The Binomial Distribution

A. It would be very tedious if, every time we had a slightly different problem, we had to determine the probability distributions from scratch. Luckily, there are enough similarities between certain types, or families, of experiments, to make it possible to develop formulas representing their general characteristics.

For example, many experiments share the common element that their outcomes can be classified into one of two events, e.g. a coin can come up heads or tails; a child can be male or female; a person can die or not die; a person can be employed or unemployed. These outcomes are often labeled as "success" or "failure." Note that there is no connotation of "goodness" here - for example, when looking at births, the statistician might label the birth of a boy as a "success" and the birth of a girl as a "failure," but the parents wouldn't necessarily see things that way. The usual notation is

p = probability of success,
q = probability of failure = 1 - p.

Note that p + q = 1. In statistical terms, **A Bernoulli trial is each repetition of an experiment involving only 2 outcomes**.

We are often interested in the result of *independent, repeated bernoulli trials*, i.e. the number of successes in repeated trials.

1. *independent* - the result of one trial does not affect the result of another trial.

2. *repeated* - conditions are the same for each trial, i.e. p and q remain constant across trials. Hayes refers to this as a stationary process. If p and q can change from trial to trial, the process is nonstationary. The term identically distributed is also often used.

B. A binomial distribution gives us the probabilities associated with independent, repeated Bernoulli trials. **In a binomial distribution the probabilities of interest are those of receiving a certain number of successes, r, in n independent trials each having only two possible outcomes and the same probability, p, of success.** So, for example, using a binomial distribution, we can determine the probability of getting 4 heads in 10 coin tosses.

How does the binomial distribution do this? Basically, a two part process is involved. First, we have to determine the probability of one possible way the event can occur, and then determine the number of different ways the event can occur. That is,

P(Event) = (Number of ways event can occur) * P(One occurrence).

Suppose, for example, we want to find the probability of getting 4 heads in 10 tosses. In this case, we'll call getting a heads a "success." Also, in this case, n = 10, the number of successes is r = 4, and the number of failures (tails) is n – r = 10 – 4 = 6. One way this can occur is if the first 4 tosses are heads and the last 6 are tails, i.e.

S S S S F F F F F F

The likelihood of this occurring is

P(S) * P(S) * P(S) * P(S) * P(F) * P(F) * P(F) * P(F) * P(F) * P(F)

More generally, if p = probability of success and q = 1 – p = probability of failure, the probability of a specific sequence of outcomes where there are r successes and n-r failures is

$$p^r q^{n-r}$$

So, in this particular case, p = q = .5, r = 4, n-r = 6, so the probability of 4 straight heads followed by 6 straight tails is $.5^4 .5^6 = 0.0009765625$ (or 1 out of 1024).

Of course, this is just one of many ways that you can get 4 heads; further, because the repeated trials are all independent and identically distributed, each way of getting 4 heads is equally likely, e.g. the sequence S S S S F F F F F F is just as likely as the sequence S F S F F S F F S F. So, we also need to know how many different combinations produce 4 heads.

Well, we could just write them all out…but life will be much simpler if we take advantage of two counting rules:

1. The number of different ways that N distinct things may be arranged in order is

$$\mathbf{N! = (1)(2)(3)...(N-1)(N), \text{ (where } 0! = 1).}$$

An arrangement in order is called a permutation, so that the total number of permutations of N objects is N!. The symbol N! is called N factorial.

EXAMPLE. Rank candidates A, B, and C in order. The possible permutations are: ABC ACB BAC BCA CBA CAB. Hence, there are 6 possible orderings. Note that 3! = (1)(2)(3) = 6.

> NOTE: <u>Appendix E, Table 6, p. 19</u> contains a Table of the factorials for the integers 1 through 50. For example, $12! = 4.79002 * 10^8$. (Or see Hayes Table 8, p. 947). Your calculator may have a factorial function labeled something like x!

