Discrete Random Variables

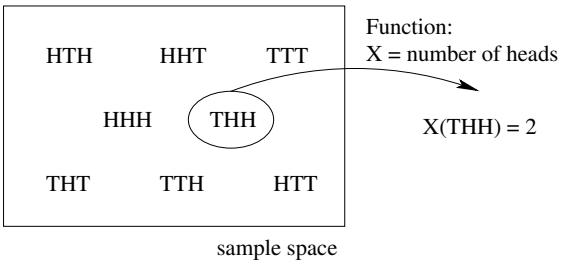
In many random experiments we are not interested in the specific outcome of an experiment, but in some characteristic of that outcome. If you play the lottery, for example, you would like the numbers chosen in the weekly drawing to match as many of the numbers you marked on your ticket as possible. You do not particularly care which of your numbers are matched, but rather *how many* matches you have.

Examples: Other variables that depend on the outcome of random experiments include

- Sum of two dice.
- The number of times you have to play the lottery until you win the jackpot.
- The amount of money that a company will have to pay out for an employee's pension plan.

Remark: There are two fundamentally different kinds of random variables. Some variables can take on only particular values (sum of two dice: $2, 3, \ldots, 12$). Others take on possible values in an interval (pension: $[\$0, \infty)$). Random variables that take on finitely many (or countable infinitely many) possible values are called DISCRETE and those that take on values in intervals are called CONTINUOUS.

Definition: A random variable (RV) is a real valued function whose domain is a sample space.



Example: Toss a fair coin 3 times. Suppose we are interested in the random variable X = number of heads. Then

$$X(\{HTH\}) = \qquad X(\{HHH\}) = \qquad X(\{TTT\}) = \qquad \text{etc.}$$

We can keep track of the outcomes and their respective probabilities in a table

x	$p(x)$

Definition: The values on the right side of the table are called the PROBABILITY MASS FUNCTION (PMF) of the random variable X, i.e.

$$p(x) = P(X = x).$$

Note: You can keep track of probability mass functions with a table such as the one above or in a bar graph. The probabilities should always sum to one! We will usually denote random variables by X, Y, Z and their respective values by x, y, z.

Definition: Any random variable whose only possible values are 0 and 1 is called a BERNOULLI random variable. It's PMF looks like this:

x	$p(x)$
0	$1-p$
1	p

Here, p is the probability that the random variable will take on the value "1". p is called a PARAMETER of the distribution (the probabilities of the outcomes depend on the value of p).

Example: A box contains one red, two orange and five black balls. You choose two balls at random without replacement. For every red ball you win \$5, for every orange ball you win \$1 and you do not win anything for a black ball. Let X be the amount of money you'll win.

(a) Write down the probability mass function of X.

x	p(x)
6	2/(8C2)
5	5/(8C2)
2	1/(8C2)
1	10/(8C2)
0	(5C2)/(8C2)

(b) What is the probability that you will win at most \$3?

P=P(X=0)+P(X=1)+P(X=2)

Definition: The CUMULATIVE DISTRIBUTION FUNCTION (CDF) $F(x)$ of a discrete RV X with PMF $p(x)$ is defined for every number x as

$$F(x) = P(X \leq x) = \sum_{y \leq x} p(y)$$

That is, $F(x)$ is the probability that X is at most x.

Example: Let X the the number of heads counted in three tosses of a fair coin. We have seen before that the PMF of X is:

x	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

(a) Find $F(2)$.

(1+3+3)/8=7/8

Poll Question 5.1

(b) Draw a graph of the CDF $F(x)$.

Note: The CDF of a discrete random variable X is a step-function. It always starts out at zero, has jumps at the possible values x and ends up at 1.

Poll Question 5.2

Expected Value of Discrete Random Variables

Question: How do you determine the "value" of a game? Is it better to play Roulette than the Lottery? We are looking for ways of describing random variables.

Definition: The expected value of a random variable X with PMF $p(x)$ is given by

$$E(X) = \mu_X = \sum_{\text{all } x} x \cdot p_X(x).$$

Note: We may interchangeably use the terms mean, average, expectation and expected value and the notations $E(X)$ or μ (for mean). The expected value is a weighted average of the possible values of X, weighted by the probabilities.

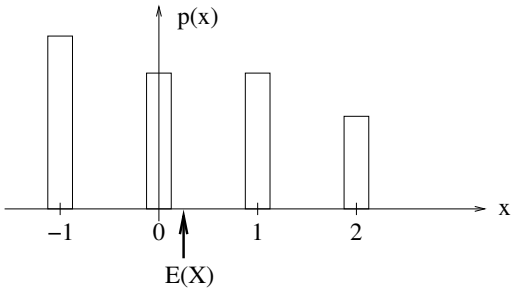
Remark: The expected value of a random variable can be understood as the long-run-average value of the random variable in repeated independent observations. If you are playing a game, and X is what you win in the game, then $E(X)$ would be your average win if you would play the game (very) many times.

