tablefileext=lot,placement=H,within=section,name=Table

Introduction to Mathematical Probability and Statistics

A Calculus-based Approach

Introduction to Mathematical Probability and Statistics

A Calculus-based Approach

John Travis
Mississippi College

January 4, 2017

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 and Chair of 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 opensouce online homework system WeBWorK.

Preface

This text is intended for a one-semester calculus-based undergraduate course in probability and statistics .

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.

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).

John Travis
Clinton, Mississippi 2015

Contents

Chapter 1

Statistical Measures

1.1 Introduction

To compute your final grade in a class your teacher will first make several assignments and examinations for you to take. These assessments ultimately will be assigned some numerical score indicating your level of success. However, you final grade can only be one value and it would make sense that the grade be a reflection of your work on these tasks. But, it can only be a single value that represents the totality of your work in the course.

In this chapter, you will consider a number of ways to use point values to represent different aspects of a provided set of data. In doing so, you will also 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 from that population–that is, a sample. Each of these values will be called a "statistical measure".

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 class grades which are used to compute your GPA.
- 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.

The following categorizes several statistical measures which will be developed in this chapter. Details on each are provided in subsequent sections.

- Tabular Methods - based on the entire population yielding a global picture
  - frequency distributions
  - relative frequency distributions
  - cummulative frequency distributions
  - Stem-and-Leaf Displays
  - Box-and-Whisker Diagrams
- Summary Methods
  - Measures of the center
    1. Mean
    2. Median
    3. Mode
  - Measures of spread
    1. Range
    2. Variance and Standard Deviation
    3. Interquartile Range
    4. Quantiles
  - Measures of Skewness - indicates the level of symmetry of the data
    1. Standard Skewness
    2. Pearson Coefficient
  - Measures of Kurtosis - indicates flatness or roundedness of the peak of the data
    1. Standard Kurtosis
    2. Coefficient of Kurtosis
  - Detection of Outliers - indicates whether abnormally large or small data distorts other techniques
    1. Z-scores
    2. Trimming
    3. Winsorizing
  - Tests for Normality - indictes if the data is bell-shaped
    1. Standard Percentages relative to standard deviations from the mean

Remark: Many of these measures above are relative and some are absolute.

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 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 but now in order with the subscript indicating the positional placement.

Definition 1.3.1 (Order Statistic)**.** From the data set $x_1, x_2, ..., x_n$, label the sorted data as $y_1, y_2, ..., y_n$ where

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

Define y_k as the kth order statistic.

Example 1.3.2 (Age of Presidents - order statistics)**.** For example, the age at inauguration for presidents from 1981-2016 gives the data $x_1 = 69, x_2 = 64, x_3 = 46, x_4 = 54, x_5 = 47$ (Reagan, Bush, Clinton, Bush, Obama). For this data, the order statistics are denoted $y_1 = 46, y_2 = 47, y_3 = 54, y_4 = 64, y_5 = 69$.

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:)**.** The smallest and largest values in the data set. Using the notation above, minimum = y_1 and the maximum = y_n

Example 1.3.4 (Age of Presidents - Minimum/Maximum)**.** Using the Presidential ages above, minimum = $y_1 = 46$ and maximum = $y_5 = 69$.

Below, you will see how to determine a value which separates the ordered data into two groups with a desired percentage on each side. Note that many utilize approximate methods for computing "percentiles" including many graphing and scientific calculators. The definition below provides for a unique measure for each unique value of p.

Definition 1.3.5 (Percentiles)**.** A percentile is a numerical value P^p at which approximately 100p

To compute the percentile value with $0 < p < 1$ for order statistics $y_1, y_2, ..., y_n$ use the formula

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

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

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

and

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

the fractional part of (n+1)p.

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 then be the common value.

Example 1.3.6 (Small Example - Quartiles)**.** Consider the following data set: 2,5,8,10. The 50th percentile should be a numerical value for which approximately 50

More precisely , 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.

Some special percentiles are provided special names...

Definition 1.3.7 (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 $Q_3 = P^{0.75}$.

Definition 1.3.8 (Deciles:)**.** Given a sorted data set, the first, second, ..., ninth deciles are the value of $D_1 = P^{0.1}, D_2 = P^{0.2}, ..., D_9 = P^{0.9}$

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 1.3.9 (5-number summary)**.** Given a set of data, the 5-number summary is a vector of the order statistics given by < minimum, Q_1, Q_2, Q_3, maximum >.

Example 1.3.10 (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 $R = x_1, x_2, ..., x_n$. The arithmetic mean is given by

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

If this data comes from sample data then we call it a sample mean and denote this value by $\overline{x}$. 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 $\overline{x}$.

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

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

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 masses would balance.

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

$$\frac{m_1 x_1 + ... + m_n x_n}{m_1 + ... + m_n} = \frac{\sum_{k=1}^{n} m_k x_k}{\sum_{k=1}^{n} m_k}.$$

Definition 1.4.2 (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.

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 date. For data sets with an even number of values, this is a average of the two data values left after crossing off these 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, note that you are considering an odd number of data values and so the median is given by 54.

Definition 1.4.3 (Midrange:)**.** 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

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

From the Presidential data, the maximum is 69 and the minimum is 46 so the midrange is 57.5, the average of these two.

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. Median ignores outliers (which might be a result of collection errors) but does not account for the relative differences between terms. Midrange is very easy to compute but ignores the relative differences for all terms but the two extremes.

Example 1.4.4 (Numerical Example of these Quantitative Measures)**.** The US Census Bureau reported the following state populations (in millions) for 2013: Spreadsheet

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

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

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

Note, 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

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

million residents.

Since the number of states is odd, the median is found by looking at the 26th order statistics. In this case, that is the 4.6 million residents of Louisiana.

1.5 Statistical Measures of Variation

These measures provide some indication of how much the data set is "spread out".

Definition 1.5.1 (Range:)**.** Using the order statistics,

$$y_n - y_1.$$

Easy to compute. Ignores the spread of all the data in between.

From the Presidential data, 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):)**.** $P^{0.75} - P^{0.25}$.

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 $x_1, x_2, ..., x_n$ with mean μ each term deviates from the mean by the value $x_k - \mu$. So, averaging these gives

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

which is always zero for any provided set of data. This cancellation makes this measure not useful. To avoid cancellation, perhaps removing negatives would help.

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 roadblock, 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 values added are non-negative but very small ones have been

made even smaller and larger ones have possibly been made much larger. To undo this scaling issue, one must take a square root to get things back into the right ball park.

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 s^2. Therefore

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

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

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

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

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

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

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

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

Theorem 1.5.3 (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. □

The Population of the individual USA states according to the 2013 Census Consider the data set

Exercise 1.5.4 (Numerical Example of these Quantitative Measures)**.**

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 and perhaps expressed as a frequency distribution. In this case, rather than considering x_k to be the kth 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 $x_1, x_2, ..., x_m$ with corresponding frequencies $f_1, f_2, ..., f_m$. 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

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

and the following grouped data formula for the variance

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

Exercise 1.6.1.

1.7 Other Statistical Point Measures

Beyond measures of the middle and of spread includes a way you can determine if data is heaped up to one side or the other of the mean. One such measure is the skewness.

Definition 1.7.1 (Skewness)**.** For sample data, the Skewness of $x_1, x_2, ..., x_n$ is given by

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

and similarly for population data but using μ, σ.

A positive skewness indicates that the positive $(x_k - \overline{x})^3$ terms overwhelm the negative terms. Therefore, this indicates data which is strung out to the right. Likewise, a negative skewness indicates data which is strung out to the left.

In addition to skewness, data might tend to be clustered around the mean and often in a "bell-shaped" manner. The kurtosis can be used o measure how closely data resembles a bell-shaped collection.

Definition 1.7.2 (Kurtosis)**.** The Kurtosis of $x_1, x_2, ..., x_n$ is given by

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

and similarly for population data again using μ, σ.

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)**.** *Skewness* =

$$\frac{1}{s^3} \left[\frac{\sum_{k=1}^{n} x_k^3}{n} - 3v\overline{x} - \overline{x}^3 \right]$$

and Kurtosis =

$$\frac{1}{s^4} \left[\frac{\sum_{k=1}^{n} x_k^4}{n} - 4\overline{x} \frac{\sum_{k=1}^{n} x_k^3}{n} + 6\overline{x}^2 v - 3\overline{x}^4 \right]$$

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

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

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,

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

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. □

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.

Also, probability mass functions for discrete variables can be graphed as a set of points but sometimes these points do not convey size very well. A visual representation of these functions needs to be addressed.

