Binomial Probability Distributions

The probability of an event is its expected proportion in the long run or in the population. For example, an event will happen half the time (such as a head showing up on the flip of a fair coin) has probability 50%. If a horse will win its race 1 in every 4 starts, its probability of a win is 25%. If 4 in 5 patients will survive five or more years, the probability of survival is 80%.

Probability can occasionally be derived by logic if we can count the number of ways a thing can happen and then determine the relative frequency of the event we are interested in. To use a familiar example, we recognize that there are 52 cards in a deck, 4 of which are Kings. Therefore, the probability of drawing a King at random = 4 ÷ 52 = .0769.

Probability converge on a relative frequency. If event A occurs x of these times, then the probability of the event will *converge* on $X \div n$ as n becomes large. For example, if we flip a coin many times, we expect to see half the flips turn up heads. This experiment is unreliable when N is small, but becomes increasingly reliable as N increases. For example, if a coin is flipped 10 times, there is no guarantee that we will observe exactly 5 heads -- the proportion of heads will range from 0 to 1. (In most cases we would expect it to be closer to .50 than to 0 or 1.) However, if the coin is flipped 100 times, chances are better that the proportions of heads will be close to .50. With 1000 flips, the proportion of heads will be an even better reflection of the true probability. And so on.

Probability can be used to quantify belief: Zero represents the certainty something will *not* occur; 1 represents the certainty that it will occur. For instance, if an experienced clinician says "you have a 50% chance of recovery," he or she believes that half of similar cases will turn up positive an half will turn up negative. Presumably, this is based on experience and knowledge, and not on a whim. The benefit of stating the subjective probability is that it can be tested and modified as experience warrants.

Again: appreciation of different conceptions probability are not mutually exclusive. All have similar properties, and all are founded on a mathematical expectation.

Selected Properties of Probability

We must address certain properties of probabilities before we can work effectively with them. Consider the following:

Range of possible probabilities: Probabilities can be no less than 0% (indicating certainty something will *not* occur) and no more than 100% (indicating certainty something *will* occur).

Notation: We speak of the probability of events. For example, event "A" might represent recovery following treatment. Then, Pr(A) would represent the probability of recovery following treatment.

Complement: The complement of an event is its "opposite," i.e., the event *not* happening. For example, if event A is recovery following treatment, then the complement of event A is not recovering following treatment. Let $\bar{A}$ represent the complement of event A.

Law of complements: Since, either A or $\bar{A}$ must occur, then $\Pr(A) + \Pr(\bar{A}) = 1$. It follows that $\Pr(\bar{A}) = 1 \quad \Pr(A)$. For example, if $\Pr(A) = .75$, then $\Pr(\bar{A}) = 1 - .75 = .25$. Sometimes we will use the notation p to represent the probability of an event and q to represent the probability of its complement. Notice that $q = 1 - p$.

Random variable: A random variable is a quantity that varies depending on chance. There are two types of random variables:

1. discrete random variables
2. continuous random variables.

Discrete random variables a countable events that can take on only a finite number of possible outcomes. For example, we might consider the number of people who recover following treatment. If n people are treated, possible outcomes could vary from 0, 1, 2, ..., n.

Continuous random variables form an unbroken chain of possible outcomes and can take on an infinite number of possibilities. For example, we can consider the weight of a baby selected at random from the population. The baby may weight 8 pounds or 9 pounds *or anything in between*. There are even an infinite number of possibilities between 8 and 9. consider Zeno's paradox. In order to get from 8 to 9, you must get halfway the distance, to 8.5. In order to get from 8.5 to 9, you must again get half the distance, to 8.75. To get from 8.75 to 9, you must first get half-way there, to 8.875. (And so on, *ad nauseam*).

Initially, we will study **binomial random variables** as an example of a discrete random variable. We will study Normal random variables as an example of a continuous random variable.

The Binomial Distributions

Consider a random event that can take on only one of two possible outcomes. Each event is a **Bernoulli trial**.

Arbitrarily, define one outcome a "success" and the other a "failure." Now take a series of n independent Bernoulli trials. This is a **binomial random variable.** Binomial random variable are extremely common in public health statistics. For example, we may be interested in the number of people in a sample of n people who respond favorably to treatment.

