The Binomial Distribution

Reading:

- Daniel, W. Biostatistics.1999 John Wiley and Sons (pp89-99)

Introduction

The Binomial Distribution is the probability distribution of a discrete random variable. The value of the random variable corresponds to the number of successes in a series of independent experiments, where each experiment consists of flipping a single coin once. For example, let X =the number of heads that results from flipping a coin n=10 times. Then the random variable X will follow a Binomial Distribution. The distribution describes the probability that each possible outcome of X will occur. The simplest Binomial distribution results from a single flip of a coin. Such an experiment is called a Bernoulli trial. The random variable corresponding to the number of successes is called a Bernoulli random variable.

Examples

Suppose we have an 'experiment' with just 2 possible outcomes. Common examples are:

- pass/fail exam
- win/lose a game
- heads/tails on coin toss
- person included in sample smokes/does not smoke
- live/die after hospitalization

Although whether you pass, or fail an exam is not a random event, we will consider it random in the context of the 'experiment'. For example, assume that the proportion of students who pass is known, say 90%. The experiment consists of randomly selecting a student. If the selected student passed, the outcome, x=1. If the selected student fails, the outcome x=0.

Even a process with many outcomes and can be simplified to fit this situation, if we focus on one particular outcome vs. "not that outcome", or group responses into 2 possible categories:

- roll a '1' / any other number on a die
- person included in sample is < age 50 / age 50+
- Birth weight < 2500 g / $\geq$2500 g

The Bernoulli Trial and Binomial Distribution

An experiment that consists of a single flip of a coin, or a single classification is called a Bernoulli trial. If the experiment is repeated (which we

call a trial), and the repetitions are independent, then the probability distribution of the random variable

X = # of successes out of n independent Bernoulli trials

is called a Binomial Distribution. Such a distribution occurs when:

1. The result of each trial is one of 2 outcomes, often referred to as a "success" and a "failure".
2. The probability P of success is the same in every trial.
3. The trials are independent – the outcome of one trial has no influence on the outcome of another trial.

Studying the Binomial Distribution

The Binomial Distribution is simply a discrete probability distribution. We can study the distribution by writing the possible outcomes in the sample space, and determining their probability. We start with a simple example where a coin is tossed twice. We then consider tossing a coin n=3 times. This leads us to try to generalize about the probabilities of an outcome if the coin was tossed n=4 times, or even more times.

Example 1: Let an experiment consist of tossing a coin n=2 times, where we assume the tosses are independent. Assume the coin is fair, so that P(H)=0.5=probability of a head on a toss.

Let us represent the outcome of the two tosses as {1^{st} toss result, 2^{nd} toss result}.

Sample Space: ({HH}, {HT}, {TH}, {TT}).

Since the tosses are independent :

P({HH})=P(H on first toss)P(H on second toss)=0.25

We define the random variable X to equal the number of heads observed. Then:

Event x	Outcomes	P(X=x)	P(X≤x)
0	{TT}	0.25	0.25
1	{HT, TH}	0.50	0.75
2	{HH}	0.25	1.00
Sum:		1	

This distribution is called a Binomial distribution with n=2, P=0.5. The last column given the cumulative distribution.

Example 2: Let an experiment consist of selecting n=2 students from a class, and observing whether they received an A on an exam. Assume whether or not the second selected student received an A does not depend on the result for the 1st selected student (the results are independent). Also assume that the probability of receiving an A is P(A)=0.2. Determine the Binomial Distribution for the number of A's.

Let us represent the outcome of the two selections as {1st selection result, 2nd selection result}. We represent the grade A by the letter A, and 'not an A' by the letter B.

Sample Space: ({AA}, {AB}, {BA}, {BB}).

Since the tosses are independent :

P({AA})=P(A on selection)P(A on second selection)=0.04
P({AB})=P(A on selection)P(B on second selection)=0.16
P({BA})=P(B on selection)P(A on second selection)=0.16
P({AA})=P(B on selection)P(B on second selection)=0.64

We define the random variable X to equal the grade A's observed. Then:

Event

x	Outcomes	P(X=x)	P(X≤x)
0	{BB}	0.64	0.64
1	{AB, BA}	0.32	0.32
2	{AA}	0.04	1.00
Sum:		1	

This distribution is called a Binomial distribution with n=2, P=0.2.