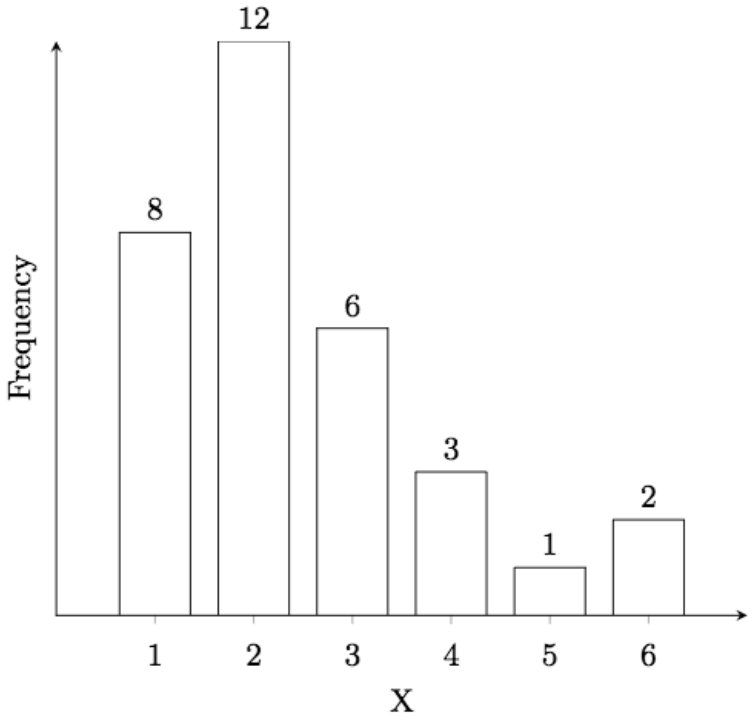
1.8.1 Histograms

Frequency Histograms - height matters

Consider the data set given by

A frequency histogram representing this data can be given by

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



Experiment with creating your own histogram by inputting your data into the interactive 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 your data. 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)
```

Stem-and-Leaf Plot - Histogram with data. 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

Table 1: Stem Plot for State Populations

Stem	Leaf
0L	06 06 07 07 08 09 10 11 13 13 14 16 19 19 21 28 29 29 30 30 31 36 39 39 44 46 48 48
0H	53 54 57 59 60 65 66 66 67 70 83 89 98 99
1L	10 16 28 29
1H	96 97
2L	
2H	64
3L	
3H	83

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.

Box and Whisker Diagram - visual order statistics. 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.

```
from pylab import boxplot,savefig,close
@interact
def _(data =
    input_box([1,2,3,4,6,7,8,9,11,15,21],label="Enter␣
    Your␣Data:")):
    B = boxplot(data, notch=True, sym='x', vert=False)
    savefig("boxplot.png")
    close()
```

1.9 Exercises

Complete the online homework "Computational Measures".

Exercise 1.9.1.

Exercise 1.9.2.

Chapter 2

Regression

2.1 Introduction

Given two distinct points, there is one line which passes exactly through both. Indeed, if the points are $(x_0, y_0), (x_0, y_0)$ then presuming the x-values are different gives the equation

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

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

$$x = x_0$$

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

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

In this section, we will presume only one independent variable x and one dependent variable y.

Consider a collection of data points

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

and a general linear function

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

It is possible that each of the given data points are exactly "interpolated" by the linear function so that

$$f(x_k) = y_k$$

for k = 1, 2, ... , n. However, in general this is unlikely since even three points are not likely to be colinear. However, you may notice that the data points exhibit a linear tendency or that the underlying physics might suggest a linear model. If so, you may find it easier to predict values of y for given values of x using a linear approximation. Here you will investigate a method for doing so called "linear regression", "least-squares", or "best-fit line".

To determine a best-fit line, you need to determine what is meant by the word "best". Here, we will derive the standard approach which interprets this to mean that the total vertical error between the line and the provided data points is minimized in some fashion. 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, you can look at taking absolute values (which is tough to deal with algebraically) or by squaring the errors. This second option will be the approach taken here. This 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

$$TSE(m, b) = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (f(x_k) - y_k)^2 = \sum_{k=0}^{n} (mx_k + b - y_k)^2.$$

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:

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

and

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

Setting equal to zero and solving gives what is known as the "normal equations":

$$m \sum_{k=1}^{n} x_k^2 + b \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} x_k y_k$$

and

$$m \sum_{k=1}^{n} x_k + b \sum_{k=1}^{n} 1 = \sum_{k=1}^{n} y_k.$$

Solving these for m and b gives the best fit line.

2.3 Higher Degree 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

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

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

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

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

$$a\sum_{k=1}^{n} x_k^4 + b\sum_{k=1}^{n} x_k^3 + c\sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k^2 y_k$$

$$a\sum_{k=1}^{n} x_k^3 + b\sum_{k=1}^{n} x_k^2 + c\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} x_k y_k$$

$$a\sum_{k=1}^{n} x_k^2 + b\sum_{k=1}^{n} x_k + c\sum_{k=1}^{n} 1 = \sum_{k=1}^{n} y_k.$$

Chapter 3

Counting and Combinatorics

3.1 Introduction

Discussion on the usefulness of having ways to count the number of elements in a set without having to explicitly listing all elements.

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 adding one more choice. This illustrates how keeping track of the number of items in a set can quickly get impossible to keep up with and to count unless we can approach this problem using a more mathematical approach.

Definition 3.1.1 (Cardinality)**.** Given a set of elements A, the number of elements in the set is known as the sets cardinality and is denoted |A|. If the set has an infinite number of elements then we set $|\mathrm{A}| = \infty$.

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

Theorem 3.1.2 (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

$$A = \left\{a_1, a_2, a_3, ..., a_{|A|}\right\}$$
$$B = \left\{b_1, b_2, b_3, ..., b_{|B|}\right\}$$

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| |B| elements then the conclusion follows. coordinates. □

Definition 3.1.3 (Factorial)**.** For any natural number n,

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

and by convention set $0! = 1$.

Example 3.1.4 (iPad security code)**.** Consider your ipad's security. To unlock the screen you need to enter your 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, the order does matter and you cannot reuse digits. In this case, 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, $10 \times 9 \times 8 \times 7 = 5040$.

Note that 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, $10 \times 10 \times 10 \times 10 = 10000$.

Example 3.1.5 (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 $4 \times 3 \times 2 \times 1 = 24$

Example 3.1.6 (National Treasure)**.** In the 2004 movie "National Treasure" Ben and Riley are attempting to guess Abagail'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 9×8 *times*... $\times 1 =$ 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 This is the answer 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?

3.2 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 3.2.1 (Permutations of everything)**.** *The number of ways to arrange 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 arrangment.

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

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

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. □

Theorem 3.2.2 (Permutations of a subset without replacement)**.** *The number of ways to arrange r items from a set of n distinct items is*

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

Proof. If $r > n$ or $r < 0$ then this is not possible and so the result would be no permuatations. 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 rth selection. This gives

$$\begin{aligned} P_r^n &= n(n-1)...(n-r+1) \\ &= n(n-1)...(n-r+1)\frac{(n-r)!}{(n-r)!} \\ &= \frac{n(n-1)...(n-r+1)(n-r)!}{(n-r)!} \\ &= \frac{n!}{(n-r)!} \end{aligned}$$

□

Theorem 3.2.3 (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,

$$n \cdot n \cdot n...n = n^r$$

□

Theorem 3.2.4 (Permutations when not all items are distinguishable (Multinomial Coefficients))**.** *If n items belong to s categories, n_1 in first, n_2 in second, ... , n_s in the last, the number of ways to pick all is*

$$\frac{n!}{n_1! \cdot n_2!...n_2!}$$

3.3 Combinations

When counting various outcomes sometimes the order of things does not matter. In this case we count each different set of outcomes a combination.

Theorem 3.3.1 (Combinations of a subset without replacement)**.** *The number of ways to arrange r items from a set of n distinct items is*

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

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,

$$P_r^n = C_r^n \cdot r!$$

Solving give the result. □

Theorem 3.3.2 (Combinations of a subset with replacement)**.** *The number of ways to arrange r items from a set of n distinct items is*

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

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 and the number of ways to rearrange this code is

$$\frac{(r+n-1)}{r!(n-1)!}.$$

□

Example 3.3.3 (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

$$\begin{aligned}\frac{13!}{3!10!} &= \frac{13\times 12\times 11}{3\times 2\times 1}\\ &= 286\end{aligned}$$

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

Definition 3.3.4 (Binomial Coefficients)**.** The value C_r^n is known as the binomial coefficient. It is denoted by $\binom{n}{r}$ and is read "n choose k".

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.

Theorem 3.3.5 (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. $$\begin{aligned}\binom{n}{r} + \binom{n}{r+1} &= \frac{n!}{r!(n-r)!} + \frac{n!}{(r+1)!(n-(r+1))!} \\ &= (r+1)\frac{n!}{(r+1)!(n-r)!} + (n-r)\frac{n!}{(r+1)!(n-r))!} \\ &= \frac{(r+1)n! + (n-r)n!}{(r+1)!(n-r)!} \\ &= \frac{(n+1)n!}{(r+1)!((n+1)-(r+1))!} \\ &= \binom{n+1}{r+1}\end{aligned}$$

□

3.4 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 for 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

This chapter uses relative frequency to motivate the definition of probability and then delves into the resulting consequences.

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.

People often make claims about being the biggest, best, most often recommended, etc. One sometimes even believes these claims. In this class, we will attempt to determine if such claims are reasonable by first introducing probability from a semi rigorous mathematical viewpoint using concepts developed in Calculus. We will use this framework to carefully discuss making such statistical inferences as above and in general to obtain accurate knowledge even when the known data is not complete.

4.2 Relative Frequency

When attempting to precisely measure uncertainty a few experiments are in order. When doing statistical experiments, a few terms and corresponding notation might be useful:

- S = Universal Set or Sample Space Experiment or Outcome Space. This is the collection of all possible outcomes.
- Random Experiment. A random experiment is a repeatable activity which 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 which 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 = 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.

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

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

```
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
```

Notice 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.

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?