2. The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is

$$\frac{N!}{r!(N-r)!} = \binom{N}{r} = \binom{N}{N-r}$$

We refer to this as "N choose r." Sometimes the number of combinations is known as a *binomial coefficient*, and sometimes the notation $_NC_r$ is used. So, in the present example,

$$\binom{N}{r} = \binom{10}{4} = \frac{N!}{r!(N-r)!} = \frac{10!}{4!(10-4)!} = \frac{10*9*8*7}{4*3*2*1} = \frac{5040}{24} = 210$$

Note that, for 10!, I stopped once I got to 7; and I didn't write out 6! in the denominator. This is because both numerator and denominator have 6! in them, so they cancel out. So, there are 210 ways you can toss a coin 10 times and get 4 heads.

EXAMPLE. Candidates A, B, C and D are running for office. Vote for two.

The possible choices are: AB AC AD BC BD CD, i.e. there are 6 possible combinations. Confirming this with the above formula, we get

$$\frac{N!}{r!(N-r)!} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{12}{2} = 6$$

EXAMPLE. There are 100 applicants for 3 job openings. The number of possible combinations is

$$\frac{N!}{r!(N-r)!} = \binom{100}{3} = \frac{100!}{3!97!} = \frac{100*99*98}{3*2} = \frac{970{,}200}{6} = 161{,}700$$

Again, note that, if you didn't take advantage of 97! appearing on both top and bottom, you'd have a much lengthier calculation.

See Appendix E, Table 7, page 20 for $_NC_r$ values for various values of N and r. (Or see Hayes, Appendix E, Table IX, p. 948). Your calculator may have a function labeled nCr or something similar.

C. So putting everything together now: we know that any specific sequence that produces 4 heads in 10 tosses has a probability of 0.0009765625. Further, we now know that there are 210 such sequences. Ergo, the probability of 4 heads in 10 tosses is 210 * 0.0009765625 = 0.205078125.

We can now write out the complete formula for the binomial distribution:

In sampling from a stationary Bernoulli process, with the probability of success equal to p, the probability of observing exactly r successes in N independent trials is

$$\binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}$$

Once again, N choose r tells you the number of sequences that will produce r successes in N tries, while $p^r q^{N-r}$ tells you what the probability of each individual sequence is.

To put it another way, the random variable X in a binomial distribution can be defined as follows:

Let Xi = 1 if the ith bernoulli trial is successful, 0 otherwise. Then,

X = ΣX_i, where the X_i's are independent and identically distributed (iid).

That is, X = the # of successes. Hence**, Any random variable X with probability function given by**

$$p(X = r; N, p) = \binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}, \quad X = 0, 1, 2, ..., N$$

is said to have a binomial distribution with parameters N and p.

EXAMPLE. In each of 4 races, the Democrats have a 60% chance of winning. Assuming that the races are independent of each other, what is the probability that:

a. The Democrats will win 0 races, 1 race, 2 races, 3 races, or all 4 races?
b. The Democrats will win at least 1 race
c. The Democrats will win a majority of the races

SOLUTION. Let X equal the number of races the Democrats win.

a. Using the formula for the binomial distribution,

$$\binom{4}{0} p^0 q^{4-0} = \frac{4!}{0!(4-0)!} .60^0 .40^4 = .40^4 = .0256,$$

$$\binom{4}{1} p^1 q^{4-1} = \frac{4!}{1!(4-1)!} .60^1 .40^3 = 4 * .60 * .40^3 = .1536,$$

$$\binom{4}{2} p^2 q^{4-2} = \frac{4!}{2!(4-2)!} .60^2 .40^2 = 6 * .60^2 .40^2 = .3456,$$

$$\binom{4}{3} p^3 q^{4-3} = \frac{4!}{3!(4-3)!} .60^3 .40^1 = 4 * .60^3 .40^1 = .3456,$$

$$\binom{4}{4} p^4 q^{4-4} = \frac{4!}{4!(4-4)!} .60^4 .40^0 = .60^4 = .1296$$

b. P(at least 1) = P(X ≥ 1) = 1 - P(none) = 1 - P(0) = .9744. Or, P(1) + P(2) + P(3) + P(4) = .9744.

c. P(Democrats will win a majority) = P(X ≥ 3) = P(3) + P(4) = .3456 + .1296 = .4752.