Example 3: Let an experiment consist of selecting n=3 records at random from a hospital emergency room, and seeing whether the patient had health insurance. Assume the selections were Bernoulli trials with the probability of having health insurance given by P=0.6. Also, let Q=0.4 = Probability that a patient does not have health insurance. Let us represent the outcome of the three selections as

{1^{st} selection result, 2^{nd} selection result, 3^{rd} selection result}.

Finally, let:

X=the number of patients with health insurance,

and let us represent the events:

Y= the patient has health insurance
N= the patient does not have health insurance.

Determine the Binomial Distribution for X.

Since the tosses are independent :
P(YYY)=P(Y) P(Y) P(Y) =PPP=(0.6)(0.6)(0.6)=0.216
P(YYN)=P(Y) P(Y) P(N) =PPQ=(0.6)(0.6)(0.4)=0.144
P(YNY)=P(Y) P(N) P(Y) =PQP=(0.6)(0.4)(0.6)=0.144
P(NYY)=P(N) P(Y) P(Y) =QPP=(0.4)(0.6)(0.6)=0.144
P(YNN)=PQQ=(0.6)(0.4)(0.4)=0.096
etc.

We first list all possible outcomes in the sample space, and the corresponding number of patients without health insurance.

x	Outcomes	P(X=x)
0	{NNN}	0.064
1	{YNN}	0.096
1	{NYN}	0.096
1	{NNY}	0.096
2	{YYN}	0.144
2	{YNY}	0.144
2	{NYY}	0.144
3	{YYY}	0.216

x	Outcomes	P(X=x)	P(X≤x)
0	{NNN}	0.064	0.064
1	{YNN},{NYN},{NNY}	3(0.096)=0.288	0.352
2	{YYN},{YNY},{NYY}	3(0.144)=0.432	0.784
3	{NNY}	0.216	1.000

Example 4a. Suppose a killing frost occurs prior to Sept. 10th about once every 10 years. In the next 5 years, how many times will there be a killing frost prior to Sept 10th? Let

X= # of years with a killing frost in the next 5 years.

Assume that X is Binomial Distributed, with n=5, P=0.10, Q=0.9. What is the distribution of X? (Let Y=Killing frost, N=no frost.)

Can we generalize the results? There are two factors making up the probabilities.

A. One factor is the probability of an outcome.

B. The other factor is the number of different outcomes that will result in the event.

A. Probability of the Outcome

For the outcome:
{YNYNN} then X=2, P(YNYNN)=PQPQQ=(PP)(QQQ)= P^2Q^3
{NNYYN} then X=2, P(NNYYN)=QQPPQ=(PP)(QQQ))= P^2Q^3

In general, with n trials, the probability of an outcome is $P^x Q^{(n-x)} = P^x (1-P)^{(n-x)}$

B. How many different outcomes can result in the event? We will show that the event can occur ${}_nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$ ways.

Combining these ideas, the <u>Binomial probabilities</u> are given by:

$$P(X=x) = {}_n C_x \left(P^x (1-P)^{(n-x)} \right) = \binom{n}{x} P^x (1-P)^{(n-x)} = \frac{n!}{x!(n-x)!} P^x (1-P)^{(n-x)}$$

Terminology

Permutation: The number of different ways that n district objects can be ordered. This number is equal to n factorial, n!.

Factorial Notation: N! is read N factorial and is equal to:
N! = N(N-1)(N-2)(N-3)…(4)(3)(2)(1). For example, 3!=3(2)(1)=6

Note: By definition 0!=1.

Combinations: The number of different ways that groups of objects can be ordered, ignoring the ordering of objects in a group.

Combination Notation: The number of combinations formed by two groups of objects is represented by $_nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$, where there are x objects in a group among n objects.

Example 1a: Suppose there are 4 chairs in the first row in a classroom. Four students enter the class, and each sits in one of the chairs. How many different ways can the students sit in the chairs? This number is the number of permutations.

Let us refer to the students as: A, B, C, and D, and list all permutations.

