sqrt{n}}{n + \sqrt{n}}$$

is a biased estimator of $p$.

Recalling that $E(X) = np$, we calculate

$$E(\hat{\Theta}) = E \left[ \frac{X + \frac{1}{2} \cdot \sqrt{n}}{n + \sqrt{n}} \right]$$

$$= \frac{E(X) + \frac{1}{2} \cdot \sqrt{n}}{n + \sqrt{n}}$$

$$= \frac{np + \frac{1}{2} \cdot \sqrt{n}}{n + \sqrt{n}}$$

$$= \frac{2np + \sqrt{n}}{2n + 2 \cdot \sqrt{n}}.$$
Were it true that \( E(\hat{\theta}) = p \), then
\[
p = \frac{2np + \sqrt{n}}{2n + 2 \cdot \sqrt{n}} \implies \sqrt{n} = (2n + 2 \cdot \sqrt{n}) \cdot p - 2np = 2 \cdot \sqrt{n} \cdot p \implies p = \frac{1}{2}.
\]
Alas, the parameter \( p \) can take on other values than just \( 1/2 \) so \( \hat{\theta} \) is, indeed, a biased estimator of \( p \).

**Definition 189.** The **bias** of an estimator \( \hat{\theta} \) of \( \theta \) is \( \text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \).

Immediately we see that an estimator \( \hat{\theta} \) of \( \theta \) is unbiased if and only if \( \text{Bias}(\hat{\theta}) = 0 \).

**Definition 190.** A statistic \( \hat{\theta} \) is said to be an **asymptotically unbiased estimator** of \( \theta \) if
\[
\lim_{n \to \infty} \text{Bias}(\hat{\theta}) = 0
\]
where \( n \) is the size of the random sample.

**Example 144.** Suppose \( X \sim B(n, p) \). Then the biased estimator
\[
\hat{\theta} = \frac{X + \frac{1}{2} \cdot \sqrt{n}}{n + \sqrt{n}}
\]
of \( p \) is an asymptotically unbiased estimator of \( p \).

First, recalling our work from Example 143, we calculate
\[
\text{Bias}(\hat{\theta}) = \frac{2np + \sqrt{n}}{2n + 2 \cdot \sqrt{n}} - p
\]
\[
= \frac{2np + \sqrt{n} - 2np - 2 \cdot \sqrt{n} \cdot p}{2n + 2 \cdot \sqrt{n}}
\]
\[
= \frac{\sqrt{n} \cdot (1 - 2p)}{\sqrt{n} \cdot (2 \cdot \sqrt{n} + 2)}
\]
\[
= \frac{1 - 2p}{2 \cdot \sqrt{n} + 2}
\]
which provides
\[
\lim_{n \to \infty} \text{Bias}(\hat{\theta}) = \lim_{n \to \infty} \frac{1 - 2p}{2 \cdot \sqrt{n} + 2} = 0.
\]
Therefore, \( \hat{\theta} \) is an asymptotically unbiased estimator of \( \theta \).

**Definition 191.** An unbiased estimator \( \hat{\theta} \) of \( \theta \) is said to be a **minimum variance** unbiased estimator if, for any other unbiased estimator \( \hat{\Phi} \) of \( \theta \),
\[
\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\Phi}).
\]

The benefit of minimum variance unbiased estimators is that they will be closer to the value of the population parameter with higher probability. To determine whether or not an unbiased estimator is of minimum variance, we will use Theorem 193.

**Definition 192.** Given an unknown parameter \( \theta \) from a population \( X \) with probability density \( f(x) \), the **Fisher information**, or simply **information**, of \( \theta \) is
\[
\mathcal{I}(\theta) = E \left[ \left( \frac{\partial}{\partial \theta} \ln(f(X)) \right)^2 \right].
\]
Theorem 193. Suppose $X$ has a probability density function $f(x; \theta)$ with parameter $\theta$ so that $f(x; \theta)$ has continuous first- and second-order partial derivatives at all but finitely many points and that the set of points for which $f(x; \theta) \neq 0$ does not depend on $\theta$.

(a) For any unbiased estimator $\hat{\Theta}$ of $\theta$,
\[
\text{Var}(\hat{\Theta}) \geq \frac{1}{n \cdot I(\theta)}
\]
where $n$ is the size of the random sample.

(b) If $\hat{\Theta}$ is an unbiased estimator of $\theta$ and
\[
\text{Var}(\hat{\Theta}) = \frac{1}{n \cdot I(\theta)}
\]
then $\hat{\Theta}$ is a minimum variance unbiased estimator for $\theta$.

Example 145. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{X}$ is a minimum variance unbiased estimator of $\mu$.

Recall that the probability density function for $X$ is given by
\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} \cdot \exp \left( -\frac{1}{2} \cdot \left( \frac{x - \mu}{\sigma} \right)^2 \right).
\]
Then
\[
\ln(f(x)) = -\frac{1}{2} \left( \frac{X - \mu}{\sigma} \right)^2 - \ln(\sigma \sqrt{2\pi})
\]
Hence,
\[
\frac{\partial}{\partial \mu} \ln(f(X)) = -\frac{2}{2} \cdot \left( \frac{X - \mu}{\sigma} \right) \cdot -\frac{1}{\sigma} = \frac{X - \mu}{\sigma^2}.
\]
It follows that
\[
I(\mu) = E \left[ \left( \frac{\partial}{\partial \mu} \ln(f(X)) \right)^2 \right] = \frac{1}{\sigma^4} \cdot E[(X - \mu)^2] = \frac{1}{\sigma^2}.
\]
Recall from Theorem 170 that
\[
\text{Var}(\bar{X}) = \frac{\sigma^2}{n}
\]
where $n$ is the sample size. We finish with the observation that
\[
\text{Var}(\bar{X}) = \frac{1}{n \cdot I(\mu)}.
\]

Example 146. With reference to Example 142, show that the unbiased estimator $\frac{n+1}{n} \cdot Y_n$ does not satisfy Theorem 193.

First, notice that $f(x) = 1/\beta$ for $0 < x < \beta$. Then
\[
\ln(f(x)) = \ln(1/\beta) = -\ln(\beta)
\]
which implies that
\[
\frac{d}{d\beta} \ln(f(x)) = \frac{-1}{\beta}.
\]
Hence,
\[
I(\beta) = \frac{1}{\beta^2}.
\]
We will now compute \( \text{Var}(Y_n) \). Recall from Example 137 that the probability density function for \( Y_n \) is given by
\[
g(y_n) = \frac{ny_n^{n-1}}{\beta^n}
\]
where \( 0 < y_n < \beta \) and that \( E(Y_n) = \frac{n\beta}{n+1} \). Since \( \text{Var}(Y_n) = E(Y_n^2) - E(Y_n)^2 \), we now endeavor to find \( E(Y_n^2) \).

Note that
\[
P(Y_n^2 < t) = P(Y_n < \sqrt{t}) = \int_0^{\sqrt{t}} \frac{ny_n^{n-1}}{\beta^n} \, dy_n = \left. \frac{y_n^n}{\beta^n} \right|_0^{\sqrt{t}} = \frac{t^{n/2}}{\beta^n}
\]
for \( 0 < t < \beta^2 \). It follows that, the probability density for \( Z = Y_n^2 \) is given by
\[
h(z) = \frac{n z^{n-2}}{2\beta^n}
\]
for \( 0 < z < \beta^2 \). Hence,
\[
E(Y_n^2) = \int_0^{\beta^2} z \cdot \frac{n z^{n-2}}{2\beta^n} \, dz = \frac{n}{2\beta^n} \cdot \int_0^{\beta^2} z^{n-2} \, dz = \frac{n}{2\beta^n} \left[ \frac{z^{n+1}}{n+1} \right]_0^{\beta^2} = \frac{n}{(n+2)\beta^n} \cdot \beta^{n+2} = \frac{n}{n+2} \cdot \beta^2.
\]

Thus,
\[
\text{Var}(Y_n) = \frac{n}{n+2} \cdot \beta^2 - \left( \frac{n}{n+1} \right)^2 \cdot \beta^2 = \frac{n}{n+2} - \frac{n^2}{(n+1)^2} \cdot \beta^2 = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \cdot \beta^2 = \frac{n^3 + 2n^2 - n - n^3 - 2n^2}{(n+2)(n+1)^2} \cdot \beta^2 = \frac{n}{(n+2)(n+1)^2} \cdot \beta^2.
\]

Moreover,
\[
\text{Var} \left( \frac{n+1}{n} \cdot Y_n \right) = \frac{(n+1)^2}{n^2} \cdot \left[ \frac{n}{(n+2)(n+1)^2} \cdot \beta^2 \right] = \frac{\beta^2}{n(n+2)}
\]
What we see here is that
\[
\text{Var} \left( \frac{n+1}{n} \cdot Y_n \right) = \frac{\beta^2}{n(n+2)} < \frac{\beta^2}{n} = \frac{1}{n \cdot I(\beta)}.
\]
The failing hypothesis is the one requiring the set of \( x \) for which \( f(x; \theta) \neq 0 \) to not be dependent on \( \theta \). For uniform distributions, there is an obvious dependence on \( \beta \).

