Introduction to Mathematical Probability and Statistics

A Calculus-based Approach

Introduction to Mathematical Probability and Statistics

A Calculus-based Approach

John Travis
Mississippi College

August 25, 2016

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 .

There are additional exercises or computer projects at the ends of many of the chapters. The computer projects usually require a knowledge of programming. All of these exercises and projects are more substantial in nature and allow the exploration of new results and theory.

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

Review of Calculus

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

1.1 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.

1.1.1 Geometric Series

Theorem 1.1.1. $S = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$

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))
```

1.1.2 Alternate Forms for the Geometric Series

Theorem 1.1.2 (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 1.1.3 (Integrating and Differentiating to get new Power Series)**.** The geometric power series is a nice function which is relatively easily differentiated and integrated. In doing so, one can obtain new power series which might also be very useful in their own right. Here we develop a few which are of special interest.

Let $f(x) = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$. Then,

$$\begin{aligned}f'(x) &= \sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2} \\ f''(x) &= \sum_{k=2}^{\infty} k(k-1)x^{k-1} = \frac{2}{(1-x)^3} \\ f^{(n)}(x) &= \sum_{k=n}^{\infty} k(k-1)...(k-n+1)x^{k-n} = \frac{n!}{(1-x)^{n+1}} \\ \int f(x)dx &= \sum_{k=0}^{\infty} \frac{x^{k+1}}{k+1} = -ln(1-x)\end{aligned}$$

Example 1.1.4 (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 1.1.5 (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 1.1.6 (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

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

1.2 Binomial Sums

The binomial series is also foundational. It is technically not a series since the sum$_iffinitebutwewon'tbotherwiththatfornow.Itisgivenby$

1.2.1

Theorem 1.2.1 (Binomial Theorem)**.** *For* $n \in \mathbb{N}$, $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}$

Proof. By induction:

Basic Step: n = 1 is trivial

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

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

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

□

1.2.2 Binomial Series

Consider $B(a,b) = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$. This finite sum$_i sknownastheBinomialSeries.$

1.2.2.1

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

1.2.3 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.

1.3 Negative Binomial Series

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

Theorem 1.3.1 (Alternate Form for Negative Binomial Series)**.** $(a+b)^{-n} = \sum_{k=0}^{\infty}(-1)^k\binom{n+k-1}{k}a^k b^{-n-k}$

Chapter 2

Representing Data

2.1 Measurement Scales

- Nominal - Mutually Exclusive and Exhaustive categories for which the numerical value has only identification significance. Ex: Male = 1, Female = -1
- Ordinal - Discrete values ranked from lowest to highest or vice versa. Ex: Class grades for GPA.
- Interval - Ordinal data where distance between data values is of significance. Ex: Heights and Weights.
- Ratio - Interval data where ratios of observations have meaning. Ex: Percentile rankings

2.2 Techniques for Representing Data

- 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. Quantiles
  - Measures of Skewness - indicates the level of symmetry of the data

1. Pearson Coefficient
2. Standard Skewness
3. Bowley's Measure

- Measures of Kurtosis - indicates flatness or roundedness of the peak of the data
  1. Standard Kurtosis
  2. Coefficient of Kurtosis
- Measures of Association for Bivariate Data - indicates the likeliness of functional correlation of the data.
  1. Pearson Correlation Coefficient
  2. Spearman Rank Correlation Cooeficient
  3. Quantile-Quantile Plots
- 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
  2. Chi-square
  3. Kolmogorov-Smirnov
  4. Lilliefors
  5. Shapiro-Wilk
- Tests for Randomness - indicates whether the data has a non-systematic pattern
  1. Runs Test
  2. Mean-Square Successive Differences

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

2.3 Measures of Position

Given a collection of data, sorting the data may provide several useful descriptors. These include:

Minimum and Maximum: Easily determine the smallest and largest values in the data set. These are the minimum and maximum respectively.

Order Statistic: Given the given data set $x_1, x_2, ..., x_n$, sort the set and call the new data set $y_1, y_2, ..., y_n$ where now we have $y_1 \leq y_2 \leq ... \leq y_n$. Then the kth order statistic is given by y_k. Therefore, the minimum $= y_1$ and the maximum $= y_n$

Percentiles. A percentile is a numerical value for which approximately a given percentage of the given data lies below that value.

Consider the following data set: 2,5,8,10. The 50th percentile should be a numerical value where approximately 50

To compute the percentile value exactly, given a percentage in the form 100p, for $0 < p < 1$, from a data set sort into the order statistics $y_1, y_2, ..., y_n$. Then the 100pth percentile is given by

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

where m is the integer part of $m = \lfloor (n+1)p \rfloor$, the integer part of (n+1)p and $r = (n+1)p - m$, the fractional part of (n+1)p. This computes a weighted average between y_m and y_{m+1} which is unique for each given p so that in the case where each of the data values are all distinct that the desired percentages are most closely met.

Example 2.3.1 (Basic Percentiles)**.** Using the data set 2,5,8,10 with n=4 values, 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. You can see that 3 also lies in this same range and has the same percentages above and below. However, it would be a slightly larger percentile value. 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.

Quartiles: Given a sorted data set, the first, second, and third quartiles are the values of $P^{0.25}, P^{0.5}$ and $P^{0.75}$.

Deciles: Given a sorted data set, the first, second, ..., ninth deciles are the value of $P^{0.1}, P^{0.2}, ..., P^{0.9}$

2.4 Measures of the Middle

Defn: 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.

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

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

Other Means:
Geometric Mean

$$GM = (x_1 x_2 ... x_n)^{1/n}$$

Harmonic Mean

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

Theorem 2.4.1 (Relative sizes of Means)**.** *$HM \leq GM \leq AM$.*

Theorem 2.4.2 (Mean Formula). $AMHM = GM^2$

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.

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

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

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.

2.5 Measures of Spread

Range. Using the order statistics,

$$y_n - y_1.$$

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

Interquartile Range. $P^{0.75} - P^{0.25}$.

Average Deviation from the Mean: Given a 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:

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

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:

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

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. If this data comes from the entire universe of possibilities then we call it a population variance and denote this value by σ^2

Standard Deviation: Using the variance, 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. Doing so gives a measure of spread called the standard deviation

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

If this data comes from the entire universe of possibilities then we call it a population standard deviation and denote this value by σ

Theorem 2.5.1 (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. □

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 |A| = ∞.

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. □

Example 3.1.3 (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 3.1.4 (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 3.1.5 (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 $9 \times 8\ times... \times 1 = 362880$ 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?

Definition 3.1.6 (Factorial)**.** For any natural number n,

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

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 n objects)**.** *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 n objects selecting r)**.** *The number of ways to arrange r items from a set of n distinct items is* $P^n_r = \frac{n!}{(n-r)!}$

Proof. 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

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

□

Theorem 3.2.3 (Permutations when Not all items are distinguishable and without replacement: (Multinomial Coefficients))**.** *If n items belong to s categories, n1 in first, n2 in second, ... , ns in the last, the number of ways to pick all is !*

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 n distinct objects selecting r 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 n distinct objects selecting r with replacement)**.** *The number of ways to arrange r items from a set of n distinct items is* $C_r^{n+r-1} = \frac{n+r-1!}{r!(n-1)!}$

Proof. blah □

Example 3.3.3. 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".

Theorem 3.3.5 (Combinations when distinguishable and with replacement)**.** $=$ *Number of ways to get unordered samples of size r from n objects.*

Lots of interesting facts about the binomial coefficients.

Chapter 4

Probability and Probability Functions

This chapter is a definitions of probability, consequences, and probability functions.

4.1 Relative Frequency

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.

4.1.1

When attempting to precisely measure this 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.