EXAMPLE. In a family of 11 children, what is the probability that there will be more boys than girls? Solve this problem WITHOUT using the complements rule.

SOLUTION. You could go through the same tedious process described above, which is what most students did when I first asked this question on an exam. You would compute P(6), P(7), P(8), P(9), P(10), and P(11).

Or, you can look at Appendix E, Table II (or Hays pp. 927-931). Here, both Hayes and I list binomial probabilities for values of N and r from 1 through 20, and for values of p that range from .05 through .50.

Thus, on page E-5, we see that for N = 11 and p = .50,

P(6) + P(7) + P(8) + P(9) + P(10) + P(11) = .2256 + .1611 + .0806 + .0269 + .0054 + .0005 = .50.

NOTE: Understanding the tables in Appendix E can make things a lot simpler for you!

EXAMPLE. [WE MAY SKIP THIS EXAMPLE IF WE RUN SHORT OF TIME, BUT YOU SHOULD STILL GO OVER IT AND MAKE SURE YOU UNDERSTAND IT]

Use Appendix E, Table II, to once again solve this problem: In each of 4 races, the Democrats have a 60% chance of winning. Assuming that the races are independent of each other, what is the probability that:

a. The Democrats will win 0 races, 1 race, 2 races, 3 races, or all 4 races?
b. The Democrats will win at least 1 race
c. The Democrats will win a majority of the races

SOLUTION. It may seem like you can't do this, since the table doesn't list p = .60. However, all you have to do is redefine success and failure. Let success = P(opponents win a race) = .40. The question can then be recast as finding the probability that

a. The opponents will win 4 races, 3 races, 2 races, 1 race, or none of the races?
b. The opponents will win 0, 1, 2, or 3 races; or, the opponents will not win all the races
c. The opponents will not win a majority of the races

We therefore look at page E-4 (or Hayes, p. 927), N = 4 and p = .40, and find that

a. P(4) = .0256, P(3) = .1536, P(2) = .3456, P(1) = .3456, and P(0) = .1296.
b. P(0) + P(1) + P(2) + P(3) = 1 - P(4) = .9744
c. P(1) + P(0) = .3456 + .1296 = .4752

In general, for p > .50: To use Table II, substitute 1 - p for p, and substitute N - r for r. Thus, for p = .60 and N = 4, the probability of 1 success can be found by looking up p = .40 and r = 3.

D. *Mean of the binomial distribution.* Recall that, for any discrete random variable, $E(X) = \Sigma xp(x)$. Therefore, $E(X_i) = \Sigma xp(x) = 0 * (1 - p) + 1 * p = p$, that is, the mean of any bernoulli trial is p, the probability of success. By applying our theorems for expectations, we find that

$$\mathbf{E(X) = E(X_1 + X_2 + ... + X_N) = E(X_1) + E(X_2) + ... + E(X_N) = Np.}$$

NOTE: Hayes provides an alternative proof in section 4.8; I think my proof is much simpler.

EXAMPLE. If we toss a fair coin 3 times (N = 3, p = .50), we expect to get 1.5 heads.

E. *Variance of the binomial distribution*. Given our above definitions, note that, for any Bernoulli trial, $X_i^2 = X_i$. Thus, $V(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1 - p) = pq$. Because trials are independent, we find that