If we have occasion to consider biased estimators of a parameter \( \theta \), instead of wishing to minimize their variance, we will find it more useful to minimize their variance about the parameter.
Definition 194. For a statistic \( \hat{\Theta} \), we define the mean square error relative to a parameter \( \theta \) to be

\[
MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2].
\]

This can be further motivated by noting that, for an unbiased estimator \( \hat{\Theta} \) of \( \theta \),

\[
Var(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2] = MSE(\hat{\Theta}).
\]

Theorem 195. For an estimator \( \hat{\Theta} \) of \( \theta \),

\[
MSE(\hat{\Theta}) = Var(\hat{\Theta}) + (Bias(\hat{\Theta}))^2.
\]

Proof. Observe that

\[
MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2] = E[(\hat{\Theta} - E(\hat{\Theta}) + E(\hat{\Theta}) - \theta)^2] = E[(\hat{\Theta} - E(\hat{\Theta}))^2] + 2(\hat{\Theta} - E(\hat{\Theta}))Bias(\hat{\Theta}) + (Bias(\hat{\Theta}))^2 = Var(\hat{\Theta}) + (Bias(\hat{\Theta}))^2.
\]

Definition 196. A statistic \( \hat{\Theta} \) is said to be a consistent estimator of \( \theta \) if

\[
\lim_{n \to \infty} P(|\hat{\Theta} - \theta| < c) = 1
\]

for any constant \( c > 0 \) where \( n \) is the sample size.

Theorem 197. If \( \hat{\Theta} \) is an unbiased estimator of \( \theta \) and \( \lim_{n \to \infty} Var(\hat{\Theta}) = 0 \) were \( n \) is the sample size, then \( \hat{\Theta} \) is a consistent estimator of \( \theta \).

Proof. As \( \hat{\Theta} \) is an unbiased estimator, \( E(\hat{\Theta}) = \theta \). For \( \sigma = Var(\hat{\Theta}) \) and \( k > 1 \),

\[
P(|\hat{\Theta} - \theta| < k\sigma) \geq 1 - \frac{1}{k^2}
\]

by Chebyshev’s Inequality. Consider \( k = \frac{c}{\sigma} \). Since

\[
\lim_{n \to \infty} \frac{c}{\sigma} = \infty
\]

for a constant \( c > 0 \), we know that \( k = c/\sigma \) is eventually bigger than 1. Hence,

\[
P(|\hat{\Theta} - \theta| < k\sigma) \geq 1 - \frac{1}{k^2}\]

\[
P\left(|\hat{\Theta} - \theta| < c\right) \geq 1 - \frac{\sigma^2}{c^2}.
\]

Therefore,

\[
\lim_{n \to \infty} P\left(|\hat{\Theta} - \theta| < c\right) \geq \lim_{n \to \infty} 1 - \frac{\sigma^2}{c^2} = 1,
\]

finishing the proof.

Example 147. For a random sample from any infinite population, the sample mean \( \bar{X} \) is a consistent estimator of the population variance \( \mu \).
Let \( \sigma^2 \) be the population variance and \( n \) represent the sample size. Recall that

\[
E(\bar{X}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.
\]

As \( \lim_{n \to \infty} \text{Var}(\bar{X}) = 0 \),

Theorem 197 guarantees that \( \bar{X} \) is a consistent estimator of \( \mu \).

**Example 148.** For random samples from a normal variable, the sample variance \( S^2 \) is a consistent estimator of the population variance \( \sigma^2 \).

As demonstrated in Theorem 173, \( S^2 \) is an unbiased estimator of \( \sigma^2 \). For a sample size of \( n \), recall that Theorem 177 asserts that

\[
\frac{n-1}{\sigma^2} \cdot S^2 \sim \chi(n-1).
\]

By Corollary 137,

\[
\text{Var} \left( \frac{n-1}{\sigma^2} \cdot S^2 \right) = 2(n-1).
\]

Hence,

\[
2(n-1) = \text{Var} \left( \frac{n-1}{\sigma^2} \cdot S^2 \right) = \frac{(n-1)^2}{\sigma^4} \cdot \text{Var}(S^2)
\]

which provides

\[
\text{Var}(S^2) = \frac{2\sigma^4}{n-1}.
\]

Thus, we see that

\[
\lim_{n \to \infty} \text{Var}(S^2) = 0
\]

so we can invoke Theorem 197 to conclude that \( S^2 \) is a consistent estimator of \( \sigma^2 \).

**Definition 198.** A statistic \( \hat{\Theta} \) is said to be a \textbf{sufficient} estimator of \( \theta \) if the conditional distribution of the joint distribution of a random sample \( X_1, X_2, \ldots, X_n \) given \( \hat{\Theta} = \theta \) is independent of the parameter \( \theta \).

Intuitively, a sufficient estimator of the parameter \( \theta \) requires no more information about the random sample to produce its estimates. In other words, no other statistic (based on the same random sample) will provide more information about the parameter \( \theta \). For example, if we sample from a normal population, once the sample mean is determined (even if the values of the random sample are unknown), no more information about \( \mu \) can be obtained from this sample as we will see more formally in Example 151.

**Example 149.** Let \( X \sim \text{Bernoulli}(p) \). Then, for a random sample \( X_1, X_2, \ldots, X_n \) of \( X \), \( \bar{X}/n \) is a sufficient estimator for \( p \).

Recall that

\[
Y = \sum_{j=1}^{n} X_j \sim B(n, p).
\]

That is, the probability distribution function for \( Y \) is

\[
g(y) = \binom{n}{y} p^y (1 - p)^{n-y}
\]
for \( y = 0, 1, 2, \ldots, n \). Using substitution, the probability distribution function for \( \bar{X} = Y/n \) is given by

\[
g(\bar{x}) = \binom{n}{n \cdot \bar{x}} p^{n \cdot \bar{x}} (1 - p)^{n - n \cdot \bar{x}}
\]

for \( \bar{x} = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1 \).

The joint probability distribution for the random sample is

\[
f(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{n} p^{x_j} (1 - p)^{1-x_j} = p^{x_1 + x_2 + \cdots + x_n} (1 - p)^{(1-x_1) + (1-x_2) + \cdots + (1-x_n)} = p^y (1 - p)^{n-y}.
\]

Then the conditional probability distribution is

\[
f(x_1, x_2, \ldots, x_n | \bar{x} = p) = \frac{f(x_1, x_2, \ldots, x_n)}{g(\bar{x})} = \frac{p^y (1 - p)^{n-y}}{\binom{n}{n \cdot \bar{x}} p^{n \cdot \bar{x}} (1 - p)^{n - n \cdot \bar{x}}} = \frac{p^n (1 - p)^{n-y}}{\binom{n}{n \cdot \bar{x}}} = \frac{1}{\binom{n}{x_1 + x_2 + \cdots + x_n}}.
\]

This does not depend on \( p \).

---

**Example 150.** Let \( X \sim \text{Bernoulli}(p) \). Then, for a random sample \( X_1, X_2, X_3 \) of \( X \), the statistic

\[
Y = \frac{X_1 + 2X_2 + 3X_3}{6}
\]

is an unbiased estimator of \( p \) but it is not sufficient.

First, notice that

\[
E(Y) = E \left[ \frac{X_1 + 2X_2 + 3X_3}{6} \right] = \frac{E(X_1) + 2 \cdot E(X_2) + 3 \cdot E(X_3)}{6} = \frac{p + 2p + 3p}{6} = p.
\]

That is, \( Y \) is an unbiased estimator of \( p \).

The joint probability distribution for \( X_1, X_2, X_3 \) is given by

\[
f(x_1, x_2, x_3) = p^{x_1 + x_2 + x_3} (1 - p)^{3-(x_1 + x_2 + x_3)}.
\]

Consider the following table of values:

<table>
<thead>
<tr>
<th>((x_1, x_2, x_3))</th>
<th>(f(x_1, x_2, x_3))</th>
<th>(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0)</td>
<td>((1-p)^3)</td>
<td>0</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>(p(1-p)^2)</td>
<td>1/2</td>
</tr>
<tr>
<td>(0,1,0)</td>
<td>(p(1-p)^2)</td>
<td>1/3</td>
</tr>
<tr>
<td>(0,1,1)</td>
<td>(p^2(1-p))</td>
<td>5/6</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>(p(1-p)^2)</td>
<td>1/6</td>
</tr>
<tr>
<td>(1,0,1)</td>
<td>(p^2(1-p))</td>
<td>2/3</td>
</tr>
<tr>
<td>(1,1,0)</td>
<td>(p^2(1-p))</td>
<td>1/2</td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>(p^3)</td>
<td>1</td>
</tr>
</tbody>
</table>
Observe that
\[ P(Y = 1/2) = p(1 - p)^2 + p^2(1 - p) = p(1 - p)(1 - p + p) = p(1 - p) \]
and that the conditional probability distribution function given \( Y = 1/2 \) is
\[
f(x_1, x_2, x_3 | Y = 1/2) = \begin{cases} 
1 - p, & (x_1, x_2, x_3) = (0, 0, 1); \\
p, & (x_1, x_2, x_3) = (1, 1, 0).
\end{cases}
\]
Moreover,
\[
f(1, 1, 0) \cdot P(Y = 1/2) = p^2(1 - p) \cdot p(1 - p) \neq p.
\]
This witnesses the dependence on \( p \).

**Theorem 199.** For a statistic \( \Theta \) and a population parameter \( \theta \), if the conditional probability distribution/density function
\[
f(x_1, x_2, \ldots, x_n | \Theta = \theta) = g(\Theta, \theta) \cdot h(x_1, x_2, \ldots, x_n)
\]
where \( g \) doesn’t depend on \( x_1, x_2, \ldots, x_n \) and \( h \) only depends on \( x_1, x_2, \ldots, x_n \), then \( \Theta \) is a sufficient estimator for \( \theta \).

Though we won’t prove Theorem 199 in its full generality, we’ll offer a proof for the discrete case.