```
@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 in this instance that 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␣Frequence␣of␣",dice[0],"␣is␣about␣
        ",rel_freq[0].n(digits=4)
    print "Relative␣Frequence␣of␣",dice[num_sides-1],"␣
        is␣about␣",rel_freq[num_sides-1].n(digits=4)
```

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.

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 leads to what is known as the "Law of Large Numbers". We would like to formalize what these relative frequencies seem to be 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.

4.3.1 Motivating the Definition

Using the ideas from our examples above, consider how you might formally define a way to measure the expectation from similar experiments. Before doing so, we need a little notation:

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

To model the behavior above, consider how we might create a definition for our expectation of a given outcome by following the ideas uncovered above. To do so, first consider a desired collection of outcomes A. If each outcome in A is equally likely then we might follow the concept behind relative frequency and consider a measure of expectation be |A|/|S|. Indeed, on a standard 6-sided die, the expectation of the outcome A=2 from the collection S = 1,2,3,4,5,6 should be |A|/|S| = 1/6.

From the example where we take the sum of two die, the outcome A=4,5 from the collection S = 2,3,4,...,12 would be

$$|A| = |(1,3),(2,2),(3,1),(1,4),(2,3),(3,2),(4,1)| = 7$$
$$|S| = |(1,1),...,(1,6),(2,1),...,(2,6),...,(6,1),...,(6,6)| = 36$$

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 the ideas

related to equally likely outcomes, we make our definition have the following basic 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

4.3.2 Probability

Based upon these we give the following:

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

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

4.3.3 Basic Probability Theorems

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

Theorem 4.3.3 (Probability of Complements)**.** *For any event A,* $P(A) + P(A^c) = 1$

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 4.3.4. $P(\emptyset) = 0$

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 4.3.5. *For events A and B with* $A \subset B, P(A) \leq P(B)$.

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

$$0 \leq P(B - A)$$
$$P(A) \leq P(A) + P(B - A)$$
$$P(A) \leq P(B)$$

□

Theorem 4.3.6. *For any event A,* $P(A) \leq 1$

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

Theorem 4.3.7. *For any sets A and B,* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

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

Hence,

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

□

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

Corollary 4.3.8. *For any sets A, B and C,*

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

Corollary 4.3.9. *For any sets A, B, C and D,*

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

4.3.4 Equally Likely Outcomes

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 is 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 4.3.10 (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 = $x_1, x_2, ..., x_{|S|}$ 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,

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

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

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

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

as desired. □

4.3.5 HOMEWORK

A. Determine the probabilities associated with the various 5-card hands.

B. 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.

C. 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.

D. Analyze the game of "craps".

4.4 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 than it 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 occurence or not of some other event. 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(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.

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 then P(Heart) = 13/52. Now, assume that one card has already been extracted and setaside. Now, prepare to extract another. If the first card drawn was a Heart, then there are only 12 Hearts left

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

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.

To do this, consider P(Heart on 2nd draw) = P([Heart on 1st draw ∩ Heart on 2nd draw] ∪ [Not Heart on 1st draw ∩ Heart on 2nd draw]) = P(Heart on 1st draw ∩ Heart on 2nd draw) + P(Not Heart on 1st draw ∩ Heart on 2nd draw) = | Heart on 1st draw ∩ Heart on 2nd draw | / | Number of ways to get two cards | + | Not Heart on 1st draw ∩ Heart on 2nd draw / | Number of ways to get two cards | = (13 12) / (52 51) + (39 13) / (52 51) = 12 / (4 51) + (3 13) / (4 51) =

Definition 4.4.1 (Conditional Probability)**.** P(B | A) = P(A ∩ B) / P(A), provided P(A)> 0.

Theorem 4.4.2. *Conditional Probability satisfies all of the requirements of regular probability.*

Proof. By definition, for any event probability must be nonnegative. Therefore $P(A \cap B) \geq 0$. Therefore, P(B | A) ≥ 0.

Further, P (S | A) = P(A ∩ S)/P(A) = P(A)/P(A) = 1. □

Theorem 4.4.3 (Multiplication Rule)**.**

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

Proof. Unravel the definition of conditional probably by taking the denominator to the other side. Also note that you can write $A \cap B = B \cap A$. □

4.4.1 HOMEWORK

A. 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. $P(A^c \cap B^c)$

B. The table below classifies students at your university according to gender and according to major.

Determine the following:

1. P(STEM major)
2. P(STEM | Female)
3. P(Female | STEM)

4. P(Female | Not STEM)

C. 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 slop.

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.

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 kth pick gives

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

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 in 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.

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

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

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

Adding these two together gives

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

For example, if k = 5,

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

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

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.

D. 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.

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
```

E. 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.

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.

F. 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?

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 kth person sits in their own set as f(k). Consider the first traveler's and your seats. Then you get the following cases:

- P(first guy sits in his own seat and you sit in yours) = $1\frac{}{k \cdot 1 P(first guy sits in your seat and you do not sit in yours) = \frac{1}{k} \cdot 0}$
- P(other k-2 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 1/2 = 1/2. f(5) = 1/5 + 3/5 1/2 = 1/2. f(6) = 1/6 + 4/6 1/2 = 1/2. Etc.

4.5 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 the following.

Theorem 4.5.1 (Bayes Theorem)**.** *Let $S = \{S_1, S_2, ..., S_m\}$ where the S_k are pairwise disjoint and $S_1 \cup S_2 \cup ... \cup S_m = S$ (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

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

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:

Exercise 4.5.2.

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="$P(B_1),P(B_2
        Conditional_Probabilities=input_box('0.02,0.01,0.03',label='$P(A|B_1)
        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_Bk_given_A = []                   # P( B_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_Bk_given_A.append(PP[k]*CP[k] )
            pA += p_Bk_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('$B_'+str(k+1)+'$',(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("$P(B_{%s}␣|␣A)$"%(str(k+1))+"$␣
                        =␣%s$"%str(p_Bk_given_A[k]/pA))

                    G +=
                        text(str(p_Bk_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_Bk_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_Bk_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=

            for k in range(n):
                H += disk((0,0), 0.6, (ends[k]*2*pi,
```

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

```
                    ends[k+1]*2*pi),
                    color=colors[mod(k,10)],alpha = 0.2)
                H +=
                    text('$B_'+str(k+1)+'|A$',(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_Bk_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')
```

4.5.1 HOMEWORK

A. 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:

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.

B. 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.

4.6 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(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 4.6.1 (Independent Events)**.** Events A and B are independent provided

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

Corollary 4.6.2 (Independence and Conditional Probability)**.** *Given independent events 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. □

Corollary 4.6.3 (Independence and Mutual Exclusivity)**.** *If events A and B are both independent and mutually exclusive, 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 4.6.4 (Successive Independent Events)**.** *Given a sequence of independent events $A_1, A_2, A_3, ...,$*

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

4.6.1 HOMEWORK

A. Given P(A) = 0.43, P(B) = 0.72, and $P(A \cap B) = 0.29$, verify that A and B are not independent.

B. 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. $P(A \cup B \cup C)$

C. Suppose for a pair of dice you want to consider the events A = rolling a 7 or 11 and B = otherwise. Rolling the dice 5 times, determine

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

D. 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

E. 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.

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 it's value, does not affect the other. The question in this case does not ask for a conditional probability.

F. 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 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?

P(4 wins) = P(4 beats 1) P(4 beats the winner of the other bracket)
P(4 wins) = (1/5) * P(4 beats 2 | 2 beats 3) + P(4 beats 3 | 3 beats 2)
P(4 wins) = 1/5 [(3/5)(2/6) + (2/5)(3/7)] = 78/1050 = 0.0742
For the other teams:
P(1 wins) = 4/5 [(3/5)(2/3) + (2/5)(3/4)] = 0.56
P(2 wins) = 3/5 [(4/5)(1/3) + (1/5)(4/6)] = 0.24
P(3 wins) = 2/5 [(4/5)(1/4) + (1/5)(4/7)] = 0.1257

Chapter 5

Probability Functions

5.1 Introduction

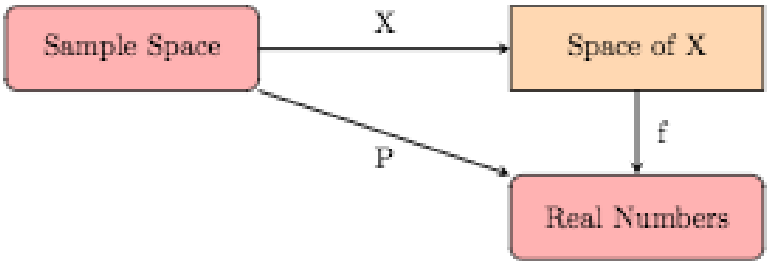
Each of the probability exercises thus far required you to utilize basic definitions and theorems to determine the answer. Starting a new problem meant starting over from scratch. This is burdensome. However, you may have noticed that some of the ways you might have created solutions for some problems 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 call the range of X the space of X: R= x : X(s)=x, 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

$$X(success) = 1$$
$$X(failure) = 0.$$

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

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

Choose

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

Then, R=2, 3, 4, 5, ..., 12.

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

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

Choose

$$X((a,b)) = \text{maxa,b.}$$

Then, R=1, 2, 3, 4, 5, 6.

Definition 5.2.5. 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.

5.2.1 HOMEWORK

A. 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.

B. 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.

C. 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.

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(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.

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 $f : R \to \mathbb{R}$ such that:

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

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 $f : R \to \mathbb{R}$ such that:

$$\forall x \in R, f(x) > 0$$
$$\int_R f(x)dx = 1$$
$$A \subset R \Rightarrow P(X \in A) = \int_A f(x)dx$$

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],labe
            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] #
```

