A Review of Some Important Discrete Distributions for Categorical Data Analysis

Kevin Quinn
Assistant Professor
Department of Political Science and
The Center for Statistics and the Social Sciences
Box 354322, Padelford Hall
University of Washington
Seattle, WA 98195-4322

September 27, 2000

1 Introduction

These notes are written to give students in CSSS/SOC/STAT 536 a quick review of some of the probabilistic concepts they will need to make use of in the class. These notes are not meant to be comprehensive but rather to be a succinct treatment of some of the key ideas.

These notes are organized as follows. The next section reviews basic probability theory. Sections 3-6 cover the Bernoulli, binomial, multinomial, and Poisson distributions respectively. These distributions will play central roles in the models examined in CSSS 536.

2 Preliminaries

While these notes are by definition a bare-bones treatment, this section in particular is an *extremely* brief review of some critical concepts. For a more detailed treatment the reader should refer to a standard text such as DeGroot (1986), Feller (1968), or Gallant (1997).

2.1 Basic Set Theoretic Notation

Let A denote a *set*. If a is a member of A we write $a \in A$. Braces are used to enumerate the members of sets. For instance, if a_1, a_2, and a_3 are the only members of A we could write $A = \{a_1, a_2, a_3\}$.

The *empty set* is defined to be the set that has no members and is written $\emptyset$.

We say that a set A is a *subset* of another set B if every member of A is also a member of B. If A is a subset of B we write $A \subset B$. For example, if $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4, 5, 6\}$ we could say that $A \subset B$ since every member of A is also in B.

The *union* of two sets A and B is defined to be the set containing all the members of just A, just B, and both A and B. We write the union of two sets A and B as $A \cup B$. For example, if $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5\}$ then $A \cup B = \{1, 2, 3, 4, 5\}$. A shorthand way to write $A_1 \cup A_2 \cup \ldots, \cup A_n$ is $\bigcup_{i=1}^{n} A_i$.

The *intersection* two sets A and B is defined as the set containing all elements that belong to both A and B. We write the intersection of two sets A and B as $A \cap B$. For example, if

$A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5\}$ then $A \cap B = \{3, 4\}$. If $A \cap B = \emptyset$ we say that A and B are *disjoint* or *mutually exclusive.*

2.2 Probability

2.2.1 Sample Spaces, Events, and Probability Functions

Consider the situation in which a coin is tossed two times and the result (heads or tails) of each toss is recorded. There are four possible outcomes of this thought experiment. They are $\{heads, heads\}$, $\{heads, tails\}$, $\{tails, heads\}$, and $\{tails, tails\}$, where the first element in brackets denotes the outcome of the first toss and the second element in brackets denotes the outcome of the second toss. Forming the set of all possible outcomes we have:

$$\big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\}.$$

This set of all possible outcomes is called the *sample space* and is often denoted Ω.

An *event* is defined as a subset of a sample space. An event can be the sample space itself, the empty set, or any proper subset of the sample space. For example, if the sample space is:

$$\Omega = \big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\},$$

then

- $\big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\}$
- $\emptyset$
- $\big\{\{heads, heads\}, \{heads, tails\}, \{tails, tails\}\big\}$
- $\{heads, tails\}$

are all events.

A *probability function* $P(\cdot)$ is a function defined over all subsets of a sample space Ω and that satisfies three properties:

1. $P(A) \geq 0$ for all A in the set of all events.
2. $P(\Omega) = 1$
3. if events $A_1, A_2, \ldots$ are disjoint then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

The first condition ensures that the probability of an event occurring is always greater than or equal to 0. The second condition states that the the probability of some element in the sample space occurring is 1. The final condition together with the fact that $P(\emptyset) = 0$ implies that if

$$\Omega = \big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\}$$

then the probability of $\{heads, heads\}$ or $\{tails, heads\}$ occurring is $P(\{heads, heads\}) + P(\{tails, heads\})$. Note well that for this calculation to be valid the events must be disjoint. The case of intersecting events is treated at the end of section 2.2.2.