**Proof.** Suppose a discrete random variable \( x \) has probability distribution function \( f(x) \) for \( x = x_1, x_2, \ldots, x_m \). Also, \( \Theta = T(X_1, X_2, \ldots, X_n) \) for a random sample \( X_1, X_2, \ldots, X_n \) from \( X \). By substitution, the marginal distribution for \( \Theta \) is given by
\[
g_0(\hat{\theta}) = \sum_{\tilde{x}: T(\tilde{x}) = \hat{\theta}} f(x_{j_1}) \cdot f(x_{j_2}) \cdots f(x_{j_n})
\]
where \( \tilde{x} = (x_{j_1}, x_{j_2}, \ldots, x_{j_n}) \). Hence,
\[
f(x_{k_1}, x_{k_2}, \ldots, x_{k_n} | \hat{\Theta} = \hat{\theta}) = \frac{f(x_{k_1}) \cdot f(x_{k_2}) \cdots f(x_{k_n})}{g_0(\hat{\theta})}
\]
for any \( (x_{k_1}, x_{k_2}, \ldots, x_{k_n}) \) so that \( T(x_{k_1}, x_{k_2}, \ldots, x_{k_n}) = \hat{\theta} \). By the hypothesis,
\[
f(x_{k_1}, x_{k_2}, \ldots, x_{k_n} | \hat{\Theta} = \hat{\theta}) = g(\hat{\Theta}, \theta) \cdot h(x_{k_1}, x_{k_2}, \ldots, x_{k_n}).
\]
\[ \ldots \text{still need to finish} \]

**Example 151.** The sample mean \( \bar{X} \) from a random sample of a normal population with a known variance \( \sigma^2 \) is a sufficient estimator for the population mean \( \mu \).

First, recall that
\[
\sum_{j=1}^{n} (x_j - \mu)^2 = \left[ \sum_{j=1}^{n} (x_j - \bar{x})^2 \right] + n \cdot (\bar{x} - \mu)^2.
\]
Then,
\[
f(x_1, x_2, \ldots, x_n | \bar{x} = \mu) = \left( \frac{1}{\sigma \cdot \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2} \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\sigma^2} \right]
\]
\[
= \left( \frac{1}{\sigma \cdot \sqrt{2\pi}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \cdot \left( n \cdot (\bar{x} - \mu)^2 + \sum_{j=1}^{n} (x_j - \bar{x})^2 \right) \right]
\]
\[
= \left( \frac{1}{\sigma \cdot \sqrt{2\pi}} \right)^n \exp \left[ -\frac{n \cdot (\bar{x} - \mu)^2}{2\sigma^2} \right] \cdot \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \bar{x})^2 \right].
\]
Let
\[ g(\bar{x}, \mu) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \cdot \exp \left[ -\frac{n \cdot (\bar{x} - \mu)^2}{2\sigma^2} \right] \]
and
\[ h(x_1, x_2, \ldots, x_n) = \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \bar{x})^2 \right]. \]
Since \( \bar{x} \) can be rewritten in terms of the \( x_j, j = 1, 2, \ldots, n \), we can apply Theorem 199 to conclude that \( \bar{X} \) is a sufficient estimator of \( \mu \).

**Example 152.** Show that the \( n^{th} \) order statistic for a random sample \( X_1, X_2, \ldots, X_n \) from a random variable which is uniformly distributed on \( [0, \beta] \) is a sufficient estimator of \( \beta \).

Observe that
\[
 f(x_1, x_2, \ldots, x_n|y_n = \beta) = \begin{cases} 
 \left( \frac{1}{\beta} \right)^n, & 0 < \min\{x_1, x_2, \ldots, x_n\}, \\
 0, & \max\{x_1, x_2, \ldots, x_n\} \leq \beta; \\
 1, & \text{otherwise.} 
\end{cases}
\]
If
\[
 g(\beta, y_n) = \begin{cases} 
 1/\beta^n, & y_n \leq \beta; \\
 0, & \text{otherwise.} 
\end{cases}
\]
and
\[
 h(x_1, x_2, \ldots, x_n) = \begin{cases} 
 1, & 0 < \min\{x_1, x_2, \ldots, x_n\}; \\
 0, & \text{otherwise} 
\end{cases}
\]
we can appeal to Theorem 199 to conclude that \( Y_n \) is a sufficient estimator for \( \beta \).

### 9.2 Producing Estimators

Having discussed estimators and some of their properties, we now turn our attention to producing them.

#### 9.2.1 The Method of Moments

**Definition 200.** The \( k^{th} \) **sample moment** of a set of observations \( x_1, x_2, \ldots, x_n \) is the mean of their \( k^{th} \) powers
\[ m'_k = \frac{1}{n} \sum_{j=1}^{n} x_j^k \]
If a population \( X \) has \( \ell \) many parameters, the **method of moments** consists of solving the system of equations
\[
\begin{align*}
 m'_1 &= \mu'_1 \\
 m'_2 &= \mu'_2 \\
 & \vdots \\
 m'_\ell &= \mu'_\ell
\end{align*}
\]
where \( \mu'_k \) are the \( k^{th} \) moments about the origin.

**Example 153.** Given a random sample of size \( n \) from a continuous random variable which is uniformly distributed on the interval \([\alpha, 1]\), use the method of moments to find an estimator for \( \alpha \).
Since there is only one unknown parameter, we need only solve
\[ \bar{x} = \frac{\alpha + 1}{2} \implies 2 \cdot \bar{x} - 1 = \alpha. \]
Hence, we have produced the estimator \(2 \cdot \bar{X} - 1\) of \(\alpha\).

**Example 154.** Given a random sample of size \(n\) from a gamma population, use the method of moments to find an estimator for the parameters \(\alpha\) and \(\beta\).

Since we’re looking for two parameters, we need to solve
\[
\begin{cases}
m'_1 = \alpha/\beta \\
m'_2 = (\alpha^2 + \alpha)/\beta
\end{cases}
\]
recalling the moments for gamma distributions from Theorem 131. Using \(\alpha = m'_1 \cdot \beta\), we write
\[
m'_2 = \frac{(m'_1)^2 \cdot \beta^2 + m'_1 \cdot \beta}{\beta} = (m'_1)^2 \cdot \beta + m'_1
\]
which implies that
\[\beta = \frac{m'_2 - m'_1}{(m'_1)^2}.
\]
Moreover,
\[\alpha = m'_1 \cdot \frac{m'_2 - m'_1}{(m'_1)^2} = \frac{m'_2 - m'_1}{m'_1}.
\]
Letting
\[Y = \frac{1}{n} \sum_{k=1}^{n} X_k^2,
\]
we can write the estimator
\[\frac{Y - \bar{X}}{\bar{X}}\]
for \(\alpha\) and the estimator
\[\frac{Y - \bar{X}}{(\bar{X})^2}\]
for \(\beta\).

### 9.2.2 The Method of Maximum Likelihood

**Definition 201.** If \(x_1, x_2, \ldots, x_n\) are the values of a random sample from a population with a parameters \(\theta_1, \theta_2, \ldots, \theta_m\), the **likelihood function** of the sample is
\[L(\theta_1, \theta_2, \ldots, \theta_m) = f(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)\]
where \(f(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)\) is the joint probability distribution/density function relative to the parameters \(\theta_1, \theta_2, \ldots, \theta_m\). The value of \(\theta_j\) which maximizes \(L(\theta_1, \theta_2, \ldots, \theta_m)\) is known as the **maximum likelihood estimator** of \(\theta_j\).
9.2. PRODUCING ESTIMATORS

**Note.** When convenient, we will use the fact that maximizing \( \ln(L(\theta)) \) provides the same locations for the maximum values of \( L(\theta) \). This follows from the fact that

\[
\frac{\partial}{\partial \theta} \ln(L(\theta)) = \frac{\partial \ln(L(\theta))}{\partial \theta} = \frac{\partial L(\theta)}{L(\theta)}
\]

whenever \( L(\theta) \neq 0 \).

**Example 155.** Given \( x \) successes in \( n \) trials of a binomial experiment, find the maximum likelihood estimator of \( p \), the probability of success in each trial.

Here,

\[
L(p) = \binom{n}{x} p^x (1-p)^{n-x} \implies \ln(L(p)) = \ln\left( \binom{n}{x} \right) + x \ln(p) + (n-x) \ln(1-p).
\]

Moreover,

\[
\frac{\partial}{\partial p} \ln(L(p)) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x(1-p) - p(n-x)}{p(1-p)} = \frac{x-pn}{p(1-p)}.
\]

To maximize, we solve

\[
\frac{x-pn}{p(1-p)} = 0 \quad \Rightarrow \quad x = pn.
\]

Therefore, the maximum likelihood estimator of \( p \) is \( \frac{X}{n} \).

**Example 156.** Given that \( x_1, x_2, \ldots, x_n \) are the values of a random sample from an exponential population of parameter \( \lambda \), find the maximum likelihood estimator of \( \lambda \).

Observe that

\[
L(\lambda) = \prod_{j=1}^{n} \lambda e^{-\lambda x_j} = \lambda^n e^{-\lambda (x_1 + x_2 + \cdots + x_n)}
\]

which provides

\[
\ln(L(\lambda)) = n \cdot \ln(\lambda) + -\lambda \cdot \sum_{j=1}^{n} x_j.
\]

Moreover,

\[
\frac{d}{d\lambda} \ln(L(\lambda)) = \frac{n}{\lambda} - \sum_{j=1}^{n} x_j.
\]

To maximize, we solve

\[
\frac{n}{\lambda} - \sum_{j=1}^{n} x_j = 0 
\]

\[
\frac{n}{\lambda} = \sum_{j=1}^{n} x_j
\]

\[
\frac{1}{\lambda} = \bar{x}
\]

which establishes \( \frac{1}{\bar{x}} \) as the maximum likelihood estimator of \( \lambda \).
Example 157. Suppose \( x_1, x_2, \ldots, x_n \) are the values of a random sample from a population which is uniformly distributed on \([0, \beta]\). Find the maximum likelihood estimator for \( \beta \).

Observe that the likelihood function here is

\[
L(\beta) = \left( \frac{1}{\beta} \right)^n = \frac{1}{\beta^n}.
\]

For \( \beta > 0 \), this function is strictly decreasing. Hence, to maximize \( L(\beta) \) given the values \( x_1, x_2, \ldots, x_n \), we take

\[
\beta = \max\{x_1, x_2, \ldots, x_n\}.
\]

That is, the maximum likelihood estimator of \( \beta \) is the \( n \)th order statistic.