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

```
    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',thi
(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 = 3.

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 $F : \mathbb{R} \to \mathbb{R}$ such that $F(x) = P(X \leq x)$

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

Example 5.3.7 (Continuous Distribution Function)**.** Consider $f(x) = x^2/3$ over R = [-1,2]. Then, for $-1 \leq x \leq 2$,

$$F(x) = \int_{-1}^{x} u^2/3du = x^3/9 + 1/9.$$

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:

Theorem 5.3.8. $F(x) = 0, \forall x < \inf(R)$

X	F(x)
$x < -1$	0
$-1 \leq x < 2$	$x^3/9 + 1/9$
$2 \leq x$	1

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 5.3.9. $F(x) = 1, \forall x \geq \sup(R)$

Proof. Let b = sup(R). Then, for

$$x \geq b, F(x) = P(X \leq x) = P(X \leq b) + P(b < Xlex) = P(Xleb) = 1$$

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 5.3.10. *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

$$\begin{aligned} \forall x_1, x_2 \in \mathbb{R} \ni x_1 < x_2 \\ F(x_2) &= \int_{-\infty}^{x_2} f(x)dx \\ &= \int_{-\infty}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx \\ &\geq \int_{-\infty}^{x_1} f(x)dx \\ &= F(x_1) \end{aligned}$$

□

Theorem 5.3.11 (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,

$$F(x) - F(x-1) = \sum_{u \leq x} f(u) - \sum_{u < x} f(u) = f(x)$$

□

Theorem 5.3.12 (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

$$\begin{aligned} F(b) - F(a) &= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx \\ &= \int_{a}^{b} f(x)dx \\ &= P(a < x \leq b) \end{aligned}$$

□

Corollary 5.3.13. *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 $\epsilon \to 0^+$ to get the result noting that □

Theorem 5.3.14 (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, $\lim_{x\to\pm\infty} f(x) = 0$ 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,

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)du \\ &= A(x) - \lim_{u\to-\infty} A(u) \end{aligned}$$

Hence, $F'(x) = A'(x) - \lim_{u\to-\infty} A'(u) = f(x)$ as desired. □

5.3.1 HOMEWORK

A. 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.

B.

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" u(x), the expected value of u(x) is given by

$$E = E[u(X)] = \sum_{x \in R} u(x)f(x)$$

provided X is discrete, or

$$E = E[u(X)] = \int_R u(x)f(x)dx$$

provided X is continuous.

Theorem 5.4.2 (Expected Value is a Linear Operator)**.**

1. *E[c] = c*
2. *E[c u(X)] = c E[u(X)]*
3. *E[u(X) + v(X)] = E[u(X)] + E[v(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:

$$\begin{aligned} E[u(X) + v(X)] &= \int_{x \in R} [u(x) + v(x)]f(x)dx \\ &= \int_{x \in R} u(x)f(x)dx + \int_{x \in R} u(x)f(x)dx \\ &= E[u(X)] + E[v(X)]. \end{aligned}$$

□

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 u(1)=10, u(2) = 5, u(3)=2, and u(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.

Example 5.4.4 (Continuous Expected Value)**.** Consider $f(x) = x^2/3$ over R = [-1,2] with value function given by $u(x) = e^x - 1$. Then, the expected value for u(x) is given by

$$E = \int_{-1}^{2} (e^x - 1) \cdot x^2/3 = -1/9 \cdot (e + 15) \cdot e^{-1} + 2/3 \cdot e^2 - 8/9 \approx 3.3129$$

X	f(x)
0	0.10
1	0.25
2	0.40
4	0.15
7	0.10

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 $= \gamma_2 = \frac{E[(x-\mu)^4]}{\sigma^4}$

Theorem 5.4.6 (Alternate Formulas for Theoretical Measures)**.**

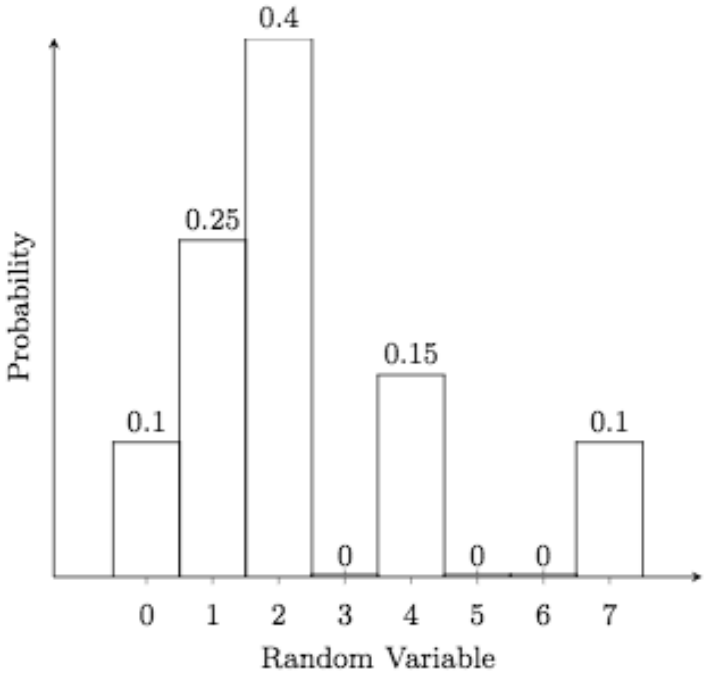
1. $\sigma^2 = E[x^2] - \mu^2 = E[X(x-1)] + \mu - \mu^2$

2. $\gamma_1 = \frac{1}{\sigma^3} \cdot \left[E[X^3] - 3\mu E[X^2] + 2\mu^3\right]$

3. $\gamma_2 = \frac{1}{\sigma^4} \cdot \left[E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4\right]$

Proof. In each case, expand the binomial inside and use the linearity of expected value. □

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



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

$$\begin{aligned}\sigma^2 &= E[X^2] - \mu^2 \\ &= \left[0^2 \cdot 0.10 + 1^2 \cdot 0.25 + 2^2 \cdot 0.40 + 4^2 \cdot 0.15 + 7^2 \cdot 0.10\right] - 2.35^2 \\ &= 0 + 0.25 + 1.60 + 2.40 + 4.90 - 5.5225 \\ &= 9.15 - 5.225 \\ &= 3.6275\end{aligned}$$

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

$$\begin{aligned}\text{Numerator} = E[X^3] - 3\mu E[X^2] + 2\mu^3 \\ = \left[0^3 \cdot 0.10 + 1^3 \cdot 0.25 + 2^3 \cdot 0.40 + 4^3 \cdot 0.15 + 7^3 \cdot 0.10\right] - 3 \cdot 2.35 \cdot 9.15 + 2 \cdot 2.35^3 \\ \approx 0 + 0.25 + 3.20 + 9.60 + 34.3 - 64.5075 + 25.96 \\ = 47.35 - 64.5075 + 25.96 \\ \approx 8.80\end{aligned}$$

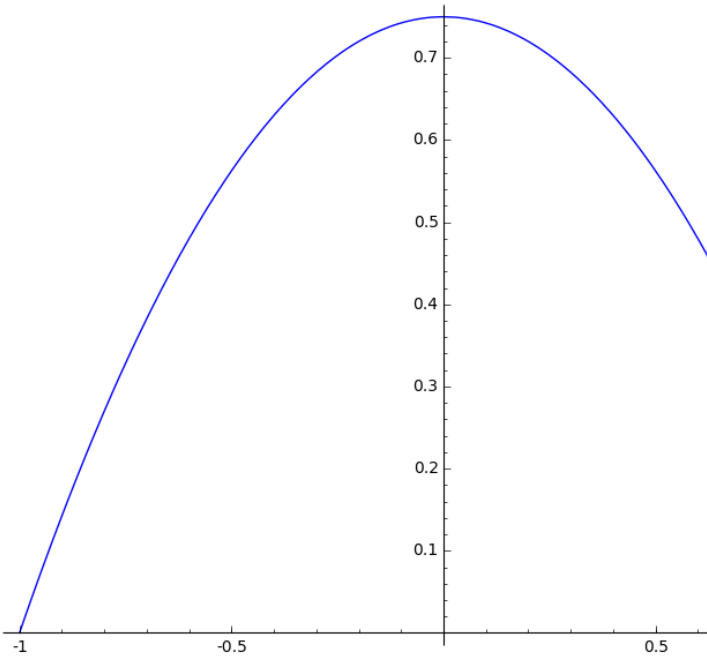
which yields a skewness of $\gamma_1 = 8.80/\sigma^3 \approx 1.27$. This indicates a slight skewness to the right of the mean. You can notice the 4 and 7 entries on the histogram illustrate a slight trailing off to the right.

Finally, for kurtosis

$$\begin{aligned}\text{Numerator} &= E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4 \\ &= \left[0^4 \cdot 0.10 + 1^4 \cdot 0.25 + 2^4 \cdot 0.40 + 4^4 \cdot 0.15 + 7^4 \cdot 0.10\right] - 4 \cdot 2.35 \cdot 47.35 + 6 \cdot 2.35^2 \cdot 9.15^2 - 3 \cdot 2.35^4 \\ &\approx 0 + 0.25 + 6.40 + 38.4 + 240.1 - 445.09 + 303.19 - 91.49 \\ &\approx 285.15 - 445.09 + 303.19 - 91.49 \\ &\approx 51.75\end{aligned}$$

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

$$\begin{aligned}\sigma^2 &= \int_{-1}^{1} x^2 \cdot \frac{3}{4} \cdot (1-x^2)dx - \mu^2 \\ &= \int_{-1}^{1} \cdot \frac{3}{4} \cdot (x^2 - x^4)dx - 0 \\ &= \frac{3}{4} \cdot (x^3/3 - x^5/5)\Big|_{-1}^{1} \\ &= \frac{3}{4} \cdot 2 \cdot (1/3 - 1/5) \\ &= \frac{3}{4} \cdot \frac{4}{15} \\ &= \frac{1}{5}\end{aligned}$$

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

$$\begin{aligned}\text{Numerator} &= E[X^3] - 3\mu E[X^2] + 2\mu^3 \\ &= \int_{-1}^{1} x^3 \cdot \frac{3}{4} \cdot (1-x^2)dx - 3E[X^2] \cdot 0 + 0^3 \\ &= \int_{-1}^{1} \cdot \frac{3}{4} \cdot (x^3 - x^5)dx \\ &= \frac{3}{4} \cdot (x^4/4 - x^6/6)\Big|_{-1}^{1} \\ &= 0\end{aligned}$$

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

$$\begin{aligned}\text{Numerator} &= E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4 \\ &= \int_{-1}^{1} x^4 \cdot \frac{3}{4} \cdot (1-x^2)dx \\ &= \frac{3}{4} \cdot (x^5/5 - x^7/7)\Big|_{-1}^{1} \\ &= \frac{3}{4} \cdot 2(1/5 - 1/7) \\ &= \frac{3}{35}\end{aligned}$$

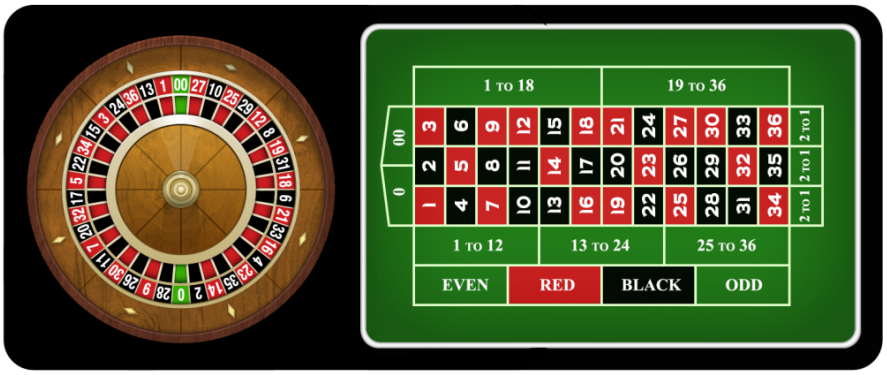
and so by dividing by $\sigma^4 = \sqrt{\frac{1}{5}}^4 = \frac{1}{25}$ gives a kurtosis of

$$\gamma_2 = \frac{3}{35} \Big/ \frac{1}{25} = \frac{75}{35} \approx 2.14.$$

Example 5.4.7 (Roulette)**.** Roulette is a gambling game popular in may 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 us

$$E[\text{win on top line of basket}] = 6 \cdot \frac{5}{38} - 1 \cdot \frac{33}{38} = -\frac{3}{38}$$

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.

Going back to Pascal's wager, let X = 0 represent disbelief when God doesn't exist and X = 1 represent disbelief when God does exist, X = 2 represent belief when God does exist, and X = 3 represent belief when God does not exist. Let p be the likelihood that God exists. Then you can compute the expected value of disbelief and the expect 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...

$$\begin{aligned} u(0) = 1,000,000, f(0) = 1 - p \\ u(1) = -\infty, f(1) = p \\ u(2) = \infty, f(2) = p \\ u(3) = 0, f(3) = 1 - p \end{aligned}$$

Then,

$$\begin{aligned} E[\text{disbelief}] &= u(0)f(0) + u(1)f(1) \\ &= 1000000 \times (1-p) - \infty \times p \\ &= -\infty \end{aligned}$$

if p>0. On the other hand,

$$\begin{aligned} E[\text{belief}] &= u(2)f(2) + u(3)f(3) \\ &= \infty \times p + 0 \times (1-p) \\ &= \infty \end{aligned}$$

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, then values of z will always represent the number of standard deviations x is from the mean and will provide "dimensionless" comparisons.

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

Assume that you have a variable with space R = 1, 2, 3, ..., 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

$$f(x) = \frac{1}{n}$$

is the probability function.