2.2.2 Marginal, Joint, and Conditional Probabilities

So far we have only considered situations where we are interested in the probability of a single event A occurring. We've denoted this $P(A)$. $P(A)$ is called a *marginal probability*. Suppose we are now in a situation where we would like to express the probability that an event A *and* an event B occur. This quantity is written as $P(A \cap B)$ or sometimes $P(AB)$ and is the *joint probability* of A and B. Suppose now that we know the event B occurs and we wish to express the probability of the event A given this fact. We write this quantity as $P(A|B)$ and call it the *conditional probability* of A given B. When $P(B) > 0$ marginal, joint, and conditional probabilities are related through the following identity:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1}$$

Of course, the identity in 1 also implies

$$P(A \cap B) = P(A|B)P(B), \tag{2}$$

and

$$P(B) = \frac{P(A \cap B)}{P(A|B)}. \tag{3}$$

The examples in section 2.4 will make these concepts more clear.

Another useful fact is that

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

regardless of whether A and B are disjoint.

2.2.3 Independence

We say that two events (A and B) are *independent* if $P(A \cap B) = P(A)P(B)$. In words, the joint probability of A and B is equal to the product of the marginal probabilities of A and B. Note that this also implies that when A and B are independent

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

Intuitively, if events A and B are independent the probability of A occurring does not depend on whether B occurs.

2.3 Random Variables

The discussion of probability has, so far, placed little structure on the the types of sample spaces, and consequently, events considered. This lack of structure makes it difficult to bring mathematical tools to bear on realistic problems. In this subsection, we introduce the concept of a random variable. This concept provides the structure which greatly simplifies many analyses.

We begin by noting that many events bear some relation to a quantifiable attribute. For example, the number of times a coin comes up heads in 10 tosses, the number of male births in a year, or the number of winning lottery tickets in a given month. The idea behind the introduction of random variables is that by mapping the original sample space onto the real numbers we can take advantage of the structure of the real number line to gain leverage on many problems.

A *random variable* is defined as a function that maps the sample space onto the real numbers. In other words, the random variable X takes each element of the sample space and assigns to it a real number. For example, if we go back to the coin tossing experiment of section 2.2.1 where the sample space is

$$\Omega = \big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\},$$

we could introduce the random variable $X(\omega)$ that gives the number of heads in each element of Ω. Accordingly, $X(\{heads, heads\}) = 2$, $X(\{heads, tails\}) = 1$, $X(\{tails, heads\}) = 1$, and $X(\{tails, tails\}) = 0$. The advantages in this particular example are minimal. However, now consider the case in which a coint is tossed 10 times. Here the sample space has $2^{10} = 1024$ unordered elements. By looking at the random variable X equal to the number of heads in the sequence we can instead focus our attention on a new sample space consisting of 11 ordered elements.

2.4 Distribution Functions and Probability (Density) Functions of Random Variables

Two related types of functions– distribution functions and probability (density) functions– are commonly used to exploit the structure that the concept of a random variable brings to problems.

A *distribution function* $F(x)$ of a random variable X is a non-decreasing function that gives the probability that $X \leq x$.

Before discussing the nature of a probability (density) functions it is necessary to discuss the differences between discrete and continuous random variables. A random variable is called *discrete* if it can take on only a finite number of values $(x_1, x_2, \ldots, x_n)$ or an infinite sequence of distinct values $(x_1, x_2, \ldots)$. A random variable X is said to be *continuous* if there exists a non-negative function $f(x)$ such that $F(x) = \int_{-\infty}^{x} f(z)dz$ for all $x \in \mathbb{R}$. In this class, we will be primarily interested in discrete random variables.