Example 158. For a normal population with mean \( \mu \) and variance \( \sigma^2 \), find the maximum likelihood estimators for both \( \mu \) and \( \sigma^2 \).

The likelihood function here is

\[
L(\mu, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \left[ -\frac{(x_j - \mu)^2}{2\sigma^2} \right] = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[ -\frac{1}{2\sigma^2} \cdot \sum_{j=1}^{n} (x_j - \mu)^2 \right].
\]

Then

\[
\ln(L(\mu, \sigma^2)) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \cdot \sum_{j=1}^{n} (x_j - \mu)^2.
\]

Note that

\[
\frac{\partial}{\partial \mu} \ln(L(\mu, \sigma^2)) = -\frac{n}{2} \cdot \frac{1}{\sigma^2} \cdot \sum_{j=1}^{n} 2(x_j - \mu) \cdot (-1)
\]
\[
= \frac{1}{\sigma^2} \cdot \sum_{j=1}^{n} (x_j - \mu)
\]
\[
= \frac{n}{\sigma^2} \cdot (\bar{x} - \mu).
\]

Then

\[
\frac{\partial}{\partial \mu} \ln(L(\mu, \sigma^2)) \text{ if and only if } \bar{x} = \mu.
\]

Similarly, note that

\[
\frac{\partial}{\partial (\sigma^2)} \ln(L(\mu, \sigma^2)) = -\frac{n}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 2\pi + \frac{1}{2\sigma^4} \cdot \sum_{j=1}^{n} (x_j - \mu)^2
\]
\[
= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \cdot \sum_{j=1}^{n} (x_j - \mu)^2.
\]

Then

\[
\frac{\partial}{\partial (\sigma^2)} \ln(L(\mu, \sigma^2)) = 0
\]

\[
\frac{1}{2\sigma^4} \cdot \sum_{j=1}^{n} (x_j - \mu)^2 = \frac{n}{2\sigma^2}
\]

\[
\frac{1}{n} \cdot \sum_{j=1}^{n} (x_j - \mu)^2 = \sigma^2.
\]
Hence,
\[ \frac{\partial}{\partial \mu} \ln(L(\mu, \sigma^2)) = \frac{\partial}{\partial (\sigma^2)} \ln(L(\mu, \sigma^2)) = 0 \]
at
\[ \left( \bar{x}, \frac{1}{n}, \sum_{j=1}^{n} (x_j - \bar{x})^2 \right) \]
which establishes \( \bar{X} \) as the maximum likelihood estimator of \( \mu \) and \( \frac{n-1}{n} \cdot S^2 \) as the maximum likelihood estimator of \( \sigma^2 \) where \( S^2 \) is the usual sample variance.

**Theorem 202.** If \( \hat{\Theta} \) is the maximum likelihood estimator of \( \theta \) and \( g \) is a continuous function, then \( g(\hat{\Theta}) \) is the maximum likelihood estimator of \( g(\theta) \).

**Remark.** Theorem 202 asserts particularly that, if \( \hat{\Theta} \) is the maximum likelihood estimator of \( \sigma^2 = \text{Var}(X) \), then \( \sqrt{\hat{\Theta}} \) is the maximum likelihood estimator of \( \sigma \).
Chapter 10

Interval Estimation

In Section 9, we were studying random variables that provide a particular value as an estimate for a population parameter. An interval estimation gives a range of values that the parameter will lie in with a particular likelihood.

**Definition 203.** A *p confidence interval* for a parameter $\theta$ is a range $(\hat{\theta}_1, \hat{\theta}_2)$ so that

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = p.$$ 

**Remark.** We sometimes speak of $\alpha$ where $p = 1 - \alpha$ when concerned with the “tail” component of the interval. The motivation for this is found in standard statistical tables where the values are given relative to the area underneath the tail end of the distribution.

10.1 Population Means

10.1.1 With Known Variance

Recall Theorem 170 which asserts that, for a normal population with mean $\mu$ and variance $\sigma^2$, the sample mean $\bar{X}$ according to a random sample of size $n$ is normally distributed with a mean of $\mu$ and variance $\sigma^2/n$. Correspondingly,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Let $z^* \in \mathbb{R}$ be so that

$$P(|Z| < z^*) = p$$

for a given $0 < p < 1$. Then, for the observed sample mean of $\bar{x}$, notice that

$$|z| < z^* \iff \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z^*$$

$$\iff |\bar{x} - \mu| < \frac{z^* \cdot \sigma}{\sqrt{n}}$$

$$\iff -\frac{z^* \cdot \sigma}{\sqrt{n}} < \bar{x} - \mu < \frac{z^* \cdot \sigma}{\sqrt{n}}$$

$$\iff \bar{x} \, - \frac{z^* \cdot \sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z^* \cdot \sigma}{\sqrt{n}}.$$ 

Then the interval

$$\left( \bar{x} - \frac{z^* \cdot \sigma}{\sqrt{n}}, \bar{x} + \frac{z^* \cdot \sigma}{\sqrt{n}} \right)$$

is a *p confidence interval* for $\mu$. 

157
Remark. For this computation to be useful, it is assumed that the population variance $\sigma^2$ is known.

Example 159. Given a sample of size 121 from a normal population with standard deviation $\sigma = 3.72$, find a 95% confidence interval for $\mu$ given that the sample mean is $\bar{x} = 17.3$.

Using the Central Limit Theorem, we can also use this procedure to build confidence intervals for the mean of a population which may not be normally distributed. This approach is only sensible when the sample size is large.

Example 160. The following data were collected.

\[
\begin{array}{cccccccccc}
34 & 58 & 75 & 34 & 59 & 87 & 61 \\
86 & 34 & 71 & 46 & 3 & 13 & 63 \\
63 & 90 & 71 & 9 & 3 & 28 & 34 \\
28 & 40 & 19 & 94 & 18 & 32 & 30 \\
18 & 33 & 46 & 92 & 18 & 32 & 30 \\
\end{array}
\]

Find a 99% confidence interval for the mean $\mu$ of the population (assuming the sample variance is the population variance).

First, note that $n = 35$, $\bar{x} \approx 46.086$, and $S \approx 26.354$. For a 99% confidence interval, we use $z = 2.575$. Hence, the confidence interval is

\[
46.086 - \frac{(2.575)(26.354)}{\sqrt{35}} \approx 34.615
\]

\[
to
\]

\[
46.086 + \frac{(2.575)(26.354)}{\sqrt{35}} \approx 57.557
\]

10.1.2 Using the Sample Variance

Alternatively, we can use the $t$-distribution to build confidence intervals recalling that

\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}}
\]

has a $t$-distribution with $n - 1$ degrees of freedom from Corollary 180. Let $t^*$ be so that

\[
P(|T| < t^*) = p
\]

and then notice that

\[
|t| < t^* \iff -t^* < \frac{\mu - \bar{x}}{S/\sqrt{n}} < t^*
\]

\[
\iff \bar{x} - \frac{t^* \cdot s}{\sqrt{n}} < \mu < \bar{x} + \frac{t^* \cdot s}{\sqrt{n}}.
\]

That is, our $p$ confidence interval for $\mu$ is given by

\[
\left(\bar{x} - \frac{t^* \cdot s}{\sqrt{n}}, \bar{x} + \frac{t^* \cdot s}{\sqrt{n}}\right).
\]

Example 161. Redo Example 160 using the $t$-distribution.
10.2 Differences between Means

Here, we will be concerned with finding confidence intervals for $\mu_1 - \mu_2$ where $\mu_1$ and $\mu_2$ are the population means from two independent populations. As when finding confidence intervals for a single population mean, our computations will be dependent upon whether we know the population variances. For large enough sample sizes, we can proceed under the assumption that the sample variance is the population variance. For smaller sample sizes, we can use the sample variances but only when it is reasonable to suppose that the two population variances are equal.

10.2.1 With Known Variances

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent. Take a sample of size $n_1$ from $X_1$ and a sample of size $n_2$ from $X_2$ and let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of those samples, respectively. Then

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

which we can use to construct confidence intervals for $\mu_1 - \mu_2$.

Example 162. We are to compare the average drying time for two inks, Cthulhu Black and Barney Purple. A sample of 30 ink tests of Cthulhu Black yielded a mean drying time of 5 seconds and a sample of 25 ink tests of Barney Purple yielded a mean drying time of 9 seconds. Previous research indicates that the variance in drying time for Cthulhu Black is $1.2$ seconds and for Barney Purple is $2.03$ seconds. Construct a $95\%$ confidence interval for the difference in average drying times.

10.2.2 With Equal but Unknown Variances

Suppose $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$ are independent. Take a sample of size $n_1$ from $X_1$ and a sample of size $n_2$ from $X_2$ and let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of those samples, respectively. Then

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1).$$

Now, to estimate $\sigma$, we use both sample variances in a particular way.

Definition 204. Suppose we have a sample of size $n_1$ from a normal population $X_1$ and a sample of size $n_2$ from a normal population $X_2$ where $X_1$ and $X_2$ are independent with equal variances. For the sample variances $S_1^2$ and $S_2^2$ taken from each of the two samples, we define the spooled estimator

$$S_p^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}$$

of $\sigma^2$.

Proposition 205. For the spooled estimator $S_p^2$ corresponding to two normal populations with equal variance $\sigma$,

$$E(S_p^2) = \sigma^2.$$

That is, the spooled estimator is an unbiased estimator of the population variance.

