

Essentials of Mathematical Probability and Statistics

Essentials of Mathematical Probability and Statistics

John Travis
Mississippi College

[3] October 30, 2018

John Travis grew up in Mississippi and had his graduate work at the University of Tennessee and Mississippi State University. As a numerical analyst, since 1988 he has been a professor of mathematics at his undergraduate alma mater Mississippi College where he currently serves as Professor of Mathematics.

John is married to Ruth Page Travis and together they have three unique daughters who all are good at mathematics.

You can find him playing racquetball or guitar but not generally at the same time. He is also an active supporter and organizer for the opensource online homework system WeBWork.

© 2015 –today John Travis

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the appendix entitled “GNU Free Documentation License.”

Preface

This text is intended for a one-semester course in probability and statistics that presumes calculus knowledge up to integration techniques. It is perhaps helpful if a student has already been exposed to sequences and series but much of what is needed is reviewed in the text.

An interactive version of this text is available at http://math.mc.edu/travis/mathbook/Probability/Essentials_Probability and an pdf version with active links is available at http://math.mc.edu/travis/mathbook/Probability/Essentials_Probability

A collection of WeBWorK online homework problems are available to correlate with the material in this text. Copies of these sets of problems are available by contacting the author. These exercises are considered an integral part when using this text although a static version of these is included in the appendix.

To successfully utilize this text, a student should review the requisite material and perhaps review the proofs and derivations if appropriate. While moving through the text, the student should review each of the examples and then attempt each of the interactive WeBWorK exercises. Whenever an interactive cell comes up, the student should play around with the cell and perhaps change the input data as appropriate to experiment. When a section is completed, a student should work the WeBWorK exercises (not part of this text) or some other exercises provided by the instructor and attempt the exercises provided in this text...many of which are famous examples or exercises that might have special significance. Some are (of course) just easy and most of the textbook exercises have solutions provided.

WeBWorK (webwork.maa.org) is an open-source online homework system for math and science courses. WeBWorK is supported by the MAA and the NSF and comes with a Open Problem Library (OPL) of over 35,000 homework problems. Problems in the OPL target most lower division undergraduate math courses and some advanced courses. Supported courses include college algebra, discrete mathematics, probability and statistics, single and multivariable calculus, differential equations, linear algebra and complex analysis.

Sage (sagemath.org) is a free, open-source, software system for advanced mathematics, which is ideal for assisting with a study of abstract algebra. Sage can be used either on your own computer, a local server, or on SageMathCloud (<https://cloud.sagemath.com>). In this text, the sage cell is used also for interactive computations related to R and octave, as needed.

John Travis

Clinton, Mississippi 2015-2019

Contents

Preface	v
1 Statistical Measures	1
1.1 Introduction	1
1.2 Measurement Scales	1
1.3 Statistical Measures of Position	2
1.4 Statistical Measures of the Middle	5
1.5 Statistical Measures of Variation	9
1.6 Adjusting Statistical Measures for Grouped Data	13
1.7 Other Statistical Point Measures	14
1.8 Visual Statistical Measures - Graphical Representation of Data	16
1.9 Summary	21
1.10 Exercises	21
2 Regression	23
2.1 Introduction	23
2.2 Linear Regression - Best Fit Line	24
2.3 Correlation	27
2.4 Higher Degree Linear Regression	29
2.5 Multi-variable Linear Regression	32
2.6 Summary	34
3 Counting and Combinatorics	35
3.1 Introduction	35
3.2 General Counting Principles	36
3.3 Permutations	38
3.4 Combinations	43
3.5 Summary	46
3.6 Exercises	46
4 Probability Theory	49
4.1 Introduction	49
4.2 Relative Frequency	49
4.3 Definition of Probability	53
4.4 Exercises	60
4.5 Conditional Probability	61
4.6 Bayes Theorem	64
4.7 Independence	70
4.8 Summary	71
4.9 More Exercises	71

5	Probability Functions	79
5.1	Introduction	79
5.2	Random Variables	79
5.3	Probability Functions	80
5.4	Expected Value	86
5.5	Standard Units	93
5.6	Summary	93
5.7	Exercises	93
6	Distributions based upon Equally likely Outcomes	95
6.1	Introduction	95
6.2	Discrete Uniform Distribution	95
6.3	Continuous Uniform Distribution	98
6.4	Hypergeometric Distribution	100
6.5	Summary	104
6.6	Exercises	104
7	Distributions based upon Bernoulli Trials	109
7.1	Introduction	109
7.2	Binomial Distribution	110
7.3	Geometric Distribution	113
7.4	Negative Binomial	116
7.5	Summary	119
7.6	Exercises	119
8	Distributions based upon Poisson Processes	123
8.1	Introduction	123
8.2	Poisson Distribution	123
8.3	Exponential Distribution	126
8.4	Gamma Distribution	128
8.5	Summary	131
8.6	Exercises	131
9	Normal Distributions	135
9.1	Introduction	135
9.2	The Normal Distribution	135
9.3	Chi-Square Distribution	138
9.4	Other "Bell Shaped" distributions	139
9.5	Normal Distribution as a Limiting Distribution	141
9.6	Central Limit Theorem	145
9.7	Summary	149
9.8	Exercises	149
10	Estimation	151
10.1	Introduction	151
10.2	Interval Estimates - Chebyshev	151
10.3	Point Estimates	152
10.4	Interval Estimates - Confidence Interval for p	153
10.5	Interval Estimates - Confidence Interval for μ	157
10.6	Interval Estimates - Confidence Interval for σ^2	158
10.7	Exercises	162

11 Hypothesis Testing	163
11.1 Introduction	163
11.2 Hypotheses and Errors	163
11.3 Hypothesis Test for one proportion	164
11.4 Hypothesis Test for one mean	166
11.5 Hypothesis Test for one variance	167
11.6 Summary	168
11.7 Exercises	168
12 Review of Calculus	169
12.1 Geometric Series	169
12.2 Binomial SumsBinomial SeriesTrinomial Series	171
12.3 Negative Binomial Series	172

Chapter 1

Statistical Measures

1.1 Introduction

To compute your final grade in a class your teacher will likely consider the scores you have earned on various assignments and examinations completed during the duration of the course. However, she ultimately will likely be required to assign some numerical score indicating your level of success in the course. One grade to rule them all. This final grade can only be one value and it would make sense that the grade be a reflection of your work on these tasks. So, what is a fair way for your teacher to complete this task?

Through this process, you will also often need to take into account whether that data set is the entire list of possibilities—known as the population—or just a subset of that population perhaps obtained by taking repeated measurements—that is, a sample.

In general, it is often useful to make decisions using quantitative data but making those decisions can be somewhat arbitrary without a mathematical basis supporting those decisions. In this chapter, you will consider a number of ways to use point values to represent a given set of data. Each of these quantitative metrics will be called a "statistical measure" and will, in some fashion, describe using one number some property of the entire data set. Such measures are part of what is known as "descriptive statistics". Later, you will learn about how other metrics can be used to predict properties of the underlying situation. Doing this is part of what is known as "inferential statistics".

So, let's go and hopefully you will in some measure enjoy the ride!

1.2 Measurement Scales

In creating statistical measures, you might want to consider one of the following general types.

- Nominal measures - In this case, data falls into mutually exclusive and exhaustive categories for which the numerical value is only used for identification purposes. For example, assigning Male = 1, Female = -1.
- Ordinal measures - In this case, data consists of discrete numerical values which can be ranked from lowest to highest or vice versa. For example, your grades in a number of classes are used to compute your GPA—which is a single number.

- Interval measures - In this case, data possesses an order and where the distance between data values is of significance. For example, heights and weights.
- Ratio measures - In this case, data can be expressed as a position in some interval and where ratios between observations have meaning. For example, percentile rankings

In the subsequent sections of this chapter, you will see that a number of different measures are available for most data sets. Determining which "correct" measure to use for describing any given data set will depend the actual situation surrounding the collection of the data.

1.3 Statistical Measures of Position

Given a collection of data, sorting the data may provide several useful descriptors. When sorting data, you can easily use something like a spreadsheet for larger data sets but in this section you will also see there are ways to perform a sort by hand. In either case, statistical measures of position generally involve very little computational work once the data is sorted and take into account only the order of the data from lowest to highest. To assist with notation, we will generally use x -values to represent the original raw data and y -values to represent that same data when ordered with the subscript indicating the positional placement.

Definition 1.3.1 Order Statistic. From the data set x_1, x_2, \dots, x_n , assume that when sorted it is denoted y_1, y_2, \dots, y_n where

$$y_1 \leq y_2 \leq \dots \leq y_n.$$

Then, y_k is known as the k th order statistic. ◇

Example 1.3.2 Age of Presidents - order statistics. The age at inauguration for presidents from 1981-2019 gives the data $x_1 = 69, x_2 = 64, x_3 = 46, x_4 = 54, x_5 = 47, x_6 = 70$ (Reagan, Bush, Clinton, Bush, Obama, Trump). For this data, the order statistics are denoted $y_1 = 46, y_2 = 47, y_3 = 54, y_4 = 64, y_5 = 69, y_6 = 71$. □

Once the data is sorted, it should be very easy for you to locate the smallest and largest values.

Definition 1.3.3 Minimum/Maximum:. For a given data set, the smallest and largest values are known as the minimum and maximum, respectively. In our notation, minimum = y_1 and the maximum = y_n . ◇

Example 1.3.4 Age of Presidents - Minimum/Maximum. Using the [President inauguration data 1.3.2](#), minimum = $y_1 = 46$ and maximum = $y_6 = 70$. □

A value that separates ordered data into two groups with a desired percentage on each side is called a percentile. There are multiple ways that have been created that achieve this goal. In this text we present two and will consistently use the first one presented below. For each, in general, a given percentile is a numerical value at which approximately a given percentage of the data is smaller.

The definition presented below provides for a unique measure for each unique value of p that corresponds to the PERCENTILE.EXC macro in Excel. This version starts by computing $(n + 1)p$ where $0 < p < 1$ and using this to linearly interpolate between two adjacent entries in the sorted list. Another

option that corresponds to PERCENTILE.INC (and PERCENTILE) in Excel is to start with $(n - 1)p + 1$ for determining how to pick the two adjacent entries and then proceeding with linear interpolation. Again, the definition below utilizes the first approach.

Definition 1.3.5 Percentiles. For $0 < s < 1$ and for order statistics y_1, y_2, \dots, y_n define the 100s-th percentile to be

$$P^s = (1 - r)y_m + ry_{m+1}$$

where m is the integer part of $(n+1)s$, namely

$$m = \lfloor (n+1)s \rfloor$$

and

$$r = (n+1)s - m,$$

the fractional part of $(n+1)s$. ◇

Checkpoint 1.3.6 Compute the following percentile values.

Consider the following data set:

39	20	36	27	51	27	12	16	12
51	39	28	51	13	27	34	49	40

Find the 15th and 88th percentiles for this data.

15th percentile = _____

88th percentile = _____

Example 1.3.7 Presidential Percentile. To compute, say, the 42nd percentile for the [President inauguration data presented earlier 1.3.2](#) consider $s = 0.42$. Since there are 6 numbers in our data set, then

$$(n+1)s = 7 \cdot 0.42 = 2.94$$

and so $m = 2$ and $r = 0.94$. Thus, the percentile will lie between $y_2 = 47$ and $y_3 = 54$ and much closer to 54 than 47. Numerically

$$P^{0.42} = 0.06 \cdot 47 + 0.94 \cdot 54 = 53.58.$$

□

The formula for percentiles determines a weighted average between y_m and y_{m+1} which is unique for distinct values of p provided each of the data values are distinct. Note that if some of the y -values are equal then some of these averages might be averages of equal numbers and will therefore be the common value.

Some special percentiles are provided special names...

Definition 1.3.8 Quartiles. Given a sorted data set, the first, second, and third quartiles are the values of $Q_1 = P^{0.25}$, $Q_2 = P^{0.5}$ and $Q_3 = P^{0.75}$. ◇

It should be noted that many graphing calculators often compute quartiles using a straight average of two adjacent entries rather than by using the formula above. This causes some difficulty and especially so when $n \bmod 4 = 2$.

Example 1.3.9 Q_1 and Q_3 when $n \bmod 4 = 2$. Suppose $n = 22 = 5(4) + 2$. Computing the first quartile as defined above gives $(n+1)p = 23(0.25) = 5.75 = 5 + 0.75 = m + r$. Therefore,

$$Q_1 = 0.25 \times y_5 + 0.75 \times y_6$$

which is a value closer to y_6 . Many graphing calculators however quickly approximate this with

$$0.5 \times y_5 + 0.5 \times y_6$$

so you should be aware of this possible difference. You should also notice that in this case $p = 0.25$ but $r = 0.75$ so these values are not required to be the same. \square

Definition 1.3.10 Deciles:. Given a sorted data set, the first, second, ..., ninth deciles are the value of $D_1 = P^{0.1}, D_2 = P^{0.2}, \dots, D_9 = P^{0.9}$ \diamond

Example 1.3.11 Small Example - Quartiles. Consider the following data set: 2,5,8,10. The 50th percentile should be a numerical value for which approximately 50

Using the [definition 1.3.5](#), the 25th percentile is computed by considering

$$(n+1)p = (4+1)0.25 = 5/4 = 1.25.$$

So, $m = 1$ and $r = 0.25$. Therefore

$$P^{0.25} = 0.75 \times 2 + 0.25 \times 5 = 2.75$$

as noted above.

Similarly, the 75th percentile is given by

$$(n+1)p = (4+1)0.75 = 15/4 = 3.75.$$

So, $m = 3$ and $r = 0.75$. Therefore

$$P^{0.75} = 0.25 \times 8 + 0.75 \times 10 = 9.5$$

It is interesting to note that 3 also lies between 2 and 5 as does 2.75 and has the same percentages above (75 percent) and below (25 percent). However, it should designate a slightly larger percentile location. Indeed, going backward:

$$\begin{aligned} 3 &= (1-r) \times 2 + r \times 5 \\ \Rightarrow r &= \frac{1}{3} \\ \Rightarrow (n+1)p &= 1 + \frac{1}{3} = \frac{4}{3} \\ \Rightarrow p &= \frac{4}{15} \approx 0.267 \end{aligned}$$

and so 3 would actually be at approximately the 26.7th percentile. \square

Checkpoint 1.3.12 In general, given a numerical value within the range of a given data set, one can determine the percentile ranking of that value by reversing the general formula for percentile and solving for p , given P^s . Determine such a formula/process for doing this in general.

For your data set 2,5,8,10, $Q_1 = 2.75$, $Q_2 = 6.5$, and $Q_3 = 9.5$.

For a given data set, a summary of these statistics is often desired in order to give the user a quick overview of the more important order statistics.

Definition 1.3.13 5-number summary. Given a set of data, the 5-number summary is a vector of the order statistics given by

$$< \text{minimum}, Q_1, Q_2, Q_3, \text{maximum} > .$$

\diamond

You can also compute these statistics automatically using the opensource statistical software known simply as "R". The following interactive cell uses the opensource software "Sage" to perform this calculation using the freely available web portal at sagemath.sagecell.org. You can change the data list if you want to use this to compute values for a different collections of numbers. The five-number-summary is displayed graphically using a "Box-Plot". Graphical representations of data will be discussed later in this chapter. You should compare the answers found using R with the values produced by our [definition 1.3.5](#)

```
data <- c( 1, 2, 5, 7, 7, -1, 3, 2)  # concatenate the
    following items into a list
paste("Quartiles:")
quantile(data)
paste("Specific_Percentiles:")
quantile(data, c(.32, .57, .98))  # find the 32nd, 57th and
    98th percentiles
paste("Box_and_Whisker_Diagram:")
boxplot(data, horizontal=TRUE)
```

Example 1.3.14 Small example - 5 number summary. Returning to our previous example, the five number summary would be

$$< 2, 2.75, 6.5, 9.5, 10 > .$$

□

1.4 Statistical Measures of the Middle

Definition 1.4.1 Arithmetic Mean. Suppose X is a discrete random variable with range $R = x_1, x_2, \dots, x_n$. The arithmetic mean is given by

$$\frac{x_1 + \dots + x_n}{n} = \frac{\sum_{k=1}^n x_k}{n}.$$

If this data comes from sample data then we call it a sample mean and denote this value by \bar{x} . If this data comes from the entire universe of possibilities then we call it a population mean and denote this value by μ . When presented with raw data, it might be good to generally presume that data comes from a sample and utilize \bar{x} . ◇

To illustrate, consider the previous data set: 2,5,8,10. The arithmetic mean is given by

$$\bar{x} = \frac{2 + 5 + 8 + 10}{4} = \frac{25}{4} = 6.25.$$

The mean is often called the centroid in the sense that if the x values were locations of objects of equal weight, then the centroid would be the point where this system of n equal masses would balance. Play around with the interactive cell below by entering your own data values into the first list.

```
x = [2, 5, 8, 10, 11]  # Put your data values in this list
x.sort()

mu = mean(x)
n = len(x)
```

```

pts = [(x[0],0.05)]
M = 0.2
for k in range(1,n):
    if x[k]==x[k-1]:
        pts.append((x[k],pts[k-1][1]+0.1))
        M += 0.1
    else:
        pts.append((x[k],0.05))
G = points(pts,size=100,figsize=[10,2])
G += polygon([(mu,0), (mu+0.2,-0.5),
              (mu-0.2,-0.5)],color='brown')
G.show(ymin=-0.5, ymax = M)

```

The values can all be provided with varying weights if desired and the result is called the weighted arithmetic mean and is given by

$$\frac{m_1x_1 + \dots + m_nx_n}{m_1 + \dots + m_n} = \frac{\sum_{k=1}^n m_kx_k}{\sum_{k=1}^n m_k}.$$

This is often how your teacher will actually compute your final grade in a class where the m_k are the relative weights for each assignment grade.

```

x = [2, 5, 8, 10]      # Put _unique_ data values in this list
w = [1, 2.5, 2.5, 4]   # Scale to be at most 10 and not tiny
                        # for good image
wsum = sum(w)

n = len(x)
pts = [(x[0],0.05)]
M = 0.2
mu = 0
for k in range(1,n):
    mu += x[k]*w[k]
    if x[k]==x[k-1]:
        pts.append((x[k],pts[k-1][1]+0.1))
        M += 0.2
    else:
        pts.append((x[k],0.05))
mu = mu/wsum
G = Graphics()
for k in range(n):
    G += point(pts[k],size=100*w[k])
P = polygon([(mu,0), (mu+0.2,-0.5),
              (mu-0.2,-0.5)],color='brown')
(G+P).show(ymin=-0.5, ymax = M)

```

Example 1.4.2 Computing class final grade. Suppose in a given class you have a daily grade of 92, exam 1 grade of 85, exam 2 grade of 87, and a final exam grade of 93. IF the daily grade counts 10 percent, the first two exams count 25 percent each and the final counts 40 percent then your final grade would be

$$\frac{0.10 \cdot 92 + 0.25 \cdot 85 + 0.25 \cdot 87 + 0.40 \cdot 93}{0.10 + 0.25 + 0.25 + 0.40} = 89.4.$$

It would then appear that you might want to do some bargaining with your teacher about how nice it would be to round that up. \square

Definition 1.4.3 Median:. A positional measure of the middle is often utilized by finding the location of the 50th percentile. This value is also called the median and indicates the value at which approximately half the sorted data lies below and half lies above. \diamond

For data sets with an odd number of values, this is the "middle" data value if one were to successively cross off pairs from the two ends of the sorted data. For data sets with an even number of values, this is a average of the two data values left after crossing off all other pairs. Using the order statistics, the median equals

$$y_{\frac{n+1}{2}}$$

if n is odd and

$$\frac{y_{\frac{n}{2}} + y_{\frac{n}{2}+1}}{2}$$

if n is even.

From the [Presidential data 1.3.2](#), note that you are considering an even number of data values and so the median is given by $(54+64)/2 = 59$.

Definition 1.4.4 Midrange:. The midrange is a mixture of the mean and median where one takes the simple average of the maximum and minimum values in the data set. Using the order statistics, this equals

$$\frac{y_1 + y_n}{2}$$

\diamond

From the [Presidential data 1.3.2](#), the maximum is 70 and the minimum is 46 so the midrange is 58, the average of these two.

There are several advantages and disadvantages associated with each of these measures. The mean utilizes all of the data values so each term is important. Utilizes them all even if some of the data values might suffer from collection errors. The median ignores outliers (which might be a result of collection errors) but does not account for the relative differences between terms. The midrange is very easy to compute but ignores the relative differences for all terms but the two extremes. A similar collection of features and drawbacks are associated with all descriptive statistics.

You can again compute many statistics automatically using R...

```
data <- c( 1, 2, 5, 7, 7, -1, 3, 2)  # concatenate the
  following items into a list
paste("Mean_=", mean(data))
paste("Median_=", median(data))
```

Example 1.4.5 USA State Population Measures of the Middle. The US Census Bureau reported the following state populations (in millions) for 2013: [Spreadsheet](#)

State	Population
Wyoming	0.6
Vermont	0.6
District of Columbia	0.6
North Dakota	0.7
Alaska	0.7
South Dakota	0.8
Delaware	0.9
Montana	1
Rhode Island	1.1
New Hampshire	1.3
Maine	1.3
Hawaii	1.4
Idaho	1.6
West Virginia	1.9
Nebraska	1.9
New Mexico	2.1
Nevada	2.8
Kansas	2.9
Utah	2.9
Arkansas	3
Mississippi	3
Iowa	3.1
Connecticut	3.6
Oklahoma	3.9
Oregon	3.9
Kentucky	4.4
Louisiana	4.6
South Carolina	4.8
Alabama	4.8
Colorado	5.3
Minnesota	5.4
Wisconsin	5.7
Maryland	5.9
Missouri	6
Tennessee	6.5
Indiana	6.6
Arizona	6.6
Massachusetts	6.7
Washington	7
Virginia	8.3
New Jersey	8.9
North Carolina	9.8
Michigan	9.9
Georgia	10
Ohio	11.6
Pennsylvania	12.8
Illinois	12.9
Florida	19.6
New York	19.7
Texas	26.4
California	38.3

Table 1.4.6: USA State Populations - 2014

Determine the minimum, maximum, midrange, and mean for this data.

Solution. Notice that these are already in order so you can presume $y_1 = 0.6$ million is the minimum and $y_{50} = 38.3$ million is the maximum. Therefore, the midrange is given by

$$\frac{0.6 + 38.3}{2} = \frac{38.9}{2} = 19.45 \text{million.}$$

In this collection of "states" data the District of Columbia is included so that the number of data items is $n=51$. The mean of this data takes a bit of arithmetic but gives

$$\bar{x} = \frac{\sum_{k=1}^{51} y_k}{51} = \frac{316.1}{51} \approx 6.20$$

million residents.

Since the number of states is odd, the median is found by looking at the 26th order statistic. In this case, that is the 4.4 million residents of Kentucky, i.e. $y_{26} = 4.4$. \square

1.5 Statistical Measures of Variation

These measures provide some indication of how much the data set is "spread out". Indeed, note that the data sets $-2, -1, 0, 1, 2$ and $-200, -100, 0, 100, 200$ have the same mean but one is much more spread out than the other. Measures of variation should catch this difference.

Definition 1.5.1 Range:. Using the order statistics,

$$y_n - y_1.$$

\diamond

It is trivial to note that the range is very easy to compute but it completely ignores all data values but the two ends.

From the [Presidential data 1.3.2](#), the maximum is 69 and the minimum is 46 so the range is 23, the difference of these two.

Definition 1.5.2 Interquartile Range (IQR):.

$$IQR = P^{0.75} - P^{0.25}.$$

\diamond

For the data set 2, 5, 8, 10, you have found that $Q_1 = 2.75$ and $Q_3 = 9.5$. Therefore,

$$IQR = 9.5 - 2.75 = 6.75.$$

Average Deviation from the Mean (Population): Given a population data set x_1, x_2, \dots, x_n with mean μ each term deviates from the mean by the value $x_k - \mu$. So, averaging these gives

$$\frac{\sum_{k=1}^n (x_k - \mu)}{n} = \frac{\sum_{k=1}^n x_k}{n} - \frac{\sum_{k=1}^n \mu}{n} = \mu - \mu = 0.$$

This metric is therefore always zero for any provided set of data since cancellation makes this not useful. So, we need to determine ways to avoid cancellation.

Average Absolute Deviation from the Mean (Population):

$$\frac{\sum_{k=1}^n |x_k - \mu|}{n}$$

which, although nicely stated, is difficult to deal with algebraically since the absolute values do not simplify well algebraically. To avoid this algebraic road-block, we can look for another way to nearly accomplish the same goal by squaring and then square rooting.

Average Squared Deviation from the Mean (Population):

$$\frac{\sum_{k=1}^n (x_k - \mu)^2}{n}$$

which will always be non-negative but can be easily expanded using algebra. Since this is a mouthful, this measure is generally called the "variance".

Using the average squared deviation from the mean, differences have been squared. Thus all of the squared differences added are non-negative but very small ones have been made even smaller and larger ones have been made relatively larger. To undo this scaling issue, one must take a square root to get things back into the right ball park.

Definition 1.5.3 Variance and Standard Deviation. The variance is the average squared deviation from the mean. If this data comes from the entire universe of possibilities then we call it a population variance and denote this value by σ^2 . Therefore

$$\sigma^2 = \frac{\sum_{k=1}^n (x_k - \mu)^2}{n}$$

The standard deviation is the square root of the variance. If this data comes from the entire universe of possibilities then we call it a population standard deviation and denote this value by σ . Therefore

$$\sigma = \sqrt{\frac{\sum_{k=1}^n (x_k - \mu)^2}{n}}.$$

If data comes from a sample of the population then we call it a sample variance and denote this value by s^2 . Since sample data tends to reflect certain "biases" then we increase this value slightly by $\frac{n}{n-1}$ to give the sample variance

$$s^2 = \frac{n}{n-1} \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n} = \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n-1}.$$

and the sample standard deviation similarly as the square root of the sample variance. \diamond

From the data 2,5,8,10, you have found that the mean is 6.25. Computing the variance then involves accumulating and averaging the squared differences of each data value and this mean. Then

$$\begin{aligned} & \frac{1}{4} ((2 - 6.25)^2 + (5 - 6.25)^2 + (8 - 6.25)^2 + (10 - 6.25)^2) \\ &= \frac{18.0625 + 1.5625 + 3.0625 + 14.0625}{4} \\ &= \frac{36.75}{4} \\ &= 9.1875. \end{aligned}$$

Theorem 1.5.4 Alternate Forms for Variance.

$$\begin{aligned} \sigma^2 &= \left(\frac{\sum_{k=1}^n x_k^2}{n} \right) - \mu^2 \\ &= \left[\frac{\sum_{k=1}^n x_k(x_k - 1)}{n} \right] + \mu - \mu^2 \end{aligned}$$

Proof.

$$\begin{aligned}
 \sigma^2 &= \frac{\sum_{k=1}^n (x_k - \mu)^2}{n} \\
 &= \frac{\sum_{k=1}^n (x_k^2 - 2x_k\mu + \mu^2)}{n} \\
 &= \frac{\sum_{k=1}^n x_k^2 - 2\mu \sum_{k=1}^n x_k + n\mu^2}{n} \\
 &= \left(\frac{\sum_{k=1}^n x_k^2}{n} \right) - \mu^2
 \end{aligned}$$

The second part is proved similarly. Using the first part of the proof above,

$$\begin{aligned}
 \sigma^2 &= \frac{\sum_{k=1}^n (x_k - \mu)^2}{n} \\
 &= \left(\frac{\sum_{k=1}^n x_k^2}{n} \right) - \mu^2 \\
 &= \left(\frac{\sum_{k=1}^n x_k(x_k - 1) + x_k}{n} \right) - \mu^2 \\
 &= \left(\frac{\sum_{k=1}^n x_k(x_k - 1)}{n} \right) + \mu - \mu^2
 \end{aligned}$$

■

Example 1.5.5 Computing means and variances by hand. In the data table below, notice that the x_k column would be the given data values but the column for x_k^2 you could easily compute.

x_k	x_k^2
1	1
-1	1
0	0
2	4
2	4
5	25

Table 1.5.6: Sample Grouped Data

So, $\sum x_k = 9$ and $\sum x_k^2 = 35$. Therefore $\bar{x} = \frac{9}{6} = \frac{3}{2}$ and $v = \frac{\sum x_k^2}{6} - (\bar{x})^2 = \left(\frac{35}{6} - \frac{3^2}{2}\right) = \frac{70-18}{12} = \frac{26}{6}$. Therefore, $s^2 = \frac{6}{5} \times v = \frac{26}{5}$. □

Use R to compute these values...

```
data <- c( 1, -1, 0, 2, 2, 5)  # concatenate the following
                             # items into a list
paste("Variance_=", var(data))
paste("Standard_Dev_=", sd(data))
paste("Inter_Quantile_Range_=", IQR(data))
paste("Box_and_Whisker_Diagram:")
boxplot(data, horizontal=TRUE)
```

Once again, the Population of the individual USA states according to the 2013 Census is considered below.

Checkpoint 1.5.7 USA State Population Measures of Variation.

Using the [US Census Bureau state populations 1.4.5](#) (in millions) for 2014 provided earlier, determine the range, quartiles, and variance for this sample

data.

Solution. Again, you should note that these are already in order so the range is quickly found to be

$$y_n - y_1 = 38.3 - 0.6 = 37.7$$

million residents.

For IQR, we first must determine the quartiles. The median (found earlier) already is the second quartile so we have $Q_2 = 4.5$ million. For the other two, the formula for computing percentiles gives you the 25th percentile

$$(n+1)p = 51(1/4) = 12.75$$

$$Q_1 = P^{0.25} = 0.25 \times 1.9 + 0.75 \times 2.1 = 2.05$$

and the 75th percentile

$$(n+1)p = 51(3/4) = 38.25$$

$$Q_3 = P^{0.75} = 0.75 \times 7 + 0.25 \times 8.3 = 7.325.$$

Hence, the IQR = 7.325 - 2.05 = 5.275 million residents.

From the computation before, again note that $n=51$ since the District of Columbia is included. The mean of this data found before was found to be approximately 6.20 million residents. So, to determine the variance you may find it easier to compute using the alternate variance formulas [Theorem 1.5.4](#).

$$v = \left(\frac{\sum_{k=1}^n y_k^2}{n} \right) - \mu^2$$

$$\approx \frac{4434.37}{51} - (6.20)^2$$

$$= 48.51$$

and so you get a sample variance of

$$s^2 \approx \frac{51}{50} \cdot 48.51 = 49.48$$

and a sample standard deviation of

$$s \approx \sqrt{49.48} \approx 7.03$$

million residents.

The state population data set has been entered for you in the R cell below...

```
data <- c (0.6,0.6,0.6,0.7,0.7,0.8,0.9,1,1.1,1.3,1.3,1.4,1.6,
1.9,1.9,2.1,2.8,2.9,2.9,3,3,3.1,3.6,3.9,3.9,4.4,4.6,
4.8,4.8,5.3,5.4,5.7,5.9,6,6.5,6.6,6.6,6.7,7,8.3,
8.9,9.8,9.9,10,11.6,12.8,12.9,19.6,19.7,26.4,38.3)
paste("Variance_=", var(data))
paste("Standard_Dev_=", sd(data))
paste("Inter_Quantile_Range_=", IQR(data))
```

1.6 Adjusting Statistical Measures for Grouped Data

As you considered the measures of the center and spread before, each data point was considered individually. Often, data may however be grouped into categories. The number of data items in each category is called the "frequency" of that outcome and the collection of these frequencies for all outcomes is called a "frequency distribution".

1.6.1 Data Grouped into Single-valued Categories

In this case, rather than considering x_k to be the k th data value can take advantage of the grouping to perhaps save a bit on arithmetic.

Indeed, let's assume that data is grouped into m categories x_1, x_2, \dots, x_m with corresponding frequencies f_1, f_2, \dots, f_m . Then, for example, when computing the mean rather than adding x_1 with itself f_1 times just compute $x_1 \times f_1$ for the first category and continuing through the remaining categories. This gives the following grouped data formula for the mean

$$\mu = \frac{x_1 f_1 + \dots + x_m f_m}{f_1 + \dots + f_m} = \frac{\sum_{k=1}^m x_k f_k}{\sum_{k=1}^m f_k}.$$

and the following grouped data formula for the variance (along with one equivalent form)

$$\sigma^2 = \frac{\sum_{k=1}^m (x_k - \mu)^2 f_k}{\sum_{k=1}^m f_k} = \frac{\sum_{k=1}^m x_k^2 f_k}{\sum_{k=1}^m f_k} - \mu^2$$

Checkpoint 1.6.1 Consider the following data set

3, 1, 2, 2, 3, 1, 3, 4, 5, 5, 1, 4, 5, 1, 2, 4, 5, 3, 2, 5, 2, 1, 2, 2, 5

Create a frequency distribution and determine the sample mean and variance.

Solution. Collecting this data into a frequency distribution gives

x_k	f_k
1	5
2	7
3	4
4	3
5	6

Table 1.6.2: Grouped Discrete Data

Therefore,

$$\bar{x} = \frac{1 \times 5 + 2 \times 7 + 3 \times 4 + 4 \times 3 + 5 \times 6}{5 + 7 + 4 + 3 + 6} = \frac{5 + 14 + 12 + 12 + 30}{25} = \frac{43}{25}$$

and

$$\begin{aligned} v &= \frac{1^2 \times 5 + 2^2 \times 7 + 3^2 \times 4 + 4^2 \times 3 + 5^2 \times 6}{5 + 7 + 4 + 3 + 6} - \left(\frac{43}{25}\right)^2 \\ &= \frac{5 + 28 + 36 + 48 + 150}{25} - \left(\frac{43}{25}\right)^2 \\ &= \frac{4826}{625} \\ &\approx 7.7216 \end{aligned}$$

and so $s^2 = \frac{25}{24} \frac{4826}{625} \approx 8.043$.

1.6.2 Data Grouped into Continuous Intervals

For measures on data grouped into intervals, it is somewhat difficult to do calculations when the data no longer exists as individual values since all you know is the frequencies of each interval. You can use "class marks"...the midpoints of each interval...as representers for all of the items that fell into that interval for computing means and variances. For positional measures, you want to approach this in the same manner as with percentiles before. That is, by doing some sort of linear interpolation on the width of each interval.

So, for medians, consider the following approach:

1. Compute frequencies f_k and cumulative frequencies F_k for each class
2. Set $m = \text{total cumulative frequency} / 2 = F_{last} / 2$
3. Determine the interval k where $m \in [F_{k-1}, F_k]$
4. Set $\text{median} = (b_k - a_k) \frac{m - F_{k-1}}{f_k} + a_k$

Example 1.6.3 Computing Median for Interval Grouped Data.

$[a_k, b_k]$	f_k
$[0, 5)$	5
$[5, 10)$	7
$[10, 20)$	4
$[20, 23)$	3
$[23, 30)$	6

Table 1.6.4: Interval Frequency Distribution

The total cumulative frequency is 25 and so $m = \frac{25}{2} = 12.5$ which lies in the $k = 3$ interval $[10, 20)$ and $F_2 = 12$. Therefore

$$\text{median} = (20 - 10) \frac{12.5 - 12}{4} + 10 = 11.25$$

□

1.7 Other Statistical Point Measures

Above, we have investigated statistical measures that help determine the middle and the spread of a given data set. There are however other metrics available that help describe the distribution of that data. Skewness is one of those metrics and describes any lack of symmetry of the data set's distribution and whether data is stretched out to one side or the other.

Definition 1.7.1 Skewness. For population data, the Skewness of x_1, x_2, \dots, x_n is given by

$$\frac{1}{\sigma^3} \frac{\sum_{k=1}^n (x_k - \mu)^3}{n}.$$

For sample data, the Skewness of x_1, x_2, \dots, x_n is given by

$$\frac{1}{s^3} \frac{\sum_{k=1}^n (x_k - \bar{x})^3}{n}.$$

◇

A positive skewness indicates that the positive $(x_k - \mu)^3$ terms (likewise $(x_k - \bar{x})^3$ terms) overwhelm the negative terms. So, a positive skewness indicates that the data set is strung out to the right. Likewise, a negative skewness indicates a data set that is strung out to the left.

Data might tend to be clustered around the mean. The "kurtosis" can be used to measure how closely data resembles a "bell-shaped" collection.

Definition 1.7.2 Kurtosis. For population data, the Kurtosis of x_1, x_2, \dots, x_n is given by

$$\frac{1}{\sigma^4} \frac{\sum_{k=1}^n (x_k - \mu)^4}{n}.$$

For sample data, the Kurtosis of x_1, x_2, \dots, x_n is given by

$$\frac{1}{s^4} \frac{\sum_{k=1}^n (x_k - \bar{x})^4}{n}.$$

◇

A kurtosis of 3 indicates that the data is perfectly bell shaped (a "normal" distribution) whereas data further away from 3 indicates data that is less bell shaped.

Theorem 1.7.3 Alternate Formulas for Skewness and Kurtosis.

$$\begin{aligned} \text{skewness} &= \frac{1}{s^3} \left[\frac{\sum_{k=1}^n x_k^3}{n} - 3v\bar{x} - \bar{x}^3 \right] \\ \text{kurtosis} &= \frac{1}{s^4} \left[\frac{\sum_{k=1}^n x_k^4}{n} - 4\bar{x} \frac{\sum_{k=1}^n x_k^3}{n} + 6\bar{x}^2 v - 3\bar{x}^4 \right] \end{aligned}$$

Proof. For skewness, expand the cubic and break up the sum. Factoring out constants (such as \bar{x}) gives

$$\begin{aligned} & \frac{\sum_{k=1}^n (x_k - \bar{x})^3}{n} \\ &= \frac{\sum_{k=1}^n x_k^3}{n} - 3\bar{x} \frac{\sum_{k=1}^n x_k^2}{n} + 3\bar{x}^2 \frac{\sum_{k=1}^n x_k}{n} - \frac{\sum_{k=1}^n \bar{x}^3}{n} \\ &= \frac{\sum_{k=1}^n x_k^3}{n} - 3\bar{x}(v + \bar{x}^2) + 3\bar{x}^3 - \bar{x}^3 \\ &= \frac{\sum_{k=1}^n x_k^3}{n} - 3\bar{x}v - \bar{x}^3 \end{aligned}$$

and divide by the cube of the standard deviation to finish. Note that the first expansion in the derivation above can be used quickly if the data is collected in a table and powers easily computed.

For kurtosis, similarly expand the quartic and break up the sum as before. Note that you can extract the value of the cubic term by solving for that term in the skewness formula above. Then,

$$\begin{aligned} & \frac{\sum_{k=1}^n (x_k - \bar{x})^4}{n} \\ &= \frac{\sum_{k=1}^n x_k^4}{n} - 4\bar{x} \frac{\sum_{k=1}^n x_k^3}{n} + 6\bar{x}^2 \frac{\sum_{k=1}^n x_k^2}{n} - 4\bar{x}^3 \frac{\sum_{k=1}^n x_k}{n} + \frac{\sum_{k=1}^n \bar{x}^4}{n} \\ &= \frac{\sum_{k=1}^n x_k^4}{n} - 4\bar{x} \frac{\sum_{k=1}^n x_k^3}{n} + 6\bar{x}^2(v + \bar{x}^2) - 4\bar{x}^4 + \bar{x}^4 \\ &= \frac{\sum_{k=1}^n x_k^4}{n} - 4\bar{x} \frac{\sum_{k=1}^n x_k^3}{n} + 6\bar{x}^2 v - 3\bar{x}^4 \end{aligned}$$

and then divide by the fourth power of the standard deviation. Note again that the first expansion in the derivation above might also be a useful shortcut. ■

Going back to a previous example...

Computing skewness and kurtosis by hand can often be better organized using a table. Below, notice that the x_k column would be the given data values but the other columns you could again easily compute.

x_k	x_k^2	x_k^3	x_k^4
1	1	1	1
-1	1	-1	1
0	0	0	0
2	4	8	16
2	4	8	16
5	25	125	625

Table 1.7.4: Computing data statistics by hand

So, $\Sigma x_k = 9$ and $\Sigma x_k^2 = 35$ as before and so $\bar{x} = \frac{3}{2}$, $v = \frac{26}{6}$, $s^2 = \frac{6}{5} \times v = \frac{26}{5}$, and so $s = \sqrt{\frac{26}{5}} = \sqrt{5.2}$. But also, $\Sigma x_k^3 = 141$ and $\Sigma x_k^4 = 659$. Use these in the formulas above to obtain skewness of

$$\left[\frac{141}{6} - 3 \cdot \frac{26}{5} \cdot \frac{3}{2} - \left(\frac{3}{2} \right)^2 \right] / s^3$$

and kurtosis of

$$\left[\frac{659}{6} - 4 \cdot \frac{3}{2} \cdot \frac{141}{6} + 6 \left(\frac{3}{2} \right)^2 \cdot \frac{26}{5} - 3 \cdot \left(\frac{3}{2} \right)^4 \right] / s^4.$$

1.8 Visual Statistical Measures - Graphical Representation of Data

Data sets can range from small to very large. Visual representations of these data sets often allow you to see trends and reveal a lot about the distribution of the data values.

1.8.1 Histograms

Frequency Histograms - height matters

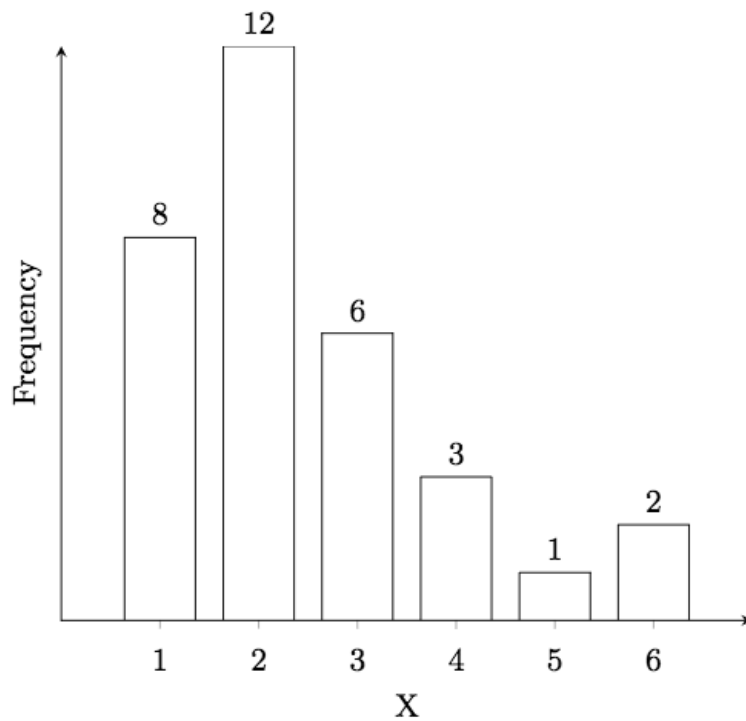
Consider the data set given by

x_k	f_k
1	8
2	12
3	6
4	3
5	1
6	2

Table 1.8.1: Basic Frequency Table

A frequency histogram representing this data looks like

Experiment with creating your own histogram by inputting your data into the interactive Sage cell below.



```
# This function is used to convert an input string into
  separate entries
def g(s): return str(s).replace(',','_').replace('(','_')
  .replace(')','_').split()

@interact
def _(freq =
  input_box("1,1,1,1,2,2,2,3,3,3,3,1,5",label="Enter_data_
  separated_by_commas")):
  freq = g(freq)
  freq = [int(k) for k in freq]
  m = min(freq)
  M = max(freq)
  bn = M-m+1
  histogram( freq, range=[m-1/2,M+1/2], bins = bn,
    align="mid", linewidth=2, edgecolor="blue",
    color="yellow").show()
```

Relative Frequency Histograms - In this case, area describes relative frequency. Notice in the interactive cell above that each bar is of width one. Therefore, frequency = area. In some instances where data may be grouped the total width of the interval may be different and so the height will need to be adjusted so that the total area of each bar corresponds to the relative frequency of that category.

Cummulative Histograms. In these a running total is presented using all values from the given point and below.

```
# This function is used to convert an input string into
  separate entries
def g(s): return str(s).replace(',','_').replace('(','_')
        .replace(')','_').split()

@interact
def _(freq =
    input_box("1,1,1,1,2,2,2,3,3,3,3,1,5",label="Enter_data_
    separated_by_commas")):
    freq = g(freq)
    freq = [int(k) for k in freq]
    top = len(freq)
    m = min(freq)
    M = max(freq)
    bn = M-m+1
    histogram( freq, range=[m-1/2,M+1/2], cumulative =
        "true", bins = bn, align="mid", linewidth=2,
        edgecolor="blue", color="yellow").show(ymax=top)
```

1.8.2 Stem and Leaf Plot

A Stem and Leaf Plot allows you to create a histogram of sorts but maintain the individual data values. To create one of these plots, you will need to consider your particular data set and create a two-step sieve for organizing the set. The first part is to create "stems" that are often associated with the highest digit(s) of each data value and the "leaves" that are often associated with the remaining digit(s) of the data value.