The *probability function* $f(x)$ of a discrete random variable X gives the probability that $X = x$. When X is discrete, $f(x)$ and $F(x)$ are related by the following identity

$$F(x) = \sum_{z \leq x} f(z).$$

The *probability density function* $f(x)$ of a continuous random variable X is the non-negative function that satisfies

$$F(x) = \int_{-\infty}^{x} f(z)dz.$$

Note that the probability density function of a continuous random variable does not have the same interpretation as the probability function of a discrete random variable. In particular, $f(x)$ does not give the probability that $X = x$ when X is continuous. To find the probability that X is in some interval (a, b) we would calculate $\int_a^b f(z)dz = F(b) - F(a)$.

The concepts of distribution function and probability (density) function will become more when we discuss specific probability distributions below.

		Y			
		1	2	3	
	1	0.22	0.04	0.09	0.35
X	2	0.15	0.10	0.20	0.45
	3	0.01	0.07	0.12	0.20
		0.38	0.21	0.41	1.00

Table 1: Table of probabilities of two random variables X and Y.

2.5 Marginal, Joint, and Conditional Distributions

Just as marginal, joint, and conditional probabilities can be defined for two arbitrary events A and B; marginal, joint, and conditional probability distributions can be defined for two random variables X and Y. For simplicity we deal only with discrete random variables.

The *marginal probability function* $f_X(x)$ of a discrete random variable X gives the probability that $X = x$ for all x. The *joint probability function* $f_{X,Y}(x, y)$ of two discrete random variables X and Y is the function that gives the probability that $X = x$ and $Y = y$ for all x and y. The *conditional probability function* $f_{X|Y}(x|y)$ of two discrete random variables gives the probability that $X = x$ given the fact that $Y = y$ for all all values of x and y.

Given two discrete random variables X and Y, the marginal probability function $f_X(x)$ of X can be calculated from the joint probability function $f_{X,Y}(x, y)$ of X and Y according to

$$f_X(x) = \sum_y f_{X,Y}(x, y).$$

The conditional probability function $f_{X|Y}(x|y)$ of two discrete random variables X and Y is given by:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

where it is assumed that $f_Y(y) > 0$. It follows that

$$f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y)$$

and

$$f_Y(y) = \frac{f_{X,Y}(x, y)}{f_{X|Y}(x|y)}.$$

Similar results hold for continuous random variables.

A concrete example will make these relationships more clear. Table 1 presents such an example in the form of probability table of two discrete random variables X and Y. Here the marginal probability function of X is in the far right hand column and the marginal probability function of Y is in the bottom column. We can see that $f_X(1) = 0.35$, $f_X(2) = 0.45$, and $f_X(3) = 0.20$. Note that these marginal probabilities sum to 1 as they must in order for $f_X(\cdot)$ to be a proper probability function. Similarly, we can see that $f_Y(1) = 0.38$, $f_Y(2) = 0.21$, and $f_Y(3) = 0.41$. Once again we see that these marginal probabilities sum to 1 as they must.

The interior cells of the table give the joint probabilities. For instance, $f_{X,Y}(1, 1) = 0.22$, $f_{X,Y}(1, 3) = 0.09$, and $f_{X,Y}(2, 3) = 0.20$.

The conditional probability function of X given Y can also be calculated from this table. This conditional probability function is obtained by finding the joint probability $f_{X,Y}(x,y)$ for some x and y and dividing by the appropriate marginal $f_Y(y)$. For instance,

$$f_{X|Y}(1|2) = \frac{f_{X,Y}(1,2)}{f_Y(2)} = \frac{0.04}{0.21} \approx 0.19,$$

$$f_{X|Y}(2|2) = \frac{f_{X,Y}(2,2)}{f_Y(2)} = \frac{0.10}{0.21} \approx 0.48,$$

and

$$f_{X|Y}(3|2) = \frac{f_{X,Y}(3,2)}{f_Y(2)} = \frac{0.07}{0.21} \approx 0.33.$$

Note that the marginal probability $f_Y(2)$ acts as a normalizing constant that rescales the joint probabilities so that they sum to 1 and in fact form a proper probability function.