From Theorem 177 we gather that

$$\frac{n_1 - 1}{\sigma^2} \cdot S_1^2 \sim \chi(n_1 - 1) \quad \text{and} \quad \frac{n_2 - 1}{\sigma^2} \cdot S_2^2 \sim \chi(n_2 - 1).$$

Moreover, by their independence,

$$Y = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{\sigma^2} \sim \chi(n_1 + n_2 - 2)$$
by Theorem 175. It can be shown that

\[
Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

and \(Y\) are independent which allows us to appeal to Theorem 179 to see that

\[
\frac{Z}{\sqrt{Y/(n_1 + n_2 - 2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

has a \(t\)-distribution with \(n_1 + n_2 - 2\) degrees of freedom.

**Example 163.** PowerCharged makes two kinds of lithium batteries: the Standard and the Stamina. Research suggests that the variances in their battery lives are equal. A sample of 19 Standard batteries provides a mean charge life of 14.3 hours and a variance of 0.32 hours. A sample of 12 Stamina batteries provides a mean charge life of 17.2 hours and a variance of 0.47 hours. Find a 95% confidence interval for the difference in their mean charge lives.

**10.3 Proportions**

Recall that, for \(Y \sim \text{Bernoulli}(p)\) and a random sample \(X_1, X_2, \ldots, X_n\) from \(Y\),

\[
X = X_1 + X_2 + \cdots + X_n \sim B(n, p).
\]

Also, recall Theorem 158 which states that, for large enough \(n\),

\[
\frac{X - np}{\sqrt{np(1 - p)}}
\]

has approximately the standard normal distribution. Then, we would use a value \(z^*\) to construct a confidence interval:

\[
\left| \frac{x - np}{\sqrt{np(1 - p)}} \right| < z^* \iff -z^* < \frac{np - x}{\sqrt{np(1 - p)}} < z^*
\]

\[
\iff x - z^* \cdot \sqrt{np(1 - p)} < np < x + z^* \cdot \sqrt{np(1 - p)}
\]

\[
\iff \frac{x}{n} - z^* \cdot \sqrt{p(1 - p)} < p < \frac{x}{n} + z^* \cdot \sqrt{p(1 - p)}
\]

For large enough sample sizes, we can use the estimate \(\hat{p} = \frac{x}{n}\) for \(p\) which gives us the confidence interval

\[
\left( \hat{p} - z^* \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z^* \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)
\]

for \(p\).

**Example 164.** Zeno’s Archer strikes a bulls-eye 312 times out of 500 attempts. Assuming that each attempt is independent and a Bernoulli variable, find a 99% confidence interval for the probability that she lands a bulls-eye.

**Example 165.** Recalling the discussion in Section 5.2.3, use a beta distribution to evaluate the interval found in Example 164.
10.4 Differences between Proportions

Now, consider $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$ which are independent. As long as $n_1$ and $n_2$ are large enough, $X_j$ is approximately normally distributed with mean $n_j p_j$ and variance $n_j p_j (1 - p_j)$ where $j = 1, 2$. It follows that

$$(-1)^{j+1} \frac{X_j}{n_j}$$

is approximately normally distributed with mean $(-1)^{j+1} \cdot p_j$ and variance $\frac{p_j (1 - p_j)}{n_j}$. Finally, we obtain that, approximately,

$$X_1 \sim N \left( p_1 - p_2, \frac{p_1 (1 - p_1)}{n_1} + \frac{p_2 (1 - p_2)}{n_2} \right)$$

which is equivalent to saying that

$$\left(\frac{X_1}{n_1} - \frac{X_2}{n_2} \right) - (p_1 - p_2) \sqrt{\frac{p_1 (1 - p_1)}{n_1} + \frac{p_2 (1 - p_2)}{n_2}} \sim N(0, 1).$$

Like before, to build a confidence interval, we let $\hat{p}_j = \frac{X_j}{n_j}$ serve as an estimator for $p_j$ find a suitable $z^*$ to establish our confidence interval as

$$\hat{p}_1 - \hat{p}_2 - z^* \cdot \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z^* \cdot \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}$$

Example 166. Zeno argues that his friend has a higher accuracy at hitting bulls-eyes than Artemis. Artemis hits 226 bulls-eyes of 310 attempts. Using Zeno’s friend’s 312 bulls-eyes out of 500 from Example 164, build a 95% confidence interval for the difference between their proportions.

10.5 Population Variances

For a random sample of size $n$ from a random variable with variance $\sigma^2$, recall that that

$$\frac{n - 1}{\sigma^2} \cdot S^2 \sim \chi(n - 1)$$

by Theorem 177 where $S^2$ is the sample variance. Then, for a given $p$, we can find $\alpha$ and $\beta$ so that

$$P \left[ \frac{n - 1}{\sigma^2} \cdot S^2 < \alpha \right] = P \left[ \beta < \frac{n - 1}{\sigma^2} \cdot S^2 \right] = \frac{1 - p}{2}.$$ 

It follows that

$$P \left[ \frac{n - 1}{\sigma^2} \cdot S^2 < \alpha \text{ or } \beta < \frac{n - 1}{\sigma^2} \cdot S^2 \right] = \frac{1 - p}{2} + \frac{1 - p}{2} = 1 - p.$$ 

Hence,

$$P \left[ \alpha < \frac{n - 1}{\sigma^2} \cdot S^2 < \beta \right] = p$$

which results in

$$P \left[ \frac{\alpha}{(n - 1) \cdot S^2} < \frac{1}{\sigma^2} < \frac{\beta}{(n - 1) \cdot S^2} \right] = p.$$ 

By taking reciprocals,

$$P \left[ (n - 1) \cdot S^2 < \sigma^2 < \frac{(n - 1) \cdot S^2}{\alpha} \right] = p$$

which demonstrates the corresponding $p$ confidence interval.

Example 167. In the context of Example 163, use the sample of 19 Standard batteries to construct a 95% confidence interval for the population variance.
10.6 Ratios of Variances

Let $S_j$ be the sample variance from a random sample of size $n_j$ of a normal variable with variance $\sigma^2_j$ for $j = 1, 2$. Also, assume that the two normal variables are independent. From Theorem 183, observe that

$$\frac{\sigma^2_2 \cdot S_1^2}{\sigma^2_1 \cdot S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

and

$$\frac{\sigma^2_1 \cdot S_2^2}{\sigma^2_2 \cdot S_1^2} \sim F(n_2 - 1, n_1 - 1).$$

Then we can build a confidence interval for $\frac{\sigma^2_1}{\sigma^2_2}$ in the following way. Find $\alpha$ and $\beta$ so that

$$P\left[\frac{\sigma^2_2 \cdot S_2^2}{\sigma^2_1 \cdot S_1^2} > \alpha\right] = P\left[\frac{\sigma^2_1 \cdot S_2^2}{\sigma^2_2 \cdot S_1^2} > \beta\right] = \frac{1 - p}{2}.$$

Then

$$P\left[\frac{\sigma^2_1}{\sigma^2_2} > \frac{\alpha \cdot S_1^2}{S_2^2}\right] = P\left[\frac{S_2^2}{\beta \cdot S_1^2} > \frac{S_1^2}{\sigma^2_2}\right] = \frac{1 - p}{2}$$

which provides

$$P\left[\frac{\sigma^2_1}{\sigma^2_2} < \frac{S_1^2}{\beta \cdot S_2^2}\right] \text{ or } \frac{\alpha \cdot S_1^2}{S_2^2} < \frac{S_1^2}{\sigma^2_2}\right] = \frac{1 - p}{2} + \frac{1 - p}{2} = 1 - p.$$

Finally, we have

$$P\left[\frac{S_1^2}{\beta \cdot S_2^2} < \frac{\sigma^2_1}{\sigma^2_2} < \frac{\alpha \cdot S_1^2}{S_2^2}\right] = p$$

which displays the relevant confidence interval.

**Example 168.** In the context of Example 163 but relaxing the assumption that Standard and Stamina battery life distributions have equal variance, build a 95% confidence interval for the ratio of the variances. Use this interval to evaluate the claim that they have equal variance.
Chapter 11

Hypothesis Testing

Hypothesis testing is used to evaluate the relative strength of claims concerning populations with unknown parameters.

**Definition 206.** Formally, a **statistical hypothesis** is a claim about a population’s distribution and the values of relevant parameters. A statistical hypothesis is said to be **simple** if it uniquely determines the population’s distribution. Any statistical hypothesis which is not simple is called **composite**.

**Example 169.** Suppose $X$ as an exponential distribution with parameter $\lambda$. Then the hypothesis $\lambda = 3$ is a simple hypothesis.

**Example 170.** Suppose $X$ has a normal distribution with unknown mean $\mu$ and variance $\sigma^2$. Then the hypothesis $\mu = 2.76$ is a composite hypothesis since the variance $\sigma^2$ is not specified.

**Example 171.** Suppose $X$ has a normal distribution with unknown mean $\mu$ and a known variance of $\sigma^2 = 3.14$. Then the hypothesis $\mu > 0$ is a composite hypothesis since it doesn’t particularly specify $\mu$.

The basic elements of a hypothesis test are
- a **null hypothesis**, $H_0$
- an **alternative hypothesis**, $H_a$
- a test statistic $\hat{\Theta}$
- a rejection region

The null hypothesis will generally be a claim about a parameter $\theta$ and the test statistic $\hat{\Theta}$ will be used as an estimator for $\theta$. The rejection region is the range of values which will determine whether we reject or fail to reject the null hypothesis. In particular, if the rejection region is $R \subseteq \mathbb{R}$ and we calculate the value of the sample statistic $\hat{\theta} \in R$, then we reject the null hypothesis.

When doing hypothesis tests, we distinguish two kinds of errors.
- A **Type I** error occurs when the null hypothesis is rejected even though it is true.
- A **Type II** error occurs when we fail to reject the null hypothesis even though it is false.

In the context of a hypothesis test, we will denote the probability of a Type I error by $\alpha$ and the probability of a Type II error by $\beta$.