Once the data set is broken down into stems and leaves, it is often simple to sort the leaves under each stem to yield an "ordered Stem and Leaf Plot". Such as mechanism is a simple two-step procedure that allows you to sort a data set by hand.

Example 1.8.2 Simple Stem and Leaf Plot. Consider the data points 25, 3, 17, 12, 22, 34, 12, 11, 16, 42, 9, 12, 17. In this case we will consider the stems to be the tens digits and the leaves to be the ones digits. This gives

Stems	Leaves
0	3 9
1	7 2 2 1 6 2 7
2	5 2
3	4
4	2

Table 1.8.3: Stem and Leaf Plot (unordered)

Then, an ordered Stem and Leaf Plot would be

Stems	Leaves
0	3 9
1	1 2 2 2 6 7 7
2	2 5
3	4
4	2

Table 1.8.4: Stem and Leaf Plot (ordered)

Notice, in each case you can extract the original data values by recombining the stem with a corresponding leaf. Indeed, for these 13 data values the median should be the 7th in the sorted list or the value in the 10's stem with leaf 6...that is, 16. \square

Example 1.8.5 Stem and Leaf Plot for State Populations. Using the state population data above, consider organizing the data but using a "two-pass sort" where you first roughly break data up into groups based upon ranges which relate to their first digit(s). In this case, let's break up into groups according to populations corresponding to 0-4 million, 5-9 million, 10-14 million, 15-19 million, 20-24 million, 25-29 million, 30-35 million, and 35-39 million. We can represent these classes by using the stems 0L, 0H, 1L, 1H, 2L, 2H, 3L, and 3H where the L and H represent the one's digits L in 0, 1, 2, 3, 4 and H in 5, 6, 7, 8, 9. Once we group the data into these smaller groups then we can write the remaining portion of the number horizontally as leaves (in this case with one decimal place for all values.) This gives a step-and-leaf plot. If we additionally sort the data in the leaves then this gives you an ordered stem-and-leaf plot. For the state population data, the ordered stem-and-leaf plot is given by Notice how it is easy to now see that most state populations are relatively small and that there are relatively few states with larger population. Also, notice that you can use this plot to relatively easily identify minimum, maximum, and other order statistics. \square

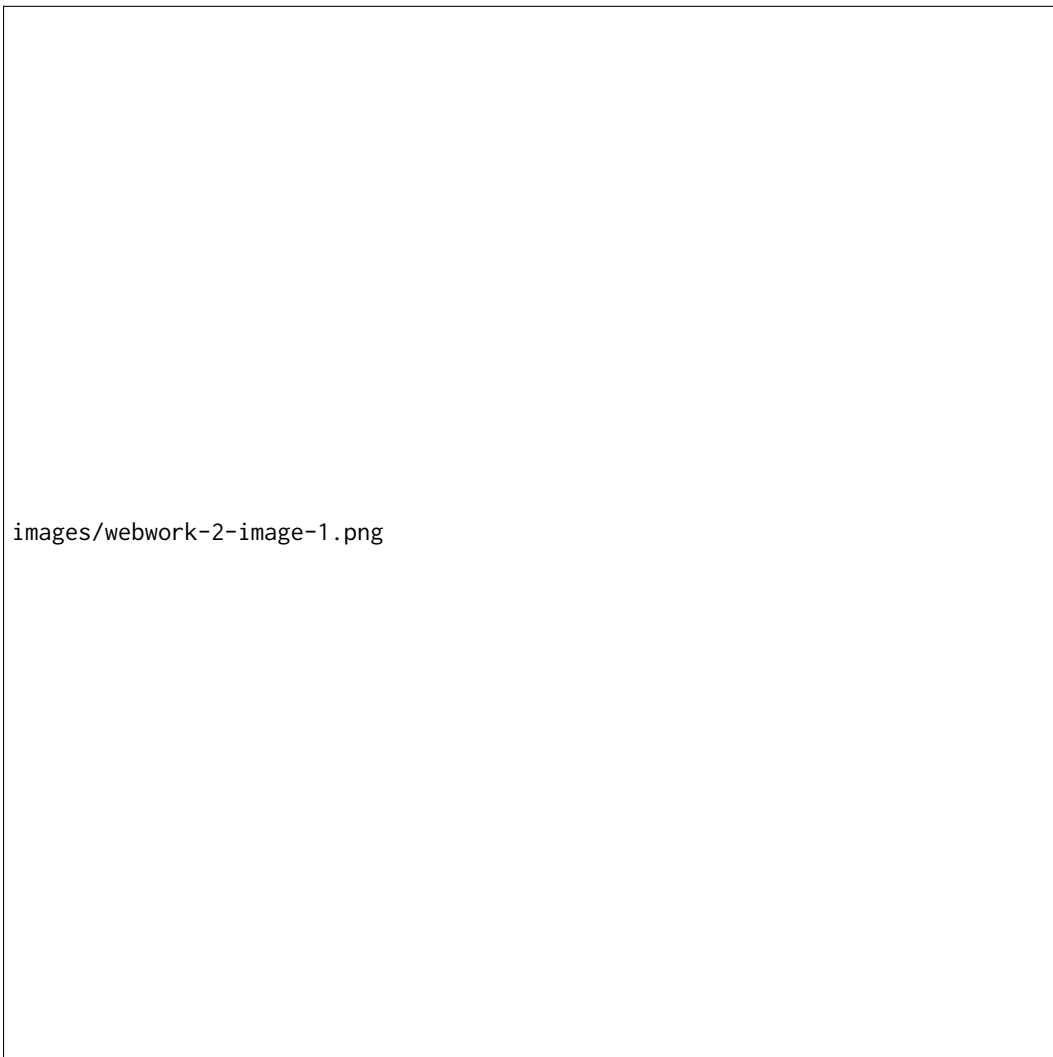
1.8.3 Box and Whisker Diagram (Box Plot)

This graphical display identifies the "5-number-summary" associated with the minimum, quartiles, and the maximum. That is, y_1, Q_1, Q_2, Q_3, y_n . These values separate the data roughly into quarters. To distinguish these quarters connect y_1 and Q_1 with a straight line (a whisker) and do the same with Q_3 and y_n . Use a box to connect Q_1 with Q_2 and the same to connect Q_2 with Q_3 . Then the boxed areas also identify the IQR.

```
data <- c (0.6,0.6,0.6,0.7,0.7,0.8,0.9,1,1.1,1.3,1.3,
1.4,1.6,1.9,1.9,2.1,2.8,2.9,2.9,3,3,3.1,
3.6,3.9,3.9,4.4,4.6,4.8,4.8,5.3,5.4,5.7,
5.9,6,6.5,6.6,6.6,6.7,7,8.3,8.9,9.8,9.9,
10,11.6,12.8,12.9,19.6,19.7,26.4,38.3)
paste("Inter_Quantile_Range_=", IQR(data))
paste("Box_and_Whisker_Diagram_-_Box_Plot: ")
boxplot(data, horizontal=TRUE)
```

Checkpoint 1.8.6 Let's use a box plot to determine some order statistics.

Consider the following box and whisker plot. Find the indicated values of the represented data.



images/webwork-2-image-1.png

Median: _____

Maximum: _____

1.8.4 Density Plots

A Density Plot can be created to visually interpret if the variable is close to normal

```
library(e1071)
par(mfrow=c(1,2)) # graph into two columns
plot(density(cars$speed), main="Density_Plot:_Speed",
      ylab="Frequency", sub=paste("Skewness:",
      round(e1071::skewness(cars$speed),2)))
```

1.9 Summary

Links to the main formulas related to descriptive statistics:[Order Statistics](#)

[Maximum and Minimum](#)

[Percentiles](#)

[Quartiles](#)

[Deciles](#)

[5-number summary](#)

[Mean](#)

[Median](#)

[Midrange](#)

[Range](#)

[Inter Quartile Range](#)

[Variance](#)

[Skewness](#)

[Kurtosis](#)

1.10 Exercises

Complete the online WebWorK homework set "Computational Measures".

Checkpoint 1.10.1 Create a data set with about 10 elements. For your data set, compute each of the measures from this chapter and present your data using a frequency histogram.

Checkpoint 1.10.2 Find a "real-world" data set (similar perhaps to the Census data presented above.) Compute each of the measures from this chapter. Interpret and present your conclusions in an electronic report which can include an excel spreadsheet.

Chapter 2

Regression

2.1 Introduction

When computing means, medians, variances, etc. in the previous chapter, you took given data and create measures that in some sense describe the data using a single value. These single values can be called "descriptive statistics" or perhaps "point estimates" that help understand the properties of the original data set. In this chapter, you will instead take a data set and create a mathematical model that can be used to predict or infer properties of the underlying problem. Statistical procedures such as in this chapter that are used to predict are often lumped into the world of "inferential statistics".

So, given a set of data points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, it is often desirable to have a nice continuous formula $y = f(x)$ that expresses the general nature of those data points. Such a formula "interpolates" the data points if

$$y_k = f(x_k),$$

that is the formula gives a graph that exactly passes through each of the given data points.

On the other hand, sometimes the data points are known to be only approximate or the complexity of the formula needed to interpolate all of the data points exactly is too large. In this case, the formula may only be required to return values that are relatively close to the data points. Such a formula is said to "approximate" and gives

$$y_k \approx f(x_k).$$

Let's consider ways to create useful models that approximate the data points.

From basic algebra, if you are given two distinct points then there is one line which passes exactly through (i.e. interpolates) both. There are many ways to create this linear model but for points $(x_0, y_0), (x_1, y_1)$,

$$y = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0$$

is the linear function which passes through both points if the x-values are distinct. If the x's are equal then

$$x = x_0$$

is linear and interpolates both data points. However, once you collect three or more points it is likely that there is no line which exactly "interpolates" all of

the points. If we desire a linear model then we must settle for a model that approximates. In this chapter, you will investigate how to create polynomial functions which in some manner approximate a collection of data point in some "best" manner.

2.2 Linear Regression - Best Fit Line

In the next few sections, we will presume only one independent variable x and one dependent variable y . Toward that end, consider a collection of data points

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$$

and a general linear function

$$f(x) = mx + b.$$

It is possible but generally unlikely that each of the given data points will be interpolated exactly by the linear function. However, you may notice that the data points exhibit a linear tendency or that the underlying physics might suggest a linear model. A "scatter plot" of a example data set is created in the interactive cell below and the provided data appears to indicate a linear trend. In general, if this is the case then you may find it easier to predict values of y for given values of x using a linear approximation. That is why this method for doing so is also often called a "best-fit line".

```
var('x')
@interact
def _(Points = input_box([(-1,1),(3,2),(4,3),(6,4)])):
    G = points(Points, size=20)
    G.show(title = "Scatter_Plot")
```

But why even bother creating a formula (a line here) to approximate data that does not satisfy that formula? Remember that you would expect collected data to vary slightly as one repeatedly collects that data in the same way that you would expect to make a slightly different score on repeated attempts at exams on the same material. Creating a formula that is close to your data gives a well-defined way to predict a y value for a given x value. This predictive behavior is illustrated in the exercise below.

Checkpoint 2.2.1 WebWork - Using an approximating line. An airline has determined that the relationship between the number of passengers on a flight and the total weight of luggage stored in the baggage compartment can be estimated by the least squares regression equation

$$y = 127 + 28x.$$

Predict the weight of luggage for a flight with 121 passengers.

Answer: _____ pounds

Solution. Since x represents weight, choose $x = 121$ and evaluate the line at that value to get the estimated total luggage weight to be

$$y = 127 + 28(121) = 3515$$

To determine this best-fit line, you need to determine what is meant by the word "best". For linear regression, to reach this goal consider the total of

all vertical deviations between the desired line and the provided data points. Indeed, this vertical error would be of the form

$$e_k = f(x_k) - y_k$$

and would be zero if $f(x)$ exactly interpolated at the given data point. Note, some of these errors will be positive and some will be negative. To avoid any possible cancellation of errors, we could consider taking absolute values (which is tough to deal with algebraically) or perhaps squaring the errors. This second option is the standard approach. This approach is similar to the approach taken earlier when developing formulas for the variance.

The best-fit line therefore will be the line $f(x) = mx + b$ so that the "total squared error" is minimized. This total squared error is given by

$$TSE(m, b) = \sum_{k=1}^n e_k^2 = \sum_{k=1}^n (f(x_k) - y_k)^2 = \sum_{k=1}^n (mx_k + b - y_k)^2.$$

For the following interactive cell, consider for the given data points various values for the slope and y-intercept and see if you can make the total squared error as small as possible. In doing so, notice the vertical distances from the line to the given data points generally decreases as this error measure gets smaller.

```
var('x')
@interact
def _(Points = input_box([(-1,1),(3,1),(4,3),(6,4)]), m =
  slider(-4,4,1/50,1), b = slider(-2,2,1/50,1)):
  G = points(Points,size=20)
  xpt = []
  ypt = []
  f = m*x + b
  TSE = 0
  for k in range(len(Points)):
    x0 = Points[k][0]
    xpt.append(x0)
    y0 = Points[k][1]
    ypt.append(y0)
    TSE += (f(x=x0) - y0)^2
  G += line([(x0,f(x=x0)),(x0,y0)],color='orange')
  G += plot(f,x,min(xpt)-0.2,max(xpt)+0.2,color='gray')
  T = 'Total_Squared_Error_=_$s$'%str(n(TSE))
  G.show(title = T)
```

Checkpoint 2.2.2 Non-functional data. Experiment in the interactive cell above using exactly two data points that have the same x-value. Such as (1,1) and (1,2). Next, add some additional data points in the same general vicinity as your original two points. What is the effect to your best-fit line of adding non-functional points?

So that we don't have to guess the best values for slope and intercept, we can appeal to calculus. Indeed, to minimize this function of the two variables m and b take partial derivatives and set them equal to zero to get the critical values:

$$TSE_m = \sum_{k=1}^n 2(mx_k + b - y_k) \cdot x_k$$

2.3 Correlation

You can plot points and plot the resulting best-fit line determined in the previous section but the question remains whether the line is any good. In particular, the real use of the line often is to subsequently predict y-values for a given x-value. However, it is very likely that the best-fit line does not even pass through any of the provided data points. So, how can something that misses every marker still be considered a good fit. To quantify this, we first need to discuss a way to measure how two variables might vary with each other.

Definition 2.3.1 Covariance. Given paired (sample) data

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$$

with corresponding means \bar{x} and \bar{y} , the covariance is given by

$$Cov(X, Y) = \sum_{k=0}^n (x_k - \bar{x})(y_k - \bar{y})/n$$

and similarly if using population data in which you would use instead the mean of the x-values μ_x and the mean of the y-values μ_y . \diamond

Theorem 2.3.2 Alternate Formula for Covariance.

$$Cov(X, Y) = \frac{\sum_{k=0}^n x_k y_k}{n} - \bar{x} \cdot \bar{y}$$

Proof.

$$\begin{aligned} Cov(X, Y) &= \sum_{k=0}^n (x_k - \bar{x})(y_k - \bar{y})/n \\ &= \sum_{k=0}^n [x_k y_k - \bar{x} \cdot y_k - \bar{y} \cdot x_k + \bar{x} \cdot \bar{y}] / n. \\ &= \sum_{k=0}^n x_k y_k / n - \bar{x} \cdot \sum_{k=0}^n y_k / n - \bar{y} \cdot \sum_{k=0}^n x_k / n + \bar{x} \cdot \bar{y} \end{aligned}$$

which simplifies to the desired result using the definition of the mean. \blacksquare

This general definition provides a general measure which is a second order term (like variance) but also maintains "units". To provide a unit-less metric, consider the following measure.

Definition 2.3.3 Correlation Coefficient. Given a collection of data points, the correlation coefficient is given by

$$r = \frac{Cov(X, Y)}{s_x s_y}$$

where s_x is the standard deviation of the x-values only and s_y is the standard deviation of the y-values only. A similar statistics for population data would instead utilize σ_x and σ_y as the respective standard deviations of the x-values and y-values. \diamond

Theorem 2.3.4 Correlation Coefficient for Linear Data. *If the points are colinear with a positive slope then $r=1$ and if the points are collinear with a negative slope then $r=-1$.*

Proof. Assume the data points are colinear with a positive slope. Then the $TSE(m_0, b_0) = 0$ for some m_0 and b_0 . For this line notice that $f(x_k) = y_k$ exactly for all data points. It is easy to show then that $\bar{y} = m_0\bar{x} + b_0$ and $s_y = |m_0|s_x$. Therefore,

$$Cov(X, Y) = \sum_{k=0}^n (x_k - \bar{x})(m_0x_k + b_0 - (m_0\bar{x} + b_0))/n = m_0s_x^2$$

Putting these together gives correlation coefficient

$$r = \frac{m_0s_x^2}{s_xm_0s_x} = 1.$$

A similar proof follows in the second case by noting that $m_0/|m_0| = -1$. ■

Definition 2.3.5 Coefficient of Determination. Given the correlation coefficient r , the coefficient of determination is given by

$$r^2.$$

This measure indicates the percentage of the variation in y that can be explained by the collection of x values. Note, if $r=1$ (or $r=-1$), then the theorem above indicates that the linear model explains the variability for all of the y -values. ◇

Checkpoint 2.3.6 WebWork. Interpreting correlation coefficients.

For each problem, select the best response.

(a) A study found a correlation of $r = -0.61$ between the gender of a worker and his or her income. You may correctly conclude

- ⊙ an arithmetic mistake was made. Correlation must be positive.
- ⊙ women earn more than men on average.
- ⊙ women earn less than men on average.
- ⊙ this is incorrect because r makes no sense here.
- ⊙ None of the above.

(b) For a biology project, you measure the weight in grams and the tail length in millimeters of a group of mice. The correlation is $r = 0.8$. If you had measured tail length in centimeters instead of millimeters, what would be the correlation? (There are 10 millimeters in a centimeter.)

- ⊙ $0.8/10 = 0.08$
- ⊙ $(0.8)(10) = 8$
- ⊙ 0.8
- ⊙ None of the above.

Checkpoint 2.3.7 WebWork. Interpreting correlation coefficients.

For each problem, select the best response.

(a) What are all the values that a correlation r can possibly take?

- ⊙ $0 \leq r \leq 1$

- ⊙ $r \geq 0$
- ⊙ $-1 \leq r \leq 1$
- ⊙ None of the above.

(b) You have data for many years on the average price of a barrel of oil and the average retail price of a gallon of unleaded regular gasoline. When you make a scatterplot, the explanatory variable on the x -axis

- ⊙ can be either oil price or gasoline price.
- ⊙ is the price of gasoline.
- ⊙ is the price of oil.
- ⊙ None of the above.

(c) In a scatterplot of the average price of a barrel of oil and the average retail price of a gallon of gasoline, you expect to see

- ⊙ a positive association.
- ⊙ a negative association.
- ⊙ very little association.
- ⊙ None of the above.

Checkpoint 2.3.8 Correlation equaling 0. Consider the data points (1,1), (1,2), (2,1), (2,2). Plot these points and consider the nature of the best fit line. Show using software that the correlation coefficient is zero. Justify why $TSE(m,b) = 1$ must be the minimum.

2.4 Higher Degree Linear Regression

Continuing in a similar fashion to the previous section, consider now an approximation using a quadratic function $f(x) = ax^2 + bx + c$. In this case, the total squared error would be of the form

$$TSE(a, b, c) = \sum_{k=0}^n (ax_k^2 + bx_k + c - y_k)^2.$$

Taking all three partials gives

$$TSE_a = \sum_{k=0}^n 2(ax_k^2 + bx_k + c - y_k) \cdot x_k^2$$

$$TSE_b = \sum_{k=0}^n 2(ax_k^2 + bx_k + c - y_k) \cdot x_k$$

$$TSE_c = \sum_{k=0}^n 2(ax_k^2 + bx_k + c - y_k) \cdot 1.$$

Once again, setting equal to zero and solving gives the normal equations for the best-fit quadratic

$$\begin{aligned} a \sum_{k=0}^n x_k^4 + b \sum_{k=0}^n x_k^3 + c \sum_{k=0}^n x_k^2 &= \sum_{k=0}^n x_k^2 y_k \\ a \sum_{k=0}^n x_k^3 + b \sum_{k=0}^n x_k^2 + c \sum_{k=0}^n x_k &= \sum_{k=0}^n x_k y_k \\ a \sum_{k=0}^n x_k^2 + b \sum_{k=0}^n x_k + c \sum_{k=0}^n 1 &= \sum_{k=0}^n y_k. \end{aligned}$$

Notice that even though you are creating the best-fit quadratic, to find that quadratic boils down to solving a (slightly larger) linear system. In other words, linear regression again. Indeed, we can also approach the derivation of regression formulas directly using a linear algebra approach. To do this, consider the equations generated by plugging in the data points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ into the quadratic model. This yields a (likely overdetermined) system of equations. Appending an error term ϵ_k for each equation gives the following matrix form:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_0^2 & x_0 & 1 \\ x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \dots & \dots & \dots \\ x_n^2 & x_n & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

which in matrix form looks something like

$$Y = XA + \epsilon.$$

Solving for ϵ and then minimizing $\epsilon^t \epsilon$ yields the same solution as above. In matrix form, after some work this becomes

$$A = (X^t X)^{-1} X^t Y$$

with the matrix A containing the three unknowns a, b, and c.

We can also use Matlab (or the opensource alternative "octave") to compute this linear algebra for us. The graph here using the sagecell is a text graph and is very rudimentary but plugging this code into Matlab or a desktop version of octave should present a very nice graph.

```
x = [-1 0 1 3 5 5]
y = [5 3 0 -1 3 6]
n = max(size(x));
for k = 1:n
    X(k,1) = x(k)^2;
    X(k,2) = x(k);
    X(k,3) = 1;
end
Y = y'; %%# transpose the set of y-values to be a column vector
A = inv(X'*X)*X'*Y;
a = A(1)
b = A(2)
c = A(3)
```



```

u=_min(x):0.1:max(x);_#_create_a_set_of_input_values_for_
    plotting
v=_a*_u.^2+_b*_u+_c;
plot(x,y,'o',u,v,'.')

```

Cutting and pasting this code into perhaps <http://octave-online.net> gives a nice, non ASCII graph. Below, we do the same thing but using Sage.

```

var('x')
xpts = vector(RR,(-1, 0, 1, 3, 5, 5))
ypts = vector(RR,(5, 3, 0, -1, 3, 6))
ones = vector(RR,(1, 1, 1, 1, 1, 1))
xpts2 = []      # accumulate the squares
pts = []        # accumulate the (x,y) pairs for plotting
                purposes
for k in range(len(xpts)):
    xpts2.append(xpts[k]^2)
    pts.append((xpts[k],ypts[k]))
xpts2 = vector(xpts2)

X = matrix(RR,[xpts2]).stack(xpts).stack(ones).transpose()
# create X
Y = matrix(RR,ypts).transpose()
Xt = X.transpose()
A = (Xt*X).inverse()*Xt*Y
[a,b,c] = [A[0][0], A[1][0], A[2][0]]
f = a*x^2+b*x+c
banner = "The_quadratic_interpolant_is_given_by_
    \(%s\) "%str(latex(f))
G = points(pts,size=20)
H = plot(f,x,min(xpts)-0.2,max(xpts)+0.2,title=banner)
show(G+H)

```

Checkpoint 2.4.1 Creating a Cubic Linear Regression Interpolant.

Modify the Sage code above to give a best-fit cubic interpolant.

Solution.

```

var('x')
xpts = vector((-1, 0, 1, 3, 5, 5))
ypts = vector((5, 3, 0, 4, 3, -1))
ones = vector((1, 1, 1, 1, 1, 1))
xpts3 = []
xpts2 = []
pts = []
for k in range(len(xpts)):
    xpts3.append(xpts[k]^3)
    xpts2.append(xpts[k]^2)
    pts.append((xpts[k],ypts[k]))
xpts3 = vector(xpts3)
xpts2 = vector(xpts2)

X = matrix([xpts3]).stack(xpts2).stack(xpts).stack(ones).transpose()
Y = matrix(ypts).transpose()
Xt = X.transpose()

A = (Xt*X).inverse()*Xt*Y
[a,b,c,d] = [A[0][0],A[1][0],A[2][0],A[3][0]]

```

```
f = a*x^3 + b*x^2 + c*x + d
banner = "The cubic interpolant is given by %s"%str(latex(f))
G = points(pts,size=20)
H = plot(f,x,min(xpts)-0.2,max(xpts)+0.2,title=banner)
show(G+H)
```

Checkpoint 2.4.2 WebWork. Doing Cubic Linear Regression...use your Sage work from above.

In some cases, the best-fitting multiple regression equation is of the form $\hat{y} = b_0 + b_1x + b_2x^2 + b_3x^3$. The graph of such an equation is called a cubic. Using the data set given below, and letting $x_1 = x$, $x_2 = x^2$, and $x_3 = x^3$, find the multiple regression equation for the cubic that best fits the given data.

x	-8	-5	-2	-1	4	6	8
y	36.2	14.7	2	-0.9	-19.9	-37.4	-61.7

The equation is $\hat{y} = \underline{\hspace{2cm}} + \underline{\hspace{2cm}}x + \underline{\hspace{2cm}}$
 $x^2 + \underline{\hspace{2cm}}x^3$.

2.5 Multi-variable Linear Regression

The regression models that we have looked at presumed a single independent variable. It is much more likely when investigating cause and effect relationships that there are perhaps many independent variables that contribute.

Let's consider a linear model with two independent variables. Indeed, a basic two-variable linear model of the form

$$z = \alpha_1x + \alpha_2y + \beta$$

can be used to approximate data points

$$(x_0, y_0, z_0), (x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_n, y_n, z_n).$$

Using a linear systems approach similar to the previous section by evaluating at these data points and appending an error term to each equation gives, in matrix form:

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} x_0 & y_0 & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where the ϵ_k terms are the deviation between the exact data point and the approximation of that point on some plane. Symbolically

$$Z = XA + \epsilon.$$

If all of the points lie on the same plane (unlikely), then $\epsilon = 0$. Otherwise, once again applying a least squares solution approach is the same as minimizing $\epsilon^t\epsilon$ and eventually gives

$$A = (X^tX)^{-1}X^tZ$$

in general. Evaluating this with X and Z as above gives the needed coefficients

$$A = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix}$$

Let's first see how this is done automatically in R using one of the built-in data sets

```
x1 <- c(1, 2, 3, 5, 5)
x2 <- c(1, 1, 2, 2, 3)
y <- c(5, 1, 4, 3, -1)
fit <- lm(y ~ x1+x2, data=(x,y,z))
summary(fit)                # basic results
coefficients(fit)           # a,b,c, for y = a x1 + b x2 + c x3
fitted(fit)                 # predicted values
residuals(fit)              # errors
```

A good example of the usefulness and limitations of multi-variate linear regression is the calculation of the "Heat Index". This measure determines a measure of discomfort relative to the ambient temperature and the relative humidity. Indeed, in warm climates a high temperature is more difficult to bear if the humidity is also high. One reason is that with high humidity the body is less effective in shedding heat through evaporation of body sweat.

The National Weather Service in 1990 published the following multiple regression equation for Heat Index (HI) relative to the ambient temperature (T) and the relative humidity (RH)

$$\begin{aligned} H = & -42.379 + 2.04901523 \cdot T + 10.14333127 \cdot R - 0.22475541 \cdot T \cdot R \\ & - 6.83783 \cdot 10^{-3} \cdot T^2 - 5.481717 \cdot 10^{-2} \cdot R^2 + 1.22874 \cdot 10^{-3} \cdot T^2 \cdot R \\ & + 8.5282 \cdot 10^{-4} \cdot T \cdot R^2 - 1.99 \cdot 10^{-6} \cdot T^2 \cdot R^2. \end{aligned}$$

Since this model utilizes a linear combination of terms and it's derivation could also be generated using a generalization of the linear regression method presented above. Details on how this equation was determined and other details are available at https://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml.

```
@interact
def _(T = (90), R = (95)):
    H = -42.379 + 2.04901523*T + 10.14333127*R \
        - 0.22475541*T*R - 6.83783*10^(-3)*T^2 \
        - 5.481717*10^(-2)*R^2 + 1.22874*10^(-3)*T^2*R \
        + 8.5282*10^(-4)*T*R^2 - 1.99*10^(-6)*T^2*R^2
    print "For T = ", T, " with humidity = ", R, " percent, Heat \
        Index = ", H
```

Below one can compute a table for various ambient Temperature readings given one value for relative humidity. Notice what happens for a relatively high humidity and relatively high temperature.

```
R = 95
for T in range(80, 121):
    H = -42.379 + 2.04901523*T + 10.14333127*R \
        - 0.22475541*T*R - 6.83783*10^(-3)*T^2 \
```

```

-5.481717*10^(-2)*R^2+1.22874*10^(-3)*T^2*R \
+8.5282*10^(-4)*T*R^2-1.99*10^(-6)*T^2*R^2
print "For T=",T,"with humidity=",R,"percent",Heat_
Index=",H

```

Indeed, you cannot roast a turkey by simply turning the oven on 120 and pumping in a lot of humidity since the turkey is not trying to cool itself anymore. Any discomfort measured on the turkey's behalf would certainly be matched by the human since the bird would be a rare bird and remain very much uncooked. The issue is that this model doesn't presume the possibility of 120F and 95

2.6 Summary

Here are the important formulas from this section: Later

Chapter 3

Counting and Combinatorics

3.1 Introduction

One of the earliest applications of mathematics you probably remember is how you could use number to count things. For many, this is what they think people do when they do mathematics. In this chapter, we will discover that it is possible to count items without actually listing them all.

Example 3.1.1 Counting by actually listing out all possibilities. Consider counting the number of ways one can arrange Peter, Paul, and Mary with the order important. Listing the possibilities:

- Peter, Paul, Mary
- Peter, Mary, Paul
- Paul, Peter, Mary
- Paul, Mary, Peter
- Mary, Peter, Paul
- Mary, Paul, Peter

So, it is easy to see that these are all of the possible outcomes and that the total number of such outcomes is 6. What happens however if we add Simone to the list?

- Simone, Peter, Paul, Mary
- Simone, Peter, Mary, Paul
- Simone, Paul, Peter, Mary
- Simone, Paul, Mary, Peter
- Simone, Mary, Peter, Paul
- Simone, Mary, Paul, Peter
- Peter, Simone, Paul, Mary
- Peter, Simone, Mary, Paul
- Paul, Simone, Peter, Mary
- Paul, Simone, Mary, Peter

- Mary, Simone, Peter, Paul
- Mary, Simone, Paul, Peter
- Peter, Paul, Simone, Mary
- Peter, Mary, Simone, Paul
- Paul, Peter, Simone, Mary
- Paul, Mary, Simone, Peter
- Mary, Peter, Simone, Paul
- Mary, Paul, Simone, Peter
- Peter, Paul, Mary, Simone
- Peter, Mary, Paul, Simone
- Paul, Peter, Mary, Simone
- Paul, Mary, Peter, Simone
- Mary, Peter, Paul, Simone
- Mary, Paul, Peter, Simone

Notice how the list quickly grows when just one more choice is added. This example illustrates how keeping track of the number of items in a set can quickly get impossible to manage unless we can use a more mathematical approach that allows you to count the number of possibilities without having to list them all.

□

3.2 General Counting Principles

Definition 3.2.1 Cardinality. Given a set of elements A , the number of elements in the set is known as its cardinality and is denoted $|A|$. If the set has an infinite number of elements then we set $|A| = \infty$. \diamond

In order to "count without counting" we establish the following foundational principle:

Theorem 3.2.2 Multiplication Principle. *Given two successive events A and B , the number of ways to perform A and then B is $|A||B|$.*

Proof. If either of the events has infinite cardinality, then it is clear that the number of ways to perform A and then B will also be infinite. So, assume that both $|A|$ and $|B|$ are finite. In order to count the successive events, enumerate the elements in each set

$$A = \{a_1, a_2, a_3, \dots, a_{|A|}\}$$

$$B = \{b_1, b_2, b_3, \dots, b_{|B|}\}$$

and consider the function $f(k, j) = (k-1)|B| + j$. This function is one-to-one and onto from the set

$$\{(k, j) : 1 \leq k \leq |A|, 1 \leq j \leq |B|\}$$

onto

$$\{s : 1 \leq s \leq |A||B|\}.$$

Since this second set has $|A| \cdot |B|$ elements then the conclusion follows. \blacksquare

Checkpoint 3.2.3 WebWork. Let's apply the Multiplication Principle.

A fair 6-sided die is rolled 10 times and the resulting sequence of 10 numbers is recorded.

How many different sequences are possible? _____

How many different sequences consist entirely of even numbers? _____

How many different sequences are possible if the first, third, and fourth numbers must be the same? _____

Checkpoint 3.2.4 WebWork.

Find how many positive integers with exactly four decimal digits, that is, positive integers between 1000 and 9999 inclusive, have the following properties:

- (a) are not divisible by either 5 or 7.
- (b) are divisible by 5 but not by 7.
- (c) are divisible by 7.
- (d) are divisible by 5 and by 7.

Definition 3.2.5 Factorial. For any natural number n ,

$$n! = n(n-1)(n-2)\dots 3 \cdot 2 \cdot 1$$

and by convention set $0! = 1$. \diamond

Example 3.2.6 iPad security code. Consider your iPad's security. To unlock the screen suppose you need to enter a four digit pass code. How easy is it to guess this pass code?

Using the standard 10 digit keypad, we first have two questions to consider?

1. Does the order in which the digits are entered matter?
2. Can you reuse a digit more than once?

For the iPad, if the order does matter and you cannot reuse digits, the number of possible codes can be determined by considering each digit as a separate event with four such events in succession providing the right code. By successively applying the multiplication principle, you find that the number of possible codes is the number of remaining available digits at each step. Namely, $10 \times 9 \times 8 \times 7 = 5040$.

On the other hand, if you were allowed to reuse the digits then the number of possible outcomes would be more since all 10 digits would be available for each event. Namely, $10 \times 10 \times 10 \times 10 = 10000$.

Now, consider how this changes if you can use a 4 or 6 digit passcode. Determine the number of possible passcodes. \square

Example 3.2.7 iPad security code with greasy fingers. Reconsider your iPad's security. In this case, you like to eat chocolate bars and have greasy fingers. When you type in your passcode your fingers leave a residue over the four numbers pressed. If someone now tries to guess your passcode, how many possible attempts are necessary?

Since there are only four numbers to pick from with order important, the number of possible passcodes remaining is $4 \times 3 \times 2 \times 1 = 24$ \square

Example 3.2.8 National Treasure. In the 2004 movie National Treasure, Ben and Riley are attempting to guess Abigail's password to enter the room with the Declaration. They are able to determine the passphrase to get into the vault room by doing a scan that detects the buttons pushed (not due to chocolate but just due to the natural oils on fingers). They notice that the

buttons pushed include the characters AEFGLORVY.

Assuming these characters are used only once each, how many possible passphrases are possible?

In this case, the order of the characters matters but all of the characters are distinct. Since we have 9 characters provided, the we can consider each character as an event with the first event as a choice from the 9, the second event as a choice from the remaining 8, etc. This gives $9 \times 8 \times \dots \times 1 = 362880$ possible passphrases.

Assuming that some of the characters could be used more than once, how many passphrases need to be considered if the total length of passphrase can be at most 12 characters?

Notice, in this case you don't know which characters might be reused and so the number of possible outcomes will be much larger. What is the answer?

You can break this problem down into distinct cases:

- Using 9 characters: The answer was computed above.
- Using 10 characters: In this case, 1 character can be used twice. To determine the number of possibilities, let's first pick which character can be doubled. There are 9 options for picking that character. Next, if we consider the two instances of that letter as distinct values then we can just count the number of ways to arrange unique 10 characters which is $10!$. However, swapping the two characters (which are actually identical) would not give a new passphrase. Since these are counted twice, let's divide these out to give $10!/2$.
- Using 11 characters: In this situation we have two unique options:
 - One character is used three times and the others just once.
Continuing as in the previous case, $11!/3!$. Two characters are used twice and the others just once.
- Using 12 characters
 1. One letter from the nine is used four times and all the others are used once.
 2. One letter is used three times, another letter is used two times, and the others are used once.
 3. Three letters are used twice and the others are used once.

With this large collection of possible outcomes, how are the movie characters able to determine the correct "VALLEYFORGE" passphrase? \square

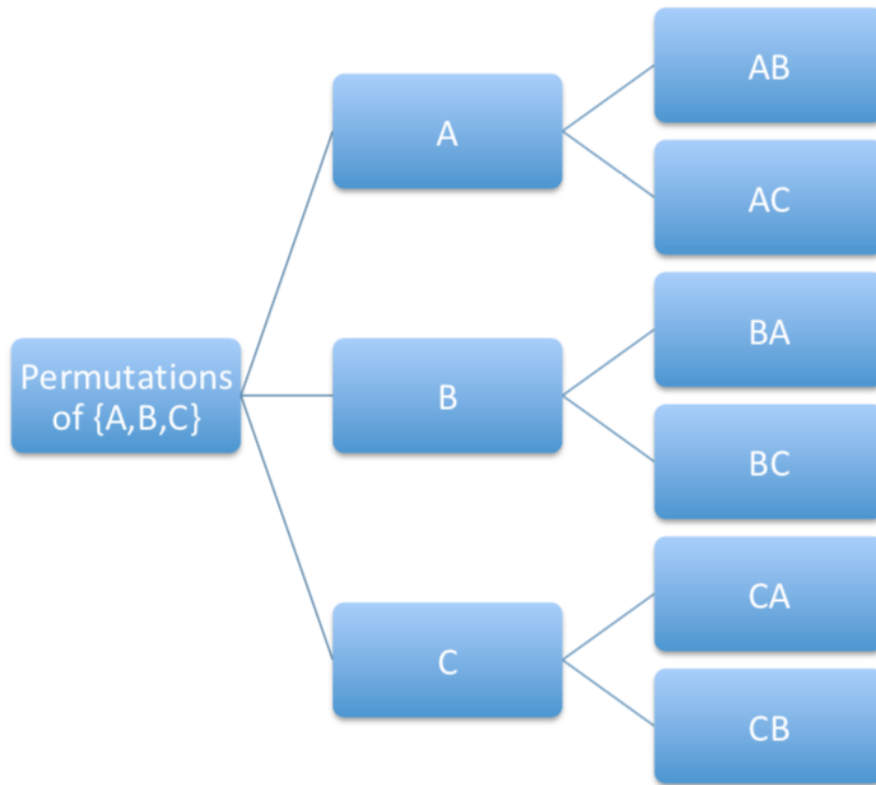
3.3 Permutations

When counting various outcomes the order of things sometimes matters. When the order of a set of elements changes we call the second a permutation (or an arrangement) of the first.

Theorem 3.3.1 Permutations of everything. *The number of permutations of n distinct items is $n!$*

Proof. Notice that if $n=1$, then there is only 1 item to arrange and that there is only one possible arrangement.

By induction, assume that any set with n elements has $n!$ arrangements and



assume that

$$A = \{a_1, a_2, \dots, a_n, a_{n+1}\}.$$

Notice that there are $n+1$ ways to choose 1 element from A and that in doing so leaves a set with n elements. Combining the induction hypothesis with the multiplication principle this gives $(n+1)n! = (n+1)!$ possible outcomes. ■

One can interpret successive ordered selections as branching through a "tree" structure. Indeed, starting with the set A, B, C one may pick any of the three but then a subsequent selection only has two possibilities for the next selection and so forth. The tree below illustrates that there are six ways to order two items from a group of size three.

Going one step further, what about ordering the letters in A, B, C, D ? You can start by picking one of the four letters, say A , and then arranging B, C , and D . Then, start with B and arrange A, C , and D and so on. This gives:

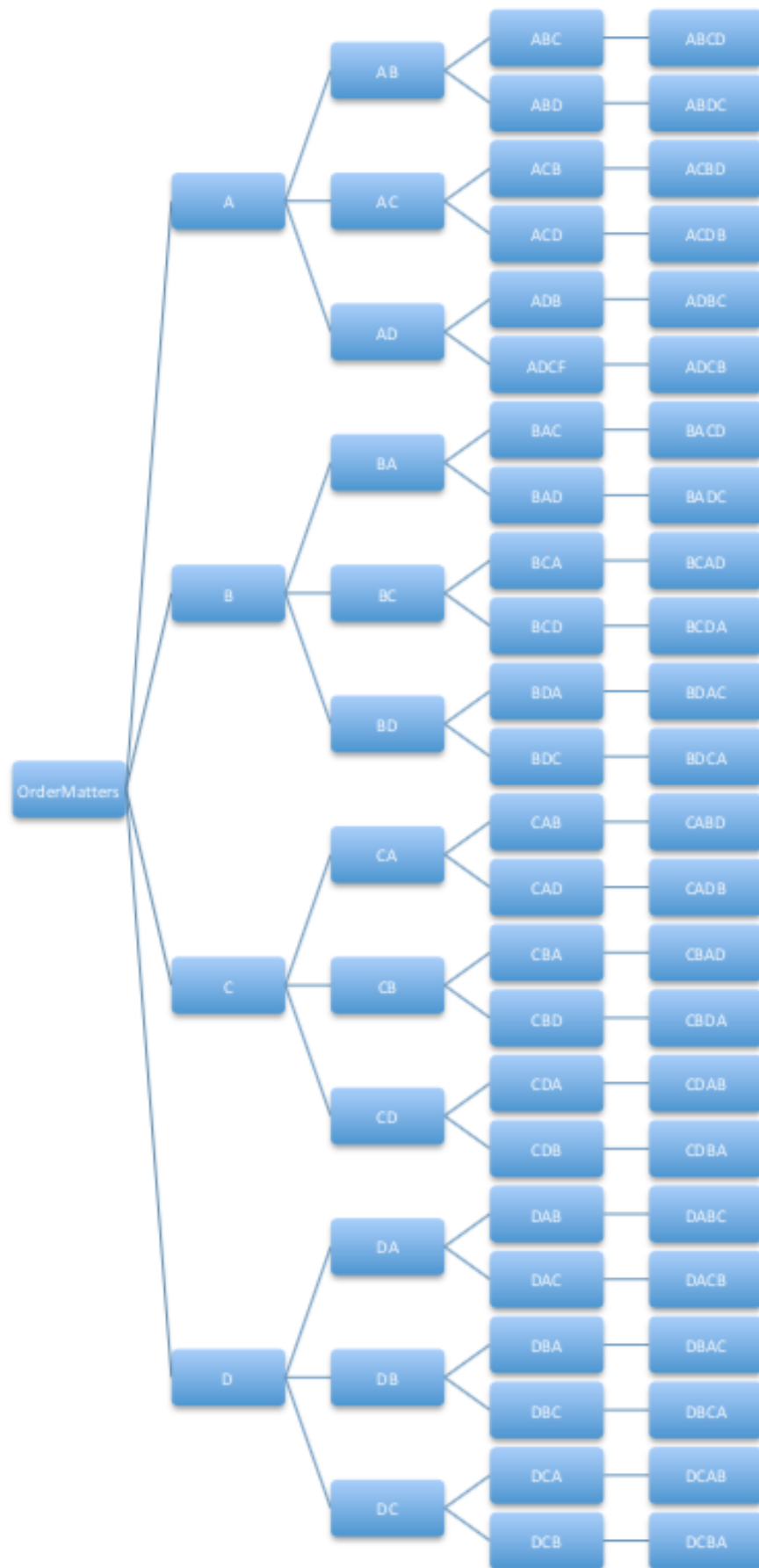
ABCD, ABDC, ACBD, ACDB, ADBC, ADCB
 BACD, BADC, BCAD, BCAB, BDAC, BDCA
 CBAD, CBDA, CABD, CADB, CDBA, CDAB
 DBCA, DBAC, DCBA, DCAB, DABC, DACB

which is 24 different options or $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. This can be viewed using a tree structure where each decision creates a new branch.

Theorem 3.3.2 Permutations of a subset without replacement. *The number of ways to arrange r items from a set of n distinct items is*

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

This is sometimes denoted also as $P(n, r)$ or P_r^n .