Illustrative Example ("Four Patients"). Suppose a treatment is successful 75% of the time. This treatment is used in 4 patients. We can assume that *on the average* 3 of the 4 patients will respond to treatment. However, it would be foolish to think 3 of every 4 patients will respond. The number of people responding favorably to treatment will vary from set of set. Sometimes, 0, patients will respond. Sometimes, 1, 2, 3, or 4 patients may respond. The number of patients out of 4 responding is a binomial random variable because it is based on n independent Bernoulli trials.

The full enumeration of probabilities for all possible outcomes of a discrete random variable is probability mass distribution for this random variable. This can be expressed in table form. For the current illustration, the binomial probability mass distribution is:

Number of successes	Probability
0	.0039
1	.0469
2	.2109
3	.4219
4	.3164

How do we interpret this table? This table states the probability of observing 0 successes out of 4treatments is .0039, the probability of observing 1 success out of 4 treatments is .0469 and so on.
We will learn how to calculate binomial probabilities.

Binomial distributions are a family of distributions, with each family member identified by two parameters. The binomial **parameters** are:

n ≡ the number of independent trials
p ≡ the probability of success for each trial

We may use the notation X~b(n, p) to denote a given binomial random variable. For the "Four Patients Illustrative example, $X \sim$ b(4, 0.75).

Combinatorics

Before calculating binomial probabilities we must first learn the **combinatorics ("choose") function**. The combinatorics function tells us how many different we can choose i items out of n. Let ${}_nC_i$ denote the number of ways to choose i items out of n. Then,

$$ {}_nC_i = \frac{n!}{i!(n-i)!} \qquad \textbf{(4.1)} $$

where **!** represents the **factorial function**. The factorial function is the product of the series of integers from n to 1. In symbols, $n! = (n)(n \quad 1)\ (n \quad 2)(n \quad 3) \ldots (1)$. For example, 3! = (3)(2)(1) = 6.As another example, 4! = (4)(3)(2)(1) = 24. By definition, 1! = 1 and 0! = 1.

Illustrative example, combinatorics #1. "How many ways are there to choose 2 items out of 3?" By formula, ${}_3C_2 = \frac{3!}{(2!)(3-2)!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1)(1)} = \frac{6}{2} = 3$. There are three ways to choose 2 items out of 3. This makes sense when you consider 3 items labeled A, B, and C. There are three different sets of two. You may choose {A, B}, {A, C}, or {B, C}.

Illustrative example, combinatorics #2."How many ways are there to choose 2 items out of 4?" By formula, ${}_4C_2 = \frac{4!}{(2!)(4-2)!} = \frac{4 \cdot 3 \cdot 2!}{(2 \cdot 1)(2!)} = \frac{4 \cdot 3}{2} = 6$. Thus, there are 6 ways to choose 2 items out of 4. This makes sense when you consider 4 items labeled A, B, C, and D. There are six sets of two: {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, and {C, D}.

Illustrative example, combinatorics #3."How many ways are there to choose 3 items out of 7?" By formula, ${}_7C_3 = \frac{7!}{(3!)(7-3)!} = \frac{7 \cdot 6 \cdot 5 \cdot 4!}{(3 \cdot 2 \cdot 1)(4!)} = 35$. There are 35 ways to choose 3 items out of 7.

Binomial Formula

We are now ready to calculate binomial probabilities. The **binomial formula** is:

$$\Pr(X=i)={}_{n}C_{i}p^{i}q^{n-i} \qquad \textbf{(4.2)}$$

where X represents the random number of successes, i is the observed number of successes, n is the number of trials, p is the probability of success for each trial, and $q = 1 \quad p$.

Illustrative Example (Four Patients, Probability of 2 successes). Recall the example that considers a treatment that is successful 75% of the time (p = .75). The treatment is used in 4 patients (n = 4). "What is the probability of seeing 2 successes in 4 patients?" ANS: n = 4, i = 2, p = .75, and q = 1 − .75 = .25. Thus, $\Pr(X = 2) = {}_4C_2\,(.75)^2\,(.25)^{4-2}$ = (6)(.5625)(.0625) = .2109.