**Definition 207.** For a hypothesis test, the rejection region is also known as the **critical region**. Given a particular null hypothesis and a critical region, the probability $\alpha$ of committing a Type I error is also known as the **level of significance**.

**Example 172.** Archaeologists unearth a mysterious coin. After tossing it a few times, the archaeologists suspect that it lands on tails more than half of the time. To test this suspicion, the CoinFlippers have constructed the following hypothesis test. The null hypothesis is that $p = 0.5$. They will toss the coin 25 times and reject the null hypothesis if they get more than 20 tails.
(a) Find the probability $\alpha$ of a Type I error.

(b) Using the alternative hypothesis that $p = 0.6$, find the probability $\beta$ of a Type II error.

(c) Using the alternative hypothesis that $p = 0.9$, find the probability $\beta$ of a Type II error.

Though we can modify the rejection region to decrease the probability of a Type I error, such a change will increase the probability of a Type II error, and vice versa.

**Example 173.** An ink retailer claims that Cthulhu Black has an average drying time of 9 seconds. The manufacturer argues that the average drying time is shorter. From a sample of 30, a mean drying time of 5 seconds with a sample variance of 1.39 seconds were obtained. Assuming that the population variance is 1.39 and using a normal distribution, find a rejection region that guarantees $\alpha = 0.05$. Then use this to evaluate the retailer’s claim.

### 11.1 The Neyman-Pearson Lemma

As we commented before, we can’t choose a rejection region which will simultaneously minimize $\alpha$ and $\beta$. Though, given a fixed $\alpha$, we may be able to choose a rejection region that relatively minimizes $\beta$.

**Definition 208.** When testing the simple hypothesis $H_0 : \theta = \theta_0$ against the simple hypothesis $H_1 : \theta = \theta_1$, the value $1 - \beta$ is referred to as the **power** of the test at $\theta = \theta_1$. A critical region for testing $H_0$ against $H_1$ is said to be the **most powerful** critical region if the power of the test is maximal at $\theta = \theta_1$.

To find the most powerful critical region, we will be using the likelihood function $L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$.

**Theorem 209** (Neyman-Pearson Lemma). Suppose $R$ is a critical region with level of significance $\alpha$ relative to a hypothesis test with two simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. If $k$ is a constant so that

$$\frac{L(\theta_0)}{L(\theta_1)} \leq k \quad \text{inside } R$$

and

$$\frac{L(\theta_0)}{L(\theta_1)} \geq k \quad \text{outside } R,$$

then $R$ is the most powerful critical region with level of significance $\alpha$ for testing $H_0$ against $H_1$.

**Example 174.** A random sample of size $n$ from a normal population with $\sigma^2 = 1$ is used to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ where $\mu_1 > \mu_0$. Use the Neyman-Pearson Lemma to find the most powerful critical region with level of significance $\alpha$.

First, note that the likelihood function is

$$L(\mu) = \prod_{j=1}^{n} \exp\left(\frac{-1}{2\sqrt{2\pi}} (x_j - \mu)^2\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \exp\left(\frac{1}{2} \cdot \sum_{j=1}^{n} (x_j - \mu)^2\right).$$

Hence,

$$\frac{L(\mu_0)}{L(\mu_1)} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \exp\left(\frac{-1}{2} \cdot \sum_{j=1}^{n} (x_j - \mu_0)^2\right) = \exp\left[\frac{1}{2} \cdot \left(\sum_{j=1}^{n} (x_j - \mu_1)^2\right) - \left(\sum_{j=1}^{n} (x_j - \mu_0)^2\right)\right].$$
We are searching for a region $R$ and $k$ so that \( \frac{L(\mu_0)}{L(\mu_1)} \leq k \) inside $R$ and \( \frac{L(\mu_0)}{L(\mu_1)} \geq k \) otherwise. Taking logarithms, we have

\[
\frac{1}{2} \cdot \left( \sum_{j=1}^{n} (x_j - \mu_1)^2 - \sum_{j=1}^{n} (x_j - \mu_0)^2 \right) \leq \ln(k)
\]

\[
\sum_{j=1}^{n} x_j^2 - 2x_j \mu_1 + \mu_1^2 - \sum_{j=1}^{n} x_j^2 - 2x_j \mu_0 + \mu_0^2 \leq 2 \cdot \ln(k)
\]

\[
2(\mu_0 - \mu_1) \cdot \sum_{j=1}^{n} x_j + n(\mu_1^2 - \mu_0^2) \leq 2 \cdot \ln(k)
\]

\[
(\mu_0 - \mu_1) \cdot \sum_{j=1}^{n} x_j \leq \frac{2 \cdot \ln(k) - n \cdot (\mu_1^2 - \mu_0^2)}{2}
\]

\[
\sum_{j=1}^{n} x_j \geq \frac{2 \cdot \ln(k) - n \cdot (\mu_1^2 - \mu_0^2)}{2 \cdot (\mu_0 - \mu_1)}
\]

\[
\bar{x} \geq \frac{2 \cdot \ln(k) - n \cdot (\mu_1^2 - \mu_0^2)}{2n \cdot (\mu_0 - \mu_1)}
\]

Applying the same to the inequality \( \frac{L(\mu_0)}{L(\mu_1)} \geq k \) results in

\[
\bar{x} \leq \frac{2 \cdot \ln(k) - n \cdot (\mu_1^2 - \mu_0^2)}{2n \cdot (\mu_0 - \mu_1)}
\]

Let

\[
k^* = \frac{2 \cdot \ln(k) - n \cdot (\mu_1^2 - \mu_0^2)}{2n \cdot (\mu_0 - \mu_1)}
\]

and then the rejection region is $\bar{x} > k^*$.

To find the value of $k^*$, we use the fact that

\[
Z = \frac{\bar{X} - \mu}{1/\sqrt{n}} \sim N(0, 1)
\]

where $\mu$ is the true population mean. Now, making a Type I error occurs when $\bar{x} > k^*$ and $\mu = \mu_0$. Particularly,

\[
P[\bar{X} > k^* \text{ and } \mu = \mu_0] = P \left[ \frac{\bar{X} - \mu_0}{1/\sqrt{n}} > \frac{k^* - \mu_0}{1/\sqrt{n}} \right] = P \left[ Z > \frac{k^* - \mu_0}{1/\sqrt{n}} \right].
\]

Let $z^*$ be so that

\[
P(Z > z^*) = \alpha
\]

and then notice that

\[
k^* = \frac{z^*}{\sqrt{n}} + \mu_0.
\]

This establishes the most powerful critical region with level of significance $\alpha$ to be $\bar{x} > \frac{z^*}{\sqrt{n}} + \mu_0$.

**Remark.** There are two important things to note in Example 174.

1. In the process of applying the Neyman-Pearson Lemma, the test statistic to be used for the hypothesis test emerged naturally.
2. The critical region found in the end actually does not depend on the alternative hypothesis at all. In fact, the critical region found can be used to test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ since the most powerful critical region will be the same for any simple alternative hypothesis $H_a : \mu = \mu_a$ where $\mu_a > \mu_0$.

Example 175. A random sample of size 15 from an exponential population with is used to test $H_0 : \lambda = 1$ against $H_1 : \lambda = 1.5$. Use the Neyman-Pearson Lemma to find the most powerful critical region with level of significance $\alpha = 0.05$. Then, find the power of the test at $\lambda = 1.5$.

First, note that the likelihood function is

$$L(\lambda) = \prod_{j=1}^{n} \lambda e^{-\lambda x_j} = \lambda^n \cdot \exp\left[-\lambda \sum_{j=1}^{n} x_j\right]$$

Then consider

$$\frac{L(1)}{L(1.5)} \leq k$$

$$\frac{\exp\left[-\sum_{j=1}^{15} x_j\right]}{1.5^{15} \cdot \exp\left[-1.5 \cdot \sum_{j=1}^{15} x_j\right]} \leq k$$

$$\left(\frac{1}{1.5}\right)^{15} \cdot \exp\left[0.5 \cdot \sum_{j=1}^{15} x_j\right] \leq k$$

$$15 \cdot \ln\left(\frac{2}{3}\right) + \frac{1}{2} \sum_{j=1}^{n} x_j \leq \ln(k)$$

$$\bar{x} \leq \frac{2 \cdot \ln(k) - 30 \cdot \ln(2/3)}{15}.$$

Applying the same to the inequality $L(1)/L(1.5) \geq k$ yields

$$\bar{x} \geq \frac{2 \cdot \ln(k) - 30 \cdot \ln(2/3)}{15}$$

so we let

$$k^* = \frac{2 \cdot \ln(k) - 30 \cdot \ln(2/3)}{15}.$$

The desired critical region is $\bar{x} < k^*$.

Now, to find the value of $k^*$, recall that

$$\sum_{j=1}^{15} X_j \sim \Gamma(15, \lambda)$$

as a consequence of Theorem 144. Hence, the probability density for $Y = \sum_{j=1}^{15} X_j = 15 \cdot \bar{X}$ is given by

$$f(y; \lambda) = \frac{\lambda^{15} \cdot y^{14} \cdot e^{-\lambda y}}{\Gamma(15)} = \frac{\lambda^{15} \cdot y^{14} \cdot e^{-\lambda y}}{14!}$$

for $y > 0$. To establish the level of significance $\alpha = 0.05$, need

$$P[\bar{X} < k^* \text{ and } \lambda = 1] = P[Y < 15 \cdot k^* \text{ and } \lambda = 1] = 0.05.$$
11.2. LIKELIHOOD RATIO TESTS

for \( t \). Referring to a gamma table, we get an approximate solution: \( t \approx 9.246 \). That is,

\[
0.05 \approx P \{ Y < 9.246 \text{ and } \lambda = 1 \} = P \left[ \bar{X} < \frac{9.246}{15} \text{ and } \lambda = 1 \right].
\]

Hence, the most powerful critical region is given by \( \bar{X} < \frac{9.246}{15} = 0.6164 \).