We say that two random variables X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y. An equivalent definition of independence states that two random variables X and Y are independent if and only if $f_{X|Y}(x|y) = f_X(x)$ for all x and y. In other words, if X and Y are independent, we cannot learn anything about the likely value of X given knowledge of the realized value of Y.

3 The Bernoulli Distribution

The Bernoulli distribution arises in situations where a random variable can take on only two distinct values, such as the toss of an idealized coin that must come up either heads or tails. Other examples of random variables that might plausibly follow the Bernoulli distribution are voting decisions in a two party race without abstention, a person's decision to enter the workforce in a given month, and a person's decision to accept or reject a marriage proposal.

Without loss of generality we specify the values of the two possible outcomes to be 0 and 1. We will call 1s "successes" and 0s "failures". This is purely a convention and does not affect any results.

The Bernoulli distribution is governed by a single parameter π. We can interpret π as the probability of a success.

The Bernoulli probability function is given by

$$f_{Bern}(x|\pi) = \begin{cases} (1-\pi) & \text{if } x = 0 \\ \pi & \text{if } x = 1 \end{cases} \tag{4}$$

Equivalently, the Bernoulli probability function can be written

$$f_{Bern}(x|\pi) = \pi^x (1-\pi)^{(1-x)}. \tag{5}$$

The Bernoulli distribution function is given by

$$F_{Bern}(x|\pi) = \begin{cases} (1-\pi) & \text{if } x = 0 \\ 1 & \text{if } x = 1. \end{cases} \tag{6}$$

The mean of a Bernoulli distributed random variable is π and its variance is $\pi(1-\pi)$.

Figures 1 and 2 display the Bernoulli probability function and distribution function respectively when $\pi = 0.8$.

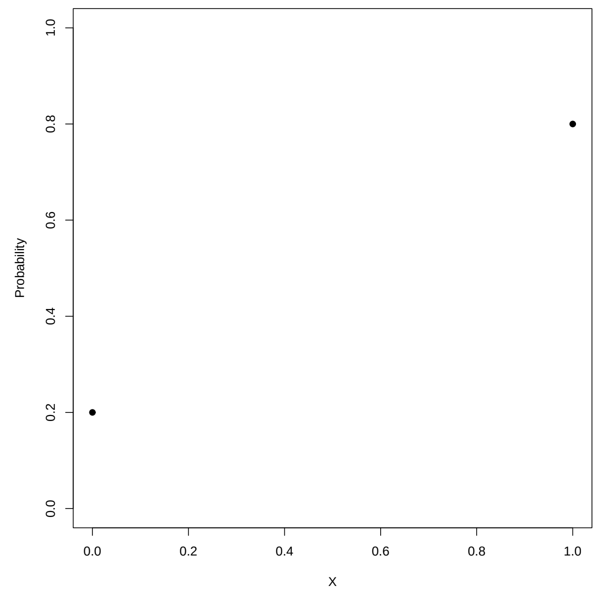

Figure 1: Bernoulli probability function for $\pi = 0.8$.

If X follows the Bernoulli distribution with probability of success π we write $X \sim Bern(\pi)$.

4 The Binomial Distribution

The binomial distribution arises naturally as the distribution of a random variable X that is the sum of n independent Bernoulli random variables $X_1, X_2, \ldots, X_n$ for which the Bernoulli probability of success π is constant. For example, if the number of heads in one toss of a coin follows the Bernoulli distribution, then the number of heads resulting from tossing this same coin under identical conditions n times will follow the binomial distribution with parameters n and π.

The binomial probability function is

$$f_{Bin}(x|n,\pi) = \binom{n}{x} \pi^x (1-\pi)^{(n-x)} \tag{7}$$

where the $\binom{n}{x}$ notation is read "n choose x" and is equal to $\frac{n!}{x!(n-x)!}$. $\binom{n}{x}$ is the number of combinations of size x that can be taken from a set of n elements.