The listing of probabilities for all possible outcomes for a discrete random variable is called a **probability mass function.**

Illustrative Example (Four Patients Example, Probability Mass Function). For a treatment that is successful 75% of the time used in 4 patients, the binomial probability mass function is:

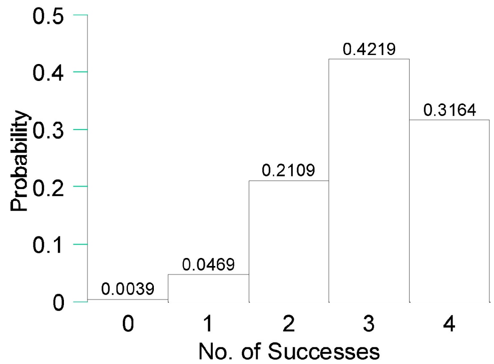
The probability of 0 successes ≡ $\Pr(X = 0) = {}_4C_0\,(.75)^0\,(.25^{4-0})$ = (1)(1)(.0039) = .0039
The probability of 1 success ≡ $\Pr(X = 1) = {}_4C_1\,(.75)^1\,(.25^{4-1})$ = (4)(.75)(.0156) = .0469
The probability of 2 successes ≡ $\Pr(X = 2) = {}_4C_2\,(.75)^2\,(.25^{4-2})$ = (6)(.5625)(.0625) = .2109
The probability of 3 successes ≡ $\Pr(X = 3) = {}_4C_3\,(.75)^3\,(.25^{4-3})$ = (4)(.4219)(.25) = .4219
The probability of 4 successes ≡ $\Pr(X = 4) = {}_4C_4\,(.75)^4\,(.25^{4-4})$ = (1)(.3164)(1) = .3164

In tabular form:

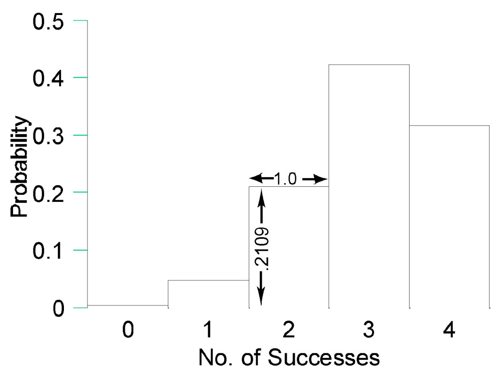
Number of success i	Probability of Occurrence $\Pr(X = i)$
0	.0039
1	.0469
2	.2109
3	.4219
4	.3164

Probability Histogram

Probability histogram. Binomial mass functions can be displayed as histograms with the Y-axis representing probability and the X-axis representing values for the random variable. Here is the probability histogram for the "Four Patients Illustration:"



Area under the bars. The area of each bar in a probability histogram represents probability. For example, the bar corresponding to 2 out of 4 successes has a width of 1.0 and height of .2109. The area of the bar = height × width = 1× .2109 = .2109, which is equal to the probability of observing 2 successes.



This "area under the bars" concept is important in practice.

Cumulative Probability

Cumulative probability. The cumulative probability of a binomial outcome is the probability of observing less than or equal to a given number of successes. For example, the cumulative probability of 2 successes is the probability of observing 2 or fewer successes, i.e., $\Pr(X \le 2)$. This is equal to the sum of the probabilities for 0, 1, or 2 successes.

Illustrative example (Four Patients Example, Cumulative Probability of Two Successes). For our illustrative example, $\Pr(X \le 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) = .0039 + .0469 \ .2109 = .2617$.

Cumulative probability distribution. The cumulative probability distribution is the compilation of cumulative probabilities for all possible outcomes. The cumulative probability distribution for the Four Patients Illustration is:

$\Pr(X \le 0) = .0039$
$\Pr(X \le 1) = [\Pr(X = 0) + \Pr(X = 1)] = 0.0039 + 0.0469 = 0.0508$
$\Pr(X \le 2) = [\Pr(X \le 1) + \Pr(X = 2)] = 0.0508 + 0.2109 = 0.2617$
$\Pr(X \le 3) = [\Pr(X \le 2) + \Pr(X = 3)] = 0.2617 + 0.4219 = 0.6836$
$\Pr(X \le 4) = [\Pr(X \le 3) + \Pr(X = 4)] = 0.6836 + 0.3164 = 1.0000$

Cumulative probability table. It is convenient to list probabilities and cumulative probabilities in tabular form. Here are the cumulative probabilities for the "Four Patient Illustration:"

No. of successes i	Probability $\Pr(X = i)$	Cumulative Probability $\Pr(X \le i)$
0	.0039	.0039
1	.0469	.0508
2	.2109	.2617
3	.4219	.6836
4	.3164	1.0000

Area under the bars to the left. Cumulative probabilities corresponds to the area under the bars to the LEFT of a given point. The shaded region in the figure below, for instance, corresponds to $\Pr(X \le 2)$ for the "Four Patients Illustration":