Proof. If $r > n$ or $r < 0$ then this is not possible and so the result would be no permutations. Otherwise, apply the multiplication principle r times noting that there are n choices for the first selection, $n-1$ choices for the second selection, and with $n-r+1$ choices for the r th selection. This gives

$$\begin{aligned} {}_nP_r &= n(n-1)\dots(n-r+1) \\ &= n(n-1)\dots(n-r+1) \frac{(n-r)!}{(n-r)!} \\ &= \frac{n(n-1)\dots(n-r+1)(n-r)!}{(n-r)!} \\ &= \frac{n!}{(n-r)!} \end{aligned} \quad \blacksquare$$

Following the tree idea from above, continue for several steps but then stop once you have gone r steps in. For example, it is easy to see that ${}_5P_2 = 20$ using a tree.

Checkpoint 3.3.3 WebWork. Let's apply the the Permutation formula.

In how many ways can 3 different novels, 4 different mathematics books, and 1 biology book be arranged on a bookshelf if

(a) the books can be arranged in any order?

Answer: _____

(b) the mathematics books must be together and the novels must be together?

Answer: _____

(c) the novels must be together but the other books can be arranged in any order?

Answer: _____

So, these are simple calculations.

Theorem 3.3.4 Permutations of a subset with replacement. *The number of ways to obtain an arrangement of r choices from a group of size n is*

$$n^r$$

Proof. Use the multiplication principle r times and see that for each choice all n objects in the universe remain available. That is,

$$n \cdot n \cdot n \dots n = n^r \quad \blacksquare$$

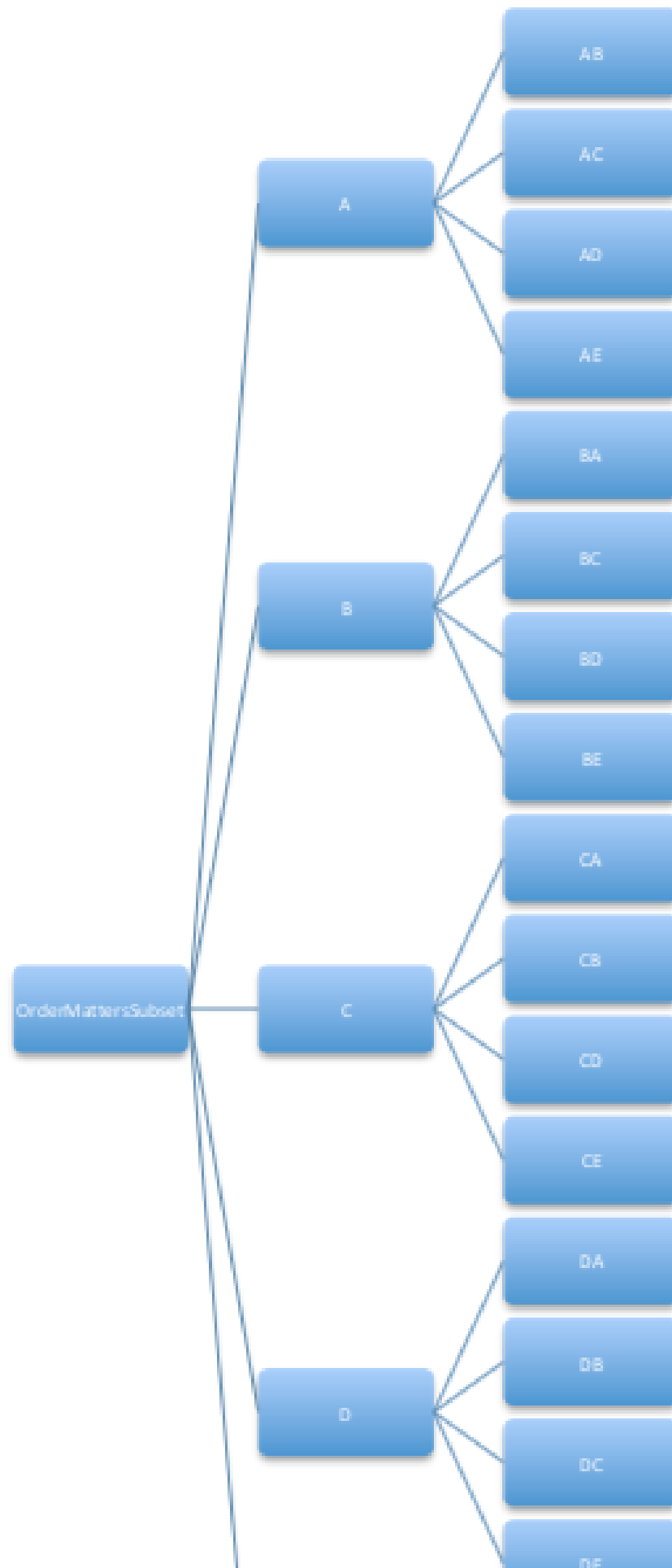
Theorem 3.3.5 Permutations when not all items are distinguishable (Multinomial Coefficients). *If n items belong to s categories, n_1 in first, n_2 in second, \dots , n_s in the last, the number of ways to pick all is*

$$\frac{n!}{n_1! \cdot n_2! \dots n_s!}$$

Proof. Enumerate all of the n data items individually with the $n_1!$ identical values first and the remaining groups in like manner to get the enumerated list

$$x_{1,1}, \dots, x_{1,n_1}, x_{2,1}, \dots, x_{2,n_2}, \dots, x_{s,1}, \dots, x_{s,n_s},$$

In this order, there are $n_1!$ ways to arrange the first group, $n_2!$ ways to arrange the second, etc. There are $n_1! \times n_2! \times \dots \times n_s!$ ways to arrange all of the categories together with groups in this order but none of those group reorders does anything since those data values are all the same. Dividing out those from the $n!$ original permutations of all items leaves one with the multinomial coefficient. \blacksquare



Checkpoint 3.3.6 WebWork. Let's apply the this new Permutation formula.

How many anagrams can be created from the word 'accommodate' if the new words do not need to be meaningful? _____

Another one bites the dust.

3.4 Combinations

When counting various outcomes sometimes the order of things does not matter. If so, each unique unordered outcome is called a combination.

Once again, consider the permutations when selecting three letters from A, B, C, D.

1.
 - A,B,C
 - A,C,B
 - B,A,C
 - B,C,A
 - C,A,B
 - C,B,A
2.
 - A,B,D
 - A,D,B
 - B,A,D
 - B,D,A
 - D,A,B
 - D,B,A
3.
 - A,C,D
 - A,D,C
 - C,A,D
 - C,D,A
 - D,A,C
 - D,C,A
4.
 - B,C,D
 - B,D,C
 - C,B,D
 - C,D,B
 - D,B,C
 - D,C,B

Notice how these 24 permutations fall into only four distinct categories if the order does not matter. Therefore, from a group of size four you can pick an unordered subset of size three in only 4 ways rather than the original 24.

In general, it would be nice to have a direct formula to determine the number of such combinations without having to explicitly list them all out.

Theorem 3.4.1 Combinations of a subset without replacement. *The*

number of ways to arrange r items from a set of n distinct items is

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

This is sometimes denoted $C(n, r)$ or C_r^n or $\binom{n}{r}$.

Proof. Consider creating a permutation of r objects from a set of size n by first picking an unordered subset of size r and then counting the number of ways to order that subset. Using our notation and the multiplication principle,

$${}_nP_r = {}_nC_r \cdot r!$$

Dividing by $r!$ gives the result. ■

Theorem 3.4.2 Combinations of a subset with replacement. *The number of ways to arrange r items from a set of n distinct items is*

$${}_{n+r-1}C_r = \binom{n+r-1}{r} = \frac{(r+n-1)!}{r!(n-1)!}$$

Proof. Label each item in your group in some defined order. Since order doesn't matter, as you repeatedly sample r times with replacement you can always write down your outcomes sorted from low to high placement. Finally, separate like values by some symbol, say "|", and consider each of the n distinct objects as indistinct $*$'s. There will be $n-1$ of these separators since there will be n to choose from. For example, if choosing $r=6$ times from the set a, b, c, d , then the outcome b, b, a, d, a, b could be collected as a, a, b, b, b, d and written in our code as `**|***|*`. Notice that shuffling around the identical $*$'s would not change the code (and similarly for the identical $|$'s) but swapping a $*$ with a $|$ would be a different outcome. Therefore, we can consider this to be a [multinomial coefficient 3.3.5](#) and the number of ways to rearrange this code is

$$\frac{(r+n-1)!}{r!(n-1)!}.$$
■

Checkpoint 3.4.3 WebWork. Let's apply the this new Combination formula.

A standard deck of cards consists of four suits (clubs, diamonds, hearts, and spades), with each suit containing 13 cards (ace, two through ten, jack, queen, and king) for a total of 52 cards in all.

How many 7-card hands will consist of exactly 4 hearts and 3 clubs?

Notice that to determine the number of outcomes required you to use the combination formula several times and then multiply the results using the multiplication principle.

Checkpoint 3.4.4 WebWork. Use the new Combination formula again.

A school dance committee is to consist of 2 freshmen, 3 sophomores, 4 juniors, and 5 seniors. If 6 freshmen, 8 sophomores, 7 juniors, and 8 seniors are eligible to be on the committee, in how many ways can the committee be chosen?

Your answer is : _____

Solution. There are $\binom{6}{2}$ ways to choose 2 freshmen for the committee, $\binom{8}{3}$ ways to choose 3 sophomores for the committee, $\binom{7}{4}$ ways to choose 4 juniors for the committee, and $\binom{8}{5}$ ways to choose 5 seniors for the committee. So by

the generalized basic principle of counting, there are a total of

$$\binom{6}{2} \cdot \binom{8}{3} \cdot \binom{7}{4} \cdot \binom{8}{5} = \frac{6!}{2!4!} \cdot \frac{8!}{3!5!} \cdot \frac{7!}{4!3!} \cdot \frac{8!}{5!3!} = 1646400$$

different possible committees.

Once again, you can see that using the formulas can be easy and also can be part of a bigger problem pasted together using the multiplication principle.

Example 3.4.5 Ipad Security. Revisiting your ipad's security, what happens if the order in which the digits are entered does not matter? If so, then you would be picking a combination of 4 digits without replacement from a group of 10 digits. Namely,

$$\begin{aligned} \frac{10!}{4!6!} &= \frac{10 \times 9 \times 8 \times 7 \times 6!}{4 \times 3 \times 2 \times 1 \times 6!} \\ &= \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} \\ &= \frac{5040}{24} \\ &= 210. \end{aligned}$$

Notice that the total number of options is much smaller when order does not matter.

Note that if you were allowed to reuse the digits then the number of possible outcomes would be

$$\begin{aligned} \frac{13!}{4!9!} &= \frac{13 \times 12 \times 11 \times 10}{4 \times 3 \times 2 \times 1} \\ &= 715 \end{aligned}$$

which once again is more since numbers are allowed to repeat. \square

Definition 3.4.6 Binomial Coefficients. The value ${}_nC_r$ is known as the binomial coefficient. It is denoted by $\binom{n}{r}$ and is read "n choose k". \diamond

Binomial coefficients have a number of interesting properties. Many of these are very useful as well in probability calculations. Several of these properties are collected below. In particular, these relationships verify that the binomial coefficients are the values found in Pascal's Triangle.

```
@interact
def _(n = slider(1,15,1,5)):
    for row in range(n+1):
        binoms = sorted(binomial_coefficients(row).items())
        given_n = []
        for k in range(row+1):
            given_n.append(binoms[k][1])
        pretty_print('%s'%given_n)
```

Theorem 3.4.7 Binomial Coefficient Formulas. For $n \in \mathbb{N}$,

1. $\binom{n}{0} = 1$
2. $\binom{n}{n} = 1$
3. $\binom{n}{1} = n$
4. $\binom{n}{n-1} = n$

5. $\binom{n}{r} = \binom{n}{n-r}$
6. $\binom{n+1}{r+1} = \binom{n}{r} + \binom{n}{r+1}$

Proof.

1. $\binom{n}{0} = \frac{n!}{0!(n-0)!} = 1$
2. $\binom{n}{n} = \frac{n!}{n!(n-n)!} = 1$
3. $\binom{n}{1} = \frac{n!}{1!(n-1)!} = n$
4. $\binom{n}{n-1} = \frac{n!}{(n-1)!(n-(n-1))!} = n$
5. $\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{(n-r)!(n-(n-r))!} = \binom{n}{n-r}$
- 6.

$$\begin{aligned}
 \binom{n}{r} + \binom{n}{r+1} &= \frac{n!}{r!(n-r)!} + \frac{n!}{(r+1)!(n-(r+1))!} \\
 &= (r+1) \frac{n!}{(r+1)!(n-r)!} // + (n-r) \frac{n!}{(r+1)!(n-r)!} \\
 &= \frac{(r+1)n! + (n-r)n!}{(r+1)!(n-r)!} \\
 &= \frac{(n+1)n!}{(r+1)!((n+1)-(r+1))!} \\
 &= \binom{n+1}{r+1}
 \end{aligned}$$

■

3.5 Summary

Here are the important results from this chapter: [Multiplication Principle 3.2.2](#)

[Factorial 3.2.5](#)

[Permutations without replacement 3.3.2](#)

[Permutations with replacement 3.3.4](#)

[Multinomial Coefficients 3.3.5](#)

[Combinations without replacement 3.4.1](#)

[Combinations with replacement 3.4.2](#)

[Binomial Coefficients 3.4.6](#)

3.6 Exercises

Complete the online homework "Counting".

A standard deck of playing cards consists of 52 cards broken up into four "suits" known as Hearts, Spades, Diamonds, and Clubs. Each suit is broken up additionally into unique cards with "face values" from 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace and generally in that order from low to high.

1. Pick two cards without replacement one after the other from this deck and determine the following number of possible outcomes:

- The number of ways to get an Ace for both cards.

- The number of ways to get an Ace for only one of the two cards.
 - The number of ways to get an Ace on the first draw and a Spade on the second draw.
2. Pick five cards without replacement one after the other from a newly shuffled full deck and determine the following number of possible outcomes:
- All cards have different faces
 - "A pair". That is, two cards have the same face but the others are from three other faces.
 - "Three of a kind". That is, three cards have the same face but the others are from two other faces.
 - "Two Pair". That is, two cards come from one face, two other cards come from a common face that is not the same as the first two cards, and the last card comes from some other face.
 - "Full House". That is, three cards have the same face and the other two come from a common face that is not the same as the first three cards.
 - "Four of a Kind". That is, four cards have the same face and the other card comes from some other face.
 - "Flush". That is, the five cards form a sequence in order of adjacent faces in the original list and from the same suit.
 - "Royal Flush". That is, a flush but only with the cards Ace, King, Queen, Jack, 10.

Completely determine the number of possible passphrases for the National Treasure example started above. Present your answer in a report form.

Chapter 4

Probability Theory

4.1 Introduction

Mathematics generally focuses on providing precise answers with absolute certainty. For example, solving an equation generates specific (and non-varying) solutions. Statistics on the other hand deals with providing precise answers to questions when there is uncertainty. It might seem impossible to provide such precise answers but the focus of this text is to show how that can be done so long as the questions are properly posed and the answers properly interpreted.

Indeed, people often make claims about being the biggest, best, most often recommended, etc. One sometimes even believes these claims based upon subjective metrics. In this chapter, we will start by looking at relative frequency and notice several properties regarding relative frequencies as the number of trials increases. We will use these examples to motivate a definition for probability and investigate the resulting consequences of that definition.

4.2 Relative Frequency

When attempting to precisely measure uncertainty one often resorts to examples or experiments that model the theoretical question of interest. Before we investigate statistical experiments, we need to create some notation that we will utilize throughout the rest of this text.

- S = Universal Set or Sample Space Experiment or Outcome Space. This is the collection of all possibilities.
- Random Experiment. A random experiment is a repeatable activity that has more than one possible outcome all of which can be specified in advance but can not be known in advance with certainty.
- Trial. Performing a Random Experiment one time and measuring the result.
- A = Event. A collection of outcomes. Generally denoted by an upper case letter such as A , B , C , etc.
- Success/Failure. When recording the result of a trial, a success for event A occurs when the outcome lies in A . If not, then the trial was a failure. There is no qualitative meaning to this term.
- Mutually Exclusive Events. Two events that share no common outcomes. Also known as disjoint events.

- $|A|$ = Frequency. In a sequence of n events, the frequency is the number of trials which resulted in a success for event A .
- $|A| / n$ = Relative Frequency. A proportion of successes to total number of trials.
- Histogram. A bar chart representation of data where area corresponds to the value being described.

To investigate these terms and to motivate our discussion of probability, consider flipping coins using the interactive cell below. Notice in this case, the sample space $S = \{ \text{Heads, Tails} \}$ and the random experiment consists of flipping a fair coin one time. Each trial results in either a Head or a Tail. Since we are measuring both Heads and Tails then we will not worry about which is a success or failure. Further, on each flip the outcomes of Heads or Tails are mutually exclusive events. We count the frequencies and compute the relative frequencies for a varying number of trials selected by you as you move the slider bar. Results are displayed using a histogram.

```
coin = ["Heads", "Tails"]
@interact
def _(num_rolls = slider([5..5000],label="Number_of_Flips")):
    rolls = [choice(coin) for roll in range(num_rolls)]
    show(rolls)
    freq = [0,0]
    for outcome in rolls:
        if (outcome=='Tails'):
            freq[0] = freq[0]+1
        else:
            freq[1] = freq[1]+1
    print("\nThe frequency of tails = " + str(freq[0])) + "
        and heads = " + str(freq[1]) + "."
    rel = [freq[0]/num_rolls, freq[1]/num_rolls]
    print("\nThe relative frequencies for Tails and
        Heads: " + str(rel))
    show(bar_chart(freq, axes=False, ymin=0))      # A
        histogram of the results
```

Question 1: What do you notice as the number of flips increases?

Question 2: Why do you rarely (if ever) get exactly the same number of Heads and Tails? Would you not "expect" that to happen?

You should have noticed that as the number of flips increases, the relative frequency of Heads (and Tails) stabilized around 0.5. This makes sense intuitively since there are two options for each individual flip and $1/2$ of those options are Heads while the other $1/2$ is Tails.

Let's try again by doing a random experiment consisting of rolling a single die one time. Note that the sample space in this case will be the outcomes $S = \{ 1, 2, 3, 4, 5, 6 \}$.

```
@interact
def _(num_rolls = slider([20..5000],label='Number_of_
    rolls'), Number_of_Sides = [4,6,8,12,20]):
    die = list((1..Number_of_Sides))
    rolls = [choice(die) for roll in range(num_rolls)]
    show(rolls)
```

```

freq = [rolls.count(outcome) for outcome in
        set(die)] # count the numbers for each outcome
print 'The_frequencies_of_each_outcome_is_' + str(freq)

print 'The_relative_frequencies_of_each_outcome:'
rel_freq = [freq[outcome-1]/num_rolls for outcome in
            set(die)] # make frequencies relative
print rel_freq
fs = []
for f in rel_freq:
    fs.append(f.n(digits=4))
print fs
show(bar_chart(freq, axes=False, ymin=0))

```

Notice for a single die there are a larger number of options (for example 6 on a regular die) but once again the relative frequencies of each outcome was close to $1/n$ (i.e. $1/6$ for the regular die) as the number of rolls increased.

In general, this suggests a rule: if there are n outcomes and each one has the same chance of occurring on a given trial then on average on a large number of trials the relative frequency of that outcome is $1/n$. In general, if a number of outcomes are "equally likely" then this is a good model for measuring the proportion of outcomes that would be expected to have any given outcome. However, it is not always true that outcomes are equally likely. Consider rolling two die and measuring their sum:

```

@interact
def _(num_rolls = slider([20..5000], label='Number_of_
    rolls'), num_sides = slider(4, 20, 1, 6, label='Number_of_
    sides')):
    die = list((1..num_sides))
    dice = list((2..num_sides*2))
    rolls = [(choice(die), choice(die)) for roll in
              range(num_rolls)]
    sums = [sum(rolls[roll]) for roll in range(num_rolls)]
    show(rolls)

    freq = [sums.count(outcome) for outcome in set(dice)] #
        count the numbers for each outcome
    print 'The_frequencies_of_each_outcome_is_' + str(freq)

    print 'The_relative_frequencies_of_each_outcome:'
    rel_freq = [freq[outcome-2]/num_rolls for outcome in
                set(dice)] # make frequencies relative
    print rel_freq
    show(bar_chart(freq, axes=False, ymin=0)) # A
        histogram of the results
    print "Relative_Frequency_of_", dice[0], "_is_about_",
        rel_freq[0].n(digits=4)
    print "Relative_Frequency_of_", dice[num_sides-1], "_is_
        about_", rel_freq[num_sides-1].n(digits=4)

```

Question 1: What do you notice as the number of rolls increases?

Question 2: What do you expect for the relative frequencies and why are they not all exactly the same?

Notice, not only are the answers not the same but they are not even close. To understand why this is different from the examples before, consider the possible outcomes from each pair of die. Since we are measuring the sum of

the dice then (for a pair of standard 6-sided dice) the possible sums are from 2 to 12. However, there is only one way to get a 2—namely from a (1,1) pair—while there are 6 ways to get a 7—namely from the pairs (1,6), (2,5), (3,4), (4,3), (5,2), and (6,1). So it might make some sense that the likelihood of getting a 7 is 6 times larger than that of getting a 2. Check to see if that is the case with your experiment above.

Play with the following several times to investigate what you might expect to get when you repeatedly receive a "hand" of 5 standard playing cards. Can you imagine how you might possibly enumerate the entire list of possible outcomes by hand? However, using this interactive cell, you can shuffle and deal 5-card hands over and over easily and then count the number of special poker outcomes.

```
var('A_C_D_H_J_K_Q_S')

suits = [S, D, C, H]
values = [2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A]

full_deck = [(value, suit) for suit in suits for value in
              values]
@interact
def _(num_hands=slider[50..5000]): # Set up
    the number of hands to create
    hands= [] # Start with a blank list.
    for i in range(num_hands): # This loops the following
        operation num_hands times.
        deck = copy(full_deck) # start over
        shuffle(deck)
        hands.append([deck.pop() for card in range(5)])
    freq_values = []
    one_pair = 0
    two_pair = 0
    three_kind = 0
    full_house = 0
    four_kind = 0
    for i in range(num_hands):
        hand = hands[i]
        hand_values = [hand[k][0] for k in range(5)]
        freq_values = [hand_values.count(value) for value in
                       set(values)]
        freq_values.sort(reverse=True)
        if freq_values[0]==4:
            four_kind=four_kind+1
        if freq_values[0]==3:
            if freq_values[1]==2:
                full_house=full_house+1
            if freq_values[1]==1:
                three_kind=three_kind+1
        if freq_values[0]==2:
            if freq_values[1]==2:
                two_pair=two_pair+1
            if freq_values[1]==1:
                one_pair=one_pair+1
    print "One_Pair_frequency=", one_pair, "_with_"
    print "relative_frequency=", one_pair/num_hands
    print "Two_Pair_frequency=", two_pair, "_with_"
    print "relative_frequency=", two_pair/num_hands
```

```

print "Three_of_a_Kind_frequency=", three_kind, "_with_
relative_frequency=", three_kind/num_hands
print "Full_House_frequency=", full_house, "_with_
relative_frequency=", full_house/num_hands
print "Four_of_a_Kind_frequency=", four_kind, "_with_
relative_frequency=", four_kind/num_hands

```

Sometimes you will find it useful to keep a running total of the relative frequencies. Such a cumulative approach is often called a distribution function.

Definition 4.2.1 Cumulative relative frequency. For a collection of ordered events $x_1 < x_2 < \dots < x_s$ with corresponding frequencies f_1, f_2, \dots, f_s , the cumulative relative frequency is the function

$$F(x) = \sum_{x_k \leq x} f_{x_k}$$

◇

Let's consider the cumulative relative frequency with the sum of dice example seen at the beginning of this chapter.

```

@interact
def _(num_rolls = slider([20..5000], label='Number_of_
rolls'), num_sides = slider(4, 20, 1, 6, label='Number_of_
sides')):
    die = list((1..num_sides))
    dice = list((2..num_sides*2))
    rolls = [(choice(die), choice(die)) for roll in
              range(num_rolls)]
    sums = [sum(rolls[roll]) for roll in range(num_rolls)]
    show(rolls)

    freq = [sums.count(outcome) for outcome in set(dice)] #
    count the numbers for each outcome
    n = len(freq)
    CF = freq
    for k in range(1, n):
        CF[k] = freq[k] + CF[k-1]

    print 'The_cumulative_relative_frequencies_of_each_
outcome:'
    Crel_freq = [CF[outcome-2]/num_rolls for outcome in
                 set(dice)] # make frequencies relative
    print Crel_freq
    show(bar_chart(CF, axes=False, ymin=0)) # A histogram
    of the results
    print "Cumulative_Relative_Frequency_of_", dice[0], "_is_
about_", Crel_freq[0].n(digits=4)
    print "Cumulative_Relative_Frequency_of_
", dice[num_sides-1], "_is_about_
", Crel_freq[num_sides-1].n(digits=4)

```

4.3 Definition of Probability

Relative frequency gives a way to measure the proportion of "successful" outcomes when doing an experimental approach. From the interactive applications above, it appears that the relative frequency does jump around as the experiment is repeated but that the amount of variation decreases as the number

of experiments increases. This is known to be true in general and is known as the "Law of Large Numbers".

We would like to formalize what these relative frequencies are approaching and will call this theoretical limit the "probability" of the outcome. In doing so, we will do our best to model our definition so that it follow the behavior of relative frequency.

To generate a general definition for probability, we need to know what is is that we measuring. In general, we will be finding the probability of sets of possible outcomes...that is, a subset of the Sample Space S . Toward that end, it is important to briefly look at some properties of sets.

Definition 4.3.1 Pairwise Disjoint Sets. $\{A_1, A_2, \dots, A_n\}$ are pairwise disjoint provided $A_k \cap A_j = \emptyset$ so long as $k \neq j$. Disjoint sets as also often called mutually exclusive. \diamond

Play around with the interactive cell below by adding and removing items in each of the three sets. Find elements so that the intersection of all three sets is empty but at least one of the paired sets are not disjoint. See if you can make all of the paired sets not disjoint but the intersection of all three disjoint. This is why we need to consider "pairwise" disjoint sets.

```
def f(s, braces=True):
    t = ', '.join(sorted(list(s)))
    if braces: return '{' + t + '}'
    return t
def g(s): return set(str(s).replace(',', '_').split())

@interact
def _(X='1,2,3', Y='2,a,3,4,apple', Z='a,b,10,apple'):
    S = [g(X), g(Y), g(Z)]
    X,Y,Z = S
    XY = X & Y
    XZ = X & Z
    YZ = Y & Z
    XYZ = XY & Z

    Txy = "_NOT_disjoint_"
    if Set(XY).is_empty():
        Txy = '_disjoint_'
    pretty_print(html("$X \cap Y = %s" % f(XY) + "%s" % Txy))
    Txz = "_NOT_disjoint_"
    if Set(XZ).is_empty():
        Txz = '_disjoint_'
    pretty_print(html("$X \cap Z = %s" % f(XZ) + "%s" % Txz))
    Tyz = "_NOT_disjoint_"
    if Set(YZ).is_empty():
        Tyz = '_disjoint_'
    pretty_print(html("$Y \cap Z = %s" % f(YZ) + "%s" % Tyz))
    Txyz = "_NOT_disjoint_"
    if Set(XYZ).is_empty():
        Txyz = '_disjoint_'
    pretty_print(html("$X \cap Y \cap Z = %s" % f(XYZ) + "%s" % Txyz))
    centers = [(cos(n*2*pi/3), sin(n*2*pi/3)) for n in
               [0,1,2]]
    scale = 1.7
    clr = ['yellow', 'blue', 'green']
    G = Graphics()
```



```

for i in range(len(S)):
    G += circle(centers[i], scale, rgbcolor=clr[i],
               fill=True, alpha=0.3)
for i in range(len(S)):
    G += circle(centers[i], scale, rgbcolor='black')

# Plot what is in one but neither other
for i in range(len(S)):
    Z = set(S[i])
    for j in range(1, len(S)):
        Z = Z.difference(S[(i+j)%3])
    G += text(f(Z, braces=False),
              (1.5*centers[i][0], 1.7*centers[i][1]),
              rgbcolor='black')

# Plot pairs of intersections
for i in range(len(S)):
    Z = (set(S[i]) & S[(i+1)%3]) - set(XYZ)
    C = (1.3*cos(i*2*pi/3 + pi/3), 1.3*sin(i*2*pi/3 +
        pi/3))
    G += text(f(Z, braces=False), C, rgbcolor='black')

# Plot intersection of all three
G += text(f(XYZ, braces=False), (0,0), rgbcolor='black')

# Show it
G.show(aspect_ratio=1, axes=False)

```

Consider how we might create a definition for the expectation of a given outcome. To do so, first consider a desired collection of outcomes A . If each outcome in A is chosen randomly then we might consider using a formula similar to relative frequency and set a measure of expectation to be $|A|/|S|$. For example, on a standard 6-sided die, the expectation of the outcome $A=2$ from the collection $S = 1,2,3,4,5,6$ could be $|A|/|S| = 1/6$.

From our example where we take the sum of two die, the outcome $A = \{4,5\}$ from the collection $S = 2,3,4,\dots,12$ would be

$$\begin{aligned}
 |A| &= |\{(1,3), (2,2), (3,1), (1,4), (2,3), (3,2), (4,1)\}| = 7 \\
 |S| &= |\{(1,1), \dots, (1,6), (2,1), \dots, (2,6), \dots, (6,1), \dots, (6,6)\}| = 36
 \end{aligned}$$

and so the expected relative frequency would be $|A|/|S| = 7/36$. Compare this theoretical value with the sum of the two outcomes from your experiment above.

We are ready to now formally give a name to the theoretical measure of expectation for outcomes from an experiment. Taking our cue from our examples, let's make our definition agree with the following relative frequency properties:

1. Relative frequency cannot be negative, since cardinality cannot be negative
2. Relative frequencies for disjoint events should sum to one
3. Relative frequencies for collections of disjoint outcomes should equal the sum of the individual relative frequencies

which leads us to the following formal definition...

Definition 4.3.2 Probability. The probability $P(A)$ of a given outcome A is a set function that satisfies:

1. (Nonnegativity) $P(A) \geq 0$
2. (Totality) $P(S) = 1$
3. (Subadditivity) If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$. In general, if A_k are pairwise disjoint then $P(\cup_k A_k) = \sum_k P(A_k)$.

◇

Checkpoint 4.3.3 Using the definition above, determine the following probabilities.

Suppose you select a letter at random from the word MISSISSIPPI.

The probability of selecting the letter S is _____

The probability of selecting the letter M is _____

The probability of selecting the letters P or I is _____

The probability of not selecting the letter I is _____

Hint. Count the number of letters in the word. When computing each probability, this is the number that goes on the bottom.

Solution. The probability of 'M's is $1/11$.

The probability of 'I's is $4/11$.

The probability of 'S's is $4/11$.

The probability of 'P's is $2/11$.

Notice when you are given complete information regarding the entire data set then determining probabilities for events can be relatively easy to compute.

Based upon this definition we can immediately establish a number of results.

Theorem 4.3.4 Probability of Complements. $P(A) + P(A^c) = 1$

Proof. Let A be any event and note that

$$A \cap A^c = \emptyset.$$

But $A \cup A^c = S$. So, by subadditivity

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

as desired. ■

Theorem 4.3.5 $P(\emptyset) = 0$

Proof. Note that $\emptyset^c = S$. So, by the theorem above,

$$1 = P(S) + P(\emptyset) \Rightarrow 1 = 1 + P(\emptyset).$$

Cancelling the 1 on both sides gives $P(\emptyset) = 0$. ■

Theorem 4.3.6 $A \subset B, P(A) \leq P(B)$

Proof. Assume sets A and B satisfy $A \subset B$. Then, notice that

$$A \cap (B - A) = \emptyset$$

and

$$B = A \cup (B - A).$$

Therefore, by subadditivity and nonnegativity

$$\begin{aligned} 0 &\leq P(B - A) \\ P(A) &\leq P(A) + P(B - A) \\ P(A) &\leq P(B) \end{aligned} \quad \blacksquare$$

Theorem 4.3.7 $P(A) \leq 1$

Proof. Notice $A \subset S$. By the theorem above $P(A) \leq P(S) = 1$ \blacksquare

Theorem 4.3.8 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof. Notice that we can write $A \cup B$ as the disjoint union

$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A).$$

We can also write disjointly

$$\begin{aligned} A &= (A - B) \cup (A \cap B) \\ B &= (A \cap B) \cup (B - A) \end{aligned}$$

Hence,

$$\begin{aligned} P(A) + P(B) - P(A \cap B) &= [P(A - B) + P(A \cap B)] \\ &\quad + [P(A \cap B) + P(B - A)] - P(A \cap B) \\ &= P(A - B) + P(A \cap B) + P(B - A) \\ &= P(A \cup B) \end{aligned} \quad \blacksquare$$

This result can be extended to more than two sets using a property known as inclusion-exclusion. The following two theorems illustrate this property and are presented without proof.

Corollary 4.3.9 For any sets A , B and C ,

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C) \end{aligned}$$

Corollary 4.3.10 For any sets A , B , C and D ,

$$\begin{aligned} P(A \cup B \cup C \cup D) &= P(A) + P(B) + P(C) + P(D) \\ &\quad - P(A \cap B) - P(A \cap C) - P(A \cap D) \\ &\quad - P(B \cap C) - P(B \cap D) - P(C \cap D) \\ &\quad + P(A \cap B \cap C) + P(A \cap B \cap D) \\ &\quad + P(A \cap C \cap D) + P(B \cap C \cap D) \\ &\quad - P(A \cap B \cap C \cap D) \end{aligned}$$

Many times, you will be dealing with making selections from a sample space where each item in the space has an equal chance of being selected. This may happen (for example) when items in the sample space are of equal size or when selecting a card from a completely shuffled deck or when coins are flipped or when a normal fair die is rolled.

It is important to notice that not all outcomes are equally likely—even in times when there are only two of them. Indeed, it is generally not an equally

likely situation when picking the winner of a football game which pits, say, the New Orleans Saints professional football team with the New Orleans Home School Saints. Even though there are only two options the probability of the professional team winning in most years ought to be much greater than the chances that the high school will prevail.

When items are equally likely (sometimes also called "randomly selected") then each individual event has the same chance of being selected as any other. In this instance, determining the probability of a collection of outcomes is relatively simple.

Theorem 4.3.11 Probability of Equally Likely Events. *If outcomes in S are equally likely, then for $A \subset S$,*

$$P(A) = \frac{|A|}{|S|}.$$

Proof. Enumerate $S = x_1, x_2, \dots, x_{|S|}$ and note $P(\{x_k\}) = c$ for some constant c since each item is equally likely. However, using each outcome as a disjoint event and the definition of probability,

$$\begin{aligned} 1 = P(S) &= P(\{x_1\} \cup \{x_2\} \cup \dots \cup \{x_{|S|}\}) \\ &= P(\{x_1\}) + P(\{x_2\}) + \dots + P(\{x_{|S|}\}) \\ &= c + c + \dots + c = |S| \times c \end{aligned}$$

and so $c = \frac{1}{|S|}$. Therefore, $P(\{x_k\}) = \frac{1}{|S|}$.

Hence, with $A = a_1, a_2, \dots, a_{|A|}$, breaking up the disjoint probabilities as above gives

$$\begin{aligned} P(A) &= P(\{a_1\} \cup \{a_2\} \cup \dots \cup \{a_{|A|}\}) \\ &= P(\{a_1\}) + P(\{a_2\}) + \dots + P(\{a_{|A|}\}) \\ &= \frac{1}{|S|} + \frac{1}{|S|} + \dots + \frac{1}{|S|} \\ &= \frac{|A|}{|S|} \end{aligned}$$

as desired. ■

```
var('A_C_D_H_J_K_Q_S')

def L(str):
    n = len(str)
    m = int(n/5)
    top = m+1
    if m == n/5:
        top = m
    for k in range(top):
        print str[5*k:5*k+5]

suits = [S, D, C, H]
values = [2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A]

deck = [(value, suit) for suit in suits for value in values]
full_deck = copy(deck) # to save a copy of the original
                        deck for later use.
```

```

L(deck)
shuffle(deck)
L(deck)
deck1 = copy(full_deck)
shuffle(deck1)

@interact
def _(auto_update=False):
    global deck1
    shuffle(deck1)
    if (Set(deck1).cardinality() < 5):
        print 'Deck_is_too_small...getting_a_new_deck'
        deck1 = copy(full_deck)
    else:
        hand = [deck1.pop() for card in range(5)]
        print "The_cards_dealt:"
        L(hand)
        print
        print "_The_remaining_cards_in_the_deck:"
        L(deck1)
        print
        print(html("\n_The_number_of_remaining_cards_in_the_
                    deck=_s"%str(Set(deck1).cardinality()))))

```

Checkpoint 4.3.12 WebWork. Let's see if you understand the relationship between frequency and relative frequency. In this exercise, presume "Probability" to be the expected fraction of outcomes you might logically expect.

A fun size bag of M

*amp*Ms has about 18 candies. You open one of the bags and discover:

3 Blues, 4 Yellows, 5 Browns, 2 Reds and 4 Greens.

The probability of choosing a brown is _____.

The odds in favor of choosing a yellow is _____

The probability of choosing either a blue or a red is _____

The odds against a green being chosen is _____

Hint. Odds in favor of an event = number of favorable outcomes / number of unfavorable outcomes.

Odds against an event = number of unfavorable outcomes / number of favorable outcomes.

So, these are simple calculations.

Checkpoint 4.3.13 WebWork. This one is a little harder and uses the binomial coefficients from Combinatorics.

(a) Count the number of ways to arrange a sample of 5 elements from a population of 10 elements. NOTE: Order is not important.

answer: _____

(b) If random sampling is to be employed, the probability that any particular sample will be selected is _____

Notice how the probabilities look similar to relative frequencies. It's just the case that you are counting ALL of the individual simple possibilities that lead to a success.

4.4 Exercises

Checkpoint 4.4.1 Poker. Determine the probabilities associated with the various 5-card hands. That is

1. $P(\text{one pair})$
2. $P(\text{two pair})$
3. $P(\text{three of a kind})$
4. $P(\text{full house})$
5. $P(\text{four of a kind})$
6. $P(\text{straight})$
7. $P(\text{flush})$
8. $P(\text{royal flush})$

Checkpoint 4.4.2 Dice. Determine the 36 possible outcomes related to the rolling a pair of fair dice. Justify why each of these outcomes is equally likely. Determine the probabilities associated with each possible sum.

Solution. Remember, when using equally likely outcomes $|A|/|S|$ assumes that the items counted for A are also in the sample space S . In this case, for example, to determine the Probability of getting a sum of (say) 4 includes the rolls (1,3), (2,2), and (3,1). These three "successes" from the 36 possible ordered pairs gives $P(4) = 3/36$. Similarly, $P(5) = |\text{dice rolls with a sum of 5}|/36 = |(1,4), (2,3), (3,2), (4,1)| / 36 = 4/36$. Continue in this manner to determine the other possibilities and then compare to the experimental sage cell seen earlier for [the sum of two dice](#) .

Checkpoint 4.4.3 Skew Dice. Suppose you have one die which only has three possible sides labeled 1, 2, or 3. Suppose a second die has twelve equally likely sides with labels 1,2,3,4,4,5,5,6,6,7,8,9. Justify that the probabilities associate with each possible sum is the same as the probabilities when using two normal 6-sided dice.

Solution. Consider the outcome space

$$S = (1, 1), (1, 2), (1, 3), (1, 4), (1, 4), (1, 5), (1, 5), (1, 6), (1, 6), (1, 7), (1, 8), (1, 9), (2, 1) \dots (3, 9)$$

Then $P(5) = |(1,4), (1,4), (2,3), (3,2)|/36 = 4/36$. Compare this to the exercise with regular dice performed above. Similarly, compute the remaining probabilities.

Checkpoint 4.4.4 Craps. Analyze the dice game known as "craps": Roll a pair of dice and consider the sum. If that sum is 7 or 11, the one who rolls wins and can roll again. If the sum is 2, 3, or 12 – known as craps – the one who rolls loses but keeps the dice. For any other outcome (called the "point"), the one who rolls continues hoping to roll the point value again before rolling a 7. If successful, then the roller wins and starts the game anew. If a 7 appears first, the roller loses and the next person gets to be the roller.

So, a win can be obtained in two ways: 7 or 11 on first roll or getting the point before the 7 thereafter. Therefore, determine the probability of a win and the probability of a loss.

Solution. ["craps"](#) .

4.5 Conditional Probability

When finding the probability of an event, sometimes you may need to consider past history and how it might affect things. Indeed, you might think that when the local station forecasts rain then the probability of it actually raining should be greater than if they forecast fair skies. At least that is the hope. :) In this section, you will develop a way to deal with the probability of some event that might change dependent upon the occurrence or not of some other event. Indeed, consider what happens when you keep on dealing a hand of five cards from a shuffled deck but without replacement. Notice how the probability of the same thing (such as $P(\text{getting a Heart on the next card})$) oscillates based upon what cards came out of the deck on previous hands.

```
print "Conditional_Events_-_successively_deal_5_cards_w/o_
replacement"

var('Ace_Clubs_Diamonds_Hearts_Jack_King_Queen_Spades')

suits = [Spades, Diamonds, Clubs, Hearts]
values = [2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace]

deck = [(value, suit) for suit in suits for value in values]
full_deck = copy(deck) # to save a copy of the original
                        deck for later use.

deck1 = copy(full_deck)
history1=[]
@interact
def _(choice=['Hearts','Spades','Diamonds','Clubs','New_
Deck'],again=['Repeat_Same_Suit']):
    global deck1, history1
    shuffle(deck1)
    if choice=='Hearts':
        suit = Hearts
    elif choice=='Spades':
        suit = Spades
    elif choice=='Diamonds':
        suit = Diamonds
    elif choice=='Clubs':
        suit = Clubs
    else:
        deck1 = copy(full_deck)
        shuffle(deck1)
        history1=[]
    if (Set(deck1).cardinality()<5):
        print "Deck_is_too_small...get_a_new_deck"
    elif choice<>'New_Deck':
        hand = [deck1.pop() for card in range(5)]
        print "Click_on_a_desired_suit_above_to_deal_out_
another_5_card_hand._The_cards_dealt:"
        print hand
        print "The_remaining_cards_in_the_deck:"
        print deck1
        num = Set(deck1).cardinality()
        print "\nThe_number_of_remaining_cards_in_the_deck=_
%s"%str(num)
    looking = []
```

```

    for card in deck1:
        if card[1]==suit:
            looking.append(card)
    prob = float(Set(looking).cardinality())/num
    history1.append(prob)

    print 'So, the remaining probability of getting a
          card from '+choice+' from the remaining cards is
          %s'%str(prob)
    list_plot(history1).show(xmin=0, xmax=9, ymin=0, ymax=1, figsize=(5,2))

```

Now, consider the case when you put the cards back in, reshuffle, and then get 5 new cards...

```

print "Independent_Events_-_Successively_deal_5_cards_but_
      WITH_replacement"

var('Ace_Clubs_Diamonds_Hearts_Jack_King_Queen_Spades')

suits = [Spades, Diamonds, Clubs, Hearts]
values = [2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace]

deck = [(value, suit) for suit in suits for value in values]
full_deck = copy(deck) # to save a copy of the original
                    deck for later use.
deck1 = copy(full_deck)

h2=[]
@interact
def
    _(choice=['Heart', 'Spade', 'Diamond', 'Club'], again=['Repeat_
    Same_Suit']):

    if choice=='Hearts':
        suit = Hearts
    elif choice=='Spades':
        suit = Spades
    elif choice=='Diamonds':
        suit = Diamonds
    else:
        suit = Clubs

    deck1 = copy(full_deck)
    shuffle(deck1)
    hand = [deck1.pop() for card in range(5)]
    print "The_cards_dealt:"
    print hand
    print "Replacing_this_hand_and_reshuffling_gives_the_
          remaining_cards_in_the_deck:"
    deck1 = copy(full_deck)
    shuffle(deck1)
    print(deck1)

    num = Set(deck1).cardinality()
    print "\nThe_number_of_remaining_cards_in_the_deck_=
          %s"%str(num)
    looking = []
    for card in deck1:

```



```

    if card[1]==suit:
        looking.append(card)
    prob = float(Set(looking).cardinality())/num
    h2.append(prob)

    print 'So, the remaining probability of getting a
          '+choice+' from the remaining cards is %s'%str(prob)
    list_plot(h2).show(xmin=0, xmax=15, ymin=0, ymax=1, figsize=(5, 2))

    print 'Independent Events - Successively deal 5 cards
          but WITH replacement'

```

Changing Sample Space - Balls: Consider a box with three balls: one Red, one White, and one Blue. Using an equally likely assumption, the probability of randomly pulling out a Red ball should be $1/3$. That is $P(\text{Red}) = 1/3$. However, suppose that for a first trial you pull out the White ball and set it aside. Attempting to pull out another ball leaves you with only two options and so the probability of randomly pulling out a Red ball is $1/2$. Notice that the probability changed for the second trial dependent on the outcome of the first trial.