The binomial distribution function is

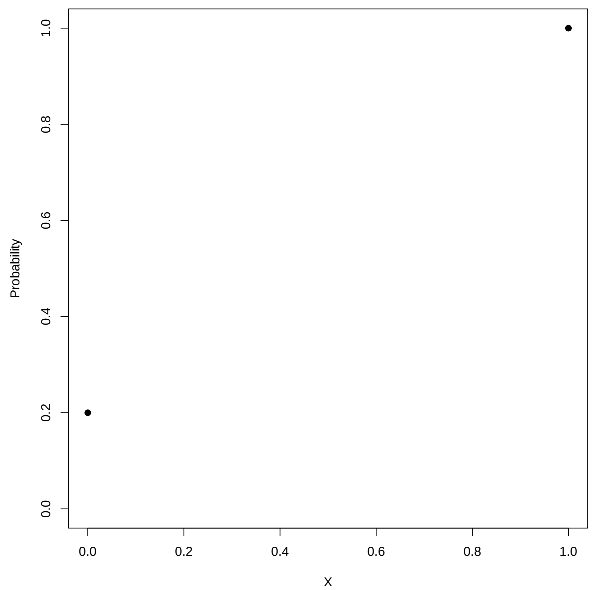

Figure 2: Bernoulli distribution function for $\pi = 0.8$.

$$F_{Bin}(x|n, \pi) = \sum_{z=0}^{x} \binom{n}{z} \pi^z (1-\pi)^{(1-z)} \tag{8}$$

To see how the binomial distribution arises from the Bernoulli distribution we consider the case of n independent draws from a Bernoulli distribution with probability of success π. From these draws $X_1, X_2, \ldots, X_n$ a new random variable $X = \sum_{i=1}^{n} X_i$ is formed. From equation 5 we know that the probability that $X_i = 1$ is π and the probability that $X_i = 0$ is $1 - \pi$. Since the X_i are assumed independent we can calculate the probability that $X_1, X_2, \ldots, X_n$ yield a particular pattern of successes and failures as the product of the marginal probabilities:

$$\begin{aligned} \Pr(X_1 = x_i, X_2 = x_2, \ldots, X_i = x_i|\pi) &= \pi^{x_1}(1-\pi)^{(1-x_1)} \\ &\times \pi^{x_2}(1-\pi)^{(1-x_2)} \\ &\times \ldots \\ &\times \pi^{x_n}(1-\pi)^{(1-x_n)} \\ &= \pi^{(\sum_{i=1}^{n} x_i)}(1-\pi)^{(n-\sum_{i=1}^{n} x_i)} \end{aligned} \tag{9}$$

Now, note that by definition, there are $\binom{n}{x}$ ways that x successes could appear in n Bernoulli draws. Since the Bernoulli draws are independently and identically distributed the probability of each of these $\binom{n}{x}$ ways has equal probability. It follows that to get the probability of x successes in n trials we need to calculate the probability that *any single* patttern of x successes would occur in n Bernoulli trials and then multiply this by the total number of ways this many successes could occur. The first part is given in equation 9 and the second is defined to be $\binom{n}{x}$. This implies that the binomial probability function for x successes in n trials with probability of success π is $f_{Bin}(x|n,\pi) = \binom{n}{x}\pi^x(1-\pi)^{(n-x)}$.

The mean of a binomially distributed random variable is $n\pi$ and its variance is $n\pi(1-\pi)$

If X follows the binomial distribution with probability of success π and sample size n we write $X \sim Bin(n,\pi)$.

