

The Binomial Distribution

Some Typical Key Questions

Suppose a person knows absolutely nothing, and simply guesses on a 12 item true-false test. What is their probability of passing the test, if 65% is a passing mark?

Suppose exactly 50% of voters favor a particular position. If you take a random sample of 100 voters, how likely is it that your sample percentage will be within 10% of the correct answer?

Steps to Understanding

What is a random variable?

What is a probability distribution function?

What is a cumulative distribution function?

What is the “expected value” of a random variable? How is it computed?

**What is the “variance of a random variable”?
How is it computed?**

What is a binomial process?

What is a binomial random variable?

What is the binomial distribution?

What is the normal approximation to the binomial distribution?

What is a random variable?

A random variable can be defined both formally and informally. An informal notion, adequate for our purposes, is that it is a random process whose outcomes are *numerically coded in a unique, consistent manner*.

An example: A woman has a sequence of 5 children. Each child is coded 1 for boy, 2 for girl.

A random variable outputs numbers, and these numbers have probabilities.

What is a probability distribution function (pdf)?

This is a function assigning probabilities to the numerical outcomes or values that the random variable takes on.

Often this is expressed as a table, or as a mathematical function rule:

Example:

x	$P(x)$
2	.50
1	.50

What is a cumulative distribution function (cdf)?

This is a function providing, for any numerical value, the probability that the random variable will be *less than or equal to that value*.

x	$P(x)$	$F(x)$
2	.50	1.00
1	.50	0.50

What is the “expected value” of a random variable?

It is the long run average numerical value taken on by the random variable.

Consider the table below. Since the random variable X takes on values $x = 2$ and $x = 1$ half the time each, its *expected value*, or long run average, is 1.5.

How is expected value computed?

There is a simple rule for *discrete random variables*, i.e., random variables with a countable number of outcomes.

Simply take the sum of cross-products of the outcomes and their probabilities.

$$E(X) = \sum_{i=1}^k x_i P(x_i)$$

Example.

x	$P(x)$	$xP(x)$
6	1/6	6/6
5	1/6	5/6
4	1/6	4/6
3	1/6	3/6
2	1/6	2/6
1	1/6	1/6
		$21/6=3.5$

Expected Value of a Function of X

If X is a random variable, a function of X will also be a random variable. To compute the expected value of the function, (a) write the function values, (b) take the sum of cross products of the function values and the probabilities for the original x values.

Example.

Below we compute $E(X^2)$

x	$P(x)$	$xP(x)$	x^2	$x^2P(x)$
6	1/6	6/6	36	36/6
5	1/6	5/6	25	25/6
4	1/6	4/6	16	16/6
3	1/6	3/6	9	9/6
2	1/6	2/6	4	4/6
1	1/6	1/6	1	1/6
		21/6=3.5		91/6

Variance of a Random Variable

The variance of X is the long run average squared deviation score. It is defined as

$$\text{Var}(X) = \sigma_X^2 = E(X - E(X))^2$$

We *could* compute the variance of X by first computing the squared deviation scores for each outcome, then taking the sum of cross products of these squared deviation scores and the probabilities. However, there is a much simpler way.

Computing Variance of a Random Variable

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

For a fair die, using results from the preceding table

$$\begin{aligned}\text{Var}(X) &= 91/6 - (7/2)^2 \\ &= 91/6 - 49/4 \\ &= 364/24 - 294/24 \\ &= 70/24 = 35/12\end{aligned}$$

The Binomial Process

Perhaps the simplest situation we encounter in probability theory is the case where only two things can happen. If one of the two things has probability p , the alternative thing must have probability $q = 1 - p$. Suppose we call the two outcomes “Success” and “Failure” and code the outcomes 1 and 0 with the random variable T .

The Binomial Process

Then T has the following probability distribution:

t	$P(t)$
1	p
0	$1 - p$

Find the *Expected Value and Variance of T*.

The Binomial Process

Suppose we have a sequence of N independent trials where, on each trial, we have an outcome on the random variable T . Then such a process has the following characteristics:

- There are N **independent** trials
- **Only two things can happen**. One, arbitrarily labeled “Success,” has probability p . The other, arbitrarily labeled “Failure,” has probability $q=1 - p$.
- Probabilities **remain constant** from trial to trial.

Binomial Process as a Model

Such a process is a good approximation to many “real world” processes. For example, consider the following:

- A woman has 5 children. They are either boys or girls.
- A factory produces 100 radios. They are either have a manufacturing defect or they do not.
- A basketball player attempts 385 free throws in a season. On each attempt, he either succeeds or fails.

Each of these is only *approximated* by a binomial process. Why? (C.P.)

The Binomial Random Variable X

Often, we are not interested in the precise sequence of outcomes in a binomial process, but rather in the number of successes in the N trial sequence. For example, a basketball player's free throw percentage is based on the number of successes in N free throws, not on the precise sequence that produces the number of successes. Consequently, we often code a sequence of binomial trials with the binomial random variable, defined as follows:

The Binomial Random Variable. Given N trials of a binomial process with probability of success p , the binomial random variable X is the number of success that occurs in the N trial sequence.

The Binomial Distribution Family $B(N,p)$

The binomial distribution is a family of distributions with two parameters --- N , the number of trials, and p , the probability of success. We refer to the binomial random variable with general notation $B(N,p)$. For example, $B(10,1/2)$ refers to a 10 trial binomial process with probability of success equal to $1/2$.

The Binomial Distribution

The $B(N,p)$ distribution has the following probability distribution function (or pdf):

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

Example. (An unfair coin)

Consider an unfair coin, with $p = \Pr(\text{Head}) = 2/3$. What is the probability that, in 4 tosses of this coin, exactly 3 heads occur?

$$\begin{aligned}\Pr(X = 3) &= \binom{4}{3} (2/3)^3 (1/3)^1 \\ &= 32/81\end{aligned}$$

The Cumulative Binomial

Suppose you play 5 rounds of a gambling game where the odds are in your favor, i.e., your probability of winning a round is $5/9$. If you bet the same amount of money in each of the 5 rounds, what is the probability that you will lose money, that is, win 2 or fewer rounds?

The answer is

$$\begin{aligned} F(2) &= \Pr(X \leq 2) \\ &= \sum_{i=0}^2 \Pr(X = i) \\ &= \sum_{i=0}^2 \binom{5}{i} \left(\frac{5}{9}\right)^i \left(\frac{4}{9}\right)^{5-i} \\ &= \frac{(1)(1)(4^5) + (5)(5)(4^4) + (10)(5^2)(4^3)}{9^5} \\ &= \frac{1024 + 6400 + 16000}{59049} = .396687 \end{aligned}$$

The Normal Approximation to the Binomial Distribution

When N is reasonably large and p is not too far from $1/2$ (and more generally, when both Np and $N(1 - p)$ are greater than or equal to 10), the binomial distribution can be approximated by a normal distribution with mean Np and variance $Np(1 - p)$.

Distribution of the Sample Proportion

The *sample proportion* \hat{p} is the proportion of people in a sample of N who fit a certain category. Since

$$\hat{p} = X / N$$

its distribution is the same shape as the binomial distribution, and can be approximated for large samples by a normal distribution with mean p and variance $p(1 - p) / N$.