Changing Sample Space - Cards: Consider a deck of 52 standard playing cards and a success occurs when a Heart is selected from the deck. When extracting one card randomly, the probability of that card being a Heart is $P(\text{Heart}) = 13/52$. Now, assume that one card has already been extracted and set aside. Now, prepare to extract another. If the first card drawn was a Heart, then there are only 12 Hearts left for the second draw. However, if the first card drawn was not a Heart, then there are 13 Hearts available for the second draw. To compute this probability correctly, one need to formulate the question so that subadditivity can be utilized.

Let H_1 be the outcome Heart on 1st draw and H_2 be the outcome Heart on 2nd draw. Then,

$$\begin{aligned}
 P(\text{Heart on 2nd draw}) &= P([H_1 \cap H_2] \cup [H_1^c \cap H_2]) \\
 &= P(H_1 \cap H_2) + P(H_1^c \cap H_2) \\
 &= \frac{|H_1 \cap H_2|}{|P(\text{Number of ways to get two cards})|} \\
 &\quad + \frac{|H_1^c \cap H_2|}{|\text{Number of ways to get two cards}|} \\
 &= \frac{13}{52} \cdot \frac{12}{51} + \frac{39}{52} \cdot \frac{13}{51} = \frac{12}{4 \cdot 51} + \frac{3 \cdot 13}{4 \cdot 51}
 \end{aligned}$$

Definition 4.5.1 Conditional Probability. For sets A and B,

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

provided $P(A) > 0$. ◇

You can read $P(B|A)$ as "the probability of B given A".

Theorem 4.5.2 Conditional Probability satisfies all of the requirements of regular probability.

Proof. By definition, for any event probability must be nonnegative. Therefore

$$P(A \cap B) \geq 0.$$

So,

$$P(B|A) = \frac{\text{positive or zero}}{\text{positive}} \geq 0.$$

Further,

$$P(S|A) = P(A \cap S)/P(A) = P(A)/P(A) = 1.$$

For the third part, we will only consider the case when there are two disjoint sets B and C. Then,

$$\begin{aligned} P(B \cup C|A) &= \frac{P(A \cap (B \cup C))}{P(A)} \\ &= \frac{P((A \cap B) \cup (A \cap C))}{P(A)} \\ &= \frac{P(A \cap B)}{P(A)} + \frac{P(A \cap C)}{P(A)} \\ &= P(B|A) + P(C|A). \end{aligned} \quad \blacksquare$$

Theorem 4.5.3 Multiplication Rule. For any sets A and B,

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

Proof. If $P(A)=0$ or $P(B)=0$, then the result is trivial. Otherwise, unravel the definition of conditional probability by taking the denominator to the other side. Also note that you can write $A \cap B = B \cap A$. \blacksquare

Checkpoint 4.5.4 WebWork. Conditional Probability sometimes makes you have to think carefully about the ways to get the desired outcome.

A bag contains 6 red marbles and 7 white marbles. Two marbles are drawn in succession without replacement. Find the probabilities of the following events:

1. The first marble drawn is red and the second is white.

Answer: _____

2. Both marbles drawn are red.

Answer: _____

See how you had to break the given question up into two disjoint pieces.

4.6 Bayes Theorem

Conditional probabilities can be computed using the methods developed above if the appropriate information is available. Some times you will however have some information available, such as $P(A|B)$ but need $P(B|A)$. The ability to "play around with history" by switching what has been presumed to occur leads to an important result known as Baye's Theorem.

Theorem 4.6.1 Bayes Theorem. Let $S = \{S_1, S_2, \dots, S_m\}$ where the S_k are pairwise disjoint and $S_1 \cup S_2 \cup \dots \cup S_m = S$ (i.e. a partition of the space S). Then for any $A \subset S$

$$P(S_j|A) = \frac{P(S_j)P(A|S_j)}{\sum_{k=1}^m P(S_k)P(A|S_k)}.$$

The conditional probability $P(S_j|A)$ is called the posterior probability of S_k .

Proof. Notice, by the definition of conditional probability and the multiplication rule

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

But using the disjointness of the partition

$$\begin{aligned} P(A) &= P((A \cap S_1) \cup (A \cap S_2) \cup \dots \cup (A \cap S_m)) \\ &= P(A \cap S_1) + P(A \cap S_2) + \dots + P(A \cap S_m) \\ &= P(S_1 \cap A) + P(S_2 \cap A) + \dots + P(S_m \cap A) \\ &= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \dots + P(S_m)P(A|S_m) \\ &= \sum_{k=1}^m P(S_k)P(A|S_k) \end{aligned}$$

Put these two expansions together to obtain the desired result. ■

To illustrate this result, from the web site <http://stattrek.com/probability/bayes-theorem.aspx> consider the following problem:

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90

Notice, all days can be classified into one of two disjoint options:

- Rainy, in which case we can deduce from the given info that $P(\text{Rain}) = 5/365$
- Not Rainy, and since this is the complement of above, $P(\text{Not Rain}) = 360/365$

In the notation of Bayes Theorem, let A represent a forecast of Rain and note you have

$$P(\text{Rain}) = P(S_1) = \frac{5}{365}$$

and

$$P(\text{Not Rain}) = P(S_2) = \frac{360}{365}.$$

Further, you are given the conditional probabilities

$$P(\text{Forecast Rain} | \text{Rain}) = P(A|S_1) = 0.9$$

$$P(\text{Forecast Rain} | \text{Not Rain}) = P(A|S_2) = 0.1$$

Notice that the question provided requests that you find the probability of Rain given that the weatherman has forecasted rain. What is given on the other hand is the reverse of that conditional probability. Using Bayes Theorem allows you to turn this around...

$$\begin{aligned} P(\text{Rain}) &= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) \\ &= \frac{5}{365} \cdot 0.9 + \frac{360}{365} \cdot 0.1 \end{aligned}$$

Hence, putting these together gives

$$P(\text{Rain} | \text{Forecast Rain}) = \frac{\frac{5}{365} \cdot 0.9}{\frac{5}{365} \cdot 0.9 + \frac{360}{365} \cdot 0.1}$$

$$\begin{aligned}
&= \frac{5 \cdot 0.9}{5 \cdot 0.9 + 360 \cdot 0.1} \\
&= \frac{45}{45 + 360} \approx 0.111
\end{aligned}$$

So, normally there is only a 5 percent chance of rain on a given day but given that the weatherman has forecast rain, the chance of rain has risen to a little more than 11 percent.

Checkpoint 4.6.2 WebWork. Let's try a Bayes Theorem example...

A biomedical research company produces 50 < percent/ > of its insulin at a plant in Kansas City, and the remainder is produced at a plant in Jefferson City. Quality control has shown that 1.25 < percent/ > of the insulin produced at the plant in Kansas City is defective, while 0.7 < percent/ > of the insulin produced at the plant in Jefferson City is defective. What is the probability that a randomly chosen unit of insulin came from the plant in Jefferson City given that it is defective?

(Hint: Draw a tree diagram first)

You have to be careful to extract the conditional probabilities from the problem.

Checkpoint 4.6.3 WebWork. Here is a more extensive Bayes Theorem example...

Data from Office on Smoking and Health, Centers for Disease Control and Prevention, indicate that 36% of adults who did not finish high school, 33% of high school graduates, 26% of adults who completed some college, and 15% of college graduates smoke. Suppose that one individual is selected at random and it is discovered that the individual smokes. Use the probabilities in the following table to calculate the probability that the individual is a college graduate.

Education	Employed	Unemployed
Not a high school graduate	0.0975	0.0080
High school graduate	0.3108	0.0128
Some college, no degree	0.1785	0.0062
Associate Degree	0.0849	0.0023
Bachelor Degree	0.1959	0.0041
Advanced Degree	0.0975	0.0015

Probability = _____

Hints: This problem has all the information you need, but not in the typical ready-to-use form. The table above can tell you the proportion of people with various levels of education in the population. Keep in mind that any degree (Associate, Bachelor, or Advanced) counts as graduating from college.

Notice that having the data expressed in tabular form sometimes makes it easier to deal with.

The interactive cell below can be used to easily compute all of the conditional probabilities associated with Bayes's Theorem. Notice how the relative size of the pie-shaped partition changes when you presume that an event in the space has already occurred.

```
# This function is used to convert an input string into
  separate entries
def g(s): return str(s).replace(' ','_').replace('(','_')
  ).replace(')','_').split()
```

```

@interact
def
    _ (Partition_Probabilities=input_box('0.35,0.25,0.40',label="$_{
P(S_1),P(S_2),..._$"),
        Conditional_Probabilities=input_box('0.02,0.01,0.03',label='$_{
P(A|S_1),P(A|S_2),..._$'),
        print_numbers=checkbox(True,label='Numerical_Results_
on_Graphs?'),
        auto_update=False):

    Partition_Probabilities = g(Partition_Probabilities)
    Conditional_Probabilities = g(Conditional_Probabilities)
    n = len(Partition_Probabilities)
    n0 = len(Conditional_Probabilities)

    # below needs to be n not equal to n0 but mathbook xml
    will not let me get the other
    if (n > n0):
        pretty_print("You_must_have_the_same_number_of_
partition_probabilities_and_conditional_
probabilities.")

    else:
        # input data streams
        now are the same size!
        colors = rainbow(n)
        accum = float(0) # to test whether
            partition probs sum to one
        ends = [0] # where the graphed
            partition sectors change in pie chart
        mid = [] # middle of each pie
            chart sector used for placement of text
        p_Sk_given_A = [] # P( S_k | A )
        pA = 0 # P(A)
        PP=[] # array to hold the
            numerical Partition Probabilities
        CP=[] # array to hold the
            numerical Conditional Probabilities
        for k in range(n):
            PP.append(float(Partition_Probabilities[k]))
            CP.append(float(Conditional_Probabilities[k]))
            p_Sk_given_A.append(PP[k]*CP[k] )
            pA += p_Sk_given_A[k]
            accum = accum + PP[k]
            ends.append(accum)
            mid.append((ends[k]+accum)/2)

#
# Marching along from 0 to 1, saving angles for each
# partition sector boundary.
# Later, we will multiple these by 2*pi to get actual
# sector boundary angles.
#
        if abs(accum-float(1))>0.0000001: # Due to
            roundoff issues, this should be close enough.
            pretty_print("Sum_of_probabilities_should_equal_
1.")

        else:
            # probability data
            is sensible

```

```

#
# Draw the Venn diagram by drawing sectors from the angles
# determined above
# First, create a circle of radius 1 to illustrate the the
# sample space S
# Then draw each sector with varying colors and print out
# their names on the edge
#
    G = circle((0,0), 1,
               rgbcolor='black',fill=False,
               alpha=0.4,aspect_ratio=True,axes=False,thickness=5)
    for k in range(n):
        G += disk((0,0), 1, (ends[k]*2*pi,
                             ends[k+1]*2*pi),
                  color=colors[mod(k,10)],alpha = 0.2)
        G +=
            text('$S_'+str(k+1)+'$',(1.1*cos(mid[k]*2*pi),
            1.1*sin(mid[k]*2*pi)), rgbcolor='black')

    G += circle((0,0), 0.6, facecolor='yellow', fill
               = True, alpha = 0.1,
               thickness=5,edgecolor='black')

# Print the probabilities corresponding to each particular
# region as a list and on the graphs
    if print_numbers:

        html("$P(A)_=_s$"%(str(pA),))
        for k in range(n):
            html("$P(S_{%s}_|_A)$"%(str(k+1))+"$_=_
            %s$"%str(p_Sk_given_A[k]/pA))

        G +=
            text(str(p_Sk_given_A[k]),(0.4*cos(mid[k]*2*pi),
            0.4*sin(mid[k]*2*pi)),
            rgbcolor='black')
        G += text(str(PP[k] -
            p_Sk_given_A[k]),(0.8*cos(mid[k]*2*pi),
            0.8*sin(mid[k]*2*pi)),
            rgbcolor='black')

# This is essentially a repeat of some of the above code
# but focused only on creating the smaller inner circle
# dealing
# with the set A so that the sectors now correspond in area
# to the Bayes Theorem probabilities

    accum = float(0)
    ends = [0]
    # where the
    graphed partition sectors change in pie chart
    mid = []
    # middle of each
    pie chart sector used for placement of text
    for k in range(n):
        accum += float(p_Sk_given_A[k]/pA)
        ends.append(accum)
        mid.append((ends[k]+accum)/2)
    H = circle((0,0), 1,

```

```

        rgbcolor='black',fill=False,
        alpha=0,aspect_ratio=True,axes=False,thickness=0)
H += circle((0,0), 0.6,
            facecolor='yellow',fill=True,
            alpha=0.1,aspect_ratio=True,axes=False,thickness=5,edgecolor='black')

    for k in range(n):
        H += disk((0,0), 0.6, (ends[k]*2*pi,
                                ends[k+1]*2*pi),
                    color=colors[mod(k,10)],alpha = 0.2)
        H +=
            text('$S_{'+str(k+1)+'|A$',(0.7*cos(mid[k]*2*pi),
            0.7*sin(mid[k]*2*pi)), rgbcolor='black')

# Now, print out the bayesian probabilities using
the smaller set A only

if print_numbers:
    for k in range(n):
        H += text(str(
            N(p_Sk_given_A[k]/pA,digits=4)
            ),(0.4*cos(mid[k]*2*pi),
            0.4*sin(mid[k]*2*pi)),
            rgbcolor='black')

G.show(title='Venn_diagram_of_partition_with_A_
in_middle')
print
H.show(title='Venn_diagram_presuming_A_has_
occured')

```

Checkpoint 4.6.4 Insured vs Accident. Your automobile insurance company uses past history to determine how to set rates by measuring the number of accidents caused by clients in various age ranges. The following table summarizes the proportion of those insured and the corresponding probabilities by age range:

Age	Proportion of Insured	Probability of Accident
16-20	0.05	0.08
21-25	0.06	0.07
26-55	0.49	0.02
55-65	0.25	0.03
over 65	0.15	0.04

Table 4.6.5: Age vs Accident Likelihood

One of your family friends insured by this company has an accident.

1. Determine the conditional probability that the driver was in the 16-20 age range.
2. Compare this to the probability that the driver was in the 18-20 age range. Discuss the difference.
3. Determine how much more the company should charge for someone in the 16-20 age range compared to someone in the 26-55 age range.

Solution. Plug the middle column into the first input box and the right column into the second input box of the [Bayes Sage Cell](#)

Checkpoint 4.6.6 Spinal bifida odds. Congratulations...your family is having a baby! As part of the prenatal care, some testing is part of the normal procedure including one for spinal bifida (which is a condition in which part of the spinal cord may be exposed.) Indeed, measurement of maternal serum AFP values is a standard tool used in obstetrical care to identify pregnancies that may have an increased risk for this disorder. You want to make plans for the new child's care and want to know how serious to take the test results. However, some times the test indicates that the child has the disorder when in actuality it does not (a false positive) and likewise may indicate that the child does not have the disorder when in fact it does (a false negative.) The combined accuracy rate for the screen to detect the chromosomal abnormalities mentioned above is approximately 85

- Approximately 85 out of every 100 babies affected by the abnormalities addressed by the screen will be identified. (Positive Positive)
 - Approximately 5
1. Given that your test came back negative, determine the likelihood that the child will actually have spinal bifida.
 2. Given that your test came back negative, determine the likelihood that the child will not have spina bifida
 3. Given that a positive test means you have a 1/100 to 1/300 chance of experiencing one of the abnormalities, determine the likelihood of spinal bifida in a randomly selected child.

4.7 Independence

You have seen when repeatedly sampling without replacement leads to a change the the likelihood of some event in successive trials. Indeed, this is what conditional probabilities above illustrate. However, when sampling with replacement you may find a different situation arises. Indeed, you easily notice that when flipping a coin, $P(\text{Heads}) = 1/2$ regardless of the outcome of any previous flip. In situations such as this where the probability of an event is not affected by the occurrence (or lack of occurrence) of some other event determining the probability of compound events can be greatly simplified.

Definition 4.7.1 Independent Events. Events A and B are independent provided

$$P(A \cap B) = P(A)P(B)$$

◇

Corollary 4.7.2 Independence and Conditional Probability. *Given independent events 4.7.1 A and B,*

$$P(B|A) = P(B)$$

and

$$P(A|B) = P(A).$$

Proof. By the multiplication rule and the definition of independence, for any events A and B

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

Therefore, if $P(A)$ is non-zero, canceling yields the first result. Switching around notation provides the second. ■

Checkpoint 4.7.3 WebWork. Independence makes combined probabilities VERY easy to compute.

For two events A and B , $P(A) = 0.3$ and $P(B) = 0.1$.

(a) If A and B are independent, then

$$P(A|B) = \underline{\hspace{2cm}}$$

$$P(A \cup B) = \underline{\hspace{2cm}}$$

$$P(A \cap B) = \underline{\hspace{2cm}}$$

(b) If A and B are dependent and $P(A|B) = 0.1$, then

$$P(A \cap B) = \underline{\hspace{2cm}}$$

$$P(B|A) = \underline{\hspace{2cm}}$$

Basically you just multiply individual probabilities together. Independence is often assumed since it makes computations easier. That said, you should remember to consider each time whether independence should or should not be assumed.

Corollary 4.7.4 Independence and Mutual Exclusivity. *If events A and B are both [independent 4.7.1](#) and [mutually exclusive 4.3.1](#), then at least one of them has zero probability.*

Proof. By independence, $P(A \cap B) = P(A) \cdot P(B)$. However, by mutually exclusivity, $A \cap B = \emptyset \Rightarrow P(A \cap B) = 0$ gives

$$P(A) \cdot P(B) = 0.$$

Hence, one or the other (or both) must be zero. ■

Corollary 4.7.5 Successive Independent Events. *Given a sequence of mutually independent events A_1, A_2, A_3, \dots ,*

$$P(\cap_{k \in R} A_k) = \prod_{k \in R} P(A_k)$$

4.8 Summary

TBA

4.9 More Exercises

Checkpoint 4.9.1 Conditional Basic computation. Given $P(A) = 0.43$, $P(B) = 0.72$, and $P(A \cap B) = 0.29$, determine

1. $P(A \cup B)$
2. $P(B|A)$
3. $P(A|B)$
4. $P(A^c \cap B^c)$

Checkpoint 4.9.2 Gender vs University Major. The table below classifies students at your university according to gender and according to major.

Enrollment	Male	Female	Totals
STEM	420	510	930
Business	320	270	590
Other	610	710	1320
Totals	1350	1490	2840

Table 4.9.3: Gender vs Major

Determine the following:

1. $P(\text{STEM major})$
2. $P(\text{STEM} \mid \text{Female})$
3. $P(\text{Female} \mid \text{STEM})$
4. $P(\text{Female} \mid \text{Not STEM})$

Checkpoint 4.9.4 Mean Tough Teacher. You are in a probability and statistics class with a teacher who has predetermined that only one student can make an A for the course. To be "fair", he places a number of slips of paper in a bowl equal to the number of students in the course with one of the slips having an A designation. Students in the course each can pick once randomly from the bowl and without replacement to see if they can get the lucky slip. Determine the following:

1. If there are 15 students in your course, determine the probabilities of getting an A in the course if you pick first and if you pick last.
2. Since the teacher likes you the most, she will give you the option of deciding whether to pick at any position. If so, determine the position that would give you the best likelihood of getting the A slip.
3. Suppose again that the teacher was feeling more generous and decided instead to allow for two A's. Determine how that changes your likelihood of winning and on what position you would like to choose.
4. Continue as above except that only one slip does not have an A on it.
5. Discuss how your choice is affected by the number of students in the course or the number of A slips included.

Solution. Using the normal equally-likely definition, $P(\text{first}) = \frac{1}{15}$.

To get the A on the last pick requires that all of the previous picks to be something else. You don't get the opportunity to pick the A if it has already been selected. So, if L stands for losing (not getting the A), then

$$P(\text{last}) = P(\text{LLLLLLLLLLLLLLA}) = \frac{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} = \frac{1}{15}.$$

Therefore, it is the same probability of getting the A whether you pick first or last. In general, to win on the k th pick gives

$$P(k\text{th}) = P(\text{LL...LA}) = \frac{14 \cdot 13 \cdot \dots \cdot (15 - k) \cdot 1}{15 \cdot 14 \cdot \dots \cdot (16 - k) \cdot (15 - k)} = \frac{1}{15}$$

Hence, it is the same probability regardless of when you get to pick.

If there are two A's possible, then the options for person k include either receiving the first of the two slips or the second. The probability for determining the first of the two is computed in a manner similar to above except that there is one more A and one less other.

$$P(\text{kth as first}) = P(\text{LL...LA}) = \frac{13 \cdot 12 \cdot \dots \cdot (15 - k) \cdot 2}{15 \cdot 14 \cdot \dots \cdot (16 - (k + 1)) \cdot (16 - k)} = \frac{2 \cdot (15 - k)}{15 \cdot 14}$$

The probability of getting the second A means exactly one of the previous k-1 selections also picked the other A. There are k-1 ways that this could happen. Computing for one of the options and multiplying by k-1 gives

$$P(\text{kth as second}) = P(\text{LL...LAA}) = (k - 1) \cdot \frac{13 \cdot 12 \cdot \dots \cdot (15 - k) \cdot 2 \cdot 1}{15 \cdot 14 \cdot \dots \cdot (16 - k) \cdot (15 - k)} = \frac{2 \cdot (k - 1)}{15 \cdot 14}.$$

Adding these two together gives

$$\begin{aligned} P(\text{getting an A when there are two}) &= \frac{2 \cdot (15 - k) + 2 \cdot (k - 1)}{15 \cdot 14} \\ &= \frac{28}{15 \cdot 14} = \frac{2}{15}. \end{aligned}$$

For example, if k = 5,

$$\begin{aligned} P(5\text{th as first}) &= P(\text{LLLLA}) \\ &= \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 2}{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11} = \frac{20}{15 \cdot 14} \end{aligned}$$

$$\begin{aligned} P(5\text{th as second}) &= P(\text{LL...LAA}) \\ &= 4 \cdot \frac{13 \cdot 12 \cdot 11 \cdot 2 \cdot 1}{15 \cdot 14 \cdot 13 \cdot 12 \cdot 11} = \frac{8}{15 \cdot 14}. \end{aligned}$$

Adding these together yields the general result. So, once again, it doesn't matter which pick you use since the likelihood of getting an A is the same for all positions.

Checkpoint 4.9.5 Shared Birthdays. In this problem, you want to consider how many people are necessary in order to have an even chance of finding two or more who share a common birthday. Toward that end, assuming a year has exactly 365 equally likely days let r be the number of people in a sample and consider the following:

1. Determine the number of different outcomes of birthdays when order matters and birthdays are allowed to be repeated.
2. Determine the number of different outcomes when birthdays are not allowed to be repeated.
3. Determine the probability that two or more of your r students have the same birthday.
4. Prepare a spreadsheet with the probabilities found above from r=2 to r=50. Determine the value of r for which this probability is closest to 0.5.
5. As best as you can, sample two groups of the size found above and gather birthday information. For each group, determine if there is a shared

birthday or not. Compare your results with others in the class to check whether the sampling validates that about half of the samples should have a shared birthday group.

Solution. The correct sample size to get past a probability of 0.5 is 23 people. You should justify this numerically by justifying the following probabilities:

#	P(Match)
1	0
2	0.0027
3	0.0082
4	0.0164
5	0.0271
6	0.0405
7	0.0562
8	0.0743
9	0.0946
10	0.1169
11	0.1411
12	0.1670
13	0.1944
14	0.2231
15	0.2529
16	0.2836
17	0.3150
18	0.3469
19	0.3791
20	0.4114
21	0.4437
22	0.4757
23	0.5073
24	0.5383
25	0.5687
26	0.5982
27	0.6269
28	0.6545
29	0.6810
30	0.7063

Checkpoint 4.9.6 Internet meme solution. This one is from an internet meme: Two fair 6-sided dice are rolled together and you are told that at least one of the dice is a 6. Given that a 6 will be removed, determine the probability that the other die is a 6.

Solution. In this case, you are presented with an outcome where the possible choices consist of (1,6), (2,6), (3,6), (4,6), (5,6), (6,6), (6,5), (6,4), (6,3), (6,2), (6,1). Each of these would satisfy the condition that at least one of the dice is a 6. From this group, the only success that satisfies being a 6, given that another 6 has already been removed, is the (6,6) outcome. Therefore, the conditional probability is $1/11$.

It is interesting to note that if the question instead was posed so that one of the dice was a 6 and it was removed, then the probability of the other dice showing a 6 would be $1/6$.

Checkpoint 4.9.7 100 people on an airplane with boarding pass issues. This is a famous problem. 100 people are in line, boarding an airplane with 100 seats, one at a time. They are in no particular order. The first person

has lost his boarding pass, so he sits in a random seat. The second person does the following:

- Goes to his seat (the one it says to go to on the boarding pass). If unoccupied, sit in it.
- If occupied, find a random seat to sit in.

Everyone else behind him does the same. What is the probability that the last person sits in his correct seat?

Solution. To get the idea, consider what happens with only 2 people, then only 3. Generalize.

The answer is $1/2$. To obtain this, you can define recursively the probability that the k th person sits in their own set as $f(k)$. Consider the first traveler's and your seats. Then you get the following cases:

- $P(\text{first guy sits in his own seat and you sit in yours}) = \frac{1}{k} \cdot 1$
- $P(\text{first guy sits in your seat and you do not sit in yours}) = \frac{1}{k} \cdot 0$
- $P(\text{other } k-2 \text{ travelers make their choices}) = (k-2) \frac{1}{k} f(k-1)$

$$f(k) = 1/k + 0 + (k-2)/k f(k-1)$$

with $f(2) = 1/2$.

For example,

- $f(3) = 1/3 + f(2)/3 = 1/3 + 1/6 = 1/2$.
- $f(4) = 1/4 + 2/4 f(3) = 1/4 + 1/2 \cdot 1/2 = 1/2$.
- $f(5) = 1/5 + 3/5 \cdot 1/2 = 1/2$.
- $f(6) = 1/6 + 4/6 \cdot 1/2 = 1/2$.

Etc.

Checkpoint 4.9.8 Basic Independence Calculations. Given $P(A) = 0.43$, $P(B) = 0.72$, and $P(A \cap B) = 0.31$, verify that A and B are not independent.

Solution. A and B are [independent by definition 4.7.1](#) provided

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

Using the provided values, notice that

$$P(A \cap B) = 0.31$$

but

$$P(A)P(B) = 0.43 \cdot 0.72 = 0.3096.$$

Since these are not equal (regardless how close) then A and B are not independent.

Checkpoint 4.9.9 Compound events and Independence. Given A, B, and C are independent events, with $P(A) = 2/5$, $P(B) = 3/4$, and $P(C) = 1/6$, determine:

1. $P(A \cap B \cap C)$
2. $P(A^c \cap B^c \cap C)$
3. $P(A \cup B \cup C)$

Solution. Extending the [definition of independent events 4.7.1](#) gives

$$P(A \cap B \cap C) = \frac{2}{5} \frac{3}{4} \frac{1}{6}.$$

By the corollary for independent events, complements also maintain a similar independence. So

$$P(A^c \cap B^c \cap C) = \frac{3}{5} \frac{1}{4} \frac{1}{6}.$$

To complete the third part, use the [inclusion/exclusion result 4.3.9](#) for dealing with three sets.

Checkpoint 4.9.10 Rolling multiple dice. For a pair of dice you want to consider the events A = rolling a 7 or 11 and B = otherwise...as in the first roll in the game of craps. Further, for notation purposes let's take ABA (for example) to mean event A occurs on the first roll, event B occurs on the second roll, and event A occurs again on the third roll...in that order only. If you roll the dice 5 times, determine

1. $P(AABBB)$
2. $P(BBBAA)$
3. The probability of getting A on exactly two rolls of the dice.

Solution. Successive rollings of a pair of dice are [independent events 4.7.1](#). Therefore,

$$P(AABBB) = P(A)P(A)P(B)P(B)P(B) = \frac{8}{36} \frac{8}{36} \frac{28}{36} \frac{28}{36} \frac{28}{36}$$

Similarly for the second part.

For the third part, notice that there will be $\binom{5}{2}$ ways to rearrange 2 A 's and 3 B 's but that each of these will have two $8/36$'s and three $28/36$'s but just in a different order. Therefore, you will get

$$10 \cdot \frac{8}{36} \frac{8}{36} \frac{28}{36} \frac{28}{36} \frac{28}{36}$$

Checkpoint 4.9.11 Redundancy. To help "insure" the success of a mission, you propose several redundant components so that the mission is a success if one or more succeed. Supposing that these separate components act independently of each other and that each component has a 75

1. The probability of failure if you utilize 2 components.
2. The probability of failure if you utilize 5 components.
3. The number of components needed to insure that the probability of success is at least 99

Checkpoint 4.9.12 Internet Meme redux. Again, from an internet meme: Two fair 6-sided dice are rolled together and you are told that at least one of the dice is a 6. A 6 is removed and you are presented with the other die. Determine the probability that it is a 6.

Solution. For this setting, notice that the outcomes from each of the two dice are independent of each other. Removing one of the dice, regardless of its value, does not affect the other. The question in this case does not ask for a conditional probability.

Checkpoint 4.9.13 Single Elimination Tournament. Consider a $n=4$ team single-elimination tournament where the teams are "seeded" from 1 (the best team) to 4 (the worst team). For this tournament, team 1 plays team 4 and team 2 plays team 3. The winner of each play each other to determine the final winner. When teams j and k play, set $P(j \text{ wins}) = \frac{k}{j+k}$ and similarly for team k . Assuming separate games are independent of each other, determine the probability that team 4 wins the tournament. What about with 8 teams? What about 64 teams?

Solution. $P(4 \text{ wins}) = P(4 \text{ beats } 1) P(4 \text{ beats the winner of the other bracket})$

$P(4 \text{ wins}) = (1/5) * P(4 \text{ beats } 2 \mid 2 \text{ beats } 3) + P(4 \text{ beats } 3 \mid 3 \text{ beats } 2)$

$P(4 \text{ wins}) = 1/5 [(3/5)(2/6) + (2/5)(3/7)] = 78/1050 = 0.0742$

For the other teams:

$P(1 \text{ wins}) = 4/5 [(3/5)(2/3) + (2/5)(3/4)] = 0.56$

$P(2 \text{ wins}) = 3/5 [(4/5)(1/3) + (1/5)(4/6)] = 0.24$

$P(3 \text{ wins}) = 2/5 [(4/5)(1/4) + (1/5)(4/7)] = 0.1257$

Chapter 5

Probability Functions

5.1 Introduction

Each of the probability exercises thus far required you to utilize basic definitions and theorems to determine the answer. Starting a new problem meant starting over from scratch. This is burdensome. However, you may have noticed that some of the ways you might have created solutions for some problems ending up looking very similar to the solutions for others. In this chapter, you will consider the framework needed for creating general solution techniques. These techniques will give a number of "distributions" which are general ways to solve a particular type of problem.

Toward that end, in this chapter you will see how to create a random variable which takes items in the sample space and assigns corresponding numerical values. From that, you will see how to create "Probability Functions" on that variable that provide the desired probability by simple function evaluation. General properties these functions possess will also be developed.

5.2 Random Variables

For a given set of events, we might have difficulty doing mathematics since the outcomes are not numerical. In order to accomodate our desire to convert to numerical measures we want to assign numerical values to all outcomes. The process of doing this creates what is known as a random variable.

Definition 5.2.1 Random Variable. Given a random experiment with sample space S , a function X mapping each element of S to a unique real number is called a random variable. For each element s from the sample space S , denote this function by

$$X(s) = x$$

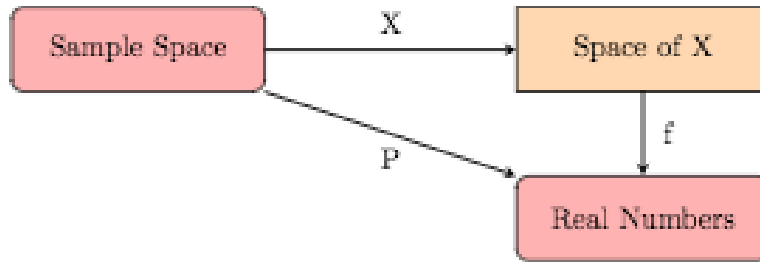
and let R be the range of X . R will be called "the space of X " and in notation

$$R = \{x : X(s) = x, \text{ for some } s \in S\}.$$

◇

We will make various restrictions on the range of the random variable to fit different generalized problems. Then, we will be able to work on a problem (which may be inherently non-numerical) by using the random variable in subsequent calculations.

Example 5.2.2 Success vs Failure. When dealing with only two outcomes,



one might use

$$S = \text{success, failure}.$$

Choose

$$X(\text{success}) = 1$$

$$X(\text{failure}) = 0.$$

Then, $R=0,1$. □

Example 5.2.3 Standard Dice Pairs. When gambling with a pair of dice, one might use S =ordered pairs of all possible rolls. Then

$$S = (a,b): a=\text{die 1 outcome, } b=\text{die 2 outcome}.$$

Choose

$$X((a,b)) = a + b.$$

Then, $R=2, 3, 4, 5, \dots, 12$. □

Example 5.2.4 Other Dice Options. When rolling dice in a board game (like RISK), one might use

$$S = (a,b): a=\text{die 1 outcome, } b=\text{die 2 outcome}$$

Choose

$$X((a,b)) = \max a,b.$$

Then, $R=1, 2, 3, 4, 5, 6$. □

Definition 5.2.5 Countable and Uncountable Sets. R contains a countable number of points if either R is finite or there is a one to one correspondence between R and the positive integers. Such a set will be called discrete. We will see that often the set R is not countable. If R consists of an interval of points (or a union of intervals), then we call X a continuous random variable. ◇

5.3 Probability Functions

In the formulas below, we will presume that we have a random variable X which maps the sample space S onto some range of real numbers R . From this set, we then can define a probability function $f(x)$ which acts on the numerical values in R and returns another real number. We attempt to do so to obtain (for

discrete values) $P(\text{sample space value } s) = f(X(s))$. That is, the probability of a given outcome s is equal to the composition which takes s to a numerical value x which is then plugged into f to get the same final values.

For example, consider a random variable which assigns a 1 when you roll a 1 on a six-sided die and 0 otherwise. Presuming each side is equally likely, $f(1) = \frac{1}{6}$ and $f(0) = \frac{5}{6}$.

Definition 5.3.1 Probability "Mass" Function. Given a discrete random variable X on a space R , a probability mass function on X is given by a function $f : R \rightarrow \mathbb{R}$ such that:

$$\begin{aligned}\forall x \in R, f(x) &> 0 \\ \sum_{x \in R} f(x) &= 1 \\ A \subset R \Rightarrow P(X \in A) &= \sum_{x \in A} f(x)\end{aligned}$$

For $x \notin R$, you can use the convention $f(x)=0$. ◇

Definition 5.3.2 Probability "Density" Function. Given a continuous random variable X on a space R , a probability density function on X is given by a function $f : R \rightarrow \mathbb{R}$ such that:

$$\begin{aligned}\forall x \in R, f(x) &> 0 \\ \int_R f(x) dx &= 1 \\ A \subset R \Rightarrow P(X \in A) &= \int_A f(x) dx\end{aligned}$$

For $x \notin R$, you can use the convention $f(x)=0$. ◇

For the purposes of this book, we will use the term "Probability Function" to refer to either of these options.

Example 5.3.3 Discrete Probability Function. Consider $f(x) = x/10$ over $R = 1,2,3,4$. Then, $f(x)$ is obviously positive for each of the values in R and certainly

$$\sum_{x \in R} f(x) = f(1) + f(2) + f(3) + f(4) = 1/10 + 2/10 + 3/10 + 4/10 = 1.$$

Therefore, $f(x)$ is a probability mass function over the space R . □

```
# Combining all of the above into one interactive cell
@interact
def _(D = input_box([1,2,3,5,6,8,9,11,12,14],label="Enter_
domain_R_(in_brackets):"),
    Probs =
        input_box([1/20,1/20,1/20,3/20,1/20,4/20,4/20,1/20,1/20,3/20],label="Enter_
corresponding_f(x)_(in_brackets):"),
    n_samples=slider(100,10000,100,100,label="Number_of_
times_to_sample_from_this_distribution:")):
    n = len(D)
    R = range(n)
    one_huh = sum(Probs)
    pretty_print('\n\nJust_to_be_certain,_we_should_check_to_
make_certain_the_probabilities_sum_to_1\n')
```

```

pretty_print(html('$\sum_{x\epsilon R} f(x) = \_
               %s$'%str(one_huh)))

G = Graphics()
if len(D)==len(Probs):
    f = zip(D,Probs)
    meanf = 0
    variancef = 0
    for k in R:
        meanf += D[k]*Probs[k]
        variancef += D[k]^2*Probs[k]
    G +=
        line([(D[k],0),(D[k],Probs[k])],color='green')
    variancef = variancef - meanf^2
    sd = sqrt(variancef)
    G += points(f,color='blue',size=50)
    G += point((meanf,0),color='yellow',size=60,zorder=3)
    G +=
        line([(meanf-sd,0),(meanf+sd,0)],color='red',thickness=5)

    g = DiscreteProbabilitySpace(D,Probs)
    pretty_print('mean = %s'%str(meanf))
    pretty_print('variance = %s'%str(variancef))

    # perhaps to add mean and variance for pmf here
else:
    print 'Domain D and Probabilities Probs must be
          lists of the same size'

# Now, let's sample from the distribution given above
# and see how a random sampling matches up

counts = [0] * len(Probs)
X = GeneralDiscreteDistribution(Probs)
sample = []

for _ in range(n_samples):
    elem = X.get_random_element()
    sample.append(D[elem])
    counts[elem] += 1
Empirical = [1.0*x/n_samples for x in counts] # random

samplemean = mean(sample)
samplevariance = variance(sample)
sampdev = sqrt(samplevariance)

E = points(zip(D,Empirical),color='orange',size=40)
E +=
    point((samplemean,0.005),color='brown',size=60,zorder=3)
E +=
    line([(samplemean-sampdev,0.005),(samplemean+sampdev,0.005)],color='orange')
(G+E).show(ymin=0,figsize=(8,5))

```

Example 5.3.4 Continuous Probability Function. Consider $f(x) = x^2/c$ for some positive real number c and presume $R = [-1,2]$. Then $f(x)$ is nonnegative (and only equals zero at one point). To make $f(x)$ a probability

density function, we must have

$$\int_{x \in R} f(x) = 1.$$

In this instance you get

$$1 = \int_{-1}^2 x^2/c = x^3/(3c)|_{-1}^2 = \frac{8}{3c} - \frac{-1}{3c} = \frac{3}{c}$$

Therefore, $f(x)$ is a probability density function over R provided $c = 3$. \square

Definition 5.3.5 Distribution Function. Given a random variable X on a space R , a probability distribution function on X is given by a function

$$F : \mathbb{R} \rightarrow \mathbb{R} \ni F(x) = P(X \leq x).$$

\diamond

Example 5.3.6 Discrete Distribution Function. Using $f(x) = x/10$ over $R = 1, 2, 3, 4$ again, note that $F(x)$ will only change at these four domain values. We get

X	F(x)
$x < 1$	0
$1 \leq x < 2$	1/10
$2 \leq x < 3$	3/10
$3 \leq x < 4$	6/10
$4 \leq x$	1

Table 5.3.7: Discrete Distribution Function Example

\square

Example 5.3.8 Continuous Distribution Function. Consider $f(x) = x^2/3$ over $R = [-1, 2]$. Then, for $-1 \leq x \leq 2$,

$$F(x) = \int_{-1}^x u^2/3 du = x^3/9 + 1/9.$$

Notice, $F(-1) = 0$ since nothing has yet been accumulated over values smaller than -1 and $F(2) = 1$ since by that time everything has been accumulated. In summary:

X	F(x)
$x < -1$	0
$-1 \leq x < 2$	$x^3/9 + 1/9$
$2 \leq x$	1

Table 5.3.9: Continuous Distribution Function Example

\square

Theorem 5.3.10 $F(x) = 0, \forall x < \inf(R)$

Proof. Let $a = \inf(R)$. Then, for $x < a$,

$$F(x) = P(X \leq x) \leq P(X < a) = 0$$

since none of the x -values in this range are in R . ■

Theorem 5.3.11 $F(x) = 1, \forall x \geq \sup(R)$

Proof. Let $b = \sup(R)$. Then, for

$$x \geq b, F(x) = P(X \leq x) = P(X \leq b) + P(b < X \leq x) = P(X \leq b) = 1$$

since all of the x -values in this range are in R and therefore will either sum over or integrate over all of R . ■

Theorem 5.3.12 F is non-decreasing

Proof. Case 1: R discrete

$$\begin{aligned} \forall x_1, x_2 \in \mathbb{Z} \ni x_1 < x_2 \\ F(x_2) &= \sum_{x \leq x_2} f(x) \\ &= \sum_{x \leq x_1} f(x) + \sum_{x_1 < x \leq x_2} f(x) \\ &\geq \sum_{x \leq x_1} f(x) = F(x_1) \end{aligned}$$

Case 2: R continuous

$$\begin{aligned} \forall x_1, x_2 \in \mathbb{R} \ni x_1 < x_2 \\ F(x_2) &= \int_{-\infty}^{x_2} f(x) dx \\ &= \int_{-\infty}^{x_1} f(x) dx + \int_{x_1}^{x_2} f(x) dx \\ &\geq \int_{-\infty}^{x_1} f(x) dx \\ &= F(x_1) \end{aligned} \quad \blacksquare$$

Theorem 5.3.13 Using Discrete Distribution Function to compute probabilities. For $x \in R, f(x) = F(x) - F(x-1)$

Proof. Assume $x \in R$ for some discrete R . Then,

$$F(x) - F(x-1) = \sum_{u \leq x} f(u) - \sum_{u < x} f(u) = f(x) \quad \blacksquare$$

Theorem 5.3.14 Using Continuous Distribution function to compute probabilities. For $a < b, (a, b) \in R, P(a < X \leq b) = F(b) - F(a)$

Proof. For a and b as noted, consider

$$\begin{aligned} F(b) - F(a) &= \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx \\ &= \int_a^b f(x) dx \\ &= P(a < x \leq b) \end{aligned} \quad \blacksquare$$

Corollary 5.3.15 For continuous distributions, $P(X = a) = 0$

Proof. We will assume that $F(x)$ is a continuous function. With that assumption, note

$$P(a - \epsilon < x \leq a) = \int_{a-\epsilon}^a f(x)dx = F(a) - F(a - \epsilon)$$

Take the limit as $\epsilon \rightarrow 0^+$ to get the result noting that \blacksquare

Theorem 5.3.16 $F(x)$ vs $f(x)$, for continuous distributions. If X is a continuous random variable, f the corresponding probability function, and F the associated distribution function, then

$$f(x) = F'(x)$$

Proof. Assume X is continuous and f and F as above. Notice, by the definition of f , $\lim_{x \rightarrow \pm\infty} f(x) = 0$ since otherwise the integral over the entire space could not be finite.

Now, let $A(x)$ be any antiderivative of $f(x)$. Then, by the Fundamental Theorem of Calculus,

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(u)du \\ &= A(x) - \lim_{u \rightarrow -\infty} A(u) \end{aligned}$$

Hence, $F'(x) = A'(x) - \lim_{u \rightarrow -\infty} A'(u) = f(x)$ as desired. \blacksquare

Definition 5.3.17 Percentiles for Random Variables. For $0 < p < 1$, the $100p^{th}$ percentile is the largest random variable value c that satisfies

$$F(c) = p.$$

For continuous random variables over an interval $R = [a, b]$, you will solve for c in the equation

$$\int_a^c f(x)dx.$$

For discrete random variables, it is unlikely that a particular percentile will land exactly on one of the elements of R but you will want to take the smallest value in R so that $F(c) \geq p$.