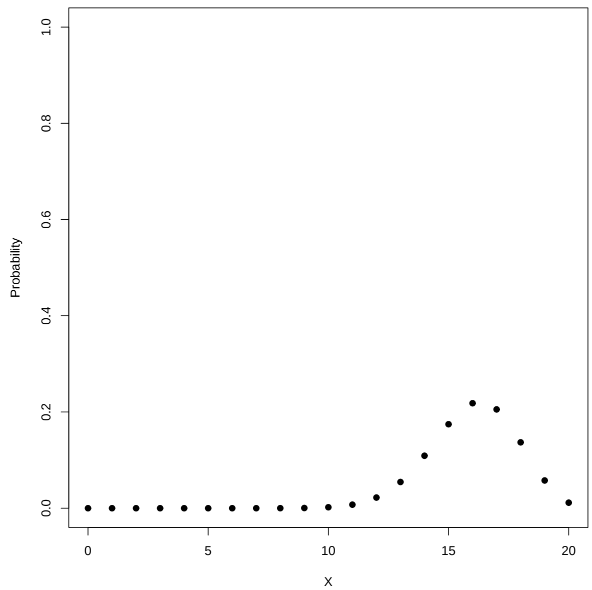

Figure 3: Binomial probability function for $n = 20$ and $\pi = 0.8$.

5 The Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution to more than two potential outcomes. Like the binomial distribution the multinomial distribution has two parameters n and $\boldsymbol{\pi}$, where n is the multinomial sample size and $\boldsymbol{\pi}$ is now a vector of the k probabilities of the k possible outcomes of a multinomial sample of size 1.

The multinomial probability function is

$$f_{Multin}(\mathbf{x}|n,\boldsymbol{\pi}) = \frac{n!}{\prod_{i=1}^{k} x!} \prod_{i=1}^{k} \pi_i^{x_i} \tag{10}$$

where it is assumed that x_i gives the number of successes of type i and $\sum_{i=1}^{k} x_i = n$.

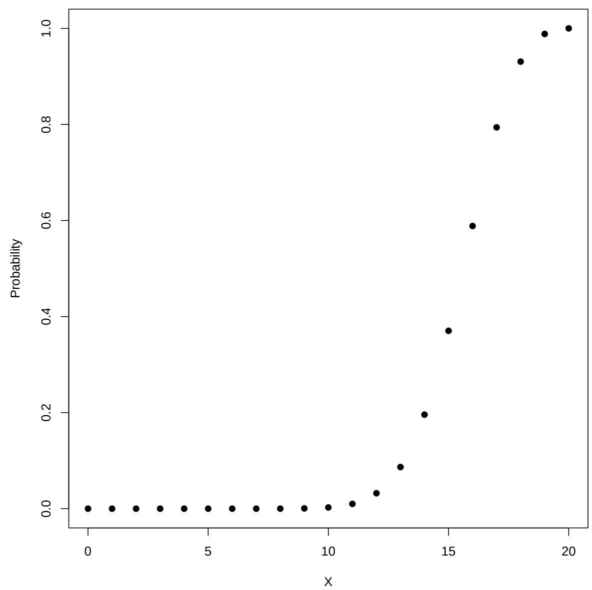

Figure 4: Binomial distribution function for $n = 20$ and $\pi = 0.8$.

The multinomial distribution is constructed using the same logic that is used to construct the binomial distribution. Since it is assumed that the n samples are independent, $\prod_{i=1}^{k} \pi_i^{x_i}$ gives the probability of x_1 successes of type 1, x_2 successes of type 2, etc, in a particular prespecified order. Since there are many ways that multinomial samples of size 1 can combine to produce a multinomial outcome of size n,[1] we have to multiply $\prod_{i=1}^{k} \pi_i^{x_i}$ by the number of different ways that the n individual trials can be specified. This number is given by $\frac{n!}{\prod_{i=1}^{k} x!}$.