```
# 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:␣$f(x)␣
    =%s$'%str(latex(f(x)))+'␣over␣the␣space␣$R␣=␣
    %s$'%str(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 6.2.1 (Properties of the Discrete Uniform Probability Function)**.**

1. *$f(x) = \frac{1}{n}$ over R = 1, 2, 3, ..., n satisfies the properties of a discrete probability function*

2. *$\mu = \frac{1+n}{2}$*

3. *$\sigma^2 = \frac{n^2-1}{12}$*

4. *$\gamma_1 = 0$*

5. *$\gamma_2 = \frac{6}{5}\frac{1+n^2}{1-n^2}$*

6. *Distribution function F(x) = fracxn for $x \in R$.*

Proof.

1. Trivially, by construction you get

$$\sum_{k=1}^{n} \frac{1}{n} = 1$$

   Also, 1/n is positive for all x values.

2. To determine the mean,

$$\begin{aligned}
\mu &= \sum_{k=1}^{n} x \cdot \frac{1}{n} \\
&= \frac{1}{n} \sum_{k=1}^{n} x \\
&= \frac{1}{n} \frac{n(n+1)}{2} \\
&= \frac{1+n}{2}
\end{aligned}$$

3. To determine the variance,

$$\begin{aligned}\sigma^2 &= \sum_{k=1}^{n} x^2 \cdot \frac{1}{n} - \mu^2 \\ &= \frac{1}{n}\sum_{k=1}^{n} x^2 - \left(\frac{1+n}{2}\right)^2 \\ &= \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \frac{1+2n+n^2}{4} \\ &= \frac{(2n^2+3n+1)}{6} - \frac{1+2n+n^2}{4} \\ &= \frac{(4n^2+6n+2)}{12} - \frac{3+6n+3n^2}{12} \\ &= \frac{(n^2-1)}{12}\end{aligned}$$

4. For skewness,

$$\begin{aligned}\gamma_1 &= \sum_{k=1}^{n} x^3 \cdot \frac{1}{n} - 3\mu\sum_{k=1}^{n} x^2 \cdot \frac{1}{n} + 2\mu^3 \\ &= \frac{n^2(n+1)^2}{4n} - 3\frac{(n(n+1))}{2}\frac{1+n}{2} + 2\left(\frac{1+n}{2}\right)^3 \\ &=\end{aligned}$$

5. For Kurtosis, use the fourth moment and simplify...the algebra is performed using Sage in the active cell below this proof.

6.

□

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 6.2.2 (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.

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 $P(u < X < v) = P(u+w < X < v+w)$. In integral notation you obtain the following:

$$\int_u^v f(x)dx = \int_{u+w}^{v+w} f(x)dx$$
$$F(v) - F(u) = F(v+w) - F(u+w)$$
$$F(u+w) - F(u) = F(v+w) - F(v)$$

which is true regardless of w so long as you stay in the domain of interest. This only happens if F is linear and therefore f must be constant. Say, f(x)=c. In many situations, the space of X will be a single interval with $R = [a, b]$. Unless otherwise noted, this will be our assumption as well.

Theorem 6.3.1 (Properties of the Continuous Uniform Probability Function)**.**

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. $\gamma_2 = \frac{9\left(a^5-5\,a^4b+10\,a^3b^2-10\,a^2b^3+5\,ab^4-b^5\right)(a-b)}{5\,(a^3-3\,a^2b+3\,ab^2-b^3)^2}$

```
# 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 6.3.2 (Occurence 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 6.3.3 (Distribution Function for Continuous Uniform)**.** *For* $x \in [a, b], F(x) = \frac{x-a}{b-a}$

Proof. For x in this range,

$$F(x) = \int_a^x \frac{1}{b-a} du = \frac{u}{b-a}\Big|_a^x = \frac{x-a}{b-a}.$$

□

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 remainder are not, let the random variable X measure the number of items from the first group in your sample with $R = \{0, 1, ..., r\}$. The resulting collection of probabilities is called a Hypergeometric Distribution.

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, ..., 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 from the sample are from the target group and the remainder are not). Breaking these up gives

$$f(x) = \frac{\binom{n_1}{x}\binom{n-n_1}{r-x}}{\binom{n}{r}}$$

Theorem 6.4.1 (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. The proof of the variance formula is similar and uses E(X(X-1)+ - 2. 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.

Consider the Hypergeometric distribution for various values of n_1, n_2, and r using the interactive cell below. Notice what happens when you start with relatively small values of n_1, n_2, and r (say, start with $n_1 = 5, n_2 = 8$, and r = 4 and then doubling then all again and again. Consider the likely skewness and kurtosis of the graph as the values get larger.