The 50th percentile (as before) is also known as the median. \diamond

Example 5.3.18 Continuous Percentile. For our earlier example with $f(x) = x^2/3$ on $R = [-1, 2]$, the 50th percentile (i.e. the median) is found by starting with $p = 0.5$ and then solving

$$F(c) = 0.5$$

or

$$c^3/9 + 1/9 = 1/2$$

or

$$c^3 + 1 = 9/2.$$

After solving for c , you find

$$\text{median} = \sqrt[3]{7/2} \approx 1.518.$$

\square

Example 5.3.19 Discrete Percentile. TBA, using one of the table examples from above. \square

5.4 Expected Value

Blaise Pascal was a 17th century mathematician and philosopher who was accomplished in many areas but may likely be best known to you for his creation of what is now known as Pascal's Triangle. As part of his philosophical pursuits, he proposed what is known as "Pascal's wager". It suggests two mutually exclusive outcomes: that God exists or that he does not. His argument is that a rational person should live as though God exists and seek to believe in God. If God does not actually exist, such a person will have only a finite loss (some pleasures, luxury, etc.), whereas they stand to receive infinite gains as represented by eternity in Heaven and avoid an infinite losses of eternity in Hell. This type of reasoning is part of what is known as "decision theory".

You may not confront such dire payouts when making your daily decisions but we need a formal method for making these determinations precise. The procedure for doing so is what we call expected value.

Definition 5.4.1 Expected Value. Given a random variable X over space R , corresponding probability function $f(x)$ and "value function" $v(x)$, the expected value of $v(x)$ is given by

$$E = E[v(X)] = \sum_{x \in R} v(x)f(x)$$

provided X is discrete, or

$$E = E[v(X)] = \int_R v(x)f(x)dx$$

provided X is continuous. \diamond

Theorem 5.4.2 Expected Value is a Linear Operator.

1. $E[c] = c$
2. $E[c v(X)] = c E[v(X)]$
3. $E[v(X) + w(X)] = E[v(X)] + E[w(X)]$

Proof. Each of these follows by utilizing the corresponding linearity properties of the summation and integration operations. For example, to verify part three in the continuous case:

$$\begin{aligned} E[v(X) + w(X)] &= \int_{x \in R} [v(x) + w(x)]f(x)dx \\ &= \int_{x \in R} v(x)f(x)dx + \int_{x \in R} w(x)f(x)dx \\ &= E[v(X)] + E[w(X)]. \end{aligned} \quad \blacksquare$$

Example 5.4.3 Discrete Expected Value. Consider $f(x) = x/10$ over $R = 1, 2, 3, 4$ where the payout is 10 euros if $x=1$, 5 euros if $x=2$, 2 euros if $x=3$ and -7 euros if $x = 4$. Then your value function would be

$$v(1) = 10, v(2) = 5, v(3) = 2, v(4) = -7.$$

Computing the expect payout gives

$$E = 10 \times 1/10 + 5 \times 2/10 + 2 \times 3/10 - 7 \times 4/10 = -2/10$$

Therefore, the expected payout is actually negative due to a relatively large negative payout associated with the largest likelihood outcome and the larger positive payout only associated with the least likely outcome. \square

Example 5.4.4 Continuous Expected Value. Consider $f(x) = x^2/3$ over $R = [-1, 2]$ with value function given by $v(x) = e^x - 1$. Then, the expected value for $v(x)$ is given by

$$E = \int_{-1}^2 (e^x - 1) \cdot x^2/3 = -1/9 \cdot (e + 15) \cdot e^{-1} + 2/3 \cdot e^2 - 8/9 \approx 3.3129$$

\square

Definition 5.4.5 Theoretical Measures. Given a random variable with probability function $f(x)$ over space R

1. The mean of $X = \mu = E[x]$
2. The variance of $X = \sigma^2 = E[(x - \mu)^2]$
3. The skewness of $X = \gamma_1 = \frac{E[(x - \mu)^3]}{\sigma^3}$
4. The kurtosis of $X = \gamma_2 = \frac{E[(x - \mu)^4]}{\sigma^4}$

\diamond

Theorem 5.4.6 Alternate Formulas for Theoretical Measures.

1. $\sigma^2 = E[x^2] - \mu^2 = E[X(x - 1)] + \mu - \mu^2$
2. $\gamma_1 = \frac{1}{\sigma^3} \cdot [E[X^3] - 3\mu E[X^2] + 2\mu^3]$
3. $\gamma_2 = \frac{1}{\sigma^4} \cdot [E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4]$

Proof. In each case, expand the binomial inside and use the linearity of expected value. \blacksquare

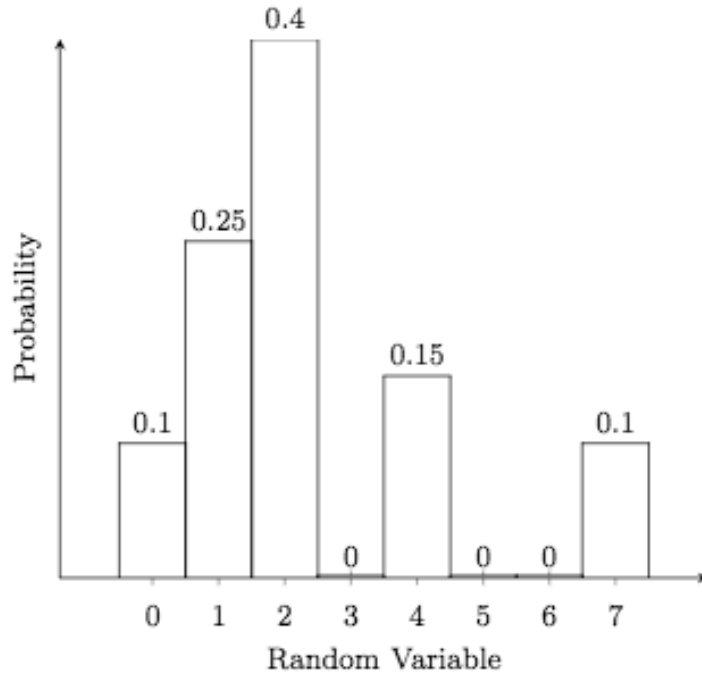
Consider the following example when computing these statistics for a discrete variable. In this case, we will utilize a variable with a relatively small space so that the summations can be easily done by hand. Indeed, consider

X	f(x)
0	0.10
1	0.25
2	0.40
4	0.15
7	0.10

Table 5.4.7: Discrete Probability Function Example

Using the definition of mean as a sum,

$$\begin{aligned} \mu &= 0 \cdot 0.10 + 1 \cdot 0.25 + 2 \cdot 0.40 + 4 \cdot 0.15 + 7 \cdot 0.10 \\ &= 0 + 0.25 + 0.80 + 0.60 + 0.70 \\ &= 2.35 \end{aligned}$$



Notice where this lies on the probability histogram for this distribution.

For the variance

$$\begin{aligned}
 \sigma^2 &= E[X^2] - \mu^2 \\
 &= [0^2 \cdot 0.10 + 1^2 \cdot 0.25 + 2^2 \cdot 0.40 + 4^2 \cdot 0.15 + 7^2 \cdot 0.10] - 2.35^2 \\
 &= 0 + 0.25 + 1.60 + 2.40 + 4.90 - 5.5225 \\
 &= 9.15 - 5.225 \\
 &= 3.6275
 \end{aligned}$$

and so the standard deviation $\sigma = \sqrt{3.6275} \approx 1.90$. Notice that 4 times this value encompasses almost all of the range of the distribution.

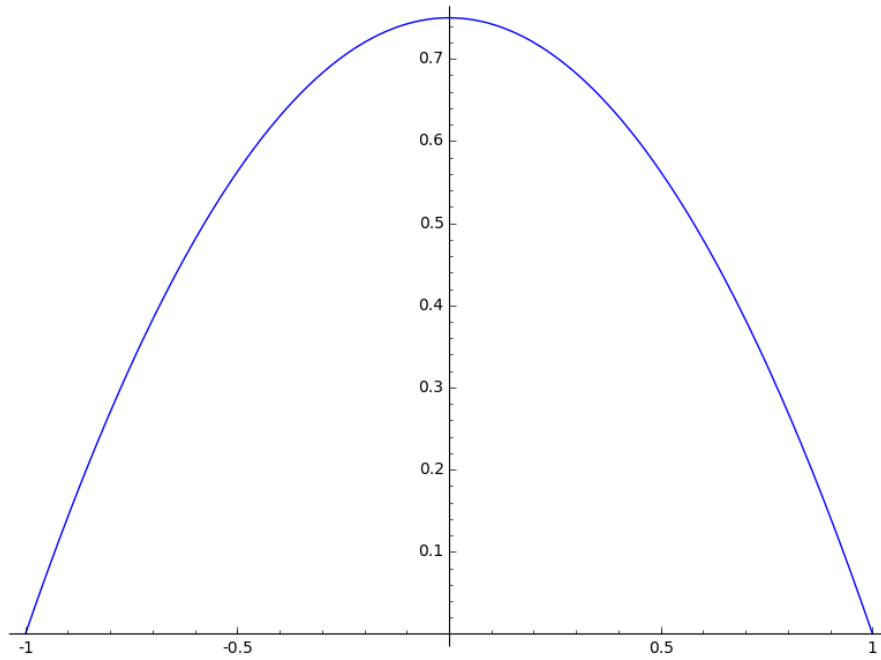
For the skewness

$$\begin{aligned}
 \text{Numerator} &= E[X^3] - 3\mu E[X^2] + 2\mu^3 \\
 &= [0^3 \cdot 0.10 + 1^3 \cdot 0.25 + 2^3 \cdot 0.40 + 4^3 \cdot 0.15 + 7^3 \cdot 0.10] - 3 \cdot 2.35 \cdot 9.15 + 2 \cdot 2.35^3 \\
 &\approx 0 + 0.25 + 3.20 + 9.60 + 34.3 - 64.5075 + 25.96 \\
 &= 47.35 - 64.5075 + 25.96 \\
 &\approx 8.80
 \end{aligned}$$

which yields a skewness of $\gamma_1 = 8.80/\sigma^3 \approx 1.27$. This indicates a slight skewness to the right of the mean. You can notice the 4 and 7 entries on the histogram illustrate a slight trailing off to the right.

Finally, for kurtosis

$$\begin{aligned}
 \text{Numerator} &= E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4 \\
 &= [0^4 \cdot 0.10 + 1^4 \cdot 0.25 + 2^4 \cdot 0.40 + 4^4 \cdot 0.15 + 7^4 \cdot 0.10] - 4 \cdot 2.35 \cdot 47.35 + 6 \cdot 2.35^2 \cdot 9.15
 \end{aligned}$$



$$\begin{aligned}
 &\approx 0 + 0.25 + 6.40 + 38.4 + 240.1 - 445.09 + 303.19 - 91.49 \\
 &\approx 285.15 - 445.09 + 303.19 - 91.49 \\
 &\approx 51.75
 \end{aligned}$$

which yields a kurtosis of $\gamma_2 = 51.75/\sigma^4 \approx 3.93$ which also notes that the data appears to have a modestly bell-shaped distribution.

Consider the following example when computing these statistics for a continuous variable. Let $f(x) = \frac{3}{4} \cdot (1 - x^2)$ over $R = [-1, 1]$.

Then for the mean

$$\begin{aligned}
 \mu &= \int_{-1}^1 x \cdot \frac{3}{4} \cdot (1 - x^2) dx \\
 &= \int_{-1}^1 \frac{3}{4} \cdot (x - x^3) dx \\
 &= \frac{3}{4} \cdot (x^2/2 - x^4/4) \Big|_{-1}^1 \\
 &= \frac{3}{4} \cdot [(1/2) - (1/4)] - [(1/2) - (1/4)] \\
 &= 0
 \end{aligned}$$

as expected since the probability function is symmetric about $x=0$.

For the variance

$$\begin{aligned}
 \sigma^2 &= \int_{-1}^1 x^2 \cdot \frac{3}{4} \cdot (1 - x^2) dx - \mu^2 \\
 &= \int_{-1}^1 \frac{3}{4} \cdot (x^2 - x^4) dx - 0 \\
 &= \frac{3}{4} \cdot (x^3/3 - x^5/5) \Big|_{-1}^1 \\
 &= \frac{3}{4} \cdot 2 \cdot (1/3 - 1/5)
 \end{aligned}$$

$$\begin{aligned}
&= \frac{3}{4} \cdot \frac{4}{15} \\
&= \frac{1}{5}
\end{aligned}$$

and taking the square root gives a standard deviation slightly less than $1/2$. Notice that four times this value encompasses almost all of the range of the distribution.

For the skewness, notice that the graph is symmetrical about the mean and so we would expect a skewness of 0. Just to check it out

$$\begin{aligned}
\text{Numerator} &= E[X^3] - 3\mu E[X^2] + 2\mu^3 \\
&= \int_{-1}^1 x^3 \cdot \frac{3}{4} \cdot (1 - x^2) dx - 3E[X^2] \cdot 0 + 0^3 \\
&= \int_{-1}^1 \frac{3}{4} \cdot (x^3 - x^5) dx \\
&= \frac{3}{4} \cdot (x^4/4 - x^6/6) \Big|_{-1}^1 \\
&= 0
\end{aligned}$$

as expected without having to actually complete the calculation by dividing by the cube of the standard deviation.

Finally, note that the probability function in this case is modestly close to a bell shaped curve so we would expect a kurtosis in the vicinity of 3. Indeed, noting that (conveniently) $\mu = 0$ gives

$$\begin{aligned}
\text{Numerator} &= E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4 \\
&= \int_{-1}^1 x^4 \cdot \frac{3}{4} \cdot (1 - x^2) dx \\
&= \frac{3}{4} \cdot (x^5/5 - x^7/7) \Big|_{-1}^1 \\
&= \frac{3}{4} \cdot 2(1/5 - 1/7) \\
&= \frac{3}{35}
\end{aligned}$$

and so by dividing by $\sigma^4 = \sqrt{\frac{1}{5}}^4 = \frac{1}{25}$ gives a kurtosis of

$$\gamma_2 = \frac{3}{35} / \frac{1}{25} = \frac{75}{35} \approx 2.14.$$

Example 5.4.8 Consider [our previous example 5.4.4](#). To compute the mean and standard deviation for this distribution,

$$\mu = \int_{-1}^2 x \cdot x^2/3 dx = \int_{-1}^2 x^3/3 dx = \frac{2^4}{12} - \frac{(-1)^4}{12} = \frac{15}{12} = \frac{5}{4}$$

and by using [the alternate formulas 5.4.6](#)

$$\begin{aligned}
\sigma^2 &= E[X^2] - \mu^2 \\
&= \int_{-1}^2 x^2 \cdot x^2/3 dx - \mu^2 \\
&= \int_{-1}^2 x^4/3 dx - \left(\frac{5}{4}\right)^2
\end{aligned}$$

$$\begin{aligned}
&= \frac{2^5}{15} - \frac{(-1)^5}{15} - \frac{25}{16} \\
&= \frac{33}{15} - \frac{25}{16} = \frac{51}{80}
\end{aligned}$$

which gives

$$\sigma = \sqrt{\frac{51}{80}} \approx 0.7984.$$

For skewness, note that in computing the variance above you also found that

$$E[X^2] = \frac{11}{5}.$$

So, once again by using [the alternate formulas 5.4.6](#)

$$E[X^3] = \int_{-1}^2 x^3 \cdot x^2/3 dx = \frac{x^6}{18} \Big|_{-1}^2 = \frac{7}{2}$$

and so

$$\gamma_1 = \frac{\frac{7}{2} - 3 \cdot \frac{5}{4} \cdot \frac{11}{5} + 2 \cdot \left(\frac{5}{4}\right)^3}{\sqrt{\frac{51}{80}}^3}$$

For kurtosis, you can reuse $E[X^3] = \frac{7}{2}$ and $E[X^2] = \frac{11}{5}$ and [the alternate formulas 5.4.6](#) to determine

$$E[(X - \mu)^4] = E[X^4] - 4\mu \cdot E[X^3] + 6\mu^2 \cdot E[X^2] - 3\mu^4$$

which is the numerator for the kurtosis. □

Example 5.4.9 Roulette. Roulette is a gambling game popular in many casinos in which a player attempts to win money from the casino by predicting the location that a ball lands on in a spinning wheel. There are two variations of this game...the American version and the European version. The difference being that the American version has one additional numbered slot on the wheel. The American version of the game will be used for the purposes of this example. A Roulette wheel consists of 38 equally-sized sectors identified with the numbers 1 through 36 plus 0 and 00. The 0 and 00 sectors are colored green and half of the remaining numbers are in sectors colored red with the remainder colored black. A steel ball is dropped onto a spinning wheel and as the wheel comes to rest the sector in which it comes to rest is noted. It is easy to determine that the probability of landing on any one of the 38 sectors is $1/38$. A picture of a typical American-style wheel and betting board is given by .

(Found at BigFishGames.com.)

Since this is a game in a casino, there must be the opportunity to bet (and likely lose) money. For the remainder of this example we will assume that you are betting 1 dollar each time. If you were to bet more then the values would scale correspondingly. However, if you place your bet on any single number and the ball ends up on the sector corresponding to that number, you win a net of 35 dollars. If the ball lands elsewhere you lose your dollar. Therefore the expected value of winning if you bet on one number is

$$E[\text{win on one}] = 35 \cdot \frac{1}{38} - 1 \cdot \frac{37}{38} = -\frac{2}{38}$$

which is a little more than a nickel loss on average.

You can bet on two numbers as well and if the ball lands on either of the two then you win a payout in this case of 17 dollars. Therefore the expected value of winning if you bet on two numbers is

$$E[\text{win on two numbers}] = 17 \cdot \frac{2}{38} - 1 \cdot \frac{36}{38} = -\frac{2}{38}.$$

Continuing, you can bet on three numbers and if the ball lands on any of the three then you win a payout of 11 dollars. Therefore the expected value of winning if you bet on three numbers is

$$E[\text{win on three numbers}] = 11 \cdot \frac{3}{38} - 1 \cdot \frac{35}{38} = -\frac{2}{38}.$$

You can bet on all reds, all blacks, all evens (ignoring 0 and 00), or all odds and get your dollar back. The expected value for any of these options is

$$E[\text{win on eighteen numbers}] = 1 \cdot \frac{18}{38} - 1 \cdot \frac{20}{38} = -\frac{2}{38}.$$

There is one special way to bet which uses the the 5 numbers 0, 00, 1, 2, 3 and pays 6 dollars. This is called the "top line of basket". Notice that the use of five numbers will make getting the same expected value as the other cases impossible using regular dollars and cents. The expected value of winning in this case is

$$E[\text{win on top line of basket}] = 6 \cdot \frac{5}{38} - 1 \cdot \frac{33}{38} = -\frac{3}{38}$$

which is of course worse and is the only normal way to bet on roulette which has a different expected value.

There are other possible ways to bet on roulette but none provide a better expected value of winning. The moral of this story is that you should never bet on the 5 number option and if you ever get ahead by winning on roulette using any of the possible options then you should probably stop quickly since over a long period of time it is expected that you will lose an average of $\frac{1}{19}$ dollars per game. \square

Going back to Pascal's wager, let

- $X = 0$ represent disbelief when God doesn't exist
- $X = 1$ represent disbelief when God does exist
- $X = 2$ represent belief when God does exist
- $X = 3$ represent belief when God does not exist

Presume that p is the likelihood that God exists. Then you can compute the expected value of disbelief and the expected value of belief by first creating a value function. Below, for argument sake we are somewhat randomly assign a value of one million to disbelief if God doesn't exist. The conclusions are the same if you choose any other finite number...

$$\begin{aligned}v(0) &= 1,000,000, f(0) = 1 - p \\v(1) &= -\infty, f(1) = p \\v(2) &= \infty, f(2) = p \\v(3) &= 0, f(3) = 1 - p\end{aligned}$$

Then,

$$\begin{aligned}E[\text{disbelief}] &= v(0)f(0) + v(1)f(1) \\&= 1000000 \times (1 - p) - \infty \times p \\&= -\infty\end{aligned}$$

if $p > 0$. On the other hand,

$$\begin{aligned}E[\text{belief}] &= v(2)f(2) + v(3)f(3) \\&= \infty \times p + 0 \times (1 - p) \\&= \infty\end{aligned}$$

if $p > 0$. So Pascal's conclusion is that if there is even the slightest chance that God exists then belief is the smart and scientific choice.

5.5 Standard Units

Any distribution variable can be converted to "standard units" using the linear translation

$$z = \frac{x - \mu}{\sigma}.$$

In doing so, values of z will always represent the number of standard deviations x is from the mean and will provide "dimensionless" comparisons.

Example 5.5.1 Consider our earlier [continuous example 5.4.4](#) in which we found $\mu = \frac{5}{4}$ and $\sigma = \sqrt{\frac{51}{80}}$. Then,

$$P(0 < X < 1) = P\left(\frac{0 - \frac{5}{4}}{\sqrt{\frac{51}{80}}} < \frac{X - \frac{5}{4}}{\sqrt{\frac{51}{80}}} < \frac{1 - \frac{5}{4}}{\sqrt{\frac{51}{80}}}\right)$$

gives the middle term is Z and the other endpoints are now in standard units that indicate the number of standard deviations from the mean rather than actual problem units. \square

5.6 Summary

TBA

5.7 Exercises

Checkpoint 5.7.1 Flipping A Fixed Number of Coins. Consider the random variable from the previous section where you flip three coins and measure the number of heads obtained. Determine $f(0)$, $f(1)$, $f(2)$, and $f(3)$ and the corresponding distribution function $F(x)$. These can be expressed in a table format. Generalize your answer to the case when you flip a n coins where n is a fixed natural number.

Checkpoint 5.7.2 Later. B.

Checkpoint 5.7.3 Flipping Fixed Number of Coins. You flip three coins and measure the number of heads obtained. Determine the space R for the corresponding random variable X . From the eight possible outcomes, determine all outcomes corresponding to $X=2$. Identify the random variable as discrete or continuous.

Checkpoint 5.7.4 Flipping Coins till success. You flip one coin repeatedly until you get a second head. Determine the space R for the corresponding random variable X . From the possibilities, determine all outcomes corresponding to $X=4$. Identify the random variable as discrete or continuous.

Checkpoint 5.7.5 Time between Accidents. Now you want to measure the time between accidents at a particular intersection in town. Determine the space R for the corresponding random variable X . Describe all outcomes corresponding to $X < 1$. Be purposeful in the problem to describe the units you are using to measure time. Identify the random variable as discrete or continuous.

Chapter 6

Distributions based upon Equally likely Outcomes

6.1 Introduction

When motivating our definition of probability you may have noticed that we modeled our definition on the relative frequency of equally-likely outcomes. In this chapter you will develop the theoretical formulas which can be used to model equally-likely outcomes.

In this chapter, you will investigate the following distributions:

1. Discrete Uniform - each of a finite collection of outcomes is equally likely and prescribed a "position" and X measures the position of an item selected randomly from the outcomes.
2. Continuous Uniform - an interval of values is possible with sub-intervals of equal length having equal probabilities and X measures a location inside that interval.
3. Hypergeometric - each of a finite collection of values are equally likely and grouped into two classes (successes vs failures) and a subset of that collection is extracted with X measuring the number of successes in the sample.

6.2 Discrete Uniform Distribution

In this section, you will investigate distributions that begin with individual outcomes that are equally likely and expand into more general settings.

Theorem 6.2.1 Discrete Uniform Distribution. *Assume outcomes in $R = 1, 2, 3, \dots, n$ are equally likely. Then, the probability function for the discrete uniform variable X is*

$$f(x) = \frac{1}{n}$$

for $x \in R$.

Proof. Assume that you have a variable with space $R = 1, 2, 3, \dots, n$ so that the likelihood of each value is equally likely. Then, the probability function

satisfies $f(x) = c$ for any $x \in R$. As before, since $\sum_{x \in R} f(x) = 1$, then

$$f(x) = \frac{1}{n}. \quad \blacksquare$$

```
# Uniform distribution over 1 .. n
pretty_print("Discrete Uniform Distribution over the set 1, 2, ..., n")
var('x')
@interact
def _(n=slider(2,10,1,2)):
    np1 = n+1
    R = range(1,np1)
    f(x) = 1/n
    pretty_print(html('Density Function:  $f(x)$  over the space  $R$ '))
    points((k,f(x=k)) for k in R).show()
    for k in R:
        pretty_print(html('f(%s)' % k) + ' = ' + '%s' % f(x=k) + ' \approx ' + '%s' % f(x=k).n(digits=5)))
```

Theorem 6.2.2 Properties of the Discrete Uniform Probability Function. $f(x) = \frac{1}{n}$ over $R = 1, 2, 3, \dots, n$ satisfies the properties of a discrete probability function and

1. $\mu = \frac{1+n}{2}$
2. $\sigma^2 = \frac{n^2-1}{12}$
3. $\gamma_1 = 0$
4. $\gamma_2 = \frac{6}{5} \frac{1+n^2}{1-n^2}$
5. Distribution function $F(x) = \frac{x}{n}$ for $x \in R$.

Proof. Trivially, by construction you get by summing over $R = \{1, 2, \dots, n\}$

$$\sum_{x=1}^n \frac{1}{n} = 1$$

Also, $1/n$ is positive for all x values.

To determine the mean,

$$\begin{aligned} \mu &= \sum_{x=1}^n x \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{x=1}^n x \\ &= \frac{1}{n} \frac{n(n+1)}{2} \\ &= \frac{1+n}{2} \end{aligned}$$

To determine the variance,

$$\sigma^2 = \sum_{x=1}^n x^2 \cdot \frac{1}{n} - \mu^2$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{x=1}^n x^2 - \left(\frac{1+n}{2} \right)^2 \\
&= \frac{1}{n} \frac{n(n+1)(2n+1)}{6} - \frac{1+2n+n^2}{4} \\
&= \frac{(2n^2+3n+1)}{6} - \frac{1+2n+n^2}{4} \\
&= \frac{(4n^2+6n+2)}{12} - \frac{3+6n+3n^2}{12} \\
&= \frac{(n^2-1)}{12}
\end{aligned}$$

For skewness,

$$\begin{aligned}
\gamma_1 &= \sum_{x=1}^n x^3 \cdot \frac{1}{n} - 3\mu \sum_{x=1}^n x^2 \cdot \frac{1}{n} + 2\mu^3 \\
&= \frac{n^2(n+1)^2}{4n} - 3 \frac{(n(n+1)(2n+1))}{2n} \frac{1+n}{2} + 2 \left(\frac{1+n}{2} \right)^3 \\
&= \frac{n^2(n+1)^2}{4n} - \frac{(n+1)^2(n(2n+1))}{4n} + \frac{(n+1)^3}{4} \\
&= \frac{(n+1)^2}{4} [n - 2n - 1 + (n+1)] \\
&= 0
\end{aligned}$$

which should be obvious since the histogram for this distribution is constantly flat.

For Kurtosis, use the fourth moment and simplify. This is tedious, the algebra is performed using Sage in the active cell below this proof. However, you might want to supply the remainder of this proof using the fact that

$$\sum_{x=1}^n x^4 = \frac{6n^5 + 15n^4 + 10n^3 - n}{30}.$$

■

Sage can also do the algebra for you to determine each of these measures. Notice, as n increases the Kurtosis approaches $\frac{6}{5}$ which indicates that there is (obviously) no tend toward central tendency over time.

```

var('x,n')
f = 1/n
mu = sum(x*f,x,1,n).factor()
pretty_print('Mean= ',mu)
mu = (1+n)/2
v = sum((x-mu)^2*f, x, 1, n)
pretty_print('Variance= ',v.factor())
stand = sqrt(v)
pretty_print('Skewness= ',(sum((x-mu)^3*f, x, 1, n)/stand^3))
kurt = sum((x-mu)^4*f, x, 1, n)/stand^4
pretty_print('Kurtosis= ',(kurt-3).factor(), ' + 3')

```

Example 6.2.3 Rolling one die. When you consider rolling a regular, fair, single 6-sided die, each side is equally likely. The sample space consists of the 6 sides, each with a unique number of physical dots. Let the random variable X correspond each side with the number corresponding to the number of dots.

Then, $R = 1, 2, 3, 4, 5, 6$. Since each side is equally likely then $f(x) = 1/6$. Further, the probability of getting an outcome in $A=2,3$ would be $f(2)+f(3) = 1/6 + 1/6 = 2/6$. \square

6.3 Continuous Uniform Distribution

Modeling the idea of "equally-likely" in a continuous world requires a slightly different perspective since there are obviously infinitely many outcomes to consider. Instead, you should consider requiring that intervals in the domain which are of equal width should have the same probability regardless of where they are in that domain. This behaviour suggests $P(u < X < v) = P(u + \Delta < X < v + \Delta)$ for reasonable values of Δ so that the interval remains inside R .

Theorem 6.3.1 For $R = [a, b]$, with $a < b$, the continuous uniform probability function is given by

$$f(x) = \frac{1}{b-a}.$$

Proof. From before, for X a continuous uniform variable, we get

$$\begin{aligned} \int_u^v f(x)dx &= \int_{u+\Delta}^{v+\Delta} f(x)dx \\ F(v) - F(u) &= F(v + \Delta) - F(u + \Delta) \\ F(u + \Delta) - F(u) &= F(v + \Delta) - F(v) \\ \frac{F(u + \Delta) - F(u)}{\Delta} &= \frac{F(v + \Delta) - F(v)}{\Delta} \end{aligned}$$

which is true regardless of Δ so long as you stay in the domain of interest. Letting $\Delta \rightarrow 0$ gives

$$F'(u) = F'(v)$$

but since F is an antiderivative of the probability function,

$$f(u) = f(v)$$

for all u and v in R . This only happens if f is constant...say, $f(x)=c$. If the space of X is a single interval with $R = [a, b]$ then

$$1 = \int_a^b c dx = c(b-a)$$

which yields $c = \frac{1}{b-a}$ as desired. \blacksquare

Example 6.3.2 Basic Continuous Uniform. On $R = [1, 2\pi]$,

$$f(x) = \frac{1}{2\pi - 1}.$$

Then, if you want to compute something like $P(2 < X < 4.5)$ integrate

$$P(2 < X < 4.5) = \int_2^{4.5} \frac{1}{2\pi - 1} dx = \frac{2.5}{2\pi - 1}$$

\square

Example 6.3.3 Continuous Uniform over two disjoint intervals.

Suppose $R = [0, 2] \cup [5, 7]$. Then, as in the theorem proof

$$1 = \int_R c dx = \int_0^2 c dx + \int_5^7 c dx = 4c.$$

Thus, $f(x) = \frac{1}{4}$. For computing probabilities, you will want to break up any resulting integrals in a similar manner. \square

Theorem 6.3.4 Properties of the Continuous Uniform Probability Function. For the Continuous Uniform Distribution over $R = [a, b]$, with $a < b$,

1. $f(x) = \frac{1}{b-a}$ satisfies the properties of a probability function over $R = [a, b]$.
2. $\mu = \frac{a+b}{2}$
3. $\sigma^2 = \frac{b^2-a^2}{12}$
4. $\gamma_1 = 0$
5. $\gamma_2 = \frac{9(a^5-5a^4b+10a^3b^2-10a^2b^3+5ab^4-b^5)(a-b)}{5(a^3-3a^2b+3ab^2-b^3)^2}$

```
# Continous uniform distribution statistics derivation
reset()
var('x,a,b')

f = 1/(b-a)

mu = integrate(x*f,x,a,b).factor()
pretty_print('Mean_=',mu)

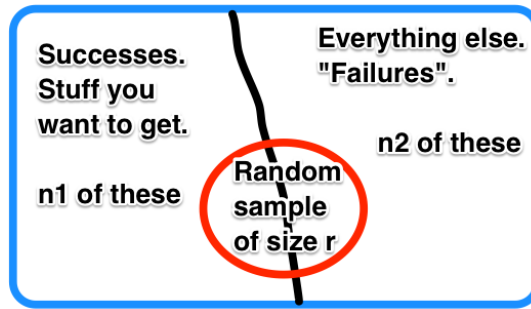
v = integrate((x-mu)^2*f, x, a, b)

pretty_print('Variance_=',v.factor())
stand = sqrt(v)
sk = (integrate((x-mu)^3*f, x, a, b)/stand^3)
pretty_print('Skewness_=',sk)
kurt = (integrate((x-mu)^4*f, x, a, b)/stand^4)
pretty_print('Kurtosis_=',kurt)

pretty_print('Several_Examples')
a1=0
for b1 in range(2,7):
    pretty_print('Using_[' ,a1 ,',',b1 ,']:')
    pretty_print('      mean_=',mu(a=a1,b=b1))
    pretty_print(' variance_=',v(a=a1,b=b1))
    pretty_print(' skewness_=',sk(a=a1,b=b1))
    pretty_print(' kurtosis_=',kurt(a=a1,b=b1))
```

Example 6.3.5 Occurrence of exactly one event randomly in a given interval. Suppose you know that only one person showed up at the counter of a local business in a given 30 minute interval of time. Then, $R=[0,30]$ given $f(x) = 1/30$.

Further, the probability that the person arrived within the first 6 minutes



would be $\int_0^6 \frac{1}{30} dx = 0.2$. □

Theorem 6.3.6 *Distribution Function for Continuous Uniform.* For $x \in [a, b]$, $F(x) = \frac{x-a}{b-a}$

Proof. For x in this range,

$$F(x) = \int_a^x \frac{1}{b-a} du = \frac{u}{b-a} \Big|_a^x = \frac{x-a}{b-a}. \quad \blacksquare$$

6.4 Hypergeometric Distribution

For the discrete uniform distribution, the presumption is that you will be making a selection one time from the collection of items. However, if you want to take a larger sample without replacement from a distribution in which originally all are equally likely then you will end up with something which will not be uniform.

Indeed, consider a collection of n items from which you want to take a sample of size r without replacement. If n_1 of the items are "desired" and the remaining $n_2 = n - n_1$ are not, let the random variable X measure the number of items from the first group in your sample with $R = \{0, 1, \dots, \min(r, n_1)\}$. The resulting collection of probabilities is called a Hypergeometric Distribution.

Theorem 6.4.1 *Hypergeometric Probability Function.* For a Hypergeometric random variable with $R = 0, 1, \dots, r$ and assuming $n_1 \geq r$ and $n - n_1 \geq r$,

$$f(x) = \frac{\binom{n_1}{x} \binom{n-n_1}{r-x}}{\binom{n}{r}}$$

Proof. For the following, we will presume that Since you are sampling without replacement and trying only measure the number of items from your desired group in the sample, then the space of X will include $R = 0, 1, \dots, r$ assuming $n_1 \geq r$ and $n - n_1 \geq r$. In the case when r is too large for either of these, the formulas below will follow noting that binomial coefficients are zero if the top is smaller than the bottom or if the bottom is negative.

So $f(x) = P(X = x) = P(x \text{ from the sample are from the target group and the remainder are not})$. Breaking these up gives

$$f(x) = \frac{\binom{n_1}{x} \binom{n-n_1}{r-x}}{\binom{n}{r}} \quad \blacksquare$$

For example, suppose that you have a bag of assorted candies but you really prefer the little dark chocolate bars. Because you are obsessive, you first empty

the whole bag onto your desk and discover that the bag contains 33 equally-sized candy bars of which 6 of them are your delightful dark chocolate bars. Putting the bars randomly back into the bag, you find that a friend's friend's friend walks into the room and grabs a handful of 5 candy bars from your bag. You are shocked and would like to know the probability that this person got 2 or more of your dark chocolate candy bars.

As a good prob/stats student you recognize that this situation fits the requirements of the hypergeometric distribution with $n_1 = 6, n_2 = 27, r = 5$ and you want $P(X \geq 2)$. You determine that it would be easier to compute the complement

$$1 - P(X \leq 1) = 1 - f(0) - f(1).$$

Therefore,

$$P(X \geq 2) = 1 - \frac{\binom{6}{0}\binom{27}{5}}{\binom{33}{5}} - \frac{\binom{6}{1}\binom{27}{4}}{\binom{33}{5}}$$

or after some simplification

$$P(X \geq 2) = 1 - \frac{13455}{39556} - \frac{8775}{19778} \approx 0.21617$$

Therefore, you have about 1 chance out of 5 that the friend got 2 or more of your bars. You can be somewhat confident that plenty of dark chocolate bars remain hoarded for yourself.

Theorem 6.4.2 Properties of the Hypergeometric Distribution.

1. $f(x) = \frac{\binom{n_1}{x}\binom{n-n_1}{r-x}}{\binom{n}{r}}$ satisfies the properties of a probability function.
2. $\mu = r \frac{n_1}{n}$
3. $\sigma^2 = r \frac{n_1}{n} \frac{n_2}{n} \frac{n-r}{n-1}$
4. $\gamma_1 = \frac{(n-2n_1)\sqrt{n-1}(n-2r)}{rn_1(n-n_1)\sqrt{n-r}(n-2)}$
5. $\gamma_2 = \frac{n(n+1)-6n(n-r)}{n_1(n-n_1)} + \frac{3r(n-r)(n+6)}{n^2} - 6$

Proof.

1.

$$\begin{aligned} \sum_{x=0}^n \binom{n}{x} y^x &= (1+y)^n, \text{ by the Binomial Theorem} \\ &= (1+y)^{n_1} \cdot (1+y)^{n_2} \\ &= \sum_{x=0}^{n_1} \binom{n_1}{x} y^x \cdot \sum_{x=0}^{n_2} \binom{n_2}{x} y^x \\ &= \sum_{x=0}^n \sum_{t=0}^r \binom{n_1}{r} \binom{n_2}{r-t} y^x \end{aligned}$$

Equating like coefficients for the various powers of y gives

$$\binom{n}{r} = \sum_{t=0}^r \binom{n_1}{r} \binom{n_2}{r-t}.$$

Dividing gives

$$1 = \sum_{x=0}^r f(x).$$

2. For the mean

$$\begin{aligned}
 \sum_{x=0}^n x \frac{\binom{n_1}{x} \binom{n-n_1}{r-x}}{\binom{n}{r}} &= \frac{1}{\binom{n}{r}} \sum_{x=1}^n \frac{n_1(n_1-1)!}{(x-1)!(n_1-x)!} \binom{n-n_1}{r-x} \\
 &= \frac{n_1}{\binom{n}{r}} \sum_{x=1}^n \frac{(n_1-1)!}{(x-1)!((n_1-1)-(x-1))!} \binom{n-n_1}{r-x} \\
 &= \frac{n_1}{\frac{n(n-1)!}{r!(n-r)!}} \sum_{x=1}^n \binom{n_1-1}{x-1} \binom{n-n_1}{r-x}
 \end{aligned}$$

Consider the following change of variables for the summation:

$$\begin{aligned}
 y &= x - 1 \\
 n_3 &= n_1 - 1 \\
 s &= r - 1 \\
 m &= n - 1
 \end{aligned}$$

Then, this becomes

$$\begin{aligned}
 \mu &= \sum_{x=0}^n x \frac{\binom{n_1}{x} \binom{n-n_1}{r-x}}{\binom{n}{r}} = r \frac{n_1}{n} \sum_{y=0}^m \frac{\binom{n_3}{y} \binom{m-n_3}{s-y}}{\binom{m}{s}} \\
 &= r \frac{n_1}{n} \cdot 1
 \end{aligned}$$

noting that the summation is in the same form as was show yields 1 above.

3. For variance, we will use an alternate form of the definition that is useful when looking for cancellation options with the numerous factorials in the hypergeometric probability function. Indeed, you can easily notice that

$$\sigma^2 = E[X^2] - \mu^2 = E[X^2 - X] + E[X] - \mu^2 = E[X(X-1)] + \mu - \mu^2.$$

Since we have $\mu = r \frac{n_1}{n}$ from above then let's focus on the first term only and use the substitutions

$$\begin{aligned}
 y &= x - 2 \\
 n_3 &= n_1 - 2 \\
 s &= r - 2 \\
 m &= n - 2
 \end{aligned}$$

to get

$$\begin{aligned}
 E[X(X-1)] &= \sum_{x=0}^n x(x-1) \frac{\binom{n_1}{x} \binom{n-n_1}{r-x}}{\binom{n}{r}} \\
 &= \sum_{x=2}^n x(x-1) \frac{\frac{n_1!}{x(x-1)(x-2)!} \frac{(n-n_1)!}{(n_1-x)!}}{\binom{n}{r}} \\
 &= \sum_{x=2}^n \frac{\frac{n_1!}{(x-2)!(n_1-x)!} \frac{n_2!}{(r-x)!(n_2-r+x)!}}{\binom{n}{r}} \\
 &= n_1 \cdot (n_1-1) \sum_{x=2}^n \frac{\frac{(n_3)!}{(x-2)!(n_3-(x-2))!} \frac{n_2!}{((r-2)-(x-2))!(n_2-(r-2)+(x-2))!}}{\binom{n}{r}}
 \end{aligned}$$

$$\begin{aligned}
&= n_1 \cdot (n_1 - 1) \sum_{y=0}^m \frac{\frac{(n_3)!}{y!(n_3-y)!} \frac{n_2!}{(s-y)!(n_2-s+y)!}}{\binom{n}{r}} \\
&= \frac{n_1 \cdot (n_1 - 1) \cdot r \cdot (r - 1)}{n(n - 1)} \sum_{y=0}^m \frac{\binom{n_3}{y} \binom{n_2}{s-y}}{\binom{m}{s}} \\
&= \frac{n_1 \cdot (n_1 - 1) \cdot r \cdot (r - 1)}{n(n - 1)}
\end{aligned}$$

where we have used the summation formula above that showed that $f(x)$ was a probability function.

Putting this together with the earlier formula gives

$$\sigma^2 = \frac{n_1 \cdot (n_1 - 1) \cdot r \cdot (r - 1)}{n(n - 1)} + r \frac{n_1}{n} - \left(r \frac{n_1}{n} \right)^2.$$

4. The proof of the variance formula is similar and uses $E(X(X-1)) = r(r-1) \frac{n_1(n_1-1)}{n(n-1)}$. The proof of skewness and kurtosis are messy and we won't bother with them for this distribution!

■

Note, if $r=1$ then you are back at a regular discrete uniform model. Indeed,

$$P(\text{desired item}) = 1 \cdot \frac{n_1}{n} = \mu.$$

which is indeed what you might expect when selecting once.

```

N1 = 10
N2 = 16
n = N1+N2
r = 5

X = 0:r      # the space R of the random variable
mu = r*N1/n  # the formula for mean of the Binomial
              Distributions
sdev = sqrt(mu*(N2/n)*(n-r)/(n-1)) # the formula for the
              standard deviation
dhyper( X, N1, N2, r ) # let's print out a bunch of actual
              probs

Phyper = dhyper(X, N1, N2, r ) # create the probability
              function over X

Psample = rhyper(10^6, N1, N2, r) # to create a histogram,
              sample a lot
Xtop=max(Psample) # for scaling the x-axis. Shift
              by 1/2 below.
hist(Psample, prob=TRUE, br=(-1:Xtop)+0.5, col="skyblue2",
     xlab="X",
     main="Hypergeometric_Probability_Function_vs_Approximating_
           'Bell_Curve'")

points(X, Phyper, pch=19, col="darkgreen") # to create
              actual (x,f(x))

```

```
Pnormal <- function(X){dnorm(X, mean=mu, sd=sdev)} # to
  overlap a bell curve
curve(Pnormal, col="red", lwd=2, add=TRUE)
```

6.5 Summary

Here is a summary of the major formulas from this chapter:

Discrete Uniform $f(x)$

Discrete Uniform statistics

Continuous Uniform $f(x)$

Continuous Uniform statistics

Hypergeometric $f(x)$

Hypergeometric statistics

6.6 Exercises

Checkpoint 6.6.1 - The Proverbial Urn Problem. You have an urn with 10 marbles of which $n_1 = 6$ are red and $N_2 = 4$ are blue. You select randomly $r = 3$ of the marbles without replacement and let X represent the number of red marbles in your sample. With $R = 0, 1, 2, 3$, determine:

- $f(x)$
- $P(2 \text{ of the } 3 \text{ are red}) = f(2)$
- $P(\text{at most } 2 \text{ of the } 3 \text{ are red}) = f(0) + f(1) + f(2)$

Hint. This is hypergeometric using

$$f(x) = \frac{\binom{6}{x} \binom{4}{3-x}}{\binom{10}{3}}.$$

Checkpoint 6.6.2 - Playing Cards. You randomly select a hand of five cards without replacement from an ordinary deck of playing cards.

- Determine the probability that four of the five are spades.
- Determine the probability that three of the five are face cards (ie, Jacks, Queens, Kings, or Aces).

Hint. This exercise is actually two different hypergeometric distributions: the first is the 13 spades vs the 39 other cards and the second is the 12 face cards vs the 40 other cards.

Checkpoint 6.6.3 - Starting Seniors. You are picking an eleven member football starting team by picking randomly from a group with 15 seniors and 35 others. Determine:

- $P(\text{all seniors})$
- $P(\text{exactly } 6 \text{ seniors})$
- the expected number of seniors on the team
- If your team has all seniors, explain whether someone could suggest that your decision on members was unfair

Hint. This is a hypergeometric distribution with $n_1 = 15, n_2 = 35, r = 11$.