When $k = 2$ the multinomial distribution reduces to the binomial distribution. To see this, suppose that $\mathbf{X} = (X_1, X_2)$ follows a multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_1, \pi_2)$. Since any realization $\mathbf{x}$ of $\mathbf{X}$ must satisfy $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} \pi_i = 1$, it must be the case that $\mathbf{X} = (X_1, n - X_1)$ and $\boldsymbol{\pi} = (\pi_1, 1 - \pi_1)$. Substituting $n - x_1$ for x_2 and $1 - \pi_1$ for π_2 in the multinomial probability function we get

$$\frac{n!}{x_1!(n - x_1)!} \pi_1^{x_1} (1 - \pi_1)^{(n - x_1)}$$

which is a binomial distribution with parameters n and π_1.

The mean of a multinomial random variable $\mathbf{X}$ is $n\boldsymbol{\pi}$, the variance of the ith element of $\mathbf{X}$ is $n\pi_i(1 - \pi_i)$, and the covariance between the ith and jth elements of $\mathbf{X}$ is $-n\pi_i\pi_j$.

If $\mathbf{X}$ follows the multinomial distribution with parameters n and $\boldsymbol{\pi}$ we write $\mathbf{X} \sim Multin(n, \boldsymbol{\pi})$.

[1] For example, (0,0,1) + (0,1,0) and (0,1,0) + (0,0,1) both produce $x = (0, 1, 1)$).

6 The Poisson Distribution

The Poisson distribution has a single parameter λ that gives both the mean and variance. The fact that mean of the Poisson distribution must equal the variance is an important point to keep in mind when modeling some types of data. The the Poisson distribution has support on the non-negative integers.

The Poisson probability function is given by

$$f_{Pois}(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{11}$$

and the Poisson distribution function is

$$F_{Pois}(x|\lambda) = \sum_{z=0}^{x} \frac{e^{-\lambda}\lambda^z}{z!} \tag{12}$$

One way to conceptualize the Poisson distribution is as an approximation to the binomial distribution when n is extremely large and π is close to 0. More precisely, as $n \to \infty$, $\pi \to 0$, and $n\pi \to \lambda$ the binomial distribution with parameters n and π approaches the Poisson distribution with parameter λ. To see this, note that we can write the binomial probability function as

$$\begin{aligned}
\binom{n}{x}\pi^x(1-\pi)^{n-x} &= \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{(n-x)} \\
&= \frac{n(n-1)\cdots(n-x+1)}{x!}\pi^x(1-\pi)^{(n-x)} \\
&= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{n\pi}{n}\right)^x\left(1-\frac{n\pi}{n}\right)^{(n-x)} \\
&= n(n-1)\cdots(n-x+1)\frac{1}{x!}\left(\frac{n\pi}{n}\right)^x\left(1-\frac{n\pi}{n}\right)^n\left(1-\frac{n\pi}{n}\right)^{-x} \\
&= n(n-1)\cdots(n-x+1)\frac{1}{x!}(n\pi)^x\left(\frac{1}{n}\right)^x\left(1-\frac{n\pi}{n}\right)^n\left(1-\frac{n\pi}{n}\right)^{-x} \\
&= \frac{n(n-1)\cdots(n-x+1)}{n^x}(1-\pi)^{-x}\frac{(n\pi)^x}{x!}\left(1-\frac{n\pi}{n}\right)^n.
\end{aligned} \tag{13}$$

Since

$$\lim_{n\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x} = 1$$

$$\lim_{\pi\to 0}(1-\pi)^{-x} = 1$$

$$\lim_{\substack{n\to\infty \\ n\pi\to\lambda}}\left(1-\frac{n\pi}{n}\right)^n = e^{-\lambda},$$

and obviously

$$\lim_{n\pi\to\lambda}\frac{(n\pi)^x}{x!} = \frac{(\lambda)^x}{x!}$$

we have the fact that in the limit as as $n \to \infty$, $\pi \to 0$, and $n\pi \to \lambda$, the binomial probability function in 13 is equal to

$$\frac{e^{-\lambda}\lambda^x}{x!}$$

which is the Poisson probability function.

