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 9, 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.

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

Overview of Features

This chapter is a review of several items needed from Calculus.

1.1 Geometric Series

Knowledge of the use of power series is very important when dealing with both probabilities and with financial mathematics. In particular, the geometric series is very useful.

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

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 $n \to \infty$,

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

□

```
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

1.3 Binomial Sums

The binomial series is also foundational. It is technically not a series since the sum if finite but we won't bother with that for now. It is given by

$$B = k = 0n\binom{n}{k}a^k b^{n-k}$$

1.3.1

Theorem 1.3.1 (Binomial Theorem)**.** *For* $n \in \mathbb{N}$, $(a + b)^n = k = 0n\binom{n}{k}a^k b^{n-k}$

Proof. By induction:

Basic Step: n = 1 is trivial

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

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

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

□

1.3.2 Binomial Series

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

1.3.2.1

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

1.3.3 Trinomial Series

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

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.

Chapter 2

Probability and Probability Functions

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

2.1 Relative Frequency Motivation

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.

2.1.1

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

Experiment Trial Success/Failure Frequency Relative Frequency Histogram

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

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 rolling a single die:

```
@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))
         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.

2.1.2 Probability DefinitionsEqually Likely Outcomes

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 2.1.1. The Cardinality of the set A is the number of elements in A. This will be denoted |A|. If a set has a infinite number of elements, then we will say it's cardinality is also infinite and write $|A| = \infty$

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 in the A and the complete set of outcomes in the set S. Then, if each outcome in A is equally likely then we would like to have the 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 should sum to one

3. Relative frequencies for collections of different outcomes should equal the sum of the individual relative frequencies

Based upon these we give the following

Definition 2.1.2. The probability P(A) of a given outcome A satisfies the following:

1. (Nonnegativity) $P(A) \geq 0$

2. (Totality) $P(S) = 1$

3. (Subadditivity) If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

The subadditivity property can be extended to cover any number of mutually exclusive sets.

2.1.3

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

Theorem 2.1.3 (Probability of Complements)**.** *For the 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 2.1.4. $P(\emptyset) = 0$

Proof. Note that $\emptyset \cap S = \emptyset$. So, by subadditivity, $1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset) = 1 + P(\emptyset)$. Cancelling the 1 on both sides gives $P(\emptyset) = 0$. □

Theorem 2.1.5. *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 2.1.6. *For any event A,* $P(A) \leq 1$

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

Theorem 2.1.7. *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 $A = (A - B) \cup (A \cap B)$ and $B = (A \cap B) \cup (B - A)$ Hence,

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

□

2.2 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.

2.3 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 2.3.1 (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:

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

Definition 2.3.2 (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:

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

Definition 2.3.3 (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 $F(x) = P(X \leq x)$

2.3.1 Properties of the Distribution Function

Theorem 2.3.4. $F(x) = 0, \forall x \leq \inf(R)$

Proof. □

Theorem 2.3.5. $F(x) = 1, \forall x \geq \sup(R)$

Proof. □

Theorem 2.3.6. *F is non-decreasing*

Proof. Case 1: R discrete

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

Case 2: R continuous

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

□

Theorem 2.3.7 (Using Discrete Distribution Function to compute probabilities)**.** *for* $x \in R, f(x) = F(x) - F(x-1)$

Theorem 2.3.8 (Using Continuous Distribution function to compute probabilities)**.** *for* $a < b, (a, b) \in R, P(a < X < b) = F(b) - F(a)$

Corollary 2.3.9. *For continuous distributions,* $P(X = a) = 0$

2.3.2 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.

Chapter 3

Binomial Distribution, Geometric, and Negative Binomial

This chapter creates the Binomial Distribution.

3.1

3.1.1 Binomial 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 successes obtained when doing a fixed number of trials n, then the resulting distribution of probabilities is called a Binomial Distribution.

3.1.1.1 Derivation of Binomial Probability Function

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

3.1.1.2 Binomial Distribution mean

$$\begin{aligned}
\mu = E[X] = x = 0nx &\binom{n}{x} p^x (1-p)^{n-x} \\
&= x = 1nx \frac{n(n-1)!}{x(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= npx = 1n \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{(n-1)-(x-1)}
\end{aligned}$$

Using the change of variables $k = x - 1$ and $m = n - 1$ yields a binomial series

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

3.1.1.3 Binomial Distribution variance

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

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

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

Chapter 4

Geometric, Negative Binomial

This chapter deals with the distributions which measure a variable value until a desired success occurs.

4.1 Geometric and Negative Binomial 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.

4.1.1 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

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

4.1.2 Properties of the Geometric DistributionGeometric Distribution sums to 1

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

4.1.3 Derivation of Geometric Mean

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

4.1.4 Derivation of Geometric Variance

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

4.1.5 Derivation of Geometric Distribution Function

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

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

4.1.6 Negative Binomial Series

Theorem 4.1.1. $\frac{1}{(a+b)^n} = k = 0\infty(-1)^k\binom{n+k-1}{k}a^k b^{-n-k}$

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

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

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

However

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

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 $x^0, x^1, x^2, ...$ 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,

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

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

4.1.7 Negative Binomial Distribution

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.

4.1.7.1 Derivation of Negative Binomial Probability Function

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

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

4.1.7.2 Negative Binomial Distribution Sums to 1

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

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