		Chair			
Number	1	2	3	4	Representation
1	A	B	C	D	ABCD
2	A	B	D	C	ABDC
3	A	C	B	D	ACBD
4	A	C	D	B	ACDB
5	A	D	B	C	ADBC
6	A	D	C	B	ADCB
7	B	A	C	D	BACD
8	B	A	D	C	BADC
9	B	C	A	D	BCAD
10	B	C	D	A	BCDA
11	B	D	A	C	BDAC
12	B	D	C	A	BDCA
13	C	B	A	D	CBAD
14	C	B	D	A	CBDA
15	C	A	B	D	CABD
16	C	A	D	B	CADB
17	C	D	A	B	CDAB
18	C	D	B	A	CDBA
19	D	B	A	C	DBCA
20	D	B	C	A	DBAC
21	D	C	A	B	DCBA
22	D	C	B	A	DCAB
23	D	A	B	C	DABC
24	D	A	C	B	DACB

There are 24 permutations, or orderings. To see how we arrive at this number, consider the number of way a different person can sit in the first chair, after taking that person out, the number of ways a person can sit in the next chair, etc. Thus.

Permutations: 4! =	4	3	2	1	=24
Chairs:	1	2	3	4	

Example 2. Suppose there are 4 chairs in the first row in a classroom. Five students enter the class, and four of the students sit in one of the chairs in the first row. How many different ways can the students sit in the chairs?

Let us refer to the students as: A, B, C, D, and E,

Using a similar argument as for example 1, consider the number of ways a different student can sit in a chair, one chair at a time.

# of possible different students:	5	4	3	2	=120
Chairs:	1	2	3	4	

We can express this as $120 = \frac{5!}{(5-4)!}$

Example 3. In the Spring 2002 class of BioEpi540, there are 7 chairs at the front of the class, and 37 students. How many different ways could the students be sitting at the front of the class?

# of possible different students:	37	36	35	34	33	32	31=51,889,178,880
Chairs:	1	2	3	4	5	6	7

We can express this as $\frac{37!}{(37-7)!}$.

If there is a total of 50 classes per year, how many years would it take for each of the student arrangements to occur once if a different arrangement occurs each class?
(over a million years)

Example 1b: Suppose there are 4 chairs in the first row in a classroom. Four students enter the class, three women (W) and 1 man (m) and each sits in one of the chairs. How many different ways can gender be arranged on the chairs?

Let us refer to the students as: A, B, C, and D,
and assume the gender is given by : W, W,W, and m

	Chair					
Number	1	2	3	4	Subjects	Genders
1	A	B	C	D	ABCD	WWWm
2	A	B	D	C	ABDC	WWmW
3	A	C	B	D	ACBD	WWWm
4	A	C	D	B	ACDB	WWmW
5	A	D	B	C	ADBC	WmWW
6	A	D	C	B	ADCB	WmWW
7	B	A	C	D	BACD	WWWm
8	B	A	D	C	BADC	WWmW
9	B	C	A	D	BCAD	WWWm
10	B	C	D	A	BCDA	WWmW
11	B	D	A	C	BDAC	WmWW
12	B	D	C	A	BDCA	WmWW
13	C	B	A	D	CBAD	WWWm
14	C	B	D	A	CBDA	WWmW
15	C	A	B	D	CABD	WWWm
16	C	A	D	B	CADB	WWmW
17	C	D	A	B	CDAB	WmWW
18	C	D	B	A	CDBA	WmWW
19	D	B	A	C	DBCA	mWWW
20	D	B	C	A	DBAC	mWWW
21	D	C	A	B	DCBA	mWWW
22	D	C	B	A	DCAB	mWWW
23	D	A	B	C	DABC	mWWW
24	D	A	C	B	DACB	mWWW

There are 4 ways gender can be arranged, with the man seated in chair 1, 2, 3, or 4. If we count all the permutations, we will over-count the district arrangements of men and women.

To see this, compare the representation for seating pattern given below:

Genders	Subjects
WWWm	ABCD
WWWm	ACBD
WWWm	BACD
WWWm	BCAD
WWWm	CABD
WWWm	CBAD

The same gender pattern was reported for 6 subject patterns. The number "6" corresponds to the number of ways the 3 women could be permuted, 6=3!.

To determine the number of ways two groups of objects, can be combined, we divide the number of permutations of the total number of objects, n!, by the number of permutations for each group, x! and (n-x)!. In this example,

# of combinations of gender = $ {}_4C_1 = \binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4(3)(2)(1)}{1(3)(2)(1)} = 4$

Example 4a. Suppose a killing frost occurs prior to Sept. 10th about once every 10 years. In the next 5 years, how many times will there be a killing frost prior to Sept 10th? Let

X= # of years with a killing frost in the next 5 years.