```
# Hypergeometric distribution over 0 .. N
# Size of classes N1 and N2 must be given as well as
    subset size r
var('x')
@interact
def _(N1=slider(1,40,1,10,label='$N_1$'),
    N2=slider(1,40,1,10,label='$N_2$'),
    r=slider(1,40,1,10,label='$r$')):
    N = N1 + N1
    R = range(r+1)
    if (r > N1)|(r > N2):
        pretty_print('When␣r␣is␣bigger␣than␣N1␣or␣N2,␣
            special␣consideration␣must␣be␣made')
    else:
        f(x) =
            binomial(N1,x)*binomial(N2,r-x)/binomial(N,r)
        pretty_print(html('Density␣Function:␣$f(x)␣
            =%s$'%str(latex(f(x)))))
        pretty_print(html('over␣the␣space␣$R␣=␣
            %s$'%str(R)))
        points((k,f(x=k)) for k in R).show()
        for k in R:
           print (html('$f(%s'%k+')␣=␣
               %s'%latex(f(x=k))+'␣\\approx␣
               %s$'%f(x=k).n(digits=5)))
```

6.5 Exercises

Exercise 6.5.1 (- The Proverbial Urn Problem)**.**

Exercise 6.5.2 (- Playing Cards)**.**

Exercise 6.5.3 (- Starting Seniors)**.**

Exercise 6.5.4 (- Old Faithful)**.**

Exercise 6.5.5 (- Continuous Uniform Random Variable Scenarios)**.**

Exercise 6.5.6 (- Continous Uniform on a different space)**.**

Exercise 6.5.7 (- Louisiana Mega Millions Lottery)**.**

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(success) = 1 and X(failure)=0. If p = P(success) then

$$f(x) = p^x \cdot (1-p)^{1-x}$$

would be a formula but which only related to two values P(failure) = f(0) = (1-p) and P(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 $R = \{0, 1, ..., n\}$, 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)))
```

Theorem 7.2.1 (Derivation of Binomial Probability Function)**.** *For R = 0, 1, ..., 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

$$P(X = x) = \binom{n}{x} P(SS...SFF...F) = \binom{n}{x} p^x (1-p)^{n-x}$$

□

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

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1$$

□

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*

$$\mu = np$$
$$\sigma^2 = np(1-p)$$
$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$$
$$\gamma_2 = \frac{1-6p(1-p)}{np(1-p)} + 3$$

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

$$\begin{aligned}
&= np \sum_{k=0}^{m} \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k} \\
&= np(p+(1-p))^m = np
\end{aligned}$$

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

$$= n(n-1)p^2 \sum_{k=0}^{m} \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k} + np - n^2p^2$$
$$= n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p)$$

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 determine 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')
```

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.

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) + ... + 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) = binomcdf(n,p,x). In this case, F(10) = 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 probably 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 $R = \{1, 2, 3, ...\}$, then the resulting distribution of probabilities is called a Geometric Distribution.

Since successive trials are independent, then the probability of the first success occurring on the mth trial presumes that the previous m-1 trials were all failures. Therefore the desired probability is given by

$$f(x) = P(X = x) = P(FF...FS) = (1 - p)^{x-1}p$$

Theorem 7.3.1 (Geometric Distribution sums to 1)**.**

$$f(x) = (1 - p)^{x-1}p$$

sums to 1 over $R = \{1, 2, ...\}$

Proof.

$$\sum_{x=1}^{\infty} f(x) = \sum_{x=1}^{\infty} (1 - p)^{x-1}p = p\sum_{j=0}^{\infty} (1 - p)^j = p\frac{1}{1 - (1 - p)} = 1$$

□

```
# 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 7.3.2 (Geometric Mean)**.** *For the geometric distribution,*

$$\mu = 1/p$$

Proof.

$$\begin{aligned}\mu = E[X] &= \sum_{k=0}^{\infty} k(1-p)^{k-1}p \\ &= p\sum_{k=1}^{\infty} k(1-p)^{k-1} \\ &= p\frac{1}{(1-(1-p))^2} \\ &= p\frac{1}{p^2} = \frac{1}{p}\end{aligned}$$

□

Theorem 7.3.3 (Geometric Variance)**.** *For the geometric distribution*

$$\sigma^2 = \frac{1-p}{p^2}$$

Proof.

$$\begin{aligned}\sigma^2 &= E[X(X-1)] + \mu - \mu^2 \\ &= \sum_{k=0}^{\infty} k(k-1)(1-p)^{k-1}p + \mu - \mu^2 \\ &= (1-p)p\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} + \frac{1}{p} - \frac{1}{p^2} \\ &= (1-p)p\frac{2}{(1-(1-p))^3} + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}\end{aligned}$$

□

Theorem 7.3.4 (Geometric Distribution Function)**.**

$$F(x) = 1 - (1-p)^x$$

Proof. Consider the accumulated probabilities over a range of values...

$$\begin{aligned}P(X \le x) &= 1 - P(X > x) \\ &= 1 - \sum_{k=x+1}^{\infty} (1-p)^{k-1}p \\ &= 1 - p\frac{(1-p)^x}{1-(1-p)} \\ &= 1 - (1-p)^x\end{aligned}$$

□

Theorem 7.3.5 (Statistics for Geometric Distribution)**.** *Mean, Variance, Skewness, Kurtosis computed by Sage.*

```
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 probably 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 rth success, $r \geq 1$, with $R = \{r, r + 1, r+2, ...\}$ 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

$$\left(\frac{1}{1+x}\right)^n = \left(\frac{1}{1-(-x)}\right)^n$$
$$= \left(\sum_{k=0}^{\infty}(-1)^k x^k\right)^n$$

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 $x^0, x^1, x^2, ...$ To detemine the final coefficients notice that the number of time 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 $x^3, x^0, ..., x^0$ and $x^2, x^1, x^0, ..., x^0$ and $x^1, x^1, x^1, x^0, ..., x^0$. 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,...,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 idential |'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,

$$\left(\frac{1}{1+x}\right)^n = \sum_{k=0}^{\infty}(-1)^k\binom{n+k-1}{k}x^k$$

Similarly, $\left(\frac{1}{1-x}\right)^n = \left(\sum_{k=0}^{\infty}x^k\right)^n = \sum_{k=0}^{\infty}\binom{n+k-1}{k}x^k$ □

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 rth success. Then the resulting distribution of probabilities is called a Negative Binomial Distribution.

Theorem 7.4.2 (Derivation of Negative Binomial Probability Function)**.**

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r,$$

for $x \in R = \{r, r+1, ...\}$.

Proof. Since successive trials are independent, then the probability of the rth 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

$$P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$$

□

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 $k = x - r$

$$\begin{aligned} &= p^r \sum_{k=0}^{\infty} \binom{r+k-1}{k}(1-p)^k \\ &= p^r \frac{1}{(1-(1-p))^r} \\ &= 1 \end{aligned}$$

□

Utilize the interactive cell below to compute 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,' %.8f'%prob,' and
            F(%s) = '%k,' %.8f'%acc)
```

Theorem 7.4.4 (Statistics for Negative Binomial Distribution)**.** *For the Negative Binomial Distribution,*

$$\mu = \frac{r}{p}$$

$$\sigma^2 = r\frac{1-p}{p^2}$$

$$\gamma_1 = \frac{2-p}{\sqrt{r(1-p)}}$$

$$\gamma_2 = \frac{p^2-6p+6}{r(1-p)} + 3$$

```
# 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 Exercises

Exercise 7.5.1 (- Gallup Consumer Confidence Polling).

Exercise 7.5.2 (- Rolling Dice).

Exercise 7.5.3 (- Collecting Kids Meal Prizes).

Exercise 7.5.4 (- Rolling Dice).

Exercise 7.5.5 (- Rolling Dice yet again).

Exercise 7.5.6 (- 2 standard deviations from the mean).

Exercise 7.5.7 (Chapter Experiment).

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(one success) = λh.

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 $R = \{0, 1, 2, ...\}$.

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:

$$\begin{aligned}
f(x) &= P(\text{X changes in [0,T]}) \\
&= \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \lim_{n\to\infty} \binom{n}{x} (\frac{\lambda T}{n})^x (1-\lambda\frac{T}{n})^{n-x} \\
&= \lim_{n\to\infty} \frac{n(n-1)...(n-x+1)}{x!} (\frac{\lambda T}{n})^x (1-\frac{\lambda T}{n})^{n-x} \\
&= \frac{(\lambda T)^x}{x!} \lim_{n\to\infty} \frac{n(n-1)...(n-x+1)}{n\cdot n\cdot ...\cdot n} (1-\lambda\frac{T}{n})^n (1-\lambda\frac{T}{n})^{-x} \\
&= \frac{(\lambda T)^x}{x!} \lim_{n\to\infty} (1-\frac{1}{n})...(1-\frac{x-1}{n})(1-\frac{\lambda T}{n})^n (1-\frac{\lambda T}{n})^{-x} \\
&= \frac{(\lambda T)^x}{x!} \lim_{n\to\infty} (1-\frac{\lambda T}{n})^n \lim_{n\to\infty} (1-\frac{\lambda T}{n})^{-x} \\
&= \frac{(\lambda T)^x}{x!} \lim_{n\to\infty} (1-\frac{\lambda T}{n})^n \cdot 1 \\
&= \frac{(\lambda T)^x}{x!} e^{-\lambda T}
\end{aligned}$$

□

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,

$$\begin{aligned}
\sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} \frac{(\lambda T)^x}{x!} e^{-\lambda T} \\
&= e^{-\lambda T} \sum_{x=0}^{\infty} \frac{(\lambda T)^x}{x!} \\
&= e^{-\lambda T} e^{\lambda T} \\
&= 1
\end{aligned}$$

□

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.

$$\begin{aligned} &\int_0^\infty \frac{1}{\mu} e^{-\frac{x}{\mu}} dx \\ &= \int_0^\infty e^{-u} dx \\ &= -e^{-u}\Big|_0^\infty = 1 \end{aligned}$$

□

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)
```