Notice that

\[
\beta = P \{ \bar{X} > 0.6164 \text{ and } \lambda = 1.5 \} = P \{ Y > 9.246 \text{ and } \lambda = 1.5 \} = \int_{9.246}^{\infty} \frac{1.5^{1.5} \cdot y^{1.5} \cdot e^{-1.5y}}{14!} \, dy \approx 0.584318.
\]

Therefore, the power of the test at \( \lambda = 1.5 \) is

\[ 1 - 0.584318 = 0.415682. \]

**Remark.** After determining that the test statistic to be used in Example 175 is the sample mean \( \bar{X} \), we could have used the Central Limit Theorem to finish the calculation as follows. Note that

\[
Z_\lambda = \frac{\bar{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{\lambda}}} = \sqrt{15} \cdot (\lambda \cdot \bar{X} - 1)
\]

under the Central Limit Theorem has approximately the standard normal distribution. To get the level of significance to be \( \alpha = 0.05 \) relative to \( H_0 : \lambda = 1 \), we can refer to a \( z \)-table to produce \( z = -1.645 \) which implies that

\[
\sqrt{15} \cdot (\bar{X} - 1) < -1.645 \implies \bar{X} < 0.575263.
\]

Still relative to the Central Limit Theorem, we would compute the power of this test at \( \lambda = 1.5 \) to be

\[
1 - P[\bar{X} > 0.575263 \text{ and } \lambda = 1.5] = 1 - P[Z_{1.5} > -0.531007] = 1 - 0.2981 = 0.7019,
\]

quite a bit higher than our original calculation.

11.2 Likelihood Ratio Tests

Likelihood ratio tests employ a strategy similar to that of the Neyman-Pearson Lemma to find powerful critical regions but allowing for composite hypotheses and multiple unknown parameters. Suppose we have a random sample of size \( n \) with values \( x_1, x_2, \ldots, x_n \) from a population with parameters \( \theta_1, \theta_2, \ldots, \theta_n \). For the sake of simplicity, we will consider only the case that we are testing hypotheses regarding \( \theta_1 \). Let \( \Omega_0 \) and \( \Omega_1 \) be disjoint sets where we are testing \( H_0 : \theta_1 \in \Omega_0 \text{ against } H_1 : \theta_1 \in \Omega_1 \). Then let

\[
\Omega_0^* = \{ (\theta_1, \theta_2, \ldots, \theta_n) : \theta_1 \in \Omega_0 \},
\]

\[
\Omega_1^* = \{ (\theta_1, \theta_2, \ldots, \theta_n) : \theta_1 \in \Omega_1 \},
\]

and \( \Omega = \Omega_0^* \cup \Omega_1^* \). Also, let \( (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{n,0}) \) be so that \( L(\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{n,0}) \) is the maximum of \( L(\theta_1, \theta_2, \ldots, \theta_n) \) on \( \Omega_0^* \), and \( (\theta_{1,1}, \theta_{2,1}, \ldots, \theta_{n,1}) \) be so that \( L(\theta_{1,1}, \theta_{2,1}, \ldots, \theta_{n,1}) \) is the maximum of \( L(\theta_1, \theta_2, \ldots, \theta_n) \) on \( \Omega \). Then the corresponding **likelihood ratio statistic** is

\[
\lambda = \frac{L(\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{n,0})}{L(\theta_{1,1}, \theta_{2,1}, \ldots, \theta_{n,1})}
\]

and the **likelihood ratio test** for testing \( H_0 : \theta_1 \in \Omega_0 \text{ against } H_1 : \theta_1 \in \Omega_1 \) uses the value \( \lambda \) of the statistic \( \Lambda \) where the rejection region is determined by \( \lambda \leq k \).
By our definitions of \((\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{n,0})\) and \((\theta_{1,1}, \theta_{2,1}, \ldots, \theta_{n,1})\), we know that \(0 \leq \lambda \leq 1\). Then, the constant \(k\) determining the rejection region will only assume values \(0 < k < 1\). The reason the rejection region is defined by \(\lambda \leq k\) is that we are capturing the relative likelihood that \(\theta_1\) is outside of the null hypothesis region \(\Omega_0\).

**Example 176.** Use the likelihood ratio test to test \(H_0 : \mu = \mu_0\) against \(H_1 : \mu \neq \mu_0\) relative to a random sample of size \(n\) from a normal population with known variance \(\sigma^2 = 1\).

Since the variance is known to be \(\sigma^2 = 1\), the likelihood function for the mean is

\[
L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \exp\left[-\frac{1}{2} \cdot \sum_{j=1}^{n} (x_j - \mu)^2\right]
\]

The null hypothesis region is just the singleton set \(\{\mu_0\}\) and the alternative region is \(\{\mu \in \mathbb{R} : \mu \neq \mu_0\}\).

Now, notice that \(\hat{\mu}_0 = \mu_0\) and \(\hat{\mu} = \bar{x}\) by Example 158. Hence,

\[
\lambda = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \exp\left[-\frac{1}{2} \cdot \sum_{j=1}^{n} (x_j - \bar{x})^2\right]
\]

\[
= \exp\left[-\frac{1}{2} \cdot \sum_{j=1}^{n} (x_j - \mu_0)^2 - (x_j - \bar{x})^2\right]
\]

\[
= \exp\left[-\frac{1}{2} \cdot \sum_{j=1}^{n} [x_j^2 - 2x_j\mu_0 + \mu_0^2 - x_j^2 + 2x_j\bar{x} - \bar{x}^2]\right]
\]

\[
= \exp\left[-\frac{1}{2} \cdot (-2n\bar{x}\mu_0 + n\mu_0^2 + 2n\bar{x} - n\bar{x})\right]
\]

\[
= \exp\left[-\frac{1}{2} \cdot (n\mu_0^2 - 2n\bar{x}\mu_0 + n\bar{x}^2)\right]
\]

\[
= \exp\left[-\frac{n}{2} \cdot (\mu_0 - \bar{x})^2\right].
\]

Now, for \(0 < k < 1\),

\[
\lambda \leq k
\]

\[
-\frac{n}{2} \cdot (\mu_0 - \bar{x})^2 \leq \ln(k)
\]

\[
(\mu_0 - \bar{x})^2 \geq \frac{-2\ln(k)}{n}
\]

\[
|\bar{x} - \mu_0| \geq \sqrt{-\frac{2\ln(k)}{n}}
\]

noting that \(\ln(k) < 0\). Let

\[
k^* = \sqrt{-\frac{2\ln(k)}{n}}.
\]

To find the critical region for the level of significance \(\alpha\), we want

\[
\alpha = P[|\bar{X} - \mu_0| \geq k^* \text{ and } \mu = \mu_0].
\]

Under the assumption \(\mu = \mu_0\), \(\bar{X} \sim N(\mu_0, 1/n)\). Let \(z^* > 0\) be so that

\[
P \left[\frac{\bar{X} - \mu_0}{1/\sqrt{n}} \leq -z^*\right] = \frac{\alpha}{2}
\]
and notice that
\[
P \left( \frac{\bar{X} - \mu_0}{1/\sqrt{n}} \geq z^* \right) = \frac{\alpha}{2}.
\]

Simplifying,
\[
P \left( \bar{X} - \mu_0 \leq -\frac{z^*}{\sqrt{n}} \right) = \frac{\alpha}{2} \quad \text{and} \quad P \left( \bar{X} - \mu_0 \geq \frac{z^*}{\sqrt{n}} \right) = \frac{\alpha}{2} \implies P \left( |\bar{X} - \mu_0| \geq \frac{z^*}{\sqrt{n}} \right) = \alpha.
\]

Therefore, we reject the null hypothesis if
\[
|\bar{x} - \mu_0| \geq \frac{z^*}{\sqrt{n}}.
\]

**Example 177.** Use the likelihood ratio test to test \( H_0 : \mu = \mu_0 \) against \( H_1 : \mu > \mu_0 \) relative to a random sample of size \( n \) from a normal population with unknown variance \( \sigma^2 \).

The likelihood function here is
\[
L(\mu, \sigma^2) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2 \right)
\]

The null hypothesis region is just the singleton set \( \{\mu_0\} \) and the alternative region is
\[
\{ \mu \in \mathbb{R} : \mu > \mu_0 \}.
\]

Hence,
\[
\Omega_0^* = \{(\mu, \sigma^2) : \mu = \mu_0\}, \quad \Omega_1^* = \{(\mu, \sigma^2) : \mu > \mu_0\},
\]

and
\[
\Omega = \{(\mu, \sigma^2) : \mu \geq \mu_0\}.
\]

First, we find the maximum value of the likelihood function on \( \Omega_0^* \). That is, we need to find \( \sigma^2 \) which maximizes
\[
L(\mu_0, \sigma^2) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu_0)^2 \right).
\]

Equivalently, we can maximize
\[
\ln[L(\mu_0, \sigma^2)] = \frac{-n}{2} \ln[2\pi\sigma^2] - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu_0)^2.
\]

By referring to Example 158, we find that
\[
\frac{\partial}{\partial \sigma^2} \ln[L(\mu_0, \sigma^2)] = 0
\]

for
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu_0)^2.
\]

So the desired ordered pair here is
\[
(\mu_0, \hat{\sigma}_0^2)
\]

where
\[
\hat{\sigma}_0^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu_0)^2.
\]
Now, we need to maximize the likelihood function over \( \Omega \). As before, we maximize
\[
\ln[L(\mu, \sigma^2)] = n \cdot \ln \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \right] - \frac{1}{2\sigma^2} \cdot \sum_{j=1}^{n} (x_j - \mu)^2.
\]