Assume that X is Binomial Distributed, with n=5, P=0.10, Q=0.9. What is the distribution of X? (Let Y=Killing frost, N=no frost.)

The binomial probabilities are given by: $P(X=x) = \frac{n!}{x!(n-x)!} P^x (1-P)^{(n-x)}$. Then

when x=0, $P(X=0) = \frac{5!}{0!(5-0)!}(0.1)^0 (0.9)^{(5)} = \frac{5!}{5!}(0.9)^{(5)} = 0.59049$

when x=1, $P(X=1) = \frac{5!}{1!(5-1)!}(0.1)^1 (0.9)^{(4)} = \frac{5!}{1!4!}(0.1)(0.9)^{(4)} = 0.32805$

when x=2, $P(X=2) = \frac{5!}{2!(5-2)!}(0.1)^2 (0.9)^{(3)} = \frac{5!}{2!3!}(0.1)^2 (0.9)^{(3)} = 0.0729$

when x=3, $P(X=3) = \frac{5!}{3!(5-3)!}(0.1)^3 (0.9)^{(2)} = \frac{5!}{3!2!}(0.1)^3 (0.9)^{(2)} = 0.0081$

when x=4, $P(X=4) = \frac{5!}{4!(5-4)!}(0.1)^4 (0.9)^1 = \frac{5!}{4!1!}(0.1)^4 (0.9)^1 = 0.00045$ and finally,

when x=5, $P(X=5)=\frac{5!}{5!(5-5)!}(0.1)^5(0.9)^0=\frac{5!}{5!0!}(0.1)^5=0.00001$

# Killing Frosts (x)	P(x)	P(X≤x)
0	0.59049	0.5905
1	0.32805	0.9185
2	0.0729	0.9914
3	0.0081	0.9995
4	0.00045	1.0000
5	0.00001	1.0000

The cumulative distribution is tabulated in Daniel, (1999), Table B., page A3.

Summary of Binomial Distribution

In general, the probability of obtaining *x* successes out of n trials, with probability P of success on each trial is:

$$P(X=x)={}_nC_xP^x(1-P)^{(n-x)}$$

${}_nC_x$ is the number of combinations, or ways of arranging n items, x of one type (successes), and (n-x) of another type (failures).

${}_nC_x=\binom{n}{x}=\frac{n!}{x!(n-x)!}$, where n!=n(n-1)(n-2)…(1) , and 0!=1

P^x is the probability of observing x successes – the probability of a success (P) on one trial, times the probability of success on another trial, times…, *x* times.

$(1-P)^{(n-x)}$ is the probability of observing (n-x) failures. The probability of a failure on any one trial is (1-P), and this will happen (n-x) times in n trials.

Example: Toss a coin n=2 times. Assume the coin is fair. Compute the Binomial Distribution.

When n=2 and p=.5, we can compute the probability of zero heads as:

$P(X=0) = {}_2C_0(.5)^0(1-.5)^{2-0} = (2!/0!2!)\ (1)\ (.5)^2 = .5^2 = .25$

${}_2C_0$ tells us the number of ways we can observe zero heads: There is just 1 way (both tails), and (2!/0!2!) =1.

We can compute the probability of observing exactly 1 head in 2 tosses as:

$P(X=1) = {}_2C_1(.5)^1(1-.5)^{2-1} = 2(.5)(.5) = .50$

${}_2C_1$ is the number of ways of observing 1 head in two tosses: HT or TH, or 2 ways. 2!/1!(2-1)! =2(1)/(1)(1) = 2.

Example : Suppose a coin is tossed n=3 times. What is the probability of obtaining exactly 2 heads?

$P(X=2) = {}_3C_2(.5)^2(1-.5)^{3-2} = 3(.5)^2(.5) = 3(.125) = .375$

${}_3C_2$ is the number of ways of observing 2 heads in three tosses: HHT, HTH, or THH, or 3 ways. 3!/2!(3-2)! =3(2)(1)/2(1)(1) = 3.

Example 2: Suppose we know that 40% of a certain population are cigarette smokers. If we take a random sample of 10 people from this population, what is the probability that we will have exactly 4 smokers in our sample?