Checkpoint 6.6.4 - Old Faithful. Ole Faithful geyser in Yellowstone National Park erupts every 91 minutes. You show up at some random time in the eruption cycle and your tour bus plans to stay for 25 minutes. Determine the likelihood that you will be able to see it erupt. Express your answer by

giving correct formulas for $f(x)$ and $F(x)$ and then determine the specific answer to this question.

Hint. This is the continuous uniform distribution over $R = [0, 91]$.

Checkpoint 6.6.5 - Uniform Scenarios. Explain how the following situations can be modeled using a continuous uniform distribution by identifying the space R and the corresponding $f(x)$ for each situation.

- The location on a prize wheel where the spun wheel will stop.
- Given a clock with only a minute hand, the current one second interval.
- The location on a automobile tire where the next puncture will occur.

Checkpoint 6.6.6 - Continuous Uniform on a different space. Determine an explicit formula for $f(x)$ and the mean and variance for a continuous uniform distribution over $R = [-2, 3] \cup [5, 6] \cup [9, 15]$.

Solution. Since you must have

$$\int_{x \in R} f(x) dx = 1$$

and since $f(x)$ must be constant than all you must do is measure the accumulated width of the intervals in R . This is $5 + 1 + 6 = 12$ and so

$$f(x) = \begin{cases} \frac{1}{12}, & -2 \leq x \leq 3 \\ \frac{1}{12}, & 5 \leq x \leq 6 \\ \frac{1}{12}, & 9 \leq x \leq 15 \\ 0, & \text{otherwise} \end{cases}$$

For the mean,

$$\begin{aligned} \int_{x \in R} x \frac{1}{12} dx &= \int_{-2}^3 \frac{x}{12} dx + \int_5^6 \frac{x}{12} dx + \int_9^{15} \frac{x}{12} dx \\ &= \frac{9-4}{24} + \frac{36-25}{24} + \frac{225-81}{24} \\ &= \frac{5+11+144}{24} = \frac{160}{24} = \frac{20}{3}. \end{aligned}$$

For the variance,

$$\begin{aligned} \int_{x \in R} x^2 \frac{1}{12} dx - \mu^2 &= \int_{-2}^3 \frac{x^2}{12} dx + \int_5^6 \frac{x^2}{12} dx + \int_9^{15} \frac{x^2}{12} dx - \mu^2 \\ &= \frac{81+8}{36} + \frac{216-125}{36} + \frac{3375-729}{36} - \left(\frac{20}{3}\right)^2 \\ &= \frac{89+91+2646}{36} - \frac{400}{9} = \frac{2826-1600}{36} \\ &= \frac{1226}{36} \approx 34.055. \end{aligned}$$

Checkpoint 6.6.7 - Louisiana Mega Millions Lottery. To play the Mega Millions Louisiana Lottery consists of picking five numbers from 1 to 75 and one yellow Mega Ball number from 1 through 15. (You can play up to five different sets of numbers on each playslip but we will just assume one play per ticket to keep things straight.) Each play costs 1 and you can pay an additional 1 to apply a "multiplier" which multiplies any non-Jackpot prize by the Multiplier number (2, 3, 4, or 5) randomly selected at the time of the drawing. On October 10, 2016 the jackpots listed were

- Match 5 plus Mega ball = Jackpot of 49,000,000 with cash value of 32,600,000
- Match only 5 = 1,000,000 *Match 4 plus Mega ball* = 5,000
- Match only 4 = 500 *Match 3 plus Mega ball* = 50
- Match only 3 = 5 *Match 2 plus Mega ball* = 5
- Match 1 plus Mega ball = 2 *Match only the Mega ball* = 1

Verify the posted odds

- Match 5 plus Mega ball = 1 in 258,890,850
- Match only 5 = 1 in 18,492,204
- Match 4 plus Mega ball = 1 in 739,688
- Match only 4 = 1 in 52,835
- Match 3 plus Mega ball = 1 in 10,720
- Match only 3 = 1 in 766
- Match 2 plus Mega ball = 1 in 473
- Match 1 plus Mega ball = 1 in 56
- Match only the Mega ball = 1 in 21

Determine the expected payout for each ticket purchased. Also, determine what the Jackpot would need to be in order for the game to be considered "fair" with an expected value of zero.

Solution. Throughout these calculations, you can presume that the first five numbers are selected independently from the Mega Ball number. However, the first five numbers are selected without replacement so computing probabilities with those does not allow for independence. This part is hypergeometric with the $n_1 = 5$ numbers you selected being the "desired" numbers and the Lottery Commission picking a subset of size $r = 5$ from the 75 possible numbers. So, your likelihood of matching all five would be

$$\frac{\binom{5}{5} \cdot \binom{70}{0}}{\binom{75}{5}} = \frac{1}{17259390}.$$

Multiplying this by the 1 chance in 15 that you also match the Mega Ball gives

$$P(\text{Match 5 plus Mega Ball}) = \frac{1}{17259390} \cdot \frac{1}{15} = \frac{1}{258,890,850}.$$

To match only 5 means you also MUST miss the Mega Ball which has probability 14/15 to give

$$\frac{1}{17259390} \cdot \frac{14}{15} = \frac{1}{17259390 \cdot \frac{15}{14}} \approx \frac{1}{18492204}.$$

Continue in this manner to determine the other odds.

For the expected earnings, first determine a value function corresponding to each outcome and apply the discrete expected value process. This gives

$$\$32600000 \cdot \frac{1}{258,890,850} + \$1000000 \cdot \frac{1}{18,492,204}$$

$$\begin{aligned}
& + \$5000 \cdot \frac{1}{739,688} + \$500 \cdot \frac{1}{52,835} \\
& + \$50 \cdot \frac{1}{10,720} + \$5 \cdot \frac{1}{766} \\
& + \$5 \cdot \frac{1}{473} + \$2 \cdot \frac{1}{56} + \$1 \cdot \frac{1}{21} \\
& \approx \$0.3013.
\end{aligned}$$

So, the expected payout is approximately 30 cents. Subtracting the cost of playing

(1) indicates that the average winnings per play of the Louisiana Lottery would be –

70 cents. So, you would be better off to take, say, 50 cents and just give it to the local school system every time you

To determine the Jackpot A needed to make this a fair game means to solve the equation

$$\begin{aligned}
& A \cdot \frac{1}{258,890,850} + \$10,000,000 \cdot \frac{1}{18,492,204} \\
& + \$5000 \cdot \frac{1}{739,688} + \$500 \cdot \frac{1}{52,835} \\
& + \$50 \cdot \frac{1}{10,720} + \$5 \cdot \frac{1}{766} \\
& + \$5 \cdot \frac{1}{473} + \$2 \cdot \frac{1}{56} + \$1 \cdot \frac{1}{21} \\
& = 1
\end{aligned}$$

for A.

Finally, to deal with the multiplier, note that all but the Jackpot payouts would be increased by the multiplier m where $m \in \{1, 2, 3, 4, 5\}$. For the cost of an extra 1 (total cost of 2 per bet) the expected payout increases as the multiplier increases but each of these decreases likelihood of winning that payout by a factor of $1/5$. In general, let $x = 1, 2, \dots, 9$ indicate the various winning options in order listed above, $f(x)$ the corresponding probabilities listed for each option, and $u(x)$ the listed payouts. Then the expected payout is given by

$$\$32,600,000 \cdot \frac{1}{258,890,850} + \sum_{m=1}^5 \sum_{x=2}^9 m \cdot u(x) f(x) / 5$$

or

$$\begin{aligned}
& \$32,600,000 \cdot \frac{1}{258,890,850} + \sum_{m=1}^5 \frac{m}{5} \sum_{x=2}^9 u(x) f(x) \\
& = \frac{\$32,600,000}{258,890,850} + \sum_{m=1}^5 \frac{m}{5} 0.17539 \\
& = 0.12592 + 3 \cdot 0.17539 \\
& = 0.65209
\end{aligned}$$

Therefore, the expected value of spending another dollar to get the multiplier effect is about –

1.35. Since this is slightly less than doubling the expected loss of 70 cents for playing without the multiplier with

it make more sense to bet 20 cents rather than betting 1 twice. Or, you can send the extra nickel to this author of this text and call it quits.

Chapter 7

Distributions based upon Bernoulli Trials

7.1 Introduction

Many practical problems involve measuring simply whether something was a success or a failure. In these situations, "success" should not be interpreted as having any moral or subjective meaning but only construed to mean that something you are looking for actually occurs.

In situation where a single trial is performed and the result is determined only to be a success or failure is called a Bernoulli event. Indeed, one could create a corresponding probability function using a random variable X over the space $R = \{0, 1\}$ mapping $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. If $p = P(\text{success})$ then

$$f(x) = p^x \cdot (1 - p)^{1-x}$$

would be a formula but which only related to two values $P(\text{failure}) = f(0) = (1-p)$ and $P(\text{Success}) = f(1) = p$.

Notice that $p=0$ means that you will always get a failure and that $p=1$ means that you will always get a success. In these cases, X would no longer be a random variable since the outcome for X could be predicted with certainty. Therefore, we will always assume that $0 < p < 1$.

The Bernoulli distribution on its own is not extremely useful but serves as a starting point for several others that are useful. Indeed, in this chapter you will investigate distributions that relate some number of successes in multiple trials to some number of independent trials. The difference between these distributions will be that one of these variables will be fixed and the other one will be variable.

In this chapter, you will investigate the following distributions:

1. Binomial - the number of trials is fixed and X measures the variable number of successes
2. Geometric - the number of successes is fixed—at 1—and X measures the variable number of trials
3. Negative Binomial - the number of successes is fixed and X measures the variable number of trials

7.2 Binomial Distribution

Consider a sequence of n independent Bernoulli trials with the likelihood of a success p on each individual trial stays constant from trial to trial with $0 < p < 1$. If we let the variable X measure the number of successes obtained when doing a fixed number of trials n with $R = \{0, 1, \dots, n\}$, then the resulting distribution of probabilities is called a Binomial Distribution.

```
# Binomial distribution over 0 .. n
# Probability of success on one independent trial = p must
  also be given
var('x')
@interact
def _(n=slider(3,50,1,3),p=slider(1/20,19/20,1/20,1/2)):
    np1 = n+1
    R = range(np1)
    f(x) =
        factorial(n)/(factorial(x)*factorial(n-x))*p^x*(1-p)^(n-x)
    pretty_print(html('Density_Function:_%f(x)_'
        =s'%str(latex(f(x))))))
    pretty_print(html('over_the_space_%R=_%s'%str(R)))
    G = points((k,f(x=k)) for k in R)
    G.show()
    R = [k for k in R]
    probs = [f(x=k) for k in R]
    # H = histogram( R, weights = probs, align="mid",
    linewidth=2, edgecolor="blue", color="yellow")
    # H.show()
    for k in R:
        pretty_print(html('$f(%s'%k+'))=_%s'%latex(f(x=k))+'_
            \approx_%s'%f(x=k).n(digits=5)))
```

You can of course get specific values and graph the Binomial Distribution using R as well...

```
n <- 10
p <- 0.3

paste('Probability_Function')
dbinom(0:n, n, p) # gives the probability function
paste('Distribution_function')
pbinom(0:n, n, p) # gives the distribution function
paste('A_random_sample')
rbinom(15, n, p) # gives a random sample of 15 items from
  b(n,p)

x <- dbinom(0:n, size=n, prob=p)
barplot(x,names.arg=0:n, main=sprintf(paste('n=',n,'_and_p=_'
  ',p)))
```

Theorem 7.2.1 Derivation of Binomial Probability Function. For $R = 0, 1, \dots, n$,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Proof. Since successive trials are independent, then the probability of X

successes occurring within n trials is given by

$$P(X = x) = \binom{n}{x} P(SS...SFF...F) = \binom{n}{x} p^x (1-p)^{n-x} \quad \blacksquare$$

Theorem 7.2.2 Verification of Binomial Distribution Formula.

$$\sum_{x \in R} f(x) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = 1.$$

Proof. Using the Binomial Theorem with $a = p$ and $b = 1-p$ yields

$$\sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1 \quad \blacksquare$$

Utilize the interactive cell below to compute $f(x)$ and $F(x)$ for the Binomial distribution

```
# Binomial calculator
@interact
def _(p=input_box(0.3,width=15),n=input_box(10,width=15)):
    R = range(n+1)
    f(x) = binomial(n,x)*p**x*(1-p)**(n-x)
    acc = 0
    for k in R:
        prob = f(x=k)
        acc = acc+prob
        pretty_print('f(%s) = %k, '%.8f'%prob, '_and_F(%s) = %k, '%.8f'%acc)
```

Theorem 7.2.3 Binomial Distribution Statistics. *For the Binomial Distribution*

$$\begin{aligned}\mu &= np \\ \sigma^2 &= np(1-p) \\ \gamma_1 &= \frac{1-2p}{\sqrt{np(1-p)}} \\ \gamma_2 &= \frac{1-6p(1-p)}{np(1-p)} + 3\end{aligned}$$

Proof. For the mean,

$$\begin{aligned}\mu &= E[X] \\ &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^n x \frac{n(n-1)!}{x(x-1)!(n-x)!} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{(n-1)-(x-1)}\end{aligned}$$

Using the change of variables $k = x - 1$ and $m = n - 1$ yields a binomial series

$$= np \sum_{k=0}^m \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k}$$

$$= np(p + (1 - p))^m = np$$

For the variance,

$$\begin{aligned}\sigma^2 &= E[X(X-1)] + \mu - \mu^2 \\ &= \sum_{x=0}^n x(x-1) \binom{n}{x} p^x (1-p)^{n-x} + np - n^2 p^2 \\ &= \sum_{x=2}^n x(x-1) \frac{n(n-1)(n-2)!}{x(x-1)(x-2)!(n-x)!} p^x (1-p)^{n-x} + np - n^2 p^2 \\ &= n(n-1)p^2 \sum_{x=2}^n \frac{(n-2)!}{(x-2)!((n-2)-(x-2))!} p^{x-2} (1-p)^{(n-2)-(x-2)} + np - n^2 p^2\end{aligned}$$

Using the change of variables $k = x - 2$ and $m = n - 2$ yields a binomial series

$$\begin{aligned}&= n(n-1)p^2 \sum_{k=0}^m \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k} + np - n^2 p^2 \\ &= n(n-1)p^2 + np - n^2 p^2 = np - np^2 = np(1-p)\end{aligned}$$

The skewness and kurtosis can be found similarly using formulas involving $E[X(X-1)(X-2)]$ and $E[X(X-1)(X-2)(X-3)]$. The complete determination is performed using Sage below. ■

The following uses Sage to symbolically confirm the general formulas for the Binomial distribution.

```
var('x,n,p')
assume(x,'integer')
f(x) = binomial(n,x)*p^x*(1-p)^(n-x)
mu = sum(x*f,x,0,n)
M2 = sum(x^2*f,x,0,n)
M3 = sum(x^3*f,x,0,n)
M4 = sum(x^4*f,x,0,n)

pretty_print('Mean= ',mu)

v = (M2-mu^2).factor()
pretty_print('Variance= ',v)
stand = sqrt(v)

sk = ((M3 - 3*M2*mu + 2*mu^3)).factor()/stand^3
pretty_print('Skewness= ',sk)

kurt = (M4 - 4*M3*mu + 6*M2*mu^2 - 3*mu^4).factor()/stand^4
pretty_print('Kurtosis= ',(kurt-3).factor(),'+3')
```

```
n = 10
p = 0.3
X = 0:n      # the space R of the random variable
mu = n*p     # the formula for mean of the Binomial
              Distributions
sdev = sqrt(n*p*(1-p)) # the formula for the standard
                        deviation
```

```

dbinom( X, n, p ) # let's print out a bunch of actual probs

Pbinom = dbinom(X, n, p ) # create the probability function
over X

Psample = rbinom(10^6, n, p) # to create a histogram,
sample a lot
Xtop=max(Psample) # for scaling the x-axis. Shift
by 1/2 below.
hist(Psample, prob=TRUE, br=(-1:Xtop)+0.5, col="skyblue2",
xlab="X",
main="Binomial_Probability_Function_vs_Approximating_'Bell_
Curve'")

points(X, Pbinom, pch=19, col="darkgreen") # to create
actual (x,f(x))

Pnormal <- function(X){dnorm(X, mean=mu, sd=sdev)} # to
overlap a bell curve
curve(Pnormal, col="red", lwd=2, add=TRUE)

```

Flipping Coins

Suppose you flip a coin exactly 20 times. Determine the probability of getting exactly 10 heads and then determine the probability of getting 10 or fewer heads. **Solution.** This is binomial with $n = 20$, $p = 1/2$ and you are looking for $f(10)$. With these values

$$f(10) = \binom{20}{10} \cdot \left(\frac{1}{2}\right)^{10} \cdot \left(\frac{1}{2}\right)^{20-10} = \frac{46189}{262144} \approx 0.176$$

Notice, the mean for this distribution is also 10 so one might expect 10 heads in general. Next, to determine the probability for 10 or fewer heads requires $F(10) = f(0) + f(1) + \dots + f(10)$. There is no "nice" formula for F but this calculation can be performed using a graphing calculator, such as the TI-84 with $F(x) = \text{binomcdf}(n,p,x)$. In this case, $F(10) = \text{binomcdf}(20,1/2,10) = 0.588$.

7.3 Geometric Distribution

Consider the situation where one can observe a sequence of independent trials where the likelihood of a success on each individual trial stays constant from trial to trial. Call this likelihood the probability of "success" and denote its value by p where $0 < p < 1$. If we let the variable X measure the number of trials needed in order to obtain the first success with $R = \{1, 2, 3, \dots\}$, then the resulting distribution of probabilities is called a Geometric Distribution.

Theorem 7.3.1 For a Geometric variable X with $R = \{1, 2, 3, \dots\}$,

$$f(x) = (1 - p)(x - 1) \cdot p$$

Proof. Since successive trials are independent, then the probability of the first success occurring on the m th trial presumes that the previous $m-1$ trials were all failures. Therefore the desired probability is given by

$$f(x) = P(X = x) = P(FF \dots FS) = (1 - p)^{x-1}p \quad \blacksquare$$

Theorem 7.3.2 Geometric Distribution sums to 1.

$$f(x) = (1-p)^{x-1}p$$

sums to 1 over $R = \{1, 2, \dots\}$

Proof.

$$\sum_{x=1}^{\infty} f(x) = \sum_{x=1}^{\infty} (1-p)^{x-1}p = p \sum_{j=0}^{\infty} (1-p)^j = p \frac{1}{1-(1-p)} = 1 \quad \blacksquare$$

```
# Geometric distribution over 0 .. n
# Probability of success on one independent trial = p must
  also be given
var('x')
# n = 50 by default. actually should be infinite
@interact
def _(p=input_box(0.1, label='p=_'), n=[25, 50, 75, 100, 200]):
    np1 = n+1
    R = range(1, np1)
    f(x) = (1-p)^(x-1)*p
    F(x) = 1 - (1-p)^x
    pretty_print(html('Density_Function:_$f(x)_'
        =s$'%str(latex(f(x)))+'_over_the_space_$R=_$'
        %s$'%str(R)))
    points((k, f(x=k)) for k in R).show(title="Probability_
        Function")
    print
    points((k, F(x=k)) for k in R).show(title="Distribution_
        Function")
    if (n == 25):
        for k in R:
            pretty_print(html('$f(%s'%k+'')_=_$'
                %s'%latex(f(x=k))+'_\\approx_'
                %s$'%f(x=k).n(digits=5)))
```

Theorem 7.3.3 Geometric Statistics Theorem. *For the geometric distribution,*

$$\mu = 1/p$$

$$\sigma^2 = \frac{1-p}{p^2}$$

$$\gamma_1 = ADD$$

$$\gamma_2 = ADD + 3$$

Proof. For the mean,

$$\begin{aligned} \mu &= E[X] = \sum_{k=0}^{\infty} k(1-p)^{k-1}p \\ &= p \sum_{k=1}^{\infty} k(1-p)^{k-1} \\ &= p \frac{1}{(1-(1-p))^2} \end{aligned}$$

$$= p \frac{1}{p^2} = \frac{1}{p}$$

For the variance,

$$\begin{aligned}\sigma^2 &= E[X(X-1)] + \mu - \mu^2 \\ &= \sum_{k=0}^{\infty} k(k-1)(1-p)^{k-1}p + \mu - \mu^2 \\ &= (1-p)p \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} + \frac{1}{p} - \frac{1}{p^2} \\ &= (1-p)p \frac{2}{(1-(1-p))^3} + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}\end{aligned}$$

■

Theorem 7.3.4 Geometric Distribution Function.

$$F(x) = 1 - (1-p)^x$$

Proof. Consider the accumulated probabilities over a range of values...

$$\begin{aligned}P(X \leq x) &= 1 - P(X > x) \\ &= 1 - \sum_{k=x+1}^{\infty} (1-p)^{k-1}p \\ &= 1 - p \frac{(1-p)^x}{1-(1-p)} \\ &= 1 - (1-p)^x\end{aligned}$$

■

Theorem 7.3.5 Statistics for Geometric Distribution. *Mean, Variance, Skewness, Kurtosis computed by Sage.*

Proof.

```
var('x,n,p')
assume(x,'integer')
f(x) = p*(1-p)^(x-1)
mu = sum(x*f,x,0,oo).full_simplify()
M2 = sum(x^2*f,x,0,oo).full_simplify()
M3 = sum(x^3*f,x,0,oo).full_simplify()
M4 = sum(x^4*f,x,0,oo).full_simplify()

pretty_print('Mean_=_',mu)

v = (M2-mu^2).factor().full_simplify()
pretty_print('Variance_=_',v)
stand = sqrt(v)

sk = (((M3 - 3*M2*mu + 2*mu^3))/stand^3).full_simplify()
pretty_print('Skewness_=_',sk)

kurt = (M4 - 4*M3*mu + 6*M2*mu^2 - 3*mu^4).factor()/stand^4
pretty_print('Kurtosis_=_',(kurt-3).factor(),'+3')
```

■

Theorem 7.3.6 The Geometric Distribution yields a memoryless model.. If X has a geometric distribution and a and b are nonnegative integers, then

$$P(X > a + b | X > b) = P(X > a)$$

Proof. Using the definition of conditional probability,

$$\begin{aligned} P(X > a + b | X > b) &= P(X > a + b \cap X > b) / P(X > b) \\ &= P(X > a + b) / P(X > b) \\ &= (1 - p)^{a+b} / (1 - p)^b \\ &= (1 - p)^a \\ &= P(X > a) \end{aligned}$$

■

7.4 Negative Binomial

Consider the situation where one can observe a sequence of independent trials where the likelihood of a success on each individual trial stays constant from trial to trial. Call this likelihood the probability of "success" and denote its value by p where $0 < p < 1$. If we let the variable X measure the number of trials needed in order to obtain the r th success, $r \geq 1$, with $R = \{r, r + 1, r + 2, \dots\}$ then the resulting distribution of probabilities is called a Geometric Distribution.

Note that $r=1$ gives the Geometric Distribution.

Theorem 7.4.1 Negative Binomial Series.

$$\frac{1}{(a+b)^n} = \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} a^k b^{-n-k}$$

Proof. First, convert the problem to a slightly different form: $\frac{1}{(a+b)^n} = \frac{1}{b^n} \frac{1}{(\frac{a}{b}+1)^n} = \frac{1}{b^n} \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} \left(\frac{a}{b}\right)^k$

So, let's replace $\frac{a}{b} = x$ and ignore for a while the term factored out. Then, we only need to show

$$\sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} x^k = \left(\frac{1}{1+x}\right)^n$$

However

$$\begin{aligned} \left(\frac{1}{1+x}\right)^n &= \left(\frac{1}{1-(-x)}\right)^n \\ &= \left(\sum_{k=0}^{\infty} (-1)^k x^k\right)^n \end{aligned}$$

This infinite sum raised to a power can be expanded by distributing terms in the standard way. In doing so, the various powers of x multiplied together will create a series in powers of x involving x^0, x^1, x^2, \dots . To determine the final coefficients notice that the number of times m^k will appear in this product depends upon the number of ways one can write k as a sum of nonnegative integers.

For example, the coefficient of x^3 will come from the n ways of multiplying the coefficients x^3, x^0, \dots, x^0 and $x^2, x^1, x^0, \dots, x^0$ and $x^1, x^1, x^1, x^0, \dots, x^0$. This

is equivalent to finding the number of ways to write the number k as a sum of nonnegative integers. The possible set of nonnegative integers is $0, 1, 2, \dots, k$ and one way to count the combinations is to separate k *'s by $n-1$ |'s. For example, if $k = 3$ then $*||**$ means $x^1x^0x^2 = x^3$. Similarly for $k = 5$ and $|***|**|$ implies $x^0x^2x^1x^2x^0 = x^5$. The number of ways to interchange the identical *'s among the identical |'s is $\binom{n+k-1}{k}$.

Furthermore, to obtain an even power of x will require an even number of odd powers and an odd power of x will require an odd number of odd powers. So, the coefficient of the odd terms stays odd and the coefficient of the even terms remains even. Therefore,

$$\left(\frac{1}{1+x}\right)^n = \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} x^k$$

$$\text{Similarly, } \left(\frac{1}{1-x}\right)^n = \left(\sum_{k=0}^{\infty} x^k\right)^n = \sum_{k=0}^{\infty} \binom{n+k-1}{k} x^k \quad \blacksquare$$

Consider the situation where one can observe a sequence of independent trials with the likelihood of a success on each individual trial p where $0 < p < 1$. For a positive integer r , let the variable X measure the number of trials needed in order to obtain the r th success. Then the resulting distribution of probabilities is called a Negative Binomial Distribution.

Theorem 7.4.2 Negative Binomial Probability Function.

$$f(x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r,$$

for $x \in R = \{r, r+1, \dots\}$.

Proof. Since successive trials are independent, then the probability of the r th success occurring on the x -th trial presumes that in the previous $x-1$ trials were $r-1$ successes and $x-r$ failures. You can arrange these indistinguishable successes (and failures) in $\binom{x-1}{r-1}$ unique ways. Therefore the desired probability is given by

$$P(X = x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r \quad \blacksquare$$

Theorem 7.4.3 Negative Binomial Distribution Sums to 1.

$$\sum_{x=r}^{\infty} \binom{x-1}{r-1} (1-p)^{x-r} p^r = 1$$

Proof.

$$\sum_{x=r}^{\infty} \binom{x-1}{r-1} (1-p)^{x-r} p^r = p^r \sum_{x=r}^{\infty} \binom{x-1}{r-1} (1-p)^{x-r}$$

and by using $k = x - r$

$$\begin{aligned} &= p^r \sum_{k=0}^{\infty} \binom{r+k-1}{k} (1-p)^k \\ &= p^r \frac{1}{(1-(1-p))^r} \\ &= 1 \quad \blacksquare \end{aligned}$$

Below, the interactive cell symbolically computes $f(x)$ and $F(x)$ for the negative binomial distribution.

```
# Negative Binomial calculator
@interact
def _(p=input_box(0.3,width=15),r=slider(1,10,1,2)):
    n = 4*(floor(r/p)+1)
    np1 = n+1
    R = range(r,np1)
    f(x) =
        (factorial(x-1)/(factorial(r-1)*factorial(x-r)))*(1-p)^(x-r)*p^r
    acc = 0
    for k in R:
        prob = f(x=k)
        acc = acc+prob
        pretty_print('f(%s) = %k, '%k, '%.8f'%prob, ' and F(%s) = %k, '%k, '%.8f'%acc)
```

Theorem 7.4.4 Statistics for Negative Binomial Distribution. *For the Negative Binomial Distribution,*

$$\mu = \frac{r}{p}$$

$$\sigma^2 = r \frac{1-p}{p^2}$$

$$\gamma_1 = \frac{2-p}{\sqrt{r(1-p)}}$$

$$\gamma_2 = \frac{p^2 - 6p + 6}{r(1-p)} + 3$$

Proof.

```
# Negative Binomial
var('x,n,p,r,alpha')
assume(x,'integer')
assume(alpha,'integer')
assume(alpha > 2)
assume(0 < p < 1)
@interact
def _(r=[2,5,10,15,alpha]):
    f(x) = binomial(x-1,r-1)*p^r*(1-p)^(x-r)
    mu = sum(x*f,x,r,oo).full_simplify()
    M2 = sum(x^2*f,x,r,oo).full_simplify()
    M3 = sum(x^3*f,x,r,oo).full_simplify()
    M4 = sum(x^4*f,x,r,oo).full_simplify()

    pretty_print('Mean = ',mu)

    v = (M2-mu^2).full_simplify()
    pretty_print('Variance = ',v)
    stand = sqrt(v)

    sk = (((M3 - 3*M2*mu +
        2*mu^3)).full_simplify()/stand^3).factor()
    pretty_print('Skewness = ',sk)
```



```
kurt = ((M4 - 4*M3*mu + 6*M2*mu^2
        -3*mu^4)/v^2).full_simplify()
pretty_print('Kurtosis_=',(kurt-3).factor(),'+3')
```

■

This is the symbolic derivation component that should be the proof of the theorem above.

```
# Negative Binomial
var('x,n,p,r,alpha')
assume(x,'integer')
assume(alpha,'integer')
assume(alpha > 2)
assume(0 < p < 1)
@interact
def _(r=[2,5,10,15,alpha]):
    f(x) = binomial(x-1,r-1)*p^r*(1-p)^(x-r)
    mu = sum(x*f,x,r,oo).full_simplify()
    M2 = sum(x^2*f,x,r,oo).full_simplify()
    M3 = sum(x^3*f,x,r,oo).full_simplify()
    M4 = sum(x^4*f,x,r,oo).full_simplify()

    pretty_print('Mean_=',mu)

    v = (M2-mu^2).full_simplify()
    pretty_print('Variance_=',v)
    stand = sqrt(v)

    sk = (((M3 - 3*M2*mu +
            2*mu^3)).full_simplify()/stand^3).factor()
    pretty_print('Skewness_=',sk)

    kurt = ((M4 - 4*M3*mu + 6*M2*mu^2
            -3*mu^4)/v^2).full_simplify()
    pretty_print('Kurtosis_=',(kurt-3).factor(),'+3')
```

7.5 Summary

Binomial Distribution

[Binomial Distribution Statistics](#)

[Geometric Distribution](#)

[Geometric Distribution Statistics](#)

[Negative Binomial Distribution](#)

[Negative Binomial Distribution Statistics](#)

7.6 Exercises

Checkpoint 7.6.1 - Gallup Consumer Confidence Polling. A January 2008 Gallup poll on consumer confidence asked the question "How would you rate economic conditions in this country today" and 22

- $P(X \text{ is at most } 5)$.
- $P(X \text{ is at least } 5)$.
- the expected number of Excellent or Good responses.

Solution. This is a Binomial distribution with $n=25$ and $p = 0.22$.

- $P(X \text{ is at most } 5) = F(5) = f(0)+f(1)+f(2)+f(3)+f(4)+f(5) = 0.51843$
- $P(X \text{ is at least } 5) = 1 - F(4) = 0.67183$
- $\mu = np = 25 \cdot .22 = 5.5$

Checkpoint 7.6.2 - Rolling Dice. You keep on rolling a pair of dice and let X be the number of rolls needed until you get a sum of 7 or 11 for the second time. Determine:

1. $P(7 \text{ or } 11 \text{ on one roll})$
2. The expected number of rolls until you get the second 7 or 11 sum.
3. $P(X = 12)$
4. $P(X \geq 4)$

Solution. For this problem, use the Negative Binomial Distribution when looking for the number of trials till the 2nd success. p is determined in the first answer.

1. $P(7 \text{ or } 11 \text{ on one roll}) = 8/36 = 2/9$, using equally likely outcomes.
2. $\mu = \frac{r}{p} = \frac{2}{2/9} = 9$
3. $P(X = 12) = \binom{11}{1} (7/9)^{10} \cdot (2/9)^2 = 0.044$
- 4.

$$\begin{aligned}
 P(X \geq 4) &= 1 - P(x \leq 3) \\
 &= 1 - [f(2) + f(3)] \\
 &= 1 - \left[\binom{1}{1} (2/9)^2 + \binom{2}{1} (7/9)^1 \cdot (2/9)^2 \right] \\
 &= 1 - 0.1262 = 0.8738.
 \end{aligned}$$

Checkpoint 7.6.3 - Collecting Kids Meal Prizes. You love to eat at Chick-Fil-A with your kids and want to collect all of the five new book titles that come randomly included with each kids meal. If the promotion with these books starts today, determine:

1. The probability that you get a book you don't have when purchasing the first children's meal.
2. The probability that you it takes more than four purchases in order to get a second title.
3. The expected total number of children's meals you would expect to purchase in order to get all five titles.

Solution.

1. One. The first meal will certainly have a book that you have not received yet.
2. This is a geometric distribution with $p=4/5$. $P(X > 4) = 1 - F(4) = (1 - 4/5)^4 = \frac{1}{625}$ which is very small. Note, in this case you would have needed to receive the same title randomly for all of the first four kids meal purchases. If this were to ever happen, please let the people at the counter know and it will be their pleasure to swap out for a new title.

3. Use the geometric distribution five times with changing values for p . For the first book $p = 1$ means you are certain to get a new title. For the second book title the probability of success is $p=4/5$; for the third book title the probability of success is $p=3/5$; for the fourth the probability is $p=2/5$; and for the last the probability is $p = 1/5$. Using the mean as $1/p$ in each case and accumulating these gives the total expected number of meals to purchase as

$$\begin{aligned}
 & 1 + \frac{1}{\frac{4}{5}} + \frac{1}{\frac{3}{5}} + \frac{1}{\frac{2}{5}} + \frac{1}{\frac{1}{5}} \\
 &= 1 + \frac{5}{4} + \frac{5}{3} + \frac{5}{2} + \frac{5}{1} \\
 &= \frac{12 + 15 + 25 + 30 + 60}{12} \\
 &= \frac{142}{12} = 11.833
 \end{aligned}$$

and so you would need 12 kids meals. If this were to happen, please be certain to donate the "extra" books to an organization that works with kids or directly to some kids that you might know.

Checkpoint 7.6.4 - Rolling Dice. Suppose you roll a standard pair of 6-sided dice 20 times and let X measure the number of outcomes which result in a sum of 7 or 11. Determine:

1. the expected number of rolls which have a sum of 7 or 11
2. $P(X=5)$
3. $P(X > 5)$
4. $P(X < 5)$

Checkpoint 7.6.5 - Rolling Dice yet again. Suppose you roll a standard pair of 6-sided dice X times until you get a sum of 7 or 11 a third time. Determine:

1. the expected number of rolls needed on average.
2. $P(X=5)$
3. $P(X > 5)$
4. $P(X < 5)$

Checkpoint 7.6.6 - 2 standard deviations from the mean. Given $p = 0.3$ determine the following:

1. For Binomial with $n = 50$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$
2. For Negative Binomial with $r = 2$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$

Solution. For Binomial with $p = 0.3$ and $n = 50$, $\mu = n \cdot p = 15$ and $\sigma^2 = n \cdot p \cdot (1 - p) = 10.5$. So, $\sigma = \sqrt{10.5}$. Therefore,

$$\begin{aligned}
 P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= P(15 - 2\sqrt{10.5} \leq X \leq 15 + 2\sqrt{10.5}) \\
 &= P(X \in \{9, 10, 11, \dots, 19, 20, 21\})
 \end{aligned}$$

Then

$$F(21) - F(8) \approx 0.97491 - 0.01825 = 0.95666$$

For Negative Binomial with $p = 0.3$ and $r = 2$, $\mu = \frac{2}{0.3} = \frac{20}{3}$ and $\sigma^2 = 2 \frac{0.7}{0.3^2} = \frac{140}{9}$ and so $\sigma \approx 3.9$. Therefore,

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = P(6.7 - 7.8 \leq X \leq 6.7 + 7.8) \\ P(X \in \{2, 3, \dots, 14, 15, 16\})$$

Then,

$$F(16) \approx 0.973888$$

Checkpoint 7.6.7 Chapter Experiment. Take a die and roll it 4 times, keeping track on paper each time you get a 6 (say). Repeat this 100 times. (Actually, you can use a standard 6-sided die or find a more exotic one with more sides.) You should have gotten anywhere from 0 of the rolls to be a 6 up to all of the rolls to be a 6. Collect the relative frequencies of each of the possible outcomes in $R = \{0, 1, 2, \dots, 100\}$ and plot. Compare several of the experimental relative frequencies with the theoretical value you would expect using the proper distribution from this chapter. Comment on how well you did.

Chapter 8

Distributions based upon Poisson Processes

8.1 Introduction

In this chapter, you will investigate the relationship between number of successes over some interval. For each, one of these quantities will be fixed and the other one variable. First, consider the following:

Definition 8.1.1 Poisson Process. A Poisson process is a course of action in which:

1. Successes in non-overlapping subintervals are independent of each other.
2. The probability of exactly one success in a sufficiently small interval of length h is proportional to h . In notation, $P(\text{one success}) = \lambda h$.
3. The probability of two or more successes in a sufficiently small interval is essentially 0.

◇

You should presume these assumptions implicitly for the distributions discussed in this chapter.

In this chapter, you will investigate the following distributions:

1. Poisson - the interval is fixed and X measures the variable number of successes.
2. Exponential - the number of successes is fixed—at 1—and X measures the variable interval length needed to get that success.
3. Gamma - the number of successes is fixed and X measures the variable interval needed to get the desired number of successes.

8.2 Poisson Distribution

Consider a Poisson Process where you start with an interval of fixed length T and where X measures the variable number of successes, or changes, within a that interval. The resulting distribution of X will be called a Poisson distribution.

Theorem 8.2.1 Poisson Probability Function. Assume X measures the number of successes in an interval $[0, T]$ within some Poisson process. Then,

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

for $R = \{0, 1, 2, \dots\}$.

Proof. For a sufficiently large natural number n , break up the given interval $[0, T]$ into n uniform parts each of width $h = T/n$. Using the properties of Poisson processes, n very large implies h will be very small and eventually small enough so that

$$P(\text{exactly one success on a given interval}) = p = \lambda \frac{T}{n}.$$

However, since there are a finite number of independent intervals each with probability p of containing a success then you can use a Binomial distribution to evaluate the corresponding probabilities so long as n is finite. Doing so yields and taking the limit as n approaches infinity gives:

$$\begin{aligned} f(x) &= P(X \text{ changes in } [0, T]) \\ &= \lim_{n \rightarrow \infty} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \lim_{n \rightarrow \infty} \binom{n}{x} \left(\frac{\lambda T}{n}\right)^x \left(1 - \frac{\lambda T}{n}\right)^{n-x} \\ &= \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda T}{n}\right)^x \left(1 - \frac{\lambda T}{n}\right)^{n-x} \\ &= \frac{(\lambda T)^x}{x!} \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-x+1)}{n \cdot n \cdot \dots \cdot n} \left(1 - \frac{\lambda T}{n}\right)^n \left(1 - \frac{\lambda T}{n}\right)^{-x} \\ &= \frac{(\lambda T)^x}{x!} \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)\dots\left(1 - \frac{x-1}{n}\right) \left(1 - \frac{\lambda T}{n}\right)^n \left(1 - \frac{\lambda T}{n}\right)^{-x} \\ &= \frac{(\lambda T)^x}{x!} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda T}{n}\right)^n \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda T}{n}\right)^{-x} \\ &= \frac{(\lambda T)^x}{x!} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda T}{n}\right)^n \cdot 1 \\ &= \frac{(\lambda T)^x}{x!} e^{-\lambda T} \end{aligned} \quad \blacksquare$$

Theorem 8.2.2 Verify Poisson Probability Function.

$$\sum_{x=0}^{\infty} \frac{(\lambda T)^x}{x!} e^{-\lambda T} = 1$$

Proof. Using the Power Series expansion for the natural exponential,

$$\begin{aligned} \sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} \frac{(\lambda T)^x}{x!} e^{-\lambda T} \\ &= e^{-\lambda T} \sum_{x=0}^{\infty} \frac{(\lambda T)^x}{x!} \\ &= e^{-\lambda T} e^{\lambda T} \\ &= 1 \end{aligned} \quad \blacksquare$$

Theorem 8.2.3 Statistics for Poisson.

$$\mu = \lambda T$$

$$\sigma^2 = \mu$$

$$\gamma_1 = \frac{1}{\sqrt{\mu}}$$

$$\gamma_2 = \frac{1}{\mu} + 3$$

Proof. Using the $f(x)$ generated in the previous theorem

$$\begin{aligned}\mu &= E[X] \\ &= \sum_{x=0}^{\infty} x \cdot \frac{(\lambda T)^x}{x!} e^{-\lambda T} \\ &= \lambda T e^{-\lambda T} \sum_{x=1}^{\infty} \frac{(\lambda T)^{x-1}}{(x-1)!} \\ &= \lambda T e^{-\lambda T} \sum_{k=0}^{\infty} \frac{(\lambda T)^k}{k!} \\ &= \lambda T e^{-\lambda T} e^{\lambda T} \\ &= \lambda T\end{aligned}$$

which confirms the use of μ in the original probability formula.

Continuing with $\mu = \lambda T$, the variance is given by

$$\begin{aligned}\sigma^2 &= E[X(X-1)] + \mu - \mu^2 \\ &= \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\mu^x}{x!} e^{-\mu} + \mu - \mu^2 \\ &= e^{-\mu} \mu^2 \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} + \mu - \mu^2 \\ &= e^{-\mu} \mu^2 \sum_{k=0}^{\infty} \frac{\mu^k}{k!} + \mu - \mu^2 \\ &= \mu^2 + \mu - \mu^2 \\ &= \mu\end{aligned}$$

To derive the skewness and kurtosis, you can depend upon Sage...see the live cell below. ■

```
var('x,mu')
assume(x,'integer')

f(x) = e^(-mu)*mu^x/factorial(x)
mu = sum(x*f,x,0,oo).factor()
M2 = sum(x^2*f,x,0,oo).factor()
M3 = sum(x^3*f,x,0,oo).factor()
M4 = sum(x^4*f,x,0,oo).factor()

pretty_print('Mean_=_',mu)
```

```

v = (M2-mu^2).factor()
pretty_print('Variance_=',v)
stand = sqrt(v)

sk = ((M3 - 3*M2*mu + 2*mu^3)).factor()/stand^3
pretty_print('Skewness_=',sk)

kurt = (M4 - 4*M3*mu + 6*M2*mu^2 - 3*mu^4).factor()/stand^4
pretty_print('Kurtosis_=',(kurt-3).factor(),'+3')

```

Approximation by binomial means you can also use Poisson to approximate Binomial for n sufficiently large.

8.3 Exponential Distribution

Once again, consider a Poisson Process where you start with an interval of variable length X so that X measures the interval needed in order to obtain a first success with $R = (0, \infty)$. The resulting distribution of X will be called an Exponential distribution.

To derive the probability function for this distribution, consider finding $f(x)$ by first considering $F(x)$. This gives

$$\begin{aligned}
 F(x) &= P(X \leq x) \\
 &= 1 - P(X > x) \\
 &= 1 - P(\text{first change occurs after an interval of length } x) \\
 &= 1 - P(\text{no changes in the interval } [0, x]) \\
 &= 1 - \frac{(\lambda x)^0 e^{-\lambda x}}{0!} \\
 &= 1 - e^{-\lambda x}
 \end{aligned}$$

where the discrete Poisson Probability Function is used to answer the probability of exactly no changes in the "fixed" interval $[0, x]$. Using this distribution function and taking the derivative yields

$$f(x) = F'(x) = \lambda e^{-\lambda x}.$$

Definition 8.3.1 Exponential Distribution Probability Function.

Given a Poisson process and a constant μ , suppose X measures the variable interval length needed until you get a first success. Then X has an exponential distribution with probability function

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}.$$

◇

Theorem 8.3.2 Verification of Exponential Probability Function.

$$\int_0^{\infty} \frac{1}{\mu} e^{-\frac{x}{\mu}} dx = 1$$

Proof.

$$\int_0^{\infty} \frac{1}{\mu} e^{-\frac{x}{\mu}} dx$$

$$\begin{aligned}
 &= \int_0^{\infty} e^{-u} dx \\
 &= -e^{-u} \Big|_0^{\infty} = 1
 \end{aligned}$$

■

Theorem 8.3.3 Distribution function for Exponential Distribution.

$$F(x) = 1 - e^{-\frac{x}{\mu}}$$

Proof. Using $f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}$, note

$$\begin{aligned}
 F(x) &= \int_0^x \frac{1}{\mu} e^{-\frac{u}{\mu}} du \\
 &= -e^{-\frac{u}{\mu}} \Big|_0^x \\
 &= 1 - e^{-\frac{x}{\mu}}
 \end{aligned}$$

■

Theorem 8.3.4 Derivation of Statistics for Exponential Distribution and Plotting.

$$\sigma^2 = \mu^2$$

$$\gamma_1 = 2$$

$$\gamma_2 = 9$$

Proof. For the mean, notice that

$$\begin{aligned}
 \text{Mean} &= \int_0^{\infty} x \cdot \frac{1}{\mu} e^{-\frac{x}{\mu}} \\
 &= [(1-x)e^{-\frac{x}{\mu}}] \Big|_0^{\infty} = \mu
 \end{aligned}$$

and so the use of μ in $f(x)$ is warranted.