```
@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.

4.2 Definition of Probability

4.2.1 Motivating the Definition

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

Definition 4.2.1. The Cardinality of the set A is the number of elements in A. This will be denoted |A| (similar to the idea of frequency of an outcome noted earlier.) If a set has a infinite number of elements, then we will say it's cardinality is also infinite and write |A| = ∞

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

4.2.2 Probability

Based upon these we give the following:

Definition 4.2.2. The probability P(A) of a given outcome A is a set function which satisfies:

1. (Nonnegativity) P(A) ≥ 0
2. (Totality) P(S) = 1
3. (Subadditivity) If A ∩ B = ∅, then P(A ∪ B) = P(A) + P(B). In general, if A_k are pairwise disjoint then $P(\cup_k A_k) = \sum_k P(A_k)$.

4.2.3 Basic Probability Theorems

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

Theorem 4.2.3 (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 4.2.4. $P(\emptyset) = 0$

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 4.2.5. *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

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

□

Theorem 4.2.6. *For any event A, $P(A) \leq 1$*

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

Theorem 4.2.7. *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

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

Hence,

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

□

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 4.2.8. *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 4.2.9. *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 Conditional Probability

4.3.1 Motivating Examples

Example 4.3.1 (Changing Sample Space - Balls)**.** Consider a box with three balls: one Red, one White, and one Blue. Using an equally likely assumption, the probability of randomly pulling out a Red ball should be 1/3. That is P(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.

Example 4.3.2 (Changing Sample Space - Cards)**.** Consider a deck of 52 standard playing cards and a success occurs when a Heart is selected from the deck. When extracting one card randomly, the probability of that card being a Heart is then P(Heart) = 13/52. Now, assume that one card has already been extracted and set aside. Now, prepare to extract another. If the first card drawn was a Heart, then there are only 12 Hearts left for the second draw. However, if the first card drawn was not a Heart, then there are 13 Hearts available for the second draw. To compute this probability correctly, one need to formulate the question so that subadditivity can be utilized.