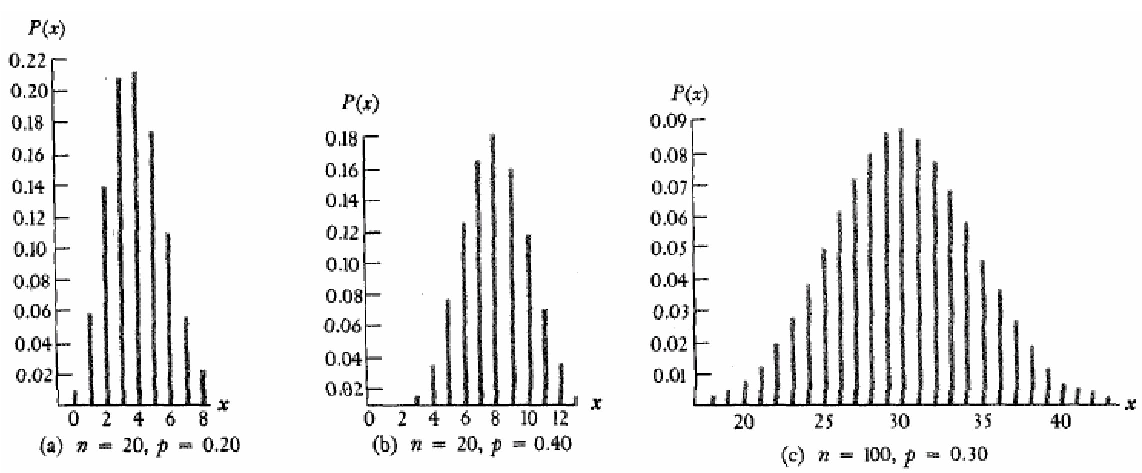
$$\mathbf{V(X) = V(X_1) + V(X_2) + ... + V(X_N) = Npq.}$$

EXAMPLE. If you toss a fair coin 3 times (N = 3, p =.50, q = .50) the variance of the expected number of heads is 3 * .50 * .50 = .75.

NOTE. These last two points mean that the mean and variance of the binomial distribution are dependent on only two parameters, N and p.

WARNING: The symbol X gets used many different ways in statistics. In this case, X = the sum of all the Bernoulli trials, but in other instances it might refer to an individual Bernoulli trial.

F. *Shape of the binomial distribution*. When p = .5, the binomial distribution is symmetrical - the mean and median are equal. Even when p <> .5, the shape of the distribution becomes more and more symmetrical the larger the value of N. This is very important, because the binomial distribution can quickly become unwieldy - as we will later see, there are approximations to the binomial that can be much easier to use when N is large.



g. A MORE COMPLETE LISTING OF THE COUNTING RULES FOR PERMUTATIONS AND COMBINATIONS (OPTIONAL).

Here is a more extensive set of counting rules that can be useful for various problems in probability. They aren't essential for our immediate purposes, so we probably won't go over them in class unless we have extra time. But, these aren't very hard, and they may come in handy for you some day, so I am including them here.

1. NUMBER OF POSSIBLE SEQUENCES FOR N TRIALS. Suppose that a series of N trials were carried out, and that on each trial any of K events might occur. Then the following rule holds:

> **If any one of K mutually exclusive and exhaustive events can occur on each of N trials, then there are K^n different sequences that may result from a set of such trials.**

EXAMPLE. If you toss a die once, any of 6 numbers can show up (K = 6). Ergo, if you toss it 3 times, any of $6^3 = 216$ sequences are possible (e.g., 111, 342, 652, etc.). [Your calculator probably has a y^x function or something similar; on mine, I press 6, then y^x, then 3, then =.]

2. SEQUENCES. Sometimes the number of possible events in the first trial of a series is different from the number possible in the second, the second different from the third, etc. That is, $K_1 \neq K_2 \neq K_3$, etc. Under such conditions,

> **If $K_1, \ldots, K_N$ are the numbers of distinct events that can occur on trials 1,..., N in a series, then the number of different sequences of N events that can occur is $(K_1)(K_2)\ldots(K_N)$.**

EXAMPLE. Two occupations and three religions yield 6 combinations of occupation and religion. Tossing a coin (2 outcomes) and tossing a die (6 outcomes) yield 12 possible outcomes.

Note that, when $K_i = K$ for all i, then rule 1 becomes a special case of rule 2. Note also that, when $K_1 = 1, K_2 = 2, \ldots, K_N = N$, then rule 3 becomes a special case of rule 2.

3. PERMUTATIONS. A rule of extreme importance in probability computations concerns the number of ways that objects may be arranged in order.