The remaining statistics are derived similarly using repeated integration by parts. The interactive Sage cell below calculates those for you automatically.

■

```

# Exponential Distribution
var('x,mu')
assume(mu>0)

f(x) = e^(-x/mu)/mu
mu = integral(x*f,x,0,oo).factor()
M2 = integral(x^2*f,x,0,oo).factor()
M3 = integral(x^3*f,x,0,oo).factor()
M4 = integral(x^4*f,x,0,oo).factor()

pretty_print('Mean_=_',mu)

v = (M2-mu^2).factor()
pretty_print('Variance_=_',v)
stand = sqrt(v)

sk = (((M3 - 3*M2*mu + 2*mu^3))/stand^3).simplify()
pretty_print('Skewness_=_',sk)

kurt = (M4 - 4*M3*mu + 6*M2*mu^2 - 3*mu^4).factor()/stand^4
pretty_print('Kurtosis_=_',(kurt-3).factor(),'+3')
```

```
@interact
def _(m = slider(1,12,1/2,2,label='mu')):
    plot(f(mu=m),x,0,30).show(ymax=1)
```

```
X = 0:20      # the space to use for calculations. Should be R
               # but since R is infinite just stop at a large
               # enough value
mu = 3        # the required parameter for Poisson
               Distributions
sdev = sqrt(mu) # the formula for the standard deviation
dpois( X, lambda=mu ) # let's print out a bunch of actual
probs

Ppoisson = dpois(X, mu) # create the probability function
over X

Psample = rpois(10^6, mu) # to create a histogram, sample a
lot
Xtop=max(Psample)         # for scaling the x-axis. Shift
by 1/2 below.
hist(Psample, prob=TRUE, br=(-1:Xtop)+0.5, col="skyblue2",
xlab="X",
main="Poisson_Probability_Function_vs_Approximating_Bell_
Curve'")

points(X, Ppoisson, pch=19, col="darkgreen") # to create
actual (x,f(x))

Pnormal <- function(X){dnorm(X, mean=mu, sd=sdev)} # to
overlap a bell curve
curve(Pnormal, col="red", lwd=2, add=TRUE)
```

Theorem 8.3.5 The Exponential Distribution yields a continuous memoryless model.. If X has an exponential distribution and a and b are nonnegative integers, then

$$P(X > a + b | X > b) = P(X > a)$$

Proof. Using the definition of conditional probability,

$$\begin{aligned} P(X > a + b | X > b) &= P(X > a + b \cap X > b) / P(X > b) \\ &= P(X > a + b) / P(X > b) \\ &= e^{-(a+b)/\mu} / e^{-b/\mu} \\ &= e^{-a/\mu} \\ &= P(X > a) \end{aligned}$$

■

8.4 Gamma Distribution

Extending the exponential distribution model developed above, consider a Poisson Process where you start with an interval of variable length X so that X measures the interval needed in order to obtain the r th success for some natural number r . Then $R = (0, \infty)$ and the resulting distribution of X will be called a Gamma distribution.

Definition 8.4.1 Gamma Function.

$$\Gamma(t) = \int_0^{\infty} u^{t-1} e^{-u} du$$

◇

Theorem 8.4.2 Gamma Function on the natural numbers. For $n \in \mathbb{N}$,

$$\Gamma(n+1) = n!$$

Proof. Letting n be a natural number and applying integration by parts one time gives

$$\begin{aligned} \Gamma(n+1) &= \int_0^{\infty} u^n e^{-u} du \\ &= -u^n \cdot e^{-u} \Big|_0^{\infty} + n \int_0^{\infty} u^{n-1} e^{-u} du \\ &= 0 - 0 + n\Gamma(n) \end{aligned}$$

Continuing using an inductive argument to obtain the final result. ■

To find the probability function for the gamma distribution, once again focus on the development of $F(x)$. Assuming r is a natural number greater than 1 and noting that X measures the interval length needed in order to achieve the r th success

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= 1 - P(X > x) \\ &= 1 - P(\text{fewer than } r \text{ successes in } [0, x]) \\ &= 1 - \left[\frac{(\lambda x)^0 e^{-\lambda x}}{0!} + \frac{(\lambda x)^1 e^{-\lambda x}}{1!} + \dots + \frac{(\lambda x)^{r-1} e^{-\lambda x}}{(r-1)!} \right] \\ &= 1 - \sum_{k=0}^{r-1} \frac{(\lambda x)^k e^{-\lambda x}}{k!} \end{aligned}$$

where the discrete Poisson probability function is used on the interval $[0, x]$. The derivative of this function however is "telescoping" and terms cancel. Indeed,

$$\begin{aligned} F'(x) &= \lambda e^{-\lambda x} / 0! \\ &\quad - \lambda e^{-\lambda x} / 1! + \lambda x \cdot \lambda e^{-\lambda x} / 1! \\ &\quad - \lambda^2 2x e^{-\lambda x} / 2! + \lambda^2 x^2 \cdot \lambda e^{-\lambda x} / 2! \\ &\quad - \lambda^3 3x^2 e^{-\lambda x} / 3! + \lambda^3 x^3 \cdot \lambda e^{-\lambda x} / 3! \\ &\quad \dots \\ &\quad - \lambda^{r-1} (r-1) x^{r-2} e^{-\lambda x} / (r-1)! + \lambda^{r-1} x^{r-1} \cdot \lambda e^{-\lambda x} / (r-1)! \\ &= \lambda^r x^{r-1} e^{-\lambda x} / (r-1)! \end{aligned}$$

where you can replace $(r-1)! = \Gamma(r)$.

Notice that for this random variable, $\mu = \lambda T$ can be obtained for the exponential distribution. For the Gamma distribution, the following takes μ to be the average interval till the first success and then modifies the corresponding Gamma parameters according to increasing values of r .

Definition 8.4.3 Gamma Distribution Probability Function. If X measures the interval until the r th success and

μ as the average interval until the 1st success, then X with probability function

$$f(x) = \frac{x^{r-1} \cdot e^{-x/\mu}}{\Gamma(r) \cdot \mu^r}$$

has a Gamma Distribution. ◇

Theorem 8.4.4 Verify Gamma Probability function.

$$\int_0^\infty \frac{x^{r-1} e^{-x/\mu}}{\Gamma(r) \mu^r} dx = 1$$

Proof. Evaluate the sage code below. ■

```
# Gamma Distribution
var('x,mu,r')
assume(mu>0)
assume(r,'integer')
assume(r>1)
f(x) = x^(r-1)*e^(-x/mu)/(gamma(r)*mu^r)
S = integral(f,x,0,oo).full_simplify()
F = '$\int_0^{\infty} \frac{x^{r-1} e^{-x/\mu}}{\Gamma(r) \mu^r} dx = %s$' % str(S)
html(F)
```

```
# Gamma Distribution Graphing
var('x,mu,r')
assume(mu>0)
assume(r,'integer')
@interact
def _(r=[2,3,6,12,24],mu=slider(1,12,1,5,label='mu')):
    f(x) = x^(r-1)*e^(-x/mu)/(gamma(r)*mu^r)
    plot(f,x,0,200).show()
```

Derivation of mean, variance, skewness, and kurtosis. Pick "alpha" for the general formulas.

```
# Gamma Distribution
var('x,mu,r,alpha')
assume(mu>0)
assume(alpha,'integer')
assume(alpha>1)
@interact
def _(r=[2,3,6,9,alpha]):
    f(x) = x^(r-1)*e^(-x/mu)/(gamma(r)*mu^r)
    mean = integral(x*f,x,0,oo).full_simplify()
    M2 = integral(x^2*f,x,0,oo).full_simplify()
    M3 = integral(x^3*f,x,0,oo).full_simplify()
    M4 = integral(x^4*f,x,0,oo).full_simplify()

    pretty_print('Mean =',mean)

    v = (M2-mean^2).factor()
    pretty_print('Variance =',v)
    stand = sqrt(v)
```

```

sk = (((M3 - 3*M2*mean +
      2*mean^3))/stand^3).full_simplify()
pretty_print('Skewness_=_',sk)

kurt = (M4 - 4*M3*mean + 6*M2*mean^2
        -3*mean^4).factor()/stand^4
pretty_print('Kurtosis_=_',(kurt-3).factor(),'+3')

```

Finally, the interactive cell below can be used to compute the distribution function for the gamma distribution for various input values. If you desire to let r get bigger than the slider allows, feel free to edit the cell above and evaluate again.

```

# Gamma Distribution Calculator
var('x,mu,r')
pretty_print('Enter_the_number_of_successes_desired,_the_
given_mean,_and_the_value_of_X_to_get_F(X)')
@interact
def _(r=slider(1,10,1,2),mu = input_box(2,label="$\mu=_
$",width=10),b=input_box(2,label="X=_",width=10)):
    f(x) =x^(r-1)*e^(-x/mu)/(gamma(r)*mu^r)
    p = integral(f,x,0,b)

    pretty_print('Probability_=_\t',p,'_which_is_
approximately_\t',p.n(digits=5))

```

8.5 Summary

Here is a summary of the major points in this chapter:TBA

8.6 Exercises

Checkpoint 8.6.1 - Home Sales. A local realty office sells on average 10 houses a week. Let X measure the number of houses the sell in the next week. Determine

1. the probability the realty office sells 12 houses next week.
2. the probability the realty office sells fewer than 10 houses next week.
3. the interval $\mu - 2\sigma \leq X \leq \mu + 2\sigma$.
4. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$.

Checkpoint 8.6.2 - Customer arrivals - Total. Customers arrive at your store on the average of 10 per hour. Assuming that the arrival of customers satisfies the properties of a Poisson process, determine:

1. the expected number of customers to arrive in a given 3 hour period.
2. the probability that fewer than 10 customers arrive in a given hour.

Solution. Using the given information, apply the Poisson distribution. For one hour $\mu = 10$ so that for three hours the expected number of customers would be triple with 30 expected customers.

With $\mu_1 = 10$,

$$P(X < 10) = F(9)$$

$$\begin{aligned}
&= \frac{10^0 e^{-10}}{0!} + \frac{10^1 e^{-10}}{1!} + \frac{10^2 e^{-10}}{2!} + \dots + \frac{10^8 e^{-10}}{8!} + \frac{10^9 e^{-10}}{9!} \\
&= e^{-10} \cdot \left(1 + 10 + \frac{100}{2} + \dots + \frac{10^8}{8!} + \frac{10^9}{9!}\right)
\end{aligned}$$

Checkpoint 8.6.3 - Customer arrivals - First. Customers arrive at your store on the average of 10 per hour. Assuming that the arrival of customers satisfies the properties of a Poisson process, determine:

1. the number of minutes expected between the arrival of each customer
2. the probability it takes more than 9 minutes before the next customer arrives.

Solution. Using the given information, apply the exponential distribution. Since 10 arrive on average in one hour then you would expect 1 to arrive in 6 minutes.

With $\mu = 6$ minutes,

$$P(X > 9) = 1 - F(9) = e^{-9/6}.$$

Checkpoint 8.6.4 - Customer arrivals - 10th. Customers arrive at your store on the average of 10 per hour. Assuming that the arrival of customers satisfies the properties of a Poisson process, determine:

1. the number of minutes expected for the arrival of three customers.
2. the probability it takes less than 20 minutes before the third customer arrives.

Solution. Using the given information, apply the gamma distribution. Since 10 arrive on average in one hour then you would expect 1 to arrive in 6 minutes and therefore 3 to arrive on average in 18 minutes.

With $\mu = 18$ minutes,

$$P(X < 20) = F(20).$$

Checkpoint 8.6.5 - Computer Network Data Traffic. Consider the arrival of requests on a server. Presume that the requests are considered as coming from an anonymous and large collection of users independently of each other on an average of 50 requests per second. If X measures the number of requests per second, determine

1. the probability that in any given second the server gets fewer than 50 requests
2. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$
3. the expected number of requests per hour.

Solution. Given the average of 50 requests per second and X measuring the number of "successes" in a given second long time interval given a Poisson distribution

$$f(x) = \frac{50^x}{x!} e^{-50}.$$

Then,

$$P(X < 50) = F(49) = \sum_{x=0}^{49} \frac{50^x}{x!} e^{-50}$$

and using the graphing calculator function $\text{poissoncdf}(50, 49) = 0.48119$.

For a time interval of one second, the mean is given to be 50 requests. Using the

formulas developed above, the standard deviation therefore is $\sqrt{50}$. Therefore

$$\begin{aligned} P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= P(50 - 2\sqrt{50} \leq X \leq 50 + 2\sqrt{50}) \\ &= P(X \in \{36, 37, 38, \dots, 62, 63, 64\}). \end{aligned}$$

Using the distribution function,

$$F(64) - F(35) \approx 0.97640 - 0.01621 = 0.96019$$

Finally, notice that the time interval has been adjusted. Since the mean formula is proportional to the interval over which X is measured, using $\mu = \lambda T$ with $\lambda = 1$ when the interval is 1 second, then when the interval is one hour, $T = 3600$ seconds. Hence, we would expect on average $50 \cdot 3600 = 180,000$ requests in one hour.

Chapter 9

Normal Distributions

9.1 Introduction

You should have noticed by now that many distributions tend to have a bell-shaped graph as parameters are allowed to increase. Indeed, the formulas for skewness γ_1 and kurtosis γ_2 approach 0 and 3 respectively for the Hypergeometric, Binomial, Negative Binomial, Poisson, and Gamma Distributions. One might wonder if this is just a happy coincidence or is something more insidious at play.

The answer by appealing to mathematics reveals that nothing sinister is going on but that it is indeed true that the eventual destiny for distributions is one that is bell-shaped. It is therefore of interest to figure out if that distribution has a nice form that can be accessed directly. The focus of this chapter is to consider this bell-shaped goal known as the "normal distribution."

We present the normal distribution by simply presenting its probability function without derivation. In order to more carefully investigate the development of the normal distribution (and the Chi-Square Distribution) you will need to study "Moment Generating Functions" and some serious mathematics. Without supplying this rigor you can still utilize the results.

9.2 The Normal Distribution

Definition 9.2.1 The Normal Distribution. Given two parameters μ and σ , a random variable X over $R = (-\infty, \infty)$ has a normal distribution provided it has a probability function given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$$

◇

The normal distribution is also sometimes referred to as the Gaussian Distribution (often by Physicists) or the Bell Curve (often by social scientists).

```
var('x,mu,sigma')
f(x) = e^(-(x-mu)/sigma)^2/2)/(sigma*sqrt(2*pi))
@interact
def
    _(m=slider(-10,10,1,0,label='$\mu$'),s=slider(1/5,5,1/10,1,label='$\sigma$')):
    titletext = "Normal_Curve_with_mean_"+str(m)+"_and_
    standard_deviation_"+str(s)
```

```

G = plot(f(mu=m, sigma=s), (x, m-5*s, m+5*s))
G +=
    point((0, 1), size=1) + point((12, 0), size=1) + point((-12, 0), size=1)
G += point((m, f(x=m, mu=m, sigma=s)), color='red', size=20)
G +=
    point((m+s, f(x=m+s, mu=m, sigma=s)), color='green', size=20)
G +=
    point((m-s, f(x=m-s, mu=m, sigma=s)), color='green', size=20)
show(G, figsize=(5, 3), title=titletext, ymin=0, ymax=1, xmin=-15, xmax=15)

```

Theorem 9.2.2 If $\mu = 0$ and $\sigma = 1$, then we say X has a standard normal distribution and often use Z as the variable name and will use $\Phi(z)$ for the standard normal distribution function. In this case, the density function reduces to

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Proof. Convert to "standard units" using the conversion

$$z = \frac{x - \mu}{\sigma} = \frac{x - 0}{1} = x. \quad \blacksquare$$

Theorem 9.2.3 Verifying the normal probability function.

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{\sigma})^2/2} dx = 1$$

Proof. Note that you can convert the integral above to (((Unresolved xref, reference "StandardUnitConversion"; check spelling or use "provisional" attribute))) standard units so that it is sufficient to show

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$$

Toward this end, consider I^2 and change the variables to get

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{u^2+v^2}{2}} dudv \end{aligned}$$

Converting to polar coordinates using

$$dudv = r dr d\theta$$

and

$$u^2 + v^2 = r^2$$

gives

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} r dr d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} -e^{-\frac{r^2}{2}} \Big|_0^{\infty} d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} 1 \cdot d\theta \\ &= \frac{1}{2\pi} \theta \Big|_0^{2\pi} = 1 \end{aligned}$$

as desired. \blacksquare

Theorem 9.2.4 Verifying the normal probability mean.

$$E[X] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} dx = \mu$$

Proof.

$$z = \frac{x - \mu}{\sigma}$$

implies by solving that

$$x = \mu + z\sigma$$

and therefore

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + z\sigma) \cdot e^{-z^2/2} dz \\ &= \mu \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz + \sigma \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot e^{-z^2/2} dz \\ &= \mu \cdot 1 + \sigma \cdot 0 \\ &= \mu \end{aligned}$$

and therefore the use of μ is warranted. ■

Theorem 9.2.5 Verifying the normal probability variance.

$$E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} dx = \sigma^2$$

Proof.

$$\begin{aligned} E[(X - \mu)^2] &= \int_{-\infty}^{\infty} (x - \mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 z^2 \cdot e^{-z^2/2} dz \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot z e^{-z^2/2} dz \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \cdot \left[-ze^{-z^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2} dz \right] \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \cdot [0 + \sqrt{2\pi}] \\ &= \sigma^2 \end{aligned}$$

using integration by parts and using the integration in the proof of the mean above. So, the use of σ is warranted. ■

Theorem 9.2.6 Properties of the Normal Distribution.

Theorem 9.2.7 Normal Distribution Maximum. *The maximum of the normal distribution probability function occurs when $x = \mu$*

Proof. Take the derivative of the probability function to get

$$\frac{\sqrt{2}(\mu - x)e^{\left(-\frac{(\mu - x)^2}{2\sigma^2}\right)}}{2\sqrt{\pi}\sigma^3}$$

which is zero only when $x = \mu$. Easily by evaluating to the left and right of this value shows that this critical value yields a maximum. ■

Theorem 9.2.8 Normal Distribution Points of Inflection. *Points of Inflection for the normal distribution probability function occurs when $x = \mu + \sigma$ and $x = \mu - \sigma$.*

Proof. Take the second derivative of the probability function to get

$$\frac{\sqrt{2}(\mu + \sigma - x)(\mu - \sigma - x)e^{\left(-\frac{\mu^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right)}}{2\sqrt{\pi}\sigma^5}$$

which is zero only when $x = \mu \pm \sigma$. Easily by evaluating to the left and right of this value shows that these critical values yield points of inflection. ■

Notice that the work needed to complete the integrals over the entire domain above was pretty serious. To determine probabilities for a given interval is however not possible in general and therefore approximations are needed. When using TI graphing calculators, you can use

$$P(a < x < b) = \text{normalcdf}(a, b, \mu, \sigma).$$

Or you can use the calculator below.

```
@interact(layout=dict(top=[['a',
    'b']],bottom=[['mu', 'sigma']]))
def _(a=input_box(-2,width=10,label='a_=_'),
    b=input_box(2,width=10,label='b_=_'),
    mu=input_box(0,width=8,label='$\mu_=_'),
    sigma=input_box(1,width=8,label='$\sigma_=_')):
    f = e^(-(x-mu)/sigma)^2/2)/(sigma*sqrt(2*pi))
    P = integral_numerical(f,a,b)[0]
    print "P("+str(a)+"_<_X_<_"+str(b)+"_)\sim_="+str(P)
```

9.3 Chi-Square Distribution

The following distribution is related to both the Normal Distribution and to the Gamma Distribution. Initially, consider a gamma distribution with probability function

$$\frac{x^{r-1} \cdot e^{-x/\mu}}{\Gamma(r) \cdot \mu^r}.$$

Replacing $\mu = 2$ and r with $r/2$ gives

$$\frac{x^{r/2-1} \cdot e^{-x/2}}{\Gamma(r/2) \cdot 2^{r/2}}$$

which is given a special name below.

Definition 9.3.1 Chi-Square Probability Function. $R = (0, \infty)$

$$f(x) = \frac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}.$$

$\chi^2(r)$

◇

```
# Chi-Square Grapher
@interact
def _(r=slider(1,20,1,3,label='r_')):
    f = x^(r/2-1)*e^(-x/2)/(gamma(r/2)*2^(r/2))
    plot(f,x,0,20).show()
```

Theorem 9.3.2 χ^2 statistics.

$$\begin{aligned}\mu &= r \\ \sigma^2 &= 2r \\ \gamma_1 &= 2\sqrt{2/r} \\ \gamma_2 &= \frac{12}{r} + 3\end{aligned}$$

Theorem 9.3.3 Relationship between Normal and χ^2 . Z_1, Z_2, \dots, Z_r

$$X = \sum_{k=1}^r Z_k^2$$

$\chi^2(r)$

It also can be difficult to compute Chi-Square probabilities manually so you will perhaps want to use a numerical approximation in this case as well. The TI graphing calculator can be used with $\chi^2\text{cdf}(a,b,r)$. Or, you can use the calculator below.

```
# Chi-Square Calculator
@interact(layout=dict(top=[['a', 'b']],bottom=[['r']]))
def _(a=input_box(0,width=10,label='a_'),
      b=input_box(2,width=10,label='b_'),
      r=input_box(2,width=8,label='r_')):
    f = x^(r/2-1)*e^(-x/2)/(gamma(r/2)*2^(r/2))
    P = numerical_integral(f,a,b)[0]
    print "P("+str(a)+"<_X_<"+str(b)+")_=_"+str(P)
```

9.4 Other "Bell Shaped" distributions

The Normal distribution discussed above is very important when doing statistical analysis. It however is not the only distribution that is symmetrical about the mean and looks like a bell. In this section, we consider two other options—one which is virtually useless and another which is very useful.

Definition 9.4.1 The Cauchy Distribution. Consider a continuous random variable on the real numbers defined by

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

A random variable with this probability function is said to be a Cauchy Distribution. \diamond

Theorem 9.4.2 The Cauchy Distribution.

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

$(-\infty, \infty)$

Proof.

$$\int_{-\infty}^{\infty} \frac{1}{1+x^2} dx = \arctan(\infty) - \arctan(-\infty) = \pi/2 - (-\pi/2) = \pi.$$

π ■

```
n <- 10
p <- 0.3

paste('Probability_Function')
dcauchy(x, location = 0, scale = 1, log = FALSE) # gives
the probability function
paste('Distribution_function')
pcauchy(q, location = 0, scale = 1, lower.tail = TRUE, log.p
= FALSE)
# gives the distribution function
paste('A_random_sample')
rcauchy(n, location = 0, scale = 1) # gives a random
sample of 15 items from b(n,p)

x=seq(-4,4,length=200)
y=dcauchy(x,dcauchy(x, location = 0, scale = 1, log = FALSE))
plot(x,y,type="l",lwd=2,col="red",ylab="p")
```

Now that we have a probability function, it is important to determine its mean and variance. It should be obvious that when doing so using the Cauchy probability function, problems quickly arise. Indeed,

$$\int_{-\infty}^{\infty} x \frac{1}{1+x^2} dx = (1/2)(\ln(|\infty|) - \ln(|-\infty|))$$

which is problematic. Further, for the variance

$$\int_{-\infty}^{\infty} x^2 \frac{1}{1+x^2} dx$$

and note that the integrand does not converge to 0 at the endpoints and therefore the integral is automatically considered divergent. Thus it is reasonable to note that the Cauchy distribution has no variance.

On the other hand, there is another bell-shaped distribution that is useful and its random variable can be created by using a mixture of a normal variable and a Cauchy variable. Indeed, suppose Z is a standard normal variable and Y is $^2(r)$ with Y and Z independent. Define a new random variable

$$T = \frac{Z}{\sqrt{(Y/r)}}.$$

Then, T is said to have a (Student) t distribution. The good news is that this distribution is useful and its statistics are presented below without proof.

Theorem 9.4.3 Student t-distribution.

$$\mu = 0$$

$$\sigma^2 = \frac{r}{r-2}.$$

```
# Display the Student's t distributions with various
# degrees of freedom and compare to the normal distribution
# Copied from www.statmethods.net

x <- seq(-4, 4, length=100)
hx <- dnorm(x)

degf <- c(1, 3, 8, 30)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=30", "normal")

plot(x, hx, type="l", lty=2, xlab="x_value",
      ylab="Density", main="Comparison_of_t_Distributions")

for (i in 1:4){
  lines(x, dt(x,degf[i]), lwd=2, col=colors[i])
}

legend("topright", inset=.05, title="Distributions",
      labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```

9.5 Normal Distribution as a Limiting Distribution

Over the past several chapters you should have noticed that many distributions have skewness and kurtosis formulae which have limiting values of 0 and 3 respectively. This means that each of those distributions which can be approximated by the normal distribution for "large" parameter values.

To see how this works, consider a "random" distribution in the following two interactive experiments. For the first graph below, a sequence of N random samples, each of size r, ranging from 0 to "Range" is generated and graphed as small data points. As the number of samples N and the sample size r increase, notice that the data seems to cover the entire range of possible values relatively uniformly. (For this scatter plot note that each row represents the data for one sample of size r. The larger the N, the greater the number of rows.) Each row is averaged and that mean value is plotted on the graph as a red circle. If you check the "Show_{Mean}" box, the mean of these circles is indicated by the green line in the middle of the plot.

For the second graph below, the means are collected and the relative frequency of each is plotted. As N increases, you should see that the results begin to show an interesting tendency. As you increase the data range, you may notice this graph has a larger number of data values. Smoothing groups this data into intervals of length two for perhaps a graph with less variability.

Consider each of the following:

- As N increases with single digit values of r, what appears to happen to the mean and range of the means? How does increasing the data range from 1-100 to 1-200 or 1-300 affect these results?

- As N increases (say, for a middle value of r), what appears to happen to the means? How does increasing the data range from 1-100 to 1-200 or 1-300 affect these results?
- As r increases (say, for a middle value of N), what appears to happen to the range of the averages? Does your conclusion actually depend upon the value of N? (Look at the graph and don't worry about the actual numerical values.) How does increasing N for the second graph affect the skewness and kurtosis of that graph? Do things change significantly as r is increased?

```

var('n,k')
from sage.finance.time_series import TimeSeries

@interact(layout=dict(top=[ 'Range' ],[ 'Show_Mean' ,
    'Smoothing' ]],
bottom=[ 'N' ],[ 'r' ]))

def _(Range=[100,200,300,500],N=slider(5,200,2,2,label="N=_
Number_of_Samples"),r=slider(3,200,1,2,label="r=_Sample_
Size"),Show_Mean=False,Smoothing=False):
    R=[1..N]      # R ranges over the number of
                  samples...will point to the list of averages
    rangemax = Range

    data = random_matrix(ZZ,N,r,x=rangemax)
    datapoints = []
    avg_values = []
    avg_string = []
    averages = []
    for n in range(N):
        temp = 0
        for k in range(r):
            datapoints += [(data[n][k],n)]
            temp += data[n][k]
        avg_values.append(round(temp/r))
        if Smoothing:
            avg_string.append(str(2*round((temp/r)/2)))
        else:
            avg_string.append(str(round(temp/r)))

        averages += [(round(temp/r),n)] # make these
        averages integers for use in grouping later
    SCAT =
        scatter_plot(datapoints,markersize=2,edgecolor='red',figsize=(10,4),axes_labels=
        Values','Sample_Number'])
    AVGS =
        scatter_plot(averages,markersize=50,edgecolor='blue',marker='o',figsize=(7,
        4))

    freqslist =
        frequency_distribution(avg_string,1).function().items()

# compute sample statistics for the raw data as well as for
the N averages
Mean_data = (sum(sum(data)))/(N*r)).n()
#   STD_data = sqrt(sum(sum( (data-Mean_data)^2
)))/(N*r)).n()

```



```

Mean_averages = mean(avg_values).n()
#   STD_averages = sqrt(variance(avg_values).n())
#   print "Data mean =",Mean_data," vs Mean of the averages
#   =",Mean_averages
#   print "Data STD = ",STD_data," vs Standard Dev of avgs
#   =", STD_averages
    if Show_Mean:
        avg_line =
            line([(Mean_data,0),(Mean_data,N-1)],rgbcolor='green',thickness=10)
        avg_text =
            text('xbar',(Mean_data,N),horizontal_alignment='right',rgbcolor='green')
    else:
        avg_line = Graphics()
        avg_text = Graphics()

# Plot a scatter plot exhibiting uniformly random data and
# the collection of averages
    print(html("The_random_data_plot_on_the_left_with_each_
row_representing_a_sample_with_size_determined_by\n"+
"the_slider_above_and_each_circle_representing_the_
average_for_that_particular_sample.\n"+
"First,_keep_sample_size_relatively_low_and_
increase_the_number_of_samples._Then,_\n"+
"watch_what_happens_when_you_slowly_increase_the_
sample_size."))

# Plot the relative frequencies of the grouped sample
# averages
    print(html("Now,_the_averages_(ie._the_circles)_from_
above_are_collected_and_counted\n"+
"with_the_relative_frequency_of_each_average_
graphed_below._For_a_relatively_large_number_
of\n"+
"samples,_notice_what_seems_to_happen_to_these_
averages_as_the_sample_size_increases."))
    if Smoothing:
        binRange = Range//2
    else:
        binRange = Range

# normed=True # if you want to have relative
# frequencies below

    his_low = 2*rangemax/7
    his_high = 5*rangemax/7

    T =
        histogram(avg_values,normed=False,bins=binRange,range=(his_low,his_high),axes_labels=
        Averages','Frequency'])
    #T =
        TimeSeries(avg_values).plot_histogram(axes_labels=['Sample
        Averages','Frequency'])

    pretty_print('Scatter_Plot_of_random_data._Horizontal_
is_number_of_samples.')
    (SCAT+AVGS+avg_line+avg_text).show()
    pretty_print('Histogram_of_Sample_Averages')
    T.show(figsize=(5,2))

```

```

var('n,k')
from sage.finance.time_series import TimeSeries

@interact(layout=dict(top=[['Range'], ['Show_Mean',
'Smoothing']],
bottom=[['N'], ['r']]))

def _(Range=[100,200,300,500],N=slider(5,200,2,2,label="N=_
Number_of_Samples"),r=slider(3,200,1,2,label="r=_Sample_
Size"),Show_Mean=False,Smoothing=False):
    R=[1..N]      # R ranges over the number of
                  # samples...will point to the list of averages
    rangemax = Range

    data = random_matrix(ZZ,N,r,x=rangemax)
    datapoints = []
    avg_values = []
    avg_string = []
    averages = []
    for n in range(N):
        temp = 0
        for k in range(r):
            datapoints += [(data[n][k],n)]
            temp += data[n][k]
        avg_values.append(round(temp/r))
        if Smoothing:
            avg_string.append(str(2*round((temp/r)/2)))
        else:
            avg_string.append(str(round(temp/r)))

        averages += [(round(temp/r),n)] # make these
        # averages integers for use in grouping later
    SCAT =
        scatter_plot(datapoints,markersize=2,edgecolor='red',figsize=(10,4),axes_labels=
        'Values','Sample_Number'])
    AVGS =
        scatter_plot(averages,markersize=50,edgecolor='blue',marker='o',figsize=(7,
        4)))

    freqslist =
        frequency_distribution(avg_string,1).function().items()

    # compute sample statistics for the raw data as well as for
    # the N averages
    Mean_data = (sum(sum(data))/(N*r)).n()
    #   STD_data = sqrt(sum(sum( (data-Mean_data)^2
    #   ))/(N*r)).n()
    Mean_averages = mean(avg_values).n()
    #   STD_averages = sqrt(variance(avg_values).n())
    #   print "Data mean =",Mean_data," vs Mean of the averages
    #   =",Mean_averages
    #   print "Data STD = ",STD_data," vs Standard Dev of avgs
    #   =", STD_averages
    if Show_Mean:
        avg_line =
            line([(Mean_data,0),(Mean_data,N-1)],rgbcolor='green',thickness=10)
        avg_text =

```

```

        text('xbar',(Mean_data,N),horizontal_alignment='right',rgbcolor='green')
    else:
        avg_line = Graphics()
        avg_text = Graphics()

# Plot a scatter plot exhibiting uniformly random data and
the collection of averages
    print(html("The_random_data_plot_on_the_left_with_each_
row_representing_a_sample_with_size_determined_by\n"+
    "the_slider_above_and_each_circle_representing_the_
average_for_that_particular_sample.\n"+
    "First,_keep_sample_size_relatively_low_and_
increase_the_number_of_samples._Then,_\n"+
    "watch_what_happens_when_you_slowly_increase_the_
sample_size."))

# Plot the relative frequencies of the grouped sample
averages
    print(html("Now,_the_averages_(ie._the_circles)_from_
above_are_collected_and_counted\n"+
    "with_the_relative_frequency_of_each_average_
graphed_below._For_a_relatively_large_number_
of\n"+
    "samples,_notice_what_seems_to_happen_to_these_
averages_as_the_sample_size_increases."))
    if Smoothing:
        binRange = Range//2
    else:
        binRange = Range

    # normed=True # if you want to have relative
frequencies below

    his_low = 2*rangemax/7
    his_high = 5*rangemax/7

    T =
        histogram(avg_values,normed=False,bins=binRange,range=(his_low,his_high),axes_labels=
Averages','Frequency'])
    #T =
        TimeSeries(avg_values).plot_histogram(axes_labels=['Sample
Averages','Frequency'])