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 rth 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 rth success

$$\begin{aligned} F(x) &= P(X \le x) \\ &= 1 - P(X > x) \\ &= 1 - P(\text{fewer than r successes in [0,x]}) \\ &= 1 - \big[\frac{(\lambda x)^0 e^{-\lambda x}}{0!} + \frac{(\lambda x)^1 e^{-\lambda x}}{1!} + ... + \frac{(\lambda x)^{r-1} e^{-\lambda x}}{(r-1)!}\big] \\ &= 1 - \sum_{k=0}^{r-1} \frac{(\lambda x)^k e^{-\lambda x}}{k!} \end{aligned}$$

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,

$$\begin{aligned}
F'(x) &= \lambda e^{-\lambda x}/0! \\
&\quad - \lambda e^{-\lambda x}/1! + \lambda x \cdot \lambda e^{-\lambda x}/1! \\
&\quad - \lambda^2 2x e^{-\lambda x}/2! + \lambda^2 x^2 \cdot \lambda e^{-\lambda x}/2! \\
&\quad - \lambda^3 3x^2 e^{-\lambda x}/3! + \lambda^3 x^3 \cdot \lambda e^{-\lambda x}/3! \\
&\quad ... \\
&\quad - \lambda^{r-1}(r-1)x^{r-2}e^{-\lambda x}/(r-1)! + \lambda^{r-1}x^{r-1} \cdot \lambda e^{-\lambda x}/(r-1)! \\
&\quad = \lambda^r x^{r-1} e^{-\lambda x}/(r-1)!
\end{aligned}$$

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 rth success and $\mu as the average interval until the 1st success$, 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 Exercises

Exercise 8.5.1 (- Home Sales).

Exercise 8.5.2 (- Customer arrivals - Total).

Exercise 8.5.3 (- Customer arrivals - First).

Exercise 8.5.4 (- Customer arrivals - 10th).

Exercise 8.5.5 (- Computer Network Data Traffic).

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 it's 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

$$z = \frac{x-\mu}{\sigma} = \frac{x-0}{1} = x.$$

□

Theorem 9.2.3 (Verifying the normal probability function)**.**

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

Proof. Note that you can convert the integral above to 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

$$dudv = rdrd\theta$$

and

$$u^2 + v^2 = r^2$$

gives

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} rdrd\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} -e^{-\frac{r^2}{2}} \Big|_0^{\infty} d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} 1 \cdot d\theta \\ &= \frac{1}{2\pi} \theta \Big|_0^{2\pi} = 1 \end{aligned}$$

as desired. □

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.

$$\begin{aligned} E[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 z^2 \cdot e^{-z^2/2} dz \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot z e^{-z^2/2} dz \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \cdot \left[-z e^{-z^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2} dz \right] \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \cdot \left[0 + \sqrt{2\pi} \right] \\ &= \sigma^2 \end{aligned}$$

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)+")␣~=␣"+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)**.** Given an natural number r, suppose X is a random variable over the space $R = (0, \infty)$ with probability function given by

$$f(x) = \frac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}.$$

Then X has a Chi-Square distribution with r degrees of freedom. This is often denoted $\chi^2(r)$.

```
# 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)**.**

$$\mu = r$$

$$\sigma^2 = 2r$$

$$\gamma_1 = 2\sqrt{2/r}$$

$$\gamma_2 = \frac{12}{r} + 3$$

Theorem 9.3.3 (Relationship between Normal and χ^2)**.** *If $Z_1, Z_2, ..., Z_r$ are r standard normal variables, then*

$$X = \sum_{k=1}^{r} Z_k^2$$

is $\chi^2(r)$.

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 χ^2cdf(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 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$_M$*ean*"*box*, *themeanofthesecirclesisindicatedbythegreenlineinthemiddleoftheplot.*

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),a
        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_hi
        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),a
        Values','Sample␣Number'])
    AVGS =
        scatter_plot(averages,markersize=50,edgecolor='blue',marker='o',figsi

    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_l
        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.

9.5 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 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 9.5.1 (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 $X_1, X_2, ..., X_n$ and set $\overline{X} = \frac{\sum X_k}{n}$.*

Then, $\overline{X}$ 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 $\overline{X}$ 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

$$W_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

with corresponding distribution function $F_n(W_n)$,

$$\lim_{n\to\infty} F_n(c) = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(c)$$

that is, the standard normal distribution function.

Example 9.5.2 (Exponential X vs Normal $\overline{X}$)**.** 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 $\overline{X}$ which has an approximate normal distribution with the same mean but with new variance $\sigma^2_{\overline{X}} = \frac{4}{32} = \frac{1}{8}$. Therefore

$$P(\overline{X} < 3.9) \approx normalcdf(0, 3.9, 4, sqrt(1/8)) = 0.2119.$$

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 $Y_1, Y_2, ..., Y_n$ with each over the space 0,1. Then,

$$X = \sum_{k=1}^{n} Y_k$$

is a Binomial variable.

Theorem 9.5.3 (Binomial as approximate Normal)**.** *Given a Binomial variable X with $\mu = np$ and $\sigma^2 = np(1-p)$, then X is approximately also normal with the same mean and variance so long as $np > 5$ and $n(1-p) > 5$.*

Proof. Using the Bernoulli variables Y_k each with mean p and variance p(1-p), note that the Central Limit Theorem applied to $\overline{X} = \frac{\sum Y_k}{n}$ gives that

$$\frac{\overline{X} - p}{\sqrt{p(1-p)/n}}$$

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 9.5.4 (Binomial as Normal)**.** Binomial becomes normal as $n \to \infty$. 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,

$$\begin{aligned} P(X = 16) &= P(15.5 < X < 16.5) \\ &\approx normalcdf(15.5, 16.5, 15, sqrt(10.5)) \\ &= 0.11697 \end{aligned}$$

Notice that these are very close.

Corollary 9.5.5 (Poisson as approximate Normal)**.** *Given a Poisson variable X with μ and $\sigma^2 = \mu$ given, then X is approximately also normal with the same mean and variance so long as $\mu > 5$.*

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 9.5.6 (Poisson as Normal)**.** Poisson becomes normal as $\mu \to \infty$. 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,

$$\begin{aligned} P(X = 19) &= P(18.5 < X < 19.5) \\ &\approx normalcdf(18.5, 19.5, 20, sqrt(20)) \\ &= 0.08683 \end{aligned}$$

Again, these are very close.

Theorem 9.5.7 (Gamma as approximate Normal)**.** *Given a Gamma variable X with mean $r\mu$ and variance $rBLOB$ given, then X is approximately also normal with the same mean and variance so long as CONDITION????.*

Example 9.5.8 (Gamma as Normal)**.** Gamma becomes normal as $r \to \infty$. 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,

$$\begin{aligned} P(90 \le X \le 100) &= \int_{90}^{100} f(x)dx \\ &= 0.59252 - 0.47536 = 0.11716. \end{aligned}$$

Using the normal distribution,

$$P(90 \le X \le 100) \approx normalcdf(90, 100, 96, 33.9411) = 0.11707.$$

Amazingly, these are also very close.

Example 9.5.9 (Uniform X vs Normal $\overline{X}$)**.** Consider a discrete uniform variable X over R = 1,2,...,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,

$$P(8 \le X < 12) = P(X \in \{8, 9, 10, 11\} = \frac{4}{20} = 1/5.$$

If instead you plan to sample from this distribution n=49 times, the Central Limit Theorem implies that you will get a random variable $\overline{X}$ which has an approximate normal distribution with the same mean but with new variance $\sigma^2_{\overline{X}} = \frac{199/20}{49} = \frac{199}{580}$. Therefore, expanding the interval to include the boundaries of the corresponding histogram areas,

$$P(8 \leq \overline{X} < 12) = P(7.5 \leq \overline{X} \leq 11.5) \approx normalcdf(7.5, 11.5, 10.5, 0.585750) \approx 0.9561.$$

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 $\overline{X}$. This usefulness is illustrated in the examples below.

9.6 Exercises

Exercise 9.6.1 (- Computing basic standard normal probabilities)**.**

Exercise 9.6.2 (- Computing basic normal probabilities)**.**

Exercise 9.6.3 (- IQ values)**.**

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 $\overline{x}$ 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)**.** *Given a random variable X with given mean μ and standard deviation σ, for $a \in \mathbb{R}^+$,*

$$P(\left|X - \mu\right| < a) > 1 - \frac{\sigma^2}{a^2}$$

Proof. Notice that the variance of a continuous variable X is given by

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\
&\geq \int_{-\infty}^{\mu-a} (x-\mu)^2 f(x)dx + \int_{\mu+a}^{\infty} (x-\mu)^2 f(x)dx \\
&\geq \int_{-\infty}^{\mu-a} a^2 f(x)dx + \int_{\mu+a}^{\infty} a^2 f(x)dx \\
&= a^2\big(\int_{-\infty}^{\mu-a} f(x)dx + \int_{\mu+a}^{\infty} f(x)dx\big) \\
&= a^2 P(X \leq \mu - a \text{or} X \geq \mu + a) \\
&= a^2 P(\big|\mu - a\big| \geq a)
\end{aligned}$$

Dividing by a^2 and taking the complement gives the result. □

Corollary 10.2.2 (Alternate Form for Chebyshev's Theorem)**.** *For positive k,*

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

Corollary 10.2.3 (Special Cases for Chebyshev's Theorem)**.** *For any distribution, it is not possible for f(x)=0 within one standard deviation of the mean. Aslo, at least 75*

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

$$\mu \approx \overline{x} = \frac{\sum x_k}{n}$$

where the values x_k are the n individual sample values.

For any continous variable and indeed for $\overline{X}$), $P(\overline{X} = \mu) = 0$. 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, symmmetry 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 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 calculators 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 calculators 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)**.**

$$\frac{\tilde{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p}) + \frac{z^2_{\alpha/2}}{4n}}{n}}}{1 + \frac{z^2_{\alpha/2}}{n}} < p < \frac{\tilde{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p}) + \frac{z^2_{\alpha/2}}{4n}}{n}}}{1 + \frac{z^2_{\alpha/2}}{n}}$$

Proof. Again, noting that $\tilde{p} = \frac{Y}{n}$, the expression above

$$\left|p - \tilde{p}\right| = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

can be simplified by squaring both sides to get

$$\left(p - \tilde{p}\right)^2 = z^2_{\alpha/2}\frac{p(1-p)}{n}.$$

Replacing $\tilde{p}$ with the relative frequency gives

$$\left(p - \frac{Y}{n}\right)^2 = z^2_{\alpha/2}\frac{p(1-p)}{n}$$

or by simplifying

$$(n + z^2_{\alpha/2})p^2 - (2Y + z^2_{/alpha/2})p + \frac{Y^2}{2} = 0.$$

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 $95(0.31296, 0.40704)$.

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...

$$z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p}) + \frac{z^2_{\alpha/2}}{4n}}{n}} = 1.96\sqrt{\frac{0.36 \cdot 0.64 + \frac{1.96^2}{1600}}{400}} \approx 0.04728$$

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.

Theorem 10.4.4 (Determining Sample Size for proportions)**.** *Given a margin of error E and preliminary relative frequency estimate $\tilde{p_0}$ the sample size needed to create the corresponding confidence interval is given by*

$$n > \left(\frac{z_{\alpha/2}}{E}\right)^2 \tilde{p_0}(1 - \tilde{p_0}).$$

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.

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 $\overline{x}$, a two-sided confidence interval for the mean with confidence level $1-\alpha$ is an interval

$$\overline{x} - E_1 < \mu < \overline{x} + E_2$$

such that

$$P(\overline{x} - E_1 < \mu < \overline{x} + E_2) = 1 - \alpha.$$

Generally, the interval is symmetrical of the form $\overline{x} \pm E$ 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.

Once again, utilize the Central Limit Theorem. Notice that the symmetrical confidence interval

$$P(\overline{x} - E < \mu < \overline{x} + E) = 1 - \alpha.$$

is equivalent to

$$P\left(\frac{-E}{\sigma/\sqrt{n}} < \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} < \frac{E}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

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 determine in a manner described in the previous section) gives a confidence interval for the mean

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

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 $\overline{x}$...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 $\overline{x}$ 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

$$F(\chi^2_{1-\alpha/2}) = F\left(\frac{(n-1)S^2}{E_1}\right) = 1-\alpha/2$$

and

$$F(\chi^2_{\alpha/2}) = F\left(\frac{(n-1)S^2}{E_2}\right) = \alpha/2.$$

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,

$$E_1 = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$$

and

$$E_2 = \frac{(n-1)S^2}{\chi^2_{\alpha/2}}$$

To determine appropriate values for $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ 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 forumaula (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 $\chi^2_{0.025} = 2.18$ and $\chi^2_{0.975} = 17.53$. 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.

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 $95\chi^2_{0.025} = 345.55$ and $\chi^2_{0.975} = 456.24$.

Therefore

$$E_1 = \frac{8 \cdot 479.5}{456.24} \approx 419.3$$

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.

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

$$F(\chi^2_\alpha) = F\left(\frac{(n-1)S^2}{E_2}\right) = \alpha$$

and

$$E_1 < \sigma^2$$

can be determined using

$$F(\chi^2_{1-\alpha}) = F\left(\frac{(n-1)S^2}{E_1}\right) = 1 - \alpha.$$

Example 10.6.4 (- One-sided Confidence intervals for σ^2)**.**

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)**.**

10.7 Exercises

Exercise 10.7.1 (- Basic Confidence interval for p)**.**

Exercise 10.7.2 (- Sample Size for confidence interval for p)**.**

Exercise 10.7.3 (- Voting projection)**.**

Exercise 10.7.4 (- Basic Confidence interval for the mean)**.**

Exercise 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

Review of Calculus

This chapter is a review of power series results from Calculus.

11.1 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 know 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.

11.1.1 Geometric Series

Theorem 11.1.1. $S = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$

Proof. Consider the partial sum

$$S_n = \sum_{k=0}^{n} x^k = 1 + x + x^2 + ... + x^n$$

$$(1-x)S_n = S_n - xS_n = 1 + x + x^2 + ... + x^n - (x + x^2 + ... + x^n + x^{n+1}) = 1 - x^{n+1}$$

$$\Rightarrow S_n = \frac{1-x^{n+1}}{1-x}$$

and so as $n \to \infty$,

$$S_n \to S = \frac{1}{1-x}$$

□

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('$S_n(x)␣=␣
            %s$'%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))
```