> **The number of different ways that N distinct things may be arranged in order is**
> **$N! = (1)(2)(3)\ldots(N-1)(N)$, (where $0! = 1$).**

An arrangement in order is called a permutation, so that the total number of permutations of N objects is N!. The symbol N! is called N factorial. The notation ${}_NP_N$ is also sometimes used for N!, for reasons which will be clear in a moment.

EXAMPLE. Rank candidates A, B, and C in order. The possible permutations are:
ABC ACB BAC BCA CBA CAB

There are 6 possible orderings. Note that 3! = (1)(2)(3) = 6.

> NOTE: Appendix E, Table 6, p. 19 contains a Table of the factorials for the integers 1 through 50. For example, $12! = 4.79002 * 10^8$. (Or see Hayes Table 8, p. 947). Your calculator may have a factorial function labeled something like x!

3B. PERMUTATIONS OF SIMILAR OBJECTS.

Suppose we have N objects, N_1 alike, N_2 alike,..., N_k alike ($\Sigma N_i = N$). Then, the number of ways of arranging these objects is

$$\frac{N!}{N_1!\,N_2!\ldots N_k!}$$

EXAMPLE. We have 4 balls, 2 red and 2 blue. The possible ways of arranging the balls are BBRR, BRBR, BRRB, RRBB, RBRB, RBBR, or 6 ways altogether. To confirm that there are 6 ways,

$$\frac{N!}{N_1!\,N_2!\ldots N_k!} = \frac{4!}{2!2!} = 6$$

EXAMPLE. If we have 6 balls, 2 red, 2 blue, and 2 green, the number of possible ways of arranging them is

$$\frac{N!}{N_1!\,N_2!\ldots N_k!} = \frac{6!}{2!2!2!} = \frac{720}{8} = 90$$

NOTE: I add this rule to Hayes's list because many of the other rules are special cases of it. When $N_i = 1$ for all i, Rules 3 (Permutations) and 3B are equivalent. When $N_i = 1$ for i = 1 to r, and $N_{r+1} = N - r$, then rules 4 (Ordered combinations) and 3B are the same. When $N_1 = r$ and $N_2 = N - r$, Rules 5 (Combinations) and 3B are the same.

4. ORDERED COMBINATIONS; or, PERMUTATIONS OF N OBJECTS TAKEN r AT A TIME.

Sometimes it is necessary to count the number of ways that r objects might be selected from among some N objects in all ($r \leq N$). Further, each different arrangement of the r objects is considered separately.

The number of ways of selecting & arranging r objects from among N distinct objects is

$$\frac{N!}{(N-r)!}$$

Verbally, we refer to this as ordered combinations of N things taken r at a time. The notation ${}_NP_r$ is sometimes used, and may appear on your calculator using similar notation. Note also that ${}_NP_N = N!$.

EXAMPLE. Candidates A, B, C, and D are running for office. Indicate your first and second choice.

There are 12 possible choices:

AB BA AC CA AD DA BC CB BD DB CD DC.

To confirm this, using counting Rule #4, we get

$$\frac{N!}{(N-r)!} = \frac{4!}{(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)} = 12$$

EXAMPLE. There are 100 applicants for 3 job openings. Indicate your first, second, and third choices.

If you try to list all the possible permutations, you will be busy for a very long time. If you try to calculate 100! and 97! by hand, you'd better plan on being in grad school a long time. And, even a calculator with a factorial function may not help you; mine just produces an error message when I give it numbers this big. If, on the other hand, you take advantage of the fact that most terms in the numerator and denominator cancel out, e.g.

$$\frac{N!}{(N-r)!} = \frac{100!}{(100-3)!} = \frac{(100)(99)(98)(97)(96)...(1)}{(97)(96)...(1)} = 100 * 99 * 98 = 970{,}200$$

you will find that it is pretty easy to figure out the number of ordered combinations.

NOTE: You could also use Rule 3B. People are divided into 4 categories: the first best ($N_1 = 1$), the 2nd best ($N_2 = 1$), the 3rd best ($N_3 = 1$) and all the rest ($N_4 = 97$). Hence, rule 3B yields 100!/(1!1!1!97!) = 100!/97!