    pretty_print('Scatter_Plot_of_random_data._Horizontal_
is_number_of_samples.')
    (SCAT+AVGS+avg_line+avg_text).show()
    pretty_print('Histogram_of_Sample_Averages')
    T.show(figsize=(5,2))

```

So, even with random data, if you are to consider the arrangement of the collected means rather than the arrangement of the actual data then the means appear to have a bell-shaped distribution as well.

9.6 Central Limit Theorem

Often, when one wants to solve various scientific problems, several assumptions will be made regarding the nature of the underlying setting and base their

conclusions on those assumptions. Indeed, if one is going to use a Binomial Distribution or a Negative Binomial Distribution, an assumption on the value of p is necessary. For Poisson and Exponential Distributions, one must know the mean. For Normal Distributions, one must assume values for both the mean and the standard deviation. Where do these values come from? Often, one may perform a preliminary study and obtain a sample statistic...such as a sample mean or a relative frequency and use these values for μ or p .

But what is the underlying distribution of these sample statistics? The Central Limit Theorem gives the answer...

The results from the previous section illustrate the tendency for bell-shaped distributions. This tendency can be described more mathematically through the following theorem. It is presented here without proof.

Theorem 9.6.1 Central Limit Theorem. *Presume X is a random variable from a distribution with known mean μ and known variance σ_x^2 . For some natural number n , sample the distribution repeatedly creating a string of random variables denoted X_1, X_2, \dots, X_n and set $\bar{X} = \frac{\sum X_k}{n}$.*

Then, \bar{X} is approximately normally distributed with mean μ and variance $\sigma^2 = \frac{\sigma_x^2}{n}$.

Often the Central Limit Theorem is stated more formally using a conversion to standard units. Indeed, the theorem indicates that the random variable \bar{X} has variance $\frac{\sigma^2}{n}$ which means as n grows this variance approaches 0. So, the limiting random variable has a zero variance and therefore is no longer a random variable. To avoid this issue, the Central Limit Theorem is often stated as:

For random variables

$$W_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

with corresponding distribution function $F_n(W_n)$,

$$\lim_{n \rightarrow \infty} F_n(c) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(c)$$

that is, the standard normal distribution function.

Example 9.6.2 Exponential X vs Normal \bar{X} . Consider an exponential variable X with mean time till first success of $\mu = 4$. Then, $\sigma = 2$ using the exponential formulas.

You can use the exponential probability function to compute probabilities dealing with X . Indeed,

$$P(X < 3.9) = F(3.9) = 1 - e^{-3.9/4} \approx 0.6228.$$

If instead you plan to sample from this distribution $n=32$ times, the Central Limit Theorem implies that you will get a random variable \bar{X} which has an approximate normal distribution with the same mean but with new variance $\sigma_{\bar{X}}^2 = \frac{4}{32} = \frac{1}{8}$. Therefore

$$P(\bar{X} < 3.9) \approx \text{normalcdf}(0, 3.9, 4, \text{sqrt}(1/8)) = 0.2119.$$

□

When converting probability problems from continuous (such as exponential or uniform) then no adjustment to the question is needed since you are approximating one area with another area. However, when converting probability problems from discrete (such as binomial or geometric) then you need to

consider how the interval would need to be adjusted so that histogram areas for the discrete problem would relate to areas under the normal curve. Generally, you will need to expand the stated interval each way by $1/2$.

The Central Limit Theorem provides that regardless of the distribution of X , the distribution of an average of X 's is approximately normally distributed. However, it also shows why X may also be approximated for some distributions using the normal distribution as certain parameters are allowed to increase. Below, you can see how Binomial and Poisson distributions can be approximated directly using the Normal distribution.

Toward that end, for $0 < p < 1$ consider a sequence of Bernoulli trials Y_1, Y_2, \dots, Y_n with each over the space $0,1$. Then,

$$X = \sum_{k=1}^n Y_k$$

is a Binomial variable.

Theorem 9.6.3 Binomial as approximate Normal. $\mu = np\sigma^2 = np(1-p)np > 5n(1-p) > 5$

Proof. Using the Bernoulli variables Y_k each with mean p and variance $p(1-p)$, note that the Central Limit Theorem applied to $\bar{X} = \frac{\sum Y_k}{n}$ gives that

$$\frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

is approximately standard normal. By multiplying top and bottom by n yields

$$\frac{\sum Y_k - np}{\sqrt{np(1-p)}}$$

is approximately standard normal. But $\sum Y_k$ actually is the sum of the number of successes in n trials and is therefore a Binomial variable. ■

Example 9.6.4 Binomial as Normal. Binomial becomes normal as $n \rightarrow \infty$. Consider $n = 50$ and $p = 0.3$. Then, $\mu = 15$ and $\sigma^2 = 10.5$. Using the binomial formulas, for example,

$$P(X = 16) = \binom{50}{16} 0.3^{16} \cdot 0.7^{34} \approx 0.11470$$

Using the normal distribution,

$$\begin{aligned} P(X = 16) &= P(15.5 < X < 16.5) \\ &\approx \text{normalcdf}(15.5, 16.5, 15, \text{sqrt}(10.5)) \\ &= 0.11697 \end{aligned}$$

Notice that these are very close. □

Corollary 9.6.5 Poisson as approximate Normal. $\mu\sigma^2 = \mu\mu > 5$

Proof. Note from before that the Poisson distribution function was derived by approximating with Binomial and letting n approach infinity. Therefore, by the previous theorem, the Poisson variable is also approximately Normal using the Poisson mean and variance rather than the binomial's. Indeed, in standard units

$$\frac{Y - \mu}{\sqrt{\mu}}$$

is approximately normal for large μ . ■

Example 9.6.6 Poisson as Normal. Poisson becomes normal as $\mu \rightarrow \infty$. Consider $\mu = 20$. Then, $\sigma^2 = \mu = 20$. Using the Poisson formulas, for example,

$$P(X = 19) = \frac{20^{19}e^{-20}}{19!} \approx 0.08883$$

Using the normal distribution,

$$\begin{aligned} P(X = 19) &= P(18.5 < X < 19.5) \\ &\approx \text{normalcdf}(18.5, 19.5, 20, \text{sqrt}(20)) \\ &= 0.08683 \end{aligned}$$

Again, these are very close. □

Theorem 9.6.7 Gamma as approximate Normal. $r\mu rBLOB$

Example 9.6.8 Gamma as Normal. Gamma becomes normal as $r \rightarrow \infty$. Assume that the average time till a first success is 12 minutes and that $r = 8$. Then, the mean for the Gamma distribution is $\mu = 12 \cdot 8 = 96$ and $\sigma^2 = 8 \cdot 12^2 = 1152$ and so $\sigma \approx 33.9411$. Using the Gamma formulas,

$$\begin{aligned} P(90 \leq X \leq 100) &= \int_{90}^{100} f(x)dx \\ &= 0.59252 - 0.47536 = 0.11716. \end{aligned}$$

Using the normal distribution,

$$P(90 \leq X \leq 100) \approx \text{normalcdf}(90, 100, 96, 33.9411) = 0.11707.$$

Amazingly, these are also very close. □

Example 9.6.9 Uniform X vs Normal \bar{X} . Consider a discrete uniform variable X over $R = 1, 2, \dots, 20$. Then, $\mu = 10.5$ and $\sigma = \frac{20^2-1^2}{20}$ using the uniform formulas.

You can use the uniform probability function to compute probabilities dealing with X. Indeed,

$$P(8 \leq X < 12) = P(X \in \{8, 9, 10, 11\}) = \frac{4}{20} = 1/5.$$

If instead you plan to sample from this distribution $n=49$ times, the Central Limit Theorem implies that you will get a random variable \bar{X} which has an approximate normal distribution with the same mean but with new variance $\sigma_{\bar{X}}^2 = \frac{199/20}{49} = \frac{199}{580}$. Therefore, expanding the interval to include the boundaries of the corresponding histogram areas,

$$P(8 \leq \bar{X} < 12) = P(7.5 \leq \bar{X} \leq 11.5) \approx \text{normalcdf}(7.5, 11.5, 10.5, 0.585750) \approx 0.9561.$$

□

As these examples illustrate, you will have increasing success in approximating the desired probabilities so long as the distribution's corresponding parameter is allowed to be "sufficiently large". The mathematical reasoning this is true is not provided but depends upon the "Central Limit Theorem" discussed in the next section.

The above theorems allow you to utilize the normal distribution to compute approximate probabilities for the variable X in the stated distributions. This

is not always true for all distributions since some do not have parameters which allow for approaching normality. However, regardless of the distribution the Central Limit Theorem always allows you to approximate probabilities if they involve an average of repeated attempts...that is, for variable \bar{X} . This usefulness is illustrated in the examples below.

9.7 Summary

Here is a summary of the major points in this chapter:TBA

9.8 Exercises

Checkpoint 9.8.1 - Computing basic standard normal probabilities.

Compute

-

$$P(Z > 0)$$

-

$$P(Z < 0.892)$$

-

$$P(Z < -0.892)$$

-

$$P(-1.45 < Z < 2.37)$$

-

$$P(-1 < Z < 1)$$

which is the probability of lying within 1 standard deviation of the mean.

-

$$P(-2 < Z < 2)$$

which is the probability of lying within 2 standard deviations of the mean.

-

$$P(-3 < Z < 3)$$

which is the probability of lying within 3 standard deviations of the mean.

- A value for a so that

$$P(Z < a) = 0.8$$

which would be the location of the 80th percentile.

Checkpoint 9.8.2 - Computing basic normal probabilities. Given

$\mu = 25$ and $\sigma = 4$ compute

$$P(X < \mu)$$

$$P(X > 26)$$

$$P(X > 22)$$

$$P(20 \leq X \leq 26)$$

Checkpoint 9.8.3 - IQ values. The Intelligence Quotient (IQ) is a measure of your ability to think and reason. Presuming that IQ scores are normally distributed with mean 100 and standard deviation 15, determine the location

of the 90th percentile. That is, the IQ score below which you will find approximately 90

Chapter 10

Estimation

10.1 Introduction

You should have noticed by now that repeatedly sampling from a given distribution will yield a variety of sample statistics such as \bar{x} as an estimate perhaps for the population mean μ or $\frac{Y}{n}$ as an estimate for the population likelihood of success p . In this section, you will see how these sample "point estimators" are actually the best possible choices.

In creating these point estimates repeatedly, you have noticed that the results will change somewhat over time. Indeed, flip a coin 20 times and you might expect 10 heads. However, in practice it is likely to 9 or 12 out of 20 and possible to get any of the other possible outcomes. This natural variation makes the point estimates noted above to almost certainly be in error. However one would expect that they should be close and the Central Limit Theorem does indicate that the distribution of sample means should be approximately normally distributed. Thus, instead of relying just on the value of the point estimate, you might want to investigate a way to determine a reasonable interval centered on the sample statistic in which you have some confidence the actual population statistic should belong. This leads to a discussion of interval estimates known as confidence intervals (using calculational tools) and statistical tolerance intervals (using order statistics).

In this chapter we first discuss how to determine appropriate methods for estimating the needed population statistics (point estimates) and then quantify how good they are (confidence intervals).

10.2 Interval Estimates - Chebyshev

An interval centered on the mean in which at least a certain proportion of the actual data must lie.

Theorem 10.2.1 Chebyshev's Theorem. $\mu\sigma a \in \mathbb{R}^+$

$$P(|X - \mu| < a) > 1 - \frac{\sigma^2}{a^2}$$

Proof. Notice that the variance of a continuous variable X is given by

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ &\geq \int_{-\infty}^{\mu-a} (x - \mu)^2 f(x) dx + \int_{\mu+a}^{\infty} (x - \mu)^2 f(x) dx\end{aligned}$$

$$\begin{aligned}
&\geq \int_{-\infty}^{\mu-a} a^2 f(x) dx + \int_{\mu+a}^{\infty} a^2 f(x) dx \\
&= a^2 \left(\int_{-\infty}^{\mu-a} f(x) dx + \int_{\mu+a}^{\infty} f(x) dx \right) \\
&= a^2 P(X \leq \mu - a \text{ or } X \geq \mu + a) \\
&= a^2 P(|\mu - a| \geq a)
\end{aligned}$$

Dividing by a^2 and taking the complement gives the result. ■

Corollary 10.2.2 Alternate Form for Chebyshev's Theorem. For positive k ,

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

Corollary 10.2.3 Special Cases for Chebyshev's Theorem.

Proof. Apply the Chebyshev Theorem with $a = \sigma$ to get

$$P(\mu - \sigma < X < \mu + \sigma) > 1 - \frac{\sigma^2}{\sigma^2} = 0$$

Apply the Chebyshev Theorem with $a = 2\sigma$ to get $1 - \frac{1}{2^2} = 0.75$ and with $k = 3\sigma$ to get $1 - \frac{1}{3^2} = \frac{8}{9} > 0.8888$. ■

Example 10.2.4 - Comparing known distribution to Chebyshev. □

10.3 Point Estimates

For Binomial, Geometric, what is p ? For exponential, what is the mean? For normal, what are the mean and standard deviation? Each of these parameters are necessary before you can compute any probability values from their respective formulas. Since they might not be given in a particular instance, they will need to be estimated in some manner.

This estimate will have to be determined likely by utilizing sampling in some form. Since such an estimate will come from partial information (i.e. a sample) then it is very likely going to only be an approximation to the exact (but unknown) value. In general, an estimator is a numerical value which is used in the place of an unknown population statistic. To determine precisely what is a "best" estimator requires a multivariate approach and is beyond the scope of this text. Indeed, to justify why each of the following are good estimators look up the topic "Maximum Likelihood Estimators".

From your previous experience with the Poisson, Exponential, and Gamma distributions, you should also remember that each required a known value for μ before proceeding with calculations. It is sensible to consider estimating the unknown population mean μ using the sample mean

$$\mu \approx \bar{x} = \frac{\sum x_k}{n}$$

where the values x_k are the n individual sample values.

For any continuous variable and indeed for \bar{X} , $P(\bar{X} = \mu) = 0$. In general, you should expect a sample statistic to be close but not precisely equal to the population statistic. Indeed, if you were so lucky as to have the sample statistic

to land on the population statistic, doing one more trial would mess things up anyway and the sample statistic would certainly change some.

In a similar manner with the Binomial, Geometric, and Negative Binomial distributions, you will remember that each required a known value for p before proceeding with any calculations. From our experiments we saw that relative frequency appeared to stabilize around what you might expect for the true proportion of success and therefore estimating the unknown proportion of success p using relative frequency

$$p \approx \tilde{p} = \frac{y}{n}$$

where y is the number of successes in a collection of n bernoulli trials. Again, notice that the relative frequency \tilde{p} is technically an average as well so the probability that a given relative frequency will like exactly on the actual value of p is again zero.

Finally, the Normal distribution requires a numerical value for σ , the population's standard deviation. It can be shown that the maximum likelihood estimator for σ^2 is the variance v found in chapter one. However, you may remember that at that time we always adjusted this value somewhat using the formula $s^2 = \frac{n}{n-1}v$ which increased the variance slightly. To uncover why you would not use the maximum likelihood estimator v requires you to look up the idea of "bias". As it turns out, v is maximum likelihood but exhibits mathematical bias whereas s^2 is slightly suboptimal with respect to likelihood but exhibits no bias. Therefore, for estimating the unknown population variance σ^2 you can use sample variance

$$\sigma^2 \approx s^2$$

and similarly sample standard deviation

$$\sigma \approx s$$

to approximate the theoretical standard deviation.

10.4 Interval Estimates - Confidence Interval for p

Sometimes selecting a value for p for a Binomial, Geometric, or Negative Binomial distribution problem can be done by using a theoretical value. Indeed, when flipping a coin it is reasonable to assume $p = 1/2$ is the probability of getting a head on one flip. Similarly, it is reasonable to assume $p = 1/6$ when you are looking for a particular side of a 6-sided die. However, many times you will want to deal with a problem in which it is not possible to determine exactly the precise value for the likelihood of success such as your true probability of making a free throw in basketball or knowing the true percentage of the electorate that will vote for your favorite candidate.

In these later situations, we found in the previous section that relative frequency $\frac{Y}{n}$ is generally a good way to estimate p . In this section, you will investigate how to measure the closeness—and thereby assure some confidence in that estimate—regarding how well the point estimate approximates the actual value of p .

Definition 10.4.1 Confidence Intervals for p . Given a point estimate \tilde{p} for p , a confidence interval for p is a range of values which contains the actual value of p with high probability. In notation, a two-sided confidence interval

for p is of the form

$$\tilde{p} - E_1 < p < \tilde{p} + E_2$$

with

$$P(\tilde{p} - E_1 < p < \tilde{p} + E_2) = 1 - \alpha$$

where α is near 0 and $E_k > 0$. One-sided confidence intervals for p can be similarly described

$$P(p < \tilde{p} + E_2) = 1 - \alpha$$

or

$$P(\tilde{p} - E_1 < p) = 1 - \alpha.$$

◇

Generally, symmetry is presumed when using a two-sided confidence interval so that $E_1 = E_2 = E$ and therefore the interval looks like

$$P(\tilde{p} - E < p < \tilde{p} + E) = 1 - \alpha.$$

In this case, E is known as the margin of error.

To determine E carefully, note that from the central limit theorem

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{\tilde{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately standard normal for large n . Presuming that $\tilde{p} \approx p$ and replacing the unknown p terms on the bottom with \tilde{p} gives

$$z = \frac{\tilde{p} - p}{\sqrt{\tilde{p}(1-\tilde{p})/n}}$$

where z is a standard normal distribution variable. So, using the central limit theorem and the standard normal distribution, you can find the value $z_{\alpha/2}$ where

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$$

$$P(-z_{\alpha/2} < \frac{\tilde{p} - p}{\sqrt{\tilde{p}(1-\tilde{p})/n}} < z_{\alpha/2}) = 1 - \alpha$$

or by rearranging the inside inequality

$$P(\tilde{p} - z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n} < p < \tilde{p} + z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n}) = 1 - \alpha.$$

Setting $E = z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n}$ gives a way to determine a confidence interval centered on $\tilde{p} = \frac{Y}{n}$ for p with "confidence level" $1 - \alpha$.

To complete the interval, one needs a specific value for $z_{\alpha/2}$. Generally, one chooses confidence levels on the order of 90

$$z_{\alpha/2} = \text{InvNorm}(1 - \frac{\alpha}{2})$$

For 90

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 0.9 = 1 - 0.1.$$

Using the symmetry of the normal distribution, this can be rewritten

$$F(z_{\frac{0.1}{2}}) = P(z < z_{\frac{0.1}{2}}) = 0.95 = 1 - \frac{0.1}{2}.$$

Using the inverse of the standard normal distribution (on the TI calculator this is $\text{InvNorm}(0.95)$) gives $z_{0.05} \approx 1.645$.

Similarly, for a 95

$$F(z_{\frac{0.05}{2}}) = P(z < z_{\frac{0.05}{2}}) = 0.975 = 1 - \frac{0.05}{2}.$$

The calculator InvNorm(0.975) gives $z_{0.025} \approx 1.960$.

For a 99

$$F(z_{\frac{0.01}{2}}) = P(z < z_{\frac{0.01}{2}}) = 0.995 = 1 - \frac{0.01}{2}.$$

The calculator InvNorm(0.995) gives $z_{0.005} \approx 2.576$.

Notice that when computing the confidence intervals above that we choose to just replace some of the p terms with \tilde{p} so that only one p term was left and could be isolated in the middle. There are other ways to deal with this. The easiest is to take the worst case scenario for the p terms in the denominator above. Indeed, the confidence interval is made wider (and therefore more likely to contain the actual p) if the square root term is as large as possible, using basic calculus it is easy to see that $p(1-p)$ is maximized when $p = 1/2$. Therefore, a second alternative is to create your confidence interval using

$$z = \frac{\tilde{p} - p}{\frac{1}{2\sqrt{n}}}$$

and therefore $E = \frac{z_{\alpha/2}}{2\sqrt{n}}$. This method should be used only when trying to create the roughest and "safest" interval.

The methods for determining a confidence interval for p above depend upon a good approximation with the Central Limit Theorem. This approximation will be fine if n is relatively large. To consider a confidence interval for p when n is small, note that the binomial random variable is discrete and so expanding the interval by a factor of $\frac{1}{2n}$ might be in order.

Another more elaborate mechanism when n is relatively large is given by the Wilson Score. This confidence interval is more complicated than just taking \tilde{p} and adding and subtracting E . This approach notes that the possible extreme values for p must satisfy (before replacing some of the p terms with \tilde{p})

Theorem 10.4.2 Wilson Score Confidence Interval for p .

$$\frac{\tilde{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p}) + \frac{z_{\alpha/2}^2}{4n}}{n}}}{1 + \frac{z_{\alpha/2}^2}{n}} < p < \frac{\tilde{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p}) + \frac{z_{\alpha/2}^2}{4n}}{n}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

Proof. Again, noting that $\tilde{p} = \frac{Y}{n}$, the expression above

$$|p - \tilde{p}| = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

can be simplified by squaring both sides to get

$$(p - \tilde{p})^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}.$$

Replacing \tilde{p} with the relative frequency gives

$$(p - \frac{Y}{n})^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}$$

or by simplifying

$$(n + z_{\alpha/2}^2)p^2 - (2Y + z_{\alpha/2}^2)p + \frac{Y^2}{n} = 0.$$

Solving for p using the quadratic formula and simplifying ultimately results in the described interval. ■

Example 10.4.3 Comparison of the three Confidence Interval methods for p. Presume that from a sample of size $n = 400$ you get $Y = 144$ successes. Determine 95

Normal Interval:

$$P(0.36 - 1.960\sqrt{0.36 \cdot 0.64}/400 < p < 0.36 + 1.960\sqrt{0.36 \cdot 0.64}/400) = 1 - \alpha.$$

or

$$P(0.36 - 1.960 \cdot 0.6 \cdot 0.8)/20 < p < 0.36 + 1.960 \cdot 0.6 \cdot 0.8)/20) = 0.95$$

or

$$P(0.36 - 0.04704 < p < 0.36 + 0.04704) = 0.95.$$

or

$$P(0.31296 < p < 0.40704) = 0.95.$$

So, there is a 95

Maximal Interval:

$$P(0.36 - 1.960\frac{1}{2\sqrt{400}} < p < 0.36 + 1.960\frac{1}{2\sqrt{400}}) = 1 - \alpha.$$

or

$$P(0.36 - 1.960\frac{1}{40} < p < 0.36 + 1.960\frac{1}{40}) = 1 - \alpha.$$

or

$$P(0.311 < p < 0.409) = 1 - \alpha.$$

Notice the interval is only slightly wider than when using \tilde{p} to estimate p in the first case.

Wilson Score Interval: Let's do this on in parts...

$$z_{\alpha/2}\sqrt{\frac{\tilde{p}(1 - \tilde{p}) + \frac{z_{\alpha/2}^2}{4n}}{n}} = 1.96\sqrt{\frac{0.36 \cdot 0.64 + \frac{1.96^2}{1600}}{400}} \approx 0.04728$$

Therefore,

$$\frac{0.36 + \frac{1.96^2}{800} - 0.04728}{1 + \frac{1.96^2}{400}} < p < \frac{0.36 + \frac{1.96^2}{800} + 0.04728}{1 + \frac{1.96^2}{400}}$$

or

$$0.3145 < p < 0.4082$$

which is slightly different than the first and slightly smaller than the second. \square

Theorem 10.4.4 Determining Sample Size for proportions. \tilde{p}_0

$$n > \left(\frac{z_{\alpha/2}}{E}\right)^2 \tilde{p}_0(1 - \tilde{p}_0).$$

Proof.

■

Example 10.4.5 Determining Sample Size for one proportion. Given a 99

$$n > \left(\frac{2.58}{0.03}\right)^2 0.35 \cdot 0.65 \approx 1682.59$$

or a sample size of at least 1683. \square

10.5 Interval Estimates - Confidence Interval for μ

As with the confidence intervals above for proportions, the Central Limit Theorem also allows you to create an interval centered on a sample mean for estimating the population mean μ .

Definition 10.5.1 Confidence Interval for One Mean. Given a sample mean \bar{x} , a two-sided confidence interval for the mean with confidence level $1 - \alpha$ is an interval

$$\bar{x} - E_1 < \mu < \bar{x} + E_2$$

such that

$$P(\bar{x} - E_1 < \mu < \bar{x} + E_2) = 1 - \alpha.$$

Generally, the interval is symmetrical of the form $\bar{x} \pm E$ with E again known as the margin of error. One-sided confidence intervals can be determined in the same manner as in the previous section. \diamond

Once again, utilize the Central Limit Theorem. Notice that the symmetrical confidence interval

$$P(\bar{x} - E < \mu < \bar{x} + E) = 1 - \alpha.$$

is equivalent to

$$P\left(\frac{-E}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{E}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

in which the middle term can be approximated using a standard normal variable and therefore this statement is approximately

$$P\left(\frac{-E}{\sigma/\sqrt{n}} < Z < \frac{E}{\sigma/\sqrt{n}}\right) = 1 - \alpha.$$

Using the symmetry of the standard normal distribution about $Z=0$ gives

$$\Phi(z_{\alpha/2}) = \Phi\left(\frac{E}{\sigma/\sqrt{n}}\right) = P\left(Z < \frac{E}{\sigma/\sqrt{n}}\right) = 1 - \frac{\alpha}{2}$$

and so to determine E again requires the inverse of the standard normal distribution function. Using an appropriate $z_{\alpha/2}$ (as determined in a manner described in the previous section) gives a confidence interval for the mean

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

with confidence level $1 - \alpha$ and margin of error

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

It should be noted that the use of the Central Limit Theorem makes the use of InvNorm an approximation. It can be shown that so long as n is larger than 30 then generally this approximation is reasonable.

Additionally, this derivation assumes that μ is not known...indeed the goal is to approximate that mean using \bar{x} ...but that σ is known. This is often not the case. It can however be shown that if n is larger than 30, replacing σ with the sample standard deviation s gives an acceptable confidence interval.

Theorem 10.5.2 Sample Size needed for μ given Margin of Error.
Given confidence level $1 - \alpha$ and margin of error E , the sample size needed to determine an appropriate confidence interval satisfies

$$n > \left(z_{\alpha/2} \frac{\sigma}{E} \right)^2$$

Proof. Solve for n in the formula for E above. Notice that n must be an integer so you will need to round up. You will also need an estimate for the sample standard deviation s by using a preliminary sample. ■

Notice, in practice you might want to take n to be a little larger than the absolute minimum value prescribed above since you are dealing with approximations (Central Limit Theorem and the use of an estimate for s rather than the actual σ .)

Example 10.5.3 Determining Sample Size for one Mean. Given a 95

$$n > \left(1.96 \cdot \frac{2}{0.1} \right)^2 \approx 1536.64$$

or a sample size of at least 1537. □

10.6 Interval Estimates - Confidence Interval for σ^2

Once again, you may need to approximate the population variance or standard deviation but only have the sample values available. One difference from the previous sections is that you are not dealing with an average of values (such as \bar{x} or \tilde{p}) but with the average of the squares of values. The Central Limit Theorem does not directly help you in this case but the following result (presented without proof) provides a solution.

Theorem 10.6.1 Relationship between Variance and χ^2 . *If S^2 is a random variable of possible sample variance values from a sample of size n , then*

$$W = \frac{(n-1)S^2}{\sigma^2}$$

is approximately $\chi^2(n-1)$.

To create a confidence interval for σ^2 first consider an interval of the form

$$E_1 < \sigma^2 < E_2$$

and determine values for the boundaries so that the likelihood of this being true is high. For this case, since the chi-square distribution only has a positive domain and is not symmetrical, you will not expect to determine a symmetrical confidence interval. Therefore, consider

$$P(E_1 < \sigma^2 < E_2) = 1 - \alpha$$

and by playing around with algebra you get

$$P\left(\frac{E_1}{(n-1)S^2} < \frac{\sigma^2}{(n-1)S^2} < \frac{E_2}{(n-1)S^2}\right) = 1 - \alpha$$

or by inverting the inequality yields

$$P\left(\frac{(n-1)S^2}{E_2} < \frac{(n-1)S^2}{\sigma^2} < \frac{(n-1)S^2}{E_1}\right) = 1 - \alpha.$$

Using the previous theorem, note that the inside variable can be replaced with a chi-square variable. If F is the distribution function for chi-square, then you get

$$F\left(\frac{(n-1)S^2}{E_1}\right) - F\left(\frac{(n-1)S^2}{E_2}\right) = 1 - \alpha.$$

For a given value of α there are many possible choices but often one often utilized is one in which

$$F(\chi_{1-\alpha/2}^2) = F\left(\frac{(n-1)S^2}{E_1}\right) = 1 - \alpha/2$$

and

$$F(\chi_{\alpha/2}^2) = F\left(\frac{(n-1)S^2}{E_2}\right) = \alpha/2.$$

Using the inverse chi-square gives values for the expression on the inside and algebra can be used to solve for each of E_1, E_2 . Indeed,

$$E_1 = \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2}$$

and

$$E_2 = \frac{(n-1)S^2}{\chi_{\alpha/2}^2}$$

To determine appropriate values for $\chi_{\alpha/2}^2$ and $\chi_{1-\alpha/2}^2$ with equal probabilities in each tail, consider using the interactive cell below:

```
# Chi-Square Calculator for confidence intervals with equal
  alpha/2 tails
var('t')
@interact(layout=dict(top=[['c'], ['n']]))
def _(c=input_box(0.95,width=10,label='Confidence_Level_=_',
),n=input_box(20,width=8,label='n_=_')):
    alpha = 1-c
    T = RealDistribution('chisquared', n)
    a = T.cum_distribution_function_inv(alpha/2)
    a1 = T.cum_distribution_function(a)
    b = T.cum_distribution_function_inv(1-alpha/2)
    b1 = T.cum_distribution_function(b)

    print 'From the Chi-Square distribution for X: '
    print 'P( ',a,' < X < ',b),')_=_',c
    print 'with'
    print 'P(X_< ',a,')_=_',a1
    print 'P(X_< ',b,')_=_',b1

    f = x^(n/2-1)*e^(-x/2)/(gamma(n/2)*2^(n/2))
    G =
        plot(f,x,0,b+(b-a)/2)+plot(f,x,a,b,thickness=5,color='green')
    G += line([(a,0),(a,f(x=a))],color='green',thickness=3)
    G += line([(b,0),(b,f(x=b))],color='green',thickness=3)
    G +=
        text(str(c.n(digits=5)),((a+b)/2,f(x=(a+b)/2)/3),color='green')
    G.show()
```

The example below uses the specific chi-square values given in the interactive cell below:

```
# Chi-Square Calculator specifics
var('t')
c=0.95
n=8
alpha = 1-c
T = RealDistribution('chisquared', n)
a = T.cum_distribution_function_inv(alpha/2)
a1 = T.cum_distribution_function(a)
b = T.cum_distribution_function_inv(1-alpha/2)
b1 = T.cum_distribution_function(b)

print 'From the Chi-Square distribution for X: '
print 'P( ', a, '<= X <= ', b, ') = ', c
print 'with'
print 'P( X <= ', a, ') = ', a1
print 'P( X <= ', b, ') = ', b1

f = x^(n/2-1)*e^(-x/2)/(gamma(n/2)*2^(n/2))
G =
    plot(f,x,0,b+(b-a)/2)+plot(f,x,a,b,thickness=5,color='green')
G += line([(a,0),(a,f(x=a))],color='green',thickness=3)
G += line([(b,0),(b,f(x=b))],color='green',thickness=3)
G +=
    text(str(c.n(digits=5)),((a+b)/2,f(x=(a+b)/2)/3),color='green')
G.show()
```

Example 10.6.2 - Two-sided Confidence interval for σ^2 and σ . Given the data 570, 561, 546, 540, 609, 580, 550, 577, 585, determine a 95 Using the computational formula (or your calculator) gives $s^2 \approx 479.5$. Also, notice for $n=9$, the resulting interval will use a Chi-square variable with 8 degrees of freedom. Using the symmetric option, gives $\chi_{0.025}^2 = 2.18$ and $\chi_{0.975}^2 = 17.53$. Therefore

$$E_1 = \frac{8 \cdot 479.5}{17.53} \approx 221.095$$

and

$$E_2 = \frac{8 \cdot 479.5}{2.18} \approx 1759.63.$$

Hence, you are 95

$$221.095 < \sigma^2 < 1759.63.$$

By taking square roots you get

$$14.87 < \sigma < 41.95.$$

Notice, this interval is relatively wide which is a result both of the number of data values being relatively small ($n=9$) and the actual data values being relatively large and spread out. \square

The example below uses the specific chi-square values given in the interactive cell below:

```
# Chi-Square Calculator specifics
var('t')
c=0.95
```

```

n=399
alpha = 1-c
T = RealDistribution('chisquared', n)
a = T.cum_distribution_function_inv(alpha/2)
a1 = T.cum_distribution_function(a)
b = T.cum_distribution_function_inv(1-alpha/2)
b1 = T.cum_distribution_function(b)

print 'From the Chi-Square distribution for X: '
print 'P( ', a, '<= X <= ', b, ') = ', c
print 'with'
print 'P( X <= ', a, ') = ', a1
print 'P( X <= ', b, ') = ', b1

```

Example 10.6.3 - Two-sided Confidence interval for the variance and standard deviation with large n.. Continuing the previous example, suppose now that you have $n=400$ data values and suppose you have computed from those a sample variance of $s^2 = 479.5$. Then, the only change in the calculation is the two chi-square statistic values. For 95

$$E_1 = \frac{8 \cdot 479.5}{456.24} \approx 419.3$$

and

$$E_2 = \frac{8 \cdot 479.5}{345.55} \approx 553.7.$$

Hence, you are 95

$$419.24 < \sigma^2 < 553.7.$$

By taking square roots you get

$$20.48 < \sigma < 23.53$$

which is a relatively tight confidence interval. Notice, these are also completely contained in the confidence intervals from the previous small n example. \square

Similar to above, another choice to estimate σ^2 is to use a one sided confidence interval. If you want to find one of these, continue as described above but just leave one endpoint off. Indeed,

$$\sigma^2 < E_2$$

can be determined using

$$F(\chi_\alpha^2) = F\left(\frac{(n-1)S^2}{E_2}\right) = \alpha$$

and

$$E_1 < \sigma^2$$

can be determined using

$$F(\chi_{1-\alpha}^2) = F\left(\frac{(n-1)S^2}{E_1}\right) = 1 - \alpha.$$

Example 10.6.4 - One-sided Confidence intervals for σ^2 . \square

Finally, to determine a confidence interval for σ , proceed using the protocols described above and simply take the square root on the resulting interval.

Example 10.6.5 - Confidence intervals for σ . \square

10.7 Exercises

Checkpoint 10.7.1 - Basic Confidence interval for p. Given $Y = 30$ successes in $n = 100$ trials, determine a 90% $\hat{p} = 0.3$ and $z_{0.05} = 1.645$ gives

$$0.3 - 1.645\sqrt{\frac{0.3 \cdot 0.7}{100}} < p < 0.3 + 1.645\sqrt{\frac{0.3 \cdot 0.7}{100}}$$

or

$$0.225 < p < 0.375.$$

Checkpoint 10.7.2 - Sample Size for confidence interval for p. Given a preliminary estimate $\tilde{p}_0 = 0.23$, determine the same size needed for determine a 95% Using $z_{0.025} = 1.96$,

$$n > \left(\frac{1.96}{0.02}\right)^2 \cdot 0.23 \cdot 0.77 \approx 1700.87$$

and so pick at least 1701 as the sample size.

Checkpoint 10.7.3 - Voting projection. Randomly polling 3200 eligible voters for governor in a particular state resulted in finding that 1590 favored your candidate. Determine an appropriate 95% Note that although the point estimate is below 50

Checkpoint 10.7.4 - Basic Confidence interval for the mean. Given a sample mean of $\bar{x} = 25.3$ with $n = 121$ and sample variance $s^2 = 12.1$, determine a 99% Using $z_{0.005} = 2.576$ and $s = \sqrt{12.1} \approx 3.4786$ gives a confidence interval

$$25.3 - 2.576 \cdot \frac{3.4786}{11} < \mu < 25.3 + 2.576 \cdot \frac{3.4786}{11}$$

or

$$24.4854 < \mu < 26.1148.$$

Checkpoint 10.7.5 - Confidence Interval Experiment.

Roll two regular pair of dice 35 times, recording the sum of the dots for each roll. Using the data from your sample, determine the corresponding sample mean and sample variance. Using this data, create a 95

Go back over your 35 rolls and count the number of 7's or 11's rolled. Determine a corresponding relative frequency for this outcome. Using this data, create a 95

Repeat this exercise but this time roll 105 times. Notice how these differ from the confidence intervals created with the smaller set. Write a paragraph describing how these compare and whether one is better or not than the other.

Chapter 11

Hypothesis Testing

11.1 Introduction

When creating confidence intervals, you started with a sample and used the sample statistic (sample mean, relative frequency, sample variance, etc.) to anchor an interval which (with high possibility) contains the corresponding population statistic μ, p, σ^2 , etc. In this section, we instead start with an educated guess for one of the population statistics μ, p, σ^2 and then statistically compare that value with the subsequently collected sample statistic.

The educated guess noted above is often called the "null hypothesis" and should be considered as a guess that one tries to disprove if possible by using a subsequent statistical sample.

11.2 Hypotheses and Errors

In formulating a hypothesis, you will be making a declarative statement (a proposition) that has an actual truth value. That is, it is either true or it is not true in real life. However, since we will assess that truth using a test sample then measuring that truth value will never be 100

In general, there are four different outcomes that are possible when testing a hypothesis:

- Your hypothesis is true and you determine that it is true.
- Your hypothesis is false but you determine that it is true.
- Your hypothesis is true but you determine that it is false.
- Your hypothesis is false and you determine that it is false.

The first and last cases are "good" since you have accurately determined the truth of the hypothesis. The second and third are however bad since you either believe something that is not true or you don't believe something that is true. We would like to minimize the likelihood of allowing these last two possibilities.

Toward that end, let's consider the case where the hypothesis is true but you determine (in error) that it is false. This is called a Type I error and we will designate the probability of this error by α . To lower the risk of a Type I error, you will want to make α smaller. In general, α is also called "the significance level" with $\alpha = 0.05$ a common choice with 0.01 and 0.10 also

sometimes used. In general, any value between 0 and 1 is ok but large values mean large likelihood for error so choosing a value closer to 0 is preferred.

In a similar manner, consider the case where the hypothesis is false but you determine (in error) that it is true. This is called a Type II error and will be denote the probability of a Type II error by β . Again, your goal is to make the risk of a Type II error smaller and therefore want β to be as small as possible.

In the following sections you will play around with minimizing Type I and Type II errors. Type I errors will be minimized by simply choosing a smaller value for α when working through the formulas. Type II errors will be minimized by talking (if possible) a larger sample size when computing the needed sample statistics.

Toward that end, you will compose in each case a *Null Hypothesis* (denoted N_0). This statement is what you will test for truth. You will also compose an *Alternate Hypothesis* (denoted N_a) that is often the logical complement (but not always) of N_0 . The null hypothesis is often a statement corresponding to the likelihood that observations occur purely by chance while the alternate hypothesis will often indicate that outcomes are not actually random but are influenced by some (possibly unknown) causes. If our sample shows that the null hypothesis N_0 is false, then we will accept the alternate N_a . If this is a bad decision then it will be true that the hypothesis is true but you will have determined that it is false...that is, made a Type I error.

To avoid ever making an Type II error, often often never "accepts" the null hypothesis N_0 even if the sample does not conclude that it is false. In other words, if you do this then you will never allow yourself to determine that N_0 is true. Seems odd but this is one way to avoid ever worrying much about type II errors. The best way to avoid this blind spot is to use relatively large test samples and in doing so you will minimize the likelihood of type II.

So, our plan of attack is to find some way to determine, from a sample, the amount of Type I error we might make. For a given problem, from the sample, you will create a specific estimate for Type I error...called a p-statistic...and compare to the chosen significance level α . Sounds easy enough?

Finally, you will notice in each of the instances presented that the form of the solution method will be very similar to the forms we used in creating confidence intervals. The difference is that for hypothesis testing we *assume* the value in the middle and test to see whether the sample statistic is in one of the tails so that we can reject rather than with confidence intervals we assume as sample statistic is in the middle and then use that sample to create boundaries within which the theoretical population statistic *must* like with high confidence.

11.3 Hypothesis Test for one proportion

In this section, you will consider the following options for null hypothesis and corresponding alternate hypothesis with respect to the unknown value p:

$$H_0 : p = p_0 H_a : p \neq p_0$$

or

$$H_0 : p \leq p_0 H_a : p > p_0$$

or

$$H_0 : p \geq p_0 H_a : p < p_0.$$

For any given problem, we choose only one of these three options. The first is called a "two-tailed" test since the alternate hypothesis can not be equal if it is

actually larger or smaller. The last two are called "one-tailed" tests since the alternate hypothesis only allows for being on one side. Note that some people will write all of these null hypothesis options using equality but the alternate hypothesis determines the number of tails.

From the Central Limit Theorem, you found that every interesting distribution eventually becomes approximately normal. This includes the binomial distribution with mean $\mu = np_0$ and variance $\sigma^2 = np_0(1 - p_0)$. Hence, the *z-statistic*

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{p - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

is approximately standard normal and so probabilities on this statistic can be computed as needed using the normal distribution.

Let's look at an example for this by considering:

$$H_0 : p = 0.20$$

vs

$$H_a : p \neq 0.20.$$

Tests like this are called *two-tailed* since there are two ways to reject the null hypothesis: we find that p should be less than 0.2 or we find that p should be greater than 0.2.

To test our hypothesis, let's now choose a significance level of $\alpha = 0.05$ and take a sample. Presuming we actually do this, let's assume that we find that out of $n=100$ sample values we get $X = 27$ successes. Hence, our actual test statistic is $p = \frac{27}{100} = 0.27$.

So, in this case we have

$$\sigma = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.2 \cdot 0.8}{100}} = 0.04$$

and the z -statistic (using the normal distribution) for the sample statistic of $p = 0.27$ is

$$z = \frac{0.27 - 0.2}{0.04} = 1.75.$$

Remember, the alternate hypothesis has two tails so to determine the P value we need to determine from the normal distribution

$$P(Z > 1.75) + P(Z < -1.75)$$

and find that this has probability approximately $0.0392 + 0.0392 = 0.0784$. However, this P value is greater than our significance level $\alpha = 0.05$ so we cannot reject the null hypothesis at the 5 percent significance level. However, if we had chosen initially to use a 10 percent significance level then we would have rejected the null hypothesis and accepted the alternate.

One tailed test for p using a "real" example...

Checkpoint 11.3.1 WebWork. Two-tailed test for p .

An article in the Washington Post on March 16, 1993 stated that nearly 45 percent of all Americans have brown eyes. A random sample of $n = 78$ C of I students found 30 with brown eyes.

We test

$$H_0 : p = .45$$

$$H_a : p \neq .45$$

(a) What is the z -statistic for this test? _____

(b) What is the P -value of the test? _____

Checkpoint 11.3.2 WebWork. One-tailed test for p . Notice, in this case you will only compute z -score probability for one tail and not both tails.

A noted psychic was tested for ESP. The psychic was presented with 220 cards face down and was asked to determine if the card was one of 5 symbols: a star, cross, circle, square, or three wavy lines. The psychic was correct in 52 cases. Let p represent the probability that the psychic correctly identifies the symbol on the card in a random trial. Assume the 220 trials can be treated as an SRS from the population of all guesses.

To see if there is evidence that the psychic is doing better than just guessing, we test

$$H_0 : p = .2$$

$$H_a : p > .2$$

(a) What is the z -statistic for this test? _____

(b) What is the P-value of the test? _____

11.4 Hypothesis Test for one mean

In this section, you will consider the following options for null hypothesis and corresponding alternate hypothesis with respect to the unknown value μ :

$$H_0 : \mu = \mu_0 H_a : \mu \neq \mu_0$$

or

$$H_0 : \mu \leq \mu_0 H_a : \mu > \mu_0$$

or

$$H_0 : \mu \geq \mu_0 H_a : \mu < \mu_0$$

Again, we choose only one of these three options and as before the first is called a "two-tailed" test and the last two are called "one-tailed" tests. Note that some people will write all of these null hypothesis options using equality but the alternate hypothesis determines the number of tails.

Once again, if the test sample size is sufficiently large and the standard deviation σ of the underlying distribution is known, one can use the normal distribution to compute probabilities. If the test sample size is relatively small or if σ is not known (and therefore is approximated by the sample standard deviation s), then the t -distribution can be utilized to compute probabilities. We will only consider using the t -distribution to compute p -values.

The Standard Error σ_e is given by

$$\sigma_e = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of the sample. For this presentation, we will assume that the actual population is relatively large relative to the sample size. In cases where this is not true, an adjustment (not presented here) will need to be made when computing σ_e .

To determine the p -value for a given sample with sample mean \bar{x} ,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is a t -variable with $n-1$ degrees of freedom. Therefore, probabilities on t can be computed using the t -distribution.

Consider a Two-tailed Hypothesis test for μ using

$$H_0 : \mu = 200 H_a : \mu \neq 200$$

using a sample of size $n = 49$ and with a resulting mean of $\bar{x} = 206$, a sample standard deviation of $s = 15$, and a significance level of $\alpha = 0.01$.

The standard error for this test is

$$\sigma_e = \frac{15}{\sqrt{49}} = \frac{15}{7}$$

and so using the t-distribution with degrees of freedom $n-1 = 48$ yields a t-statistic of

$$t = \frac{\bar{x} - 200}{\sigma_e} = \frac{206 - 200}{\frac{15}{7}} = \frac{14}{5} = 2.80.$$

To compute the p-value,

$$P(t > 2.80) + P(t < -2.80) \approx 0.0037 + 0.0037 = 0.0074.$$

Since this p-value is less than our significance level $\alpha = 0.01$ then you can reject the null hypothesis and accept the alternate.

Consider One-tailed Hypothesis test for μ using an interesting application:

Suppose that a manufacturer bottling a delicious beverage and the label indicates that the bottle contains 16 fluid ounces. Since providing the customer too little product might cause a significant negative reaction relative to the modest additional cost of providing a little too much, consider the following hypothesis test:

$$H_0 : \mu = 16 \quad H_a : \mu > 200.$$

To test this hypothesis at significance level $\alpha = 0.05$, you randomly pull out 20 bottles from the production line and accurately measure the amount of produce in each bottle. If the resulting average of these measurements is $\bar{x} = 16.05$ ounces with a standard deviation of $s = 0.08$ ounces, determine if you can safely make it known that more product is actually delivered in general to each consumer.

The standard error for this test is

$$\sigma_e = \frac{0.08}{\sqrt{20}} \approx 0.01789$$

and using the t-distribution with $n-1 = 19$ degrees of freedom yields a t-statistic of

$$t = \frac{\bar{x} - 16}{\sigma_e} = \frac{16.05 - 16}{0.01789} \approx 2.795.$$

To compute the p-value,

$$P(t > 2.795) \approx 0.0058.$$

Since this p-value is less than our significance level $\alpha = 0.05$ (by a lot) then you can reject the null hypothesis and accept the alternate. It is safe therefore to say that customers can expect at least 16 ounces! However, note that some folks will still be stiffed since the standard deviation of 0.08 certainly means that some bottles have less than 16.05-0.08 ounces of beverage.

11.5 Hypothesis Test for one variance

Proceeding in a similar manner, we can also perform hypothesis testing on variances using the χ^2 -distribution to determine probabilities.

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_a : \sigma^2 \neq \sigma_0^2$$

or

$$H_0 : \sigma^2 \geq \sigma_0^2 \quad H_a : \sigma^2 < \sigma_0^2$$

or

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad H_a : \sigma^2 > \sigma_0^2$$

$$\text{test statistic} = T = (n-1) \frac{s^2}{\sigma_0^2}$$

For two-tailed, reject if

$$T > \chi_{1-\alpha/2, n-1}^2 \quad \text{or} \quad T < \chi_{\alpha/2, n-1}^2$$

and for one-tailed to the right if

$$T > \chi_{1-\alpha, n-1}^2$$

and for one-tailed to the left if

$$T < \chi_{\alpha, n-1}^2.$$

Technically, for the two-tailed test you could pick T-values so that the total probability sums to α in any fashion but generally this probability is split evenly between the two tails as noted above.

11.6 Summary

TBA

11.7 Exercises

TBA

Chapter 12

Review of Calculus

This chapter is a review of power series results from Calculus.

12.1 Geometric Series

Knowledge of the use of power series is very important when dealing with both probability functions.

$$S = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$$

as is its extension known as the negative binomial series ($n \in \mathbb{N}$).

$$NB = \sum_{k=0}^{\infty} (-1)^k \binom{-n+k-1}{k} x^k b^{-n-k} = \frac{1}{(x+b)^n}$$

In this section, we review this series, develop its properties, and explore some of its extensions.

12.1.1 Geometric Series

Theorem 12.1.1 $S = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$

Proof. Consider the partial sum

$$\begin{aligned} S_n &= \sum_{k=0}^n x^k = 1 + x + x^2 + \dots + x^n \\ (1-x)S_n &= S_n - xS_n = 1 + x + x^2 + \dots + x^n - (x + x^2 + \dots + x^n + x^{n+1}) = 1 - x^{n+1} \\ \Rightarrow S_n &= \frac{1 - x^{n+1}}{1 - x} \end{aligned}$$

and so as $n \rightarrow \infty$,

$$S_n \rightarrow S = \frac{1}{1-x}$$

■

The interactive activity below shows how well the partial sums approximate $\frac{1}{1-x}$ as the number of terms increases.

```

var('x,n,k')
f = 1/(1-x)
@interact
def _(n = slider(2,20,1,2)):
    Sn = sum(x^k,k,0,n)
    pretty_print(html('$S_n(x) = \sum_{k=0}^n x^k$' + str(latex(Sn))))
    G = plot(f,x,-1,0.9,color='black')
    G += plot(Sn,x,-1,0.9,color='blue')
    G += plot(abs(f-Sn),x,-1,0.9,color='red')
    G.show(title="Partial Sums (blue) vs Infinite Series (black) and Error (red)",figsize=(5,4))

```

12.1.2 Alternate Forms for the Geometric Series

Theorem 12.1.2 Generalized Geometric Series. $k \in \mathbb{N}, \sum_{k=M}^{\infty} x^k = \frac{x^M}{1-x}$

Proof.

$$\begin{aligned}
 \sum_{k=M}^{\infty} x^k &= x^M \sum_{k=0}^{\infty} x^k \\
 &= x^M \frac{1}{1-x} \\
 &= \frac{x^M}{1-x}
 \end{aligned}$$

■

Example 12.1.3 Integrating and Differentiating to get new Power Series. The geometric power series is a nice function which is relatively easily differentiated and integrated. In doing so, one can obtain new power series which might also be very useful in their own right. Here we develop a few which are of special interest.

Let $f(x) = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$. Then,

$$\begin{aligned}
 f'(x) &= \sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2} \\
 f''(x) &= \sum_{k=2}^{\infty} k(k-1)x^{k-2} = \frac{2}{(1-x)^3} \\
 f^{(n)}(x) &= \sum_{k=n}^{\infty} k(k-1)\dots(k-n+1)x^{k-n} = \frac{n!}{(1-x)^{n+1}} \\
 \int f(x)dx &= \sum_{k=0}^{\infty} \frac{x^{k+1}}{k+1} = -\ln(1-x)
 \end{aligned}$$

□

Example 12.1.4 Playing with the base.

$$\begin{aligned}
 \sum_{k=0}^{\infty} a^k x^k &= \sum_{k=0}^{\infty} (ax)^k \\
 &= \frac{1}{1-ax}, |x| < \frac{1}{a}
 \end{aligned}$$

or perhaps

$$\sum_{k=0}^{\infty} (x-b)^k = \frac{1}{1-(x-b)}, |x-b| < 1$$

□

Example 12.1.5 Application: Converting repeating decimals to fractional form. Consider this example:

$$\begin{aligned} 2.48484848\dots &= 2 + 0.48 + 0.0048 + 0.000048 + \dots \\ &= 2 + 0.48(1 + 0.01 + 0.0001 + \dots) = 2 + 0.48 \sum_{k=0}^{\infty} (0.01)^k \end{aligned}$$

Therefore, applying the Geometric Series

$$\begin{aligned} 2.48484848\dots &= 2 + 0.48 \frac{1}{1-0.01} \\ &= 2 + 0.48 \frac{100}{99} = 2 + \frac{48}{99} \end{aligned}$$

□

Example 12.1.6 Playing around with repeating decimals. Certainly most students would agree that $0.333333\dots = \frac{1}{3}$. So, what about $0.999999\dots$? Simply follow the pattern above

$$\begin{aligned} 0.999999\dots &= 0.9 + 0.09 + 0.009 + 0.0009 + \dots = 0.9(1 + 0.1 + 0.1^2 + 0.1^3 + \dots) \\ &= 0.9 \frac{1}{1-0.1} = 0.9 \frac{1}{0.9} = 1 \end{aligned}$$

□

12.2 Binomial SumsBinomial SeriesTrinomial Series

The binomial series is also foundational. It is technically not a series since the sum is finite but we won't bother with that for now. It is given by

Proof. By induction:

Basic Step: $n = 1$ is trivial

Inductive Step: Assume the statement is true as given for some $n \geq 1$. Show

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

$$\begin{aligned} (a+b)^{n+1} &= (a+b)(a+b)^n \\ &= (a+b) \sum_{k=0}^n \binom{n}{k} a^k b^{n-k} \\ &= \sum_{k=0}^n \binom{n}{k} a^{k+1} b^{n-k} + \sum_{k=0}^n \binom{n}{k} a^k b^{n-k+1} \\ &= \sum_{k=0}^{n-1} \binom{n}{k} a^{k+1} b^{n-k} + a^{n+1} + b^{n+1} + \sum_{k=1}^n \binom{n}{k} a^k b^{n-k+1} \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^n \binom{n}{j-1} a^j b^{n-(j-1)} + a^{n+1} + b^{n+1} + \sum_{k=1}^n \binom{n}{k} a^k b^{n+1-k} \\
&= b^{n+1} + \sum_{k=1}^n \left[\binom{n}{k-1} + \binom{n}{k} \right] a^k b^{n+1-k} + a^{n+1} \\
&= b^{n+1} + \sum_{k=1}^n \binom{n+1}{k} a^k b^{n+1-k} + a^{n+1} \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} a^k b^{n+1-k} \quad \blacksquare
\end{aligned}$$

Consider $B(a, b) = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$. This finite sum is known as the *Binomial Series*.

Show that $B(a, b) = (a + b)^n$

Show that $B(1, 1) = 2^n$

Show that $B(-1, 1) = 0$

Show that $B(p, 1-p) = 1$

Easily, $B(x, 1) = \sum_{k=0}^n \binom{n}{k} a^k$

$$(a + b + c)^n = \sum_{k_1+k_2+k_3=n} \binom{n}{k_1, k_2, k_3} a^{k_1} b^{k_2} c^{k_3}$$

where $\binom{n}{k_1, k_2, k_3} = \frac{n!}{k_1! k_2! k_3!}$. This can be generalized to any number of terms to give what is known as a multinomial series.

12.3 Negative Binomial Series

$$(a + b)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} a^k b^{-n-k}$$

Theorem 12.3.1 Alternate Form for Negative Binomial Series. $(a + b)^{-n} = \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} a^k b^{-n-k}$