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) =

4.3.2 Conditional Probability

4.3.2.1 Definition

Definition 4.3.3 (Conditional Probability)**.** P(B | A) = P(A ∩ B) / P(A)

Theorem 4.3.4. *Conditional Probability satisfies all of the requirements of regular probability.*

4.3.2.2 Bayes

Theorem 4.3.5 (Bayes Theorem)**.** *Stub*

4.4 Independence

Definition 4.4.1 (Independent Events)**.** Events A and B are independent provided $P(A \cap B) = P(A)P(B)$

4.5 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 4.5.1 (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

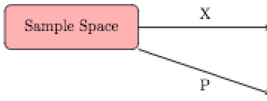
We will make various restrictions on the range of the random variable to fit different generalized problems. Then, we will be able to work on a problem (which may be inherently non-numerical) by using the random variable in subsequent calculations.

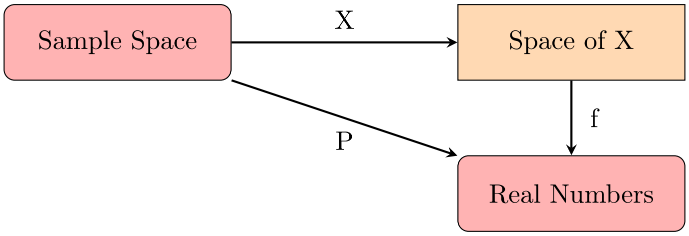
Example 4.5.2 (Success vs Failure)**.** When dealing with only two outcomes, one might use S = success, failure. Choose X(success)=1, X(failure)=0. Then, R=0,1.

Example 4.5.3 (Standard Dice Pairs)**.** When gambling with a pair of dice, one might use S=ordered pairs of all possible rolls = (a,b): a=die 1 outcome, b=die 2 outcome. Choose X((a,b)) = a+b. Then, R=2, 3, 4, 5, ..., 12.

Example 4.5.4 (Other Dice Options)**.** When rolling dice in a board game (like RISK), one might use S=(a,b): a=die 1 outcome, b=die 2 outcome Choose X((a,b)) = maxa,b. Then, R=1, 2, 3, 4, 5, 6

Definition 4.5.5. 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.





4.6 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 4.6.1 (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:

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

Definition 4.6.2 (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:

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

Definition 4.6.3 (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 $F(x) = P(X \leq x)$

4.6.1 Properties of the Distribution Function

Theorem 4.6.4. $F(x) = 0, \forall x \leq \inf(R)$

Proof. □

Theorem 4.6.5. $F(x) = 1, \forall x \geq \sup(R)$

Proof. □

Theorem 4.6.6. *F is non-decreasing*

Proof. Case 1: R discrete

$$\begin{aligned}
&\forall x_1, x_2 \in \mathbb{Z} \ni x_1 < x_2\\
&\qquad F(x_2) = \sum_{x \leq x_2} f(x)\\
&\qquad\qquad = \sum_{x \leq x_1} f(x) + \sum_{x_1 < x \leq x_2} f(x)\\
&\qquad\qquad \geq \sum_{x \leq x_1} f(x) = F(x_1)
\end{aligned}$$

Case 2: R continuous

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

□

Theorem 4.6.7 (Using Discrete Distribution Function to compute probabilities)**.** *for* $x \in R, f(x) = F(x) - F(x-1)$

Theorem 4.6.8 (Using Continuous Distribution function to compute probabilities)**.** *for* $a < b, (a,b) \in R, P(a < X < b) = F(b) - F(a)$

Corollary 4.6.9. *For continuous distributions, P(X = a) = 0*

4.6.2 Standard Units

Any distribution variable can be converted to "standard units" using the linear translation $z = \dfrac{x-\mu}{\sigma}$. 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.

4.7 Expected Value

Chapter 5

Binomial, Geometric, and Negative Binomial Distributions

Distributions relating number of successes to number of trials with one of these variable and the other fixed.

5.1 Binomial Distribution

Consider the situation where one can observe a sequence of n independent trials with 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 successes obtained when doing a fixed number of trials n, then the resulting distribution of probabilities is called a Binomial Distribution.