11.1.2 Alternate Forms for the Geometric Series

Theorem 11.1.2 (Generalized Geometric Series)**.** *For* $k \in \mathbb{N}, \sum_{k=M}^{\infty} x^k = \frac{x^M}{1-x}$

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 11.1.3 (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,

$$f'(x) = \sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$$

$$f''(x) = \sum_{k=2}^{\infty} k(k-1)x^{k-1} = \frac{2}{(1-x)^3}$$

$$f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1)...(k-n+1)x^{k-n} = \frac{n!}{(1-x)^{n+1}}$$

$$\int f(x)dx = \sum_{k=0}^{\infty} \frac{x^{k+1}}{k+1} = -ln(1-x)$$

Example 11.1.4 (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 11.1.5 (Application: Converting repeating decimals to fractional form)**.** Consider this example:

$$\begin{aligned}2.48484848... &= 2+0.48+0.0048+0.000048+... \\ &= 2+0.48(1+0.01+0.0001+...) = 2+0.48\sum_{k=0}^{\infty}(0.01)^k\end{aligned}$$

Therefore, applying the Geometric Series

$$\begin{aligned}2.48484848... &= 2+0.48\frac{1}{1-0.01} \\ &= 2+0.48\frac{100}{99} = 2+\frac{48}{99}\end{aligned}$$

Example 11.1.6 (Playing around with repeating decimals)**.** Certainly most students would agree that $0.333333... = \frac{1}{3}$. So, what about 0.999999...? Simply follow the pattern above

$$\begin{aligned}0.999999... &= 0.9+0.09+0.009+0.0009+... = 0.9(1+0.1+0.1^2+0.1^3+... \\ &= 0.9\frac{1}{1-0.1} = 0.9\frac{1}{0.9} = 1\end{aligned}$$

11.2 Binomial Sums

The binomial series is also foundational. It is technically not a series since the $\text{sum}_i finite but we won't bother with that for now. It is given by$

11.2.1

Theorem 11.2.1 (Binomial Theorem)**.** *For* $n \in \mathbb{N}$, $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$

Proof. By induction:

Basic Step: n = 1 is trivial

Inductive Step: Assume the statement is true as given for some $n \geq 1$.

Show $(a+b)^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} a^k b^{n+1-k}$

$$\begin{aligned}
(a+b)^{n+1} &= (a+b)(a+b)^n \\
&= (a+b)\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} a^{k+1} b^{n-k} + \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k+1} \\
&= \sum_{k=0}^{n-1} \binom{n}{k} a^{k+1} b^{n-k} + a^{n+1} + b^{n+1} + \sum_{k=1}^{n} \binom{n}{k} a^k b^{n-k+1} \\
&= \sum_{j=1}^{n} \binom{n}{j-1} a^j b^{n-(j-1)} + a^{n+1} + b^{n+1} + \sum_{k=1}^{n} \binom{n}{k} a^k b^{n+1-k} \\
&= b^{n+1} + \sum_{k=1}^{n} \left[\binom{n}{k-1} + \binom{n}{k} \right] a^k b^{n+1-k} + a^{n+1} \\
&= b^{n+1} + \sum_{k=1}^{n} \binom{n+1}{k} a^k b^{n+1-k} + a^{n+1} \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} a^k b^{n+1-k}
\end{aligned}$$

□

11.2.2 Binomial Series

Consider $B(a,b) = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$. This finite $\text{sum}_i s \, \text{known as the} \, Binomial Series$.

11.2.2.1

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, $B(x,1) = \sum_{k=0}^{n} \binom{n}{k} a^k$

11.2.3 Trinomial Series

$$(a+b+c)^n = \sum_{k_1+k_2+k_3=n} \binom{n}{k_1, k_2, k_3} a^{k_1} b^{k_2} c^{k_3}$$

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 know as a multinomial series.

11.3 Negative Binomial Series

$(a+b)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} a^k b^{-n-k}$

Theorem 11.3.1 (Alternate Form for Negative Binomial Series)**.** $(a+b)^{-n} = \sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} a^k b^{-n-k}$