If a random variable X follows the Poisson distribution with parameter λ we write $X \sim Pois(\lambda)$.

Two useful facts are worth noting about sums of independent Poisson random variables. If $X_1, X_2, \ldots, X_k$ are independent random variables that are distributed $X_i \sim Pois(\lambda_i), \quad i = 1, 2, \ldots, k$ it can be shown that

$$\sum_{i=1}^{k} X_i \sim Pois\left(\sum_{i=1}^{k} \lambda_i\right).$$

Second, conditional on the sum of the realized counts $n = \sum_{i=1}^{k} x_i$, it can be shown that the vector $(X_1, X_2, \ldots, X_k)$ follows the multinomial distribution:

$$(X_1, X_2, \ldots, X_k)|n \sim Multin(n, \ (\pi_1, \pi_2, \ldots, \pi_k))$$

where $\pi_i = \lambda_i / (\sum_{j=1}^{k} \lambda_j), \quad i = 1, 2, \ldots, k$.

Figure 5 shows the Poisson probability function when $\lambda = 5$ and figure 6 shows the Poisson distribution function when $\lambda = 5$.

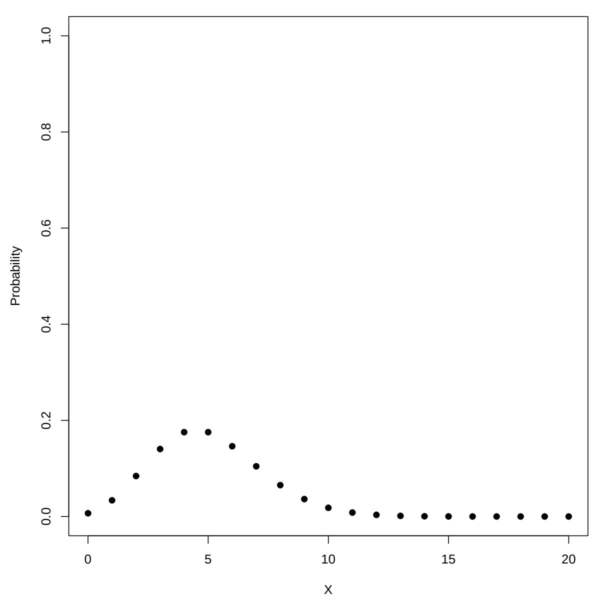

Figure 5: Poisson probability function for $\lambda = 5$.

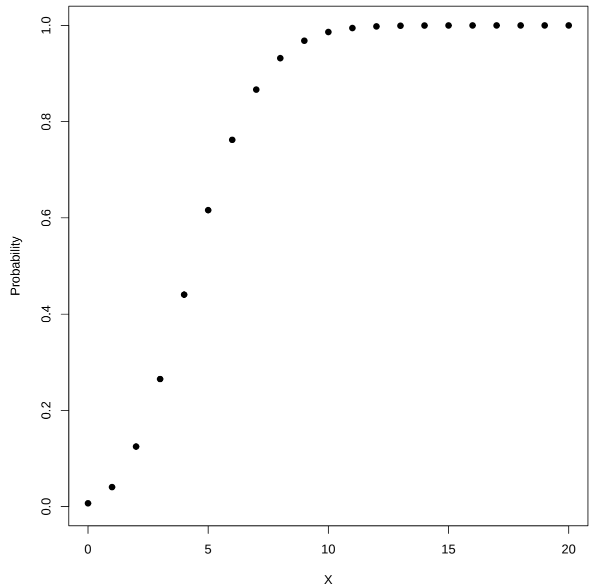

Figure 6: Poisson distribution function for $\lambda = 5$.

References

DeGroot, Morris H. 1986. *Probability and Statistics*. Reading, MA: Addison Wesley, second edition.

Feller, William. 1968. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, third edition.

Gallant, A. Ronald. 1997. *An Introduction to Econometric Theory*. Princeton: Princeton University Press.