5. *COMBINATIONS*. Often, we are not interested in the order of events, but only in the number of ways r things could be selected from N things, irrespective of order. From rule 4, we know that the total number of ways of selecting r things from N and ordering them is N!/(N-r)! From rule 3, we know that each set of r objects has r! orderings. Therefore,

The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is

$$\frac{N!}{r!(N-r)!} = \binom{N}{r} = \binom{N}{N-r}$$

We refer to this as "N choose r." Sometimes the number of combinations is known as a *binomial coefficient*, and sometimes the notation $_NC_r$ is used. Note that the number of permutations is r! times larger than the number of combinations.

EXAMPLE. Candidates A, B, C and D are running for office. Vote for two.

The possible choices are:

AB AC AD BC BD CD, i.e. there are 6 possible combinations. Confirming this with rule #5, we get

$$\frac{N!}{r!(N-r)!} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{12}{2} = 6$$

EXAMPLE. There are 100 applicants for 3 job openings. The number of possible combinations is

$$\frac{N!}{r!(N-r)!} = \binom{100}{3} = \frac{100!}{3!97!} = \frac{970{,}200}{6} = 161{,}700$$

NOTE: You could also use rule 3B. People are divided into 2 groups: the 3 best ($N_1 = 3$) and all the rest ($N_2 = 97$). Hence, rule 3B yields 100!/(3!97!)

See Appendix E, Table 7, page 20 for $_NC_r$ values for various values of N and r. (Or see Hayes, Appendix E, Table IX, p. 948). Your calculator may have a function labeled nCr or something similar.

6. Binomial distribution. **In sampling from a stationary Bernoulli process, with the probability of success equal to p, the probability of observing exactly r successes in N independent trials is**

$$\binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}$$

7. Binomially distributed variables. **Let Xi = 1 if the ith bernoulli trial is successful, 0 otherwise. If $X = \Sigma X_i$, where the X_i's are independent and identically distributed (iid), then X has a binomial distribution, and E(X) = Np, V(X) = Npq.**

SUMMARY OF HIGHLIGHTS

1. *NUMBER OF POSSIBLE SEQUENCES FOR N TRIALS.* **If any one of K mutually exclusive and exhaustive events can occur on each of N trials, then there are K^n different sequences that may result from a set of such trials.**

2. *SEQUENCES.* **If $K_1,...,K_N$ are the numbers of distinct events that can occur on trials 1,..., N in a series, then the number of different sequences of N events that can occur is $(K_1)(K_2)...(K_N)$.**

3. *PERMUTATIONS* **The number of different ways that N distinct things may be arranged in order is**

$$N! = (1)(2)(3)...(N-1)(N), \text{ (where } 0! = 1).$$

3B. *PERMUTATIONS OF SIMILAR OBJECTS.* **Suppose we have N objects, N_1 alike, N_2 alike,..., N_k alike ($\Sigma N_i = N$). Then, the number of ways of arranging these objects is**

$$\frac{N!}{N_1!\, N_2! \ldots N_k!}$$

4. *ORDERED COMBINATIONS; or, PERMUTATIONS OF N OBJECTS TAKEN r AT A TIME.* **The number of ways of selecting & arranging r objects from among N distinct objects is**

$$\frac{N!}{(N-r)!}$$

5. *COMBINATIONS.* **The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is**

$$\frac{N!}{r!(N-r)!} = \binom{N}{r} = \binom{N}{N-r}$$

6. *Binomial distribution.* **In sampling from a stationary Bernoulli process, with the probability of success equal to p, the probability of observing exactly r successes in N independent trials is**

$$\binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}$$

7. *Binomially distributed variables.* **Let Xi = 1 if the ith bernoulli trial is successful, 0 otherwise. If $X = \Sigma X_i$, where the X_i's are independent and identically distributed (iid), then X has a binomial distribution, and E(X) = Np, V(X) = Npq.**