If we assume that the probability that any individual in the population is a smoker to be P=.40, then the probability that x=4 smokers out of n=10 subjects selected is:

$P(X=4) = {}_{10}C_4(.4)^4(1-.4)^{10-4} = {}_{10}C_4(.4)^4(.6)^6 = 210(.0256)(.04666)$
$= .2508$

or the probability of obtaining exactly 4 smokers in the sample is about 25%.
Note:
${}_{10}C_4$=10!/4!(10-4)! = 10(9)(8)(7)…1 / 4(3)(2)(1)(6)(5)(4)…1= 210

We can compute the probability of observing zero smokers out of 10 subjects selected at random, exactly 1 smoker, and so on, and display the results in a table, as given, below.

Binomial Probability Distribution for n=10, p=.4

x	P(X=x)	P(X≤x)
0	.0060	.0060
1	.0404	.0464
2	.1209	.1673
3	.2150	.3823
4	.2508	.6331
5	.2007	.8338
6	.1114	.9452
7	.0452	.9877
8	.0106	.9983
9	.0016	.9999
10	.0001	1.000

The third column, $P(X \leq x)$, gives the cumulative probability. For example the probability of selecting 3 or fewer smokers into the sample of 10 subjects is $P(X \leq 3) = .3823$, or about 38%. The advantage of having a formula to define a distribution, is that we can specify the complete distribution for any n (size of the trial) and p (probability of success) that we desire.

A table of cumulative binomial probabilities is given in the appendix of your text, Daniel, Table B, lists Cumulative Binomial Distributions. On page A-9, you will find the cumulative distributions for n=10, and in the lower right corner, specifically for p=.40, which matches the third column of our example, above. Note that only the cumulative distribution is listed. To compute the probability of observing exactly 4 smokers in a sample of n=10, p=.4, we need to take a difference:

$$P(X=4) = P(X \leq 4) - P(X \leq 3) = .6331 - .3823 = .2508 ,$$

just what we computed before.

We can also ask questions such as: What is the probability of observing more than 5 smokers in a sample of 10?

$$P(X>5) = 1 - P(X \leq 5) = 1 - .8338 = .1662,$$

or about a 16.6% chance of having more than 5 smokers in a sample of n=10, from a population with p=40% smokers.

Using the Computer to get Binomial Probabilities

Parameters for the Binomial distribution are n and P, since they provide the necessary information to specify a distribution. There are, in fact, an infinite number of binomial distributions, each determined by specifying n and p. Clearly not all of these can be tabulated – we can go to the computer to determine the probability of any specified outcome of interest. Most statistical software will compute binomial probabilities. See page 98-99 in Daniel for an illustration of how to calculate Binomial Probabilities in Minitab.

The Mean and Variance of a Binomial Distribution

Once n and P are specified, we can compute the proportion of success, $\hat{P} = \frac{x}{n}$, and the mean and variance of the distribution. These are given by :

$\mu = P$ and $\text{var}(\hat{P}) = \frac{P(1-P)}{n}$

Example 1: Toss a coin twice and observe the number of heads. What is the mean or "expected" number of heads?

Let X=# heads observed, n=2, and P=.5, the probability of observing heads on a single toss.

Then $\mu = P = 0.5$ and $\text{var}(\hat{P}) = \frac{P(1-P)}{n} = \frac{0.5(0.5)}{2} = 0.125$

Example 3: 70% of a certain population has been immunized for polio. If a sample of size 50 is taken, what is the "expected total number", in the sample who have been immunized?

X=# immunized, n=50, P=.70

$n\mu = nP = 50(.70) = 35$

This tells us that "on the average" we expect to see 35 immunized subjects in a sample of 50 from this population.

Note that the probability of observing *exactly* 35 is not large: P(X=35) = $_{50}C_{35}(.7)^{35}(1-.7)^{50-35}$ = .122, or about 12%

The Poisson Distribution

One other commonly used probability distribution for discrete data is the Poisson Distribution. This distribution is applicable for counts of events in time or space, for example:

number of patients arriving at an emergency department in a day
number of new cases of HIV diagnosed at a clinic in a month

In such cases, we might take a sample of days and observe the number of patients arriving at the emergency department on each day, or a sample of months and observe the number of new cases of HIV diagnosed at the clinic. We are observing a count or number of events, rather than a yes/no or success/failure outcome for each subject or trial, as in the binomial.

We will not discuss this distribution further, except to note that it is covered in Section 4.4 in your text, Daniel, for those of you who might come across data of this type.