5.1.1 Derivation of Binomial Probability Function

Since successive trials are independent, then the probability of X successes occurring within n trials is given by $P(X = x) = \binom{n}{x} P(SS...SFF...F) = \binom{n}{x} p^x (1-p)^{n-x}$

5.1.2 Binomial Distribution 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

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

5.1.3 Binomial Distribution variance

$$
\begin{aligned}
\sigma^2 = E[X(X-1)] + \mu - \mu^2 &= \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x (1-p)^{n-x} + np - n^2p^2 \\
&= \sum_{x=2}^{n} x(x-1)\frac{n(n-1)(n-2)!}{x(x-1)(x-2)!(n-x)!} p^x (1-p)^{n-x} + np - n^2p^2 \\
&= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!((n-2)-(x-2))!} p^{x-2} (1-p)^{(n-2)-(x-2)} + np - n^2
\end{aligned}
$$

Using the change of variables $k = x - 2$ and $m = n - 2$ yields a binomial series

$$
\begin{aligned}
&= n(n-1)p^2 \sum k = 0^m \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k} + np - n^2p^2 \\
&= n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p)
\end{aligned}
$$

5.2 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, then the resulting distribution of probabilities is called a Geometric Distribution.

5.2.1 Derivation of Geometric Probability Function

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

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

5.2.2 Properties of the Geometric DistributionGeometric Distribution sums to 1

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

5.2.3 Derivation of Geometric Mean

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

5.2.4 Derivation of Geometric Variance

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

5.2.5 Derivation of Geometric Distribution Function

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

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

5.3 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, $r \geq 1$ then the resulting distribution of probabilities is called a Geometric Distribution.

Note that r=1 gives the Geometric Distribution.

5.3.1 Negative Binomial Series

Theorem 5.3.1. $\dfrac{1}{(a+b)^n} = \displaystyle\sum_{k=0}^{\infty} (-1)^k \binom{n+k-1}{k} a^k b^{-n-k}$

Proof. First, convert the problem to a slightly different form: $\frac{1}{(a+b)^n} = \frac{1}{b^n}\frac{1}{(\frac{a}{b}+1)^n} = \frac{1}{b^n}\sum_{k=0}^{\infty}(-1)^k\binom{n+k-1}{k}\left(\frac{a}{b}\right)^k$

So, let's replace $\frac{a}{b} = x$ and ignore for a while the term factored out. Then, we only need to show

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

However

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

This infinite sum raised to a power can be expanded by distributing terms in the standard way. In doing so, the various powers of x multiplied together will create a series in powers of x involving $x^0, x^1, x^2, \ldots$. 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, $\left(\frac{1}{1-x}\right)^n = \left(\sum_{k=0}^{\infty} x^k\right)^n = \sum_{k=0}^{\infty}\binom{n+k-1}{k}x^k$ □

5.3.2 Negative Binomial Distribution FormulasNegative Binomial Distribution Sums to 1

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.

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

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

$m = r\infty\binom{m-1}{r-1}(1-p)^{m-r}p^r = p^r m = r\infty\binom{m-1}{r-1}(1-p)^{m-r}$
and by using $k = m - r$

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

Chapter 6

Poisson, Exponential, and Gamma Distributions

6.1 Poisson Distribution

6.2 Exponential Distribution

6.3 Gamma Distribution

Chapter 7

Normal Distributions

7.1 Properties of the Normal Distribution

You have seen that most distributions become "bell shaped" as certain parameters are allowed to increase. The question might arise regarding whether this always must happen or is it just a happy coincidence. The amazing answer is that if you interpret the question in the correct way then this is always true.

Definition 7.1.1 (The Normal Distribution)**.** Given two parameters μ and σ a random variable X with density function

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

Theorem 7.1.2. *If $\mu = 0$ and $\sigma = 1$, then we say X has a standard normal distribution and often use Z as the variable name. In this case, the density function reduces to*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{z^2/2}$$

Proof. Convert to "standard units" using the conversion $z = \frac{x-\mu}{\sigma} = \frac{x-0}{1} = x$. □

7.2 Theorems

7.3 Chi-Square Distribution

7.4 Central Limit Theorem

7.4.1

Theorem

7.4.2

Approximating distributions Limiting distributions

Chapter 8

Estimating Data using Intervals

8.1 Chebyshev

An interval centered on the mean in which at least a certain proportion of the actual data must lie.

Theorem 8.1.1 (Chebyshev's Theorem). *Given a random variable X with given mean and standard deviation , for k>1 at least 2 1 1 k of the observations lie within k standard deviations from the mean.*

8.2 Measures of Spread

Measures of spread: • Average Deviation from the Mean – always zero for any distribution • Average Absolute Deviation from the Mean – difficult to deal with algebra when absolute values are used • Average Squared Deviation from the Mean – always non-negative and good with algebra and calculus

Definition 8.2.1 (Variance). The variance is a measure of spread found by using the average squared deviation from the mean 2 = ( ))(2 1 k n k k xfx =   if this value exists and is also denoted by Var(X). The positive square root of the variance is called the standard deviation and is denoted by .