Note that
\[
\frac{\partial}{\partial \mu} \ln[L(\mu, \sigma^2)] = \frac{1}{2\sigma^2} \cdot \sum_{j=1}^{n} 2(x_j - \mu) = \frac{n \cdot (\bar{x} - \mu)}{\sigma^2}
\]
which equals zero only when \( \bar{x} = \mu \). In a similar way to above,
\[
\frac{\partial}{\partial \sigma^2} \ln[L(\mu, \sigma^2)] = 0
\]
when
\[
\sigma^2 = \frac{1}{n} \cdot \sum_{j=1}^{n} (x_j - \mu)^2.
\]

Since we are maximizing over the region \( \mu \geq \mu_0 \), we let
\[
\hat{\mu} = \begin{cases} \bar{x}, & \bar{x} > \mu_0, \\ \mu_0, & \bar{x} \leq \mu_0. \end{cases}
\]

Then the maximum over \( \Omega \) is at the pair
\[
(\hat{\mu}, \hat{\sigma}^2).
\]
where
\[
\hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{j=1}^{n} (x_j - \hat{\mu})^2.
\]

Observe that
\[
L(\mu_0, \sigma_0^2) = \left( \frac{1}{\sigma_0 \cdot \sqrt{2\pi}} \right)^n \cdot \exp \left[ -\frac{1}{2\sigma_0^2} \cdot \sum_{j=1}^{n} (x_j - \mu_0)^2 \right] = \left( \frac{1}{\sigma_0 \cdot \sqrt{2\pi}} \right)^n \cdot \exp \left( -\frac{n}{2} \right)
\]
and, similarly, that
\[
L(\hat{\mu}, \hat{\sigma}^2) = \left( \frac{1}{\hat{\sigma} \cdot \sqrt{2\pi}} \right)^n \cdot \exp \left( -\frac{n}{2} \right).
\]

Moreover,
\[
\frac{L(\mu_0, \sigma_0^2)}{L(\hat{\mu}, \hat{\sigma}^2)} = \frac{\left( \frac{1}{\sigma_0 \cdot \sqrt{2\pi}} \right)^n \cdot \exp \left( -\frac{n}{2} \right)}{\left( \frac{1}{\hat{\sigma} \cdot \sqrt{2\pi}} \right)^n \cdot \exp \left( -\frac{n}{2} \right)} = \left( \frac{\hat{\sigma}}{\sigma_0} \right)^n = \left( \frac{\sum_{j=1}^{n} (x_j - \hat{\mu})^2}{\sum_{j=1}^{n} (x_j - \mu_0)^2} \right)^{n/2}.
\]

Now,
\[
\lambda = \frac{L(\mu_0, \sigma_0^2)}{L(\hat{\mu}, \hat{\sigma}^2)} = \begin{cases} \left( \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \mu_0)^2} \right)^{n/2}, & \bar{x} > \mu_0, \\ 1, & \bar{x} \leq \mu_0. \end{cases}
\]
11.2. LIKELIHOOD RATIO TESTS

For $0 < k < 1$, notice that

\[
\left( \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \mu_0)^2} \right)^{n/2} \leq k
\]

since it is assumed that $\bar{x} > \mu_0$.

Assuming that $E(X) = \mu_0$, recall that

\[
\frac{X - \mu_0}{S/\sqrt{n}}
\]

has a $t$-distribution with $n - 1$ degrees of freedom by Corollary 180. Hence, let $t^*$ be so that

\[
P \left[ t^* \leq \frac{X - \mu_0}{S/\sqrt{n}} \right] = \alpha
\]

which establishes the critical region with level of significance $\alpha$ to be determined by

\[
t^* \leq \frac{\bar{x} - \mu_0}{S/\sqrt{n}}.
\]

**Theorem 210** (Wilks’ Theorem). Suppose $\theta_1, \theta_2, \ldots, \theta_m$ are the unknown parameters of a particular distribution. Up to reordering the parameters, suppose the null hypothesis specifies the first $k$ parameters; i.e.

\[H_0 : \theta_1 = \theta_{1,0}, \theta_2 = \theta_{2,0}, \ldots, \theta_k = \theta_{k,0}.\]
Suppose the alternative hypothesis is

\[ H_1 : \theta_1 \neq \theta_{1,0}, \theta_2 \neq \theta_{2,0}, \ldots, \theta_{\ell} \neq \theta_{\ell,0} \]

for some \( \ell \leq k \). Then, for the likelihood ratio statistic \( \Lambda \), the distribution of \(-2 \ln(\Lambda)\) approaches a chi-square distribution with \( \ell \) degrees of freedom as the sample size \( n \) goes to infinity.

In a likelihood ratio test, the rejection region is determined by \( \lambda \leq k \) which is equivalent to \(-2 \ln(\lambda) \geq -2 \ln(k)\). That is, we can use a chi-squared table to find the desired critical region.

**Example 178.** Suppose \( X \sim \text{Pois}(\theta_1) \) and \( Y \sim \text{Pois}(\theta_2) \) are independent. We take a random sample of size \( n_1 = 100 \) from \( X \) and a random sample of size \( n_2 = 150 \) from \( Y \). The sample from \( X \) yields \( \bar{x} = 21 \) and the sample from \( Y \) yields \( \bar{y} = 24 \). Use the likelihood ratio test to test \( H_0 : \theta_1 = \theta_2 \) against \( H_1 : \theta_1 \neq \theta_2 \) with \( \alpha = 0.01 \).

By independence, notice that the likelihood for both \( \theta_1 \) and \( \theta_2 \) is given by

\[
L(\theta_1, \theta_2) = \left( \prod_{j=1}^{100} e^{-\theta_1} \cdot \frac{\theta_1^{x_j}}{x_j!} \right) \cdot \left( \prod_{j=1}^{150} e^{-\theta_2} \cdot \frac{\theta_2^{y_j}}{y_j!} \right)
\]

\[ = e^{-100\theta_1+150\theta_2} \cdot \frac{\theta_1^{\sum_{j=1}^{100} x_j} \cdot \theta_2^{\sum_{j=1}^{150} y_j}}{x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!} \]

\[ = e^{-100\theta_1+150\theta_2} \cdot \frac{\theta_1^{\bar{x}} \cdot \theta_2^{\bar{y}}}{x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!} \]

If \( \theta_1 = \theta_2 = \theta \), then

\[
L(\theta, \theta) = e^{-250\theta} \cdot \frac{\theta^{100 \bar{x} + 150 \bar{y}}}{x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!}.
\]

To maximize, we look at

\[
\ln [L(\theta, \theta)] = -250\theta + (100 \cdot \bar{x} + 150 \cdot \bar{y}) \cdot \ln(\theta) - \ln(x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!)
\]

which provides

\[
\frac{\partial}{\partial \theta} L(\theta, \theta) = -250 + \frac{100 \cdot \bar{x} + 150 \cdot \bar{y}}{\theta}.
\]

Then,

\[
\frac{\partial}{\partial \theta} L(\theta, \theta) = 0
\]

whenever

\[
\theta = \frac{100 \cdot \bar{x} + 150 \cdot \bar{y}}{250}.
\]

More generally, notice that

\[
\ln(L(\theta_1, \theta_2)) = -(100\theta_1 + 150\theta_2) + 100 \cdot \bar{x} \cdot \ln(\theta_1) + 150 \cdot \bar{y} \cdot \ln(\theta_2) - \ln(x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!)
\]

which provides

\[
\frac{\partial}{\partial \theta_1} \ln(L(\theta_1, \theta_2)) = -100 + \frac{100 \cdot \bar{x}}{\theta_1}
\]

and

\[
\frac{\partial}{\partial \theta_2} \ln(L(\theta_1, \theta_2)) = -150 + \frac{150 \cdot \bar{y}}{\theta_2}.
\]

Then we see that

\[
\frac{\partial}{\partial \theta_1} \ln(L(\theta_1, \theta_2)) = \frac{\partial}{\partial \theta_2} \ln(L(\theta_1, \theta_2)) = 0
\]

for \( \theta_1 = \bar{x} \) and \( \theta_2 = \bar{y} \).
Observe that
\[
\lambda = \frac{L \left( \frac{100 \bar{x} + 150 \bar{y}}{250}, \frac{100 \bar{x} + 150 \bar{y}}{250} \right)}{L(\bar{x}, \bar{y})}
\]
\[
= \frac{e^{-\left(100 \bar{x} + 150 \bar{y}\right)} \cdot \left(\frac{100 \bar{x} + 150 \bar{y}}{250}\right)^{100 \bar{x} + 150 \bar{y}} \cdot x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!}{e^{-(100 \bar{x} + 150 \bar{y})} \cdot \bar{x}^{100} \cdot \bar{y}^{150} \cdot \bar{x}!x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!} \cdot \bar{x}^{100} \cdot \bar{y}^{150} \cdot \bar{x}!x_1!x_2! \cdots x_{100}!y_1!y_2! \cdots y_{150}!}
\]

which provides
\[
-2 \cdot \ln(\lambda) \approx 23.911707
\]
by the fact that \( \bar{x} = 21 \) and \( \bar{y} = 24 \).

Notice that we can rewrite the hypotheses as \( H_0 : \theta_1 = \theta, \theta_2 = \theta \) and \( H_1 : \theta_1 \neq \theta, \theta_2 = \theta \) which allows use to invoke Wilks' Theorem to assert that \(-2 \ln(\Lambda)\) approximately has a chi-square distribution with 1 degree of freedom. Referring to a chi-square table, we see that
\[
P[-2 \cdot \ln(\lambda) \geq 6.635] \approx 0.01.
\]
This determines the critical region at the \( \alpha = 0.01 \) level of significance to be \(-2 \cdot \ln(\lambda) \geq 6.635\). Since we computed \(-2 \cdot \ln(\lambda) \approx 23.911707\), we reject the null hypothesis \( H_0 \).