Example: Let X be a random variable with PMF

x	$p(x)$
-1	$1/3$
0	$1/4$
1	$1/4$
2	$1/6$

Then $E(X) =$

Note: $E(X)$ can be understood as the (physical) "Center of Gravity" in the histogram picture of the PMF, if you consider the probability to be mass (literally).



Instead of $E(X)$ we can also compute the expected value of a function of X:

PROPOSITION: If X is a discrete random variable with PMF $p(x)$, and $h(x)$ is any real valued function of X, then

$$E[h(X)] = \sum_{\text{all } x} h(x) \cdot p_X(x).$$

PROPOSITION: Expectation is a linear operator:

$$E(aX) = aE(X)$$

$$E(X + b) = E(X) + b$$

$$\Rightarrow E(aX + b) = aE(X) + b$$

PROPOSITION: Let X and Y be two random variables, then

$$E(X + Y) = E(X) + E(Y)$$

The expectation of a sum is the sum of the expectations!

Poll Question 5.3

Example: In a simple game, two fair coins are tossed and the payoff is to be determined from the outcome. You may choose one of the following payoff strategies:

(a) Payoff: Win \$1 for each head.
Lose \$2 for two tails.

(b) Payoff: Win \$1 if the coins are both tails.
Win \$2 if the coins are different.
Lose \$2 if the coins are both heads.

Let X denote your winnings if you play this game once.

(a) Write down the probability mass function $p(x)$ for X for either strategy.

(a)

2	1/4
1	1/2
-2	1/4

E(X)=1/2

(b)

2	1/2
1	1/4
-2	1/4

E(X)=3/4

(b) Which strategy has the higher expected winnings?

Poll Question 5.4

Variance of Discrete Random Variables

Now we know that we can use the expected win to describe the value of a game. But can one measure the risk involved in a game or in an investment strategy?

Example: Suppose you are offered two investment opportunities:

(a) You invest \$1 and will gain \$1 with probability 0.5.

(b) You invest \$1 and will gain \$1,000,000 with probability 0.000001.

The values of the games are the same. But clearly, they are not the same game. We need another quantity to describe random variables other than their expectation.

Definition: Let X be a discrete random variable with probability mass function $p(x)$. Then the variance of X is defined as

$$V(X) = \sum_x (x-\mu)^2 p(x) = E\left[(X-\mu)^2\right]$$

It measures the mean squared distance of observations on X from their mean.

Remarks:

- $V(X)$ is always non-negative ($V(X) \geq 0$).
- Sometimes, we'll abbreviate $\sigma_X^2 = V(X)$.
- $V(X)$ is a measure of the spread of the random variable. If $V(X) = 0$, then the spread is zero, i.e. all the probability is concentrated in one point (nothing random anymore).
- The variance is not measured in the same units that the random variable is measured in (disadvantage!)

PROPOSITION: (Very useful for computing variances!)

$$V(X) = E(X^2) - E(X)^2$$

Definition: The standard deviation of a random variable X is defined to be

$$\sigma_X = \sqrt{V(X)} = \sqrt{\sigma_X^2}$$

Other than the variance, the standard deviation is measured in the same units (e.g. \$, minutes, yards) that X is measured in.

PROPOSITION: Variance is *not* a linear operator! Let X be a random variable and a, b constants. Then

$$V(aX + b) = a^2 V(X).$$

Poll Question 5.5

Example: Let X be a discrete random variable with probability mass function

x	$p_X(x)$
0	0.1
1	0.2
2	0.4
3	0.3

(a) Find $E(X)$. 0.2+0.8+0.9=1.9

(b) Find $V(X)$. E(X^2)=0*0.1+1*0.2+4*0.4+9*0.3=4.5
V(X)=4.5-1.9*1.9=0.89

(c) Find $V(2X - 3)$.
=4*0.89=3.56

The Binomial Distribution

Many discrete random variables fall into similar categories. For example, from a mathematical point of view, the number of heads in four tosses of a fair coin and the number of girls in a family with four children can be described by the same model. In order to not have to come up with the PMF, the expected value and the variance from scratch in each new case, we will now study some of these models in a general setting.

DEFINITION: The random variable which takes on only values 1 and 0 (with probabilities p and $1-p$, respectively) is called a BERNOULLI random variable with parameter p.

NOTATION: $X \sim \text{Bernoulli}(p)$.

PROPOSITION: The expected value of variance of a Bernoulli random variable are

$$E(X) = 1 \cdot p + 0 \cdot (1-p) = p, \quad E(X^2) = 1^2 \cdot p + 0^2 \cdot (1-p) = p$$
$$\text{Var}(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

Other experiments consist of a sequence of several BERNOULLI TRIALS:

(1) The same Bernoulli experiment is repeated n times.

(2) Each time, the outcome is either a success (1 or S) or a failure (0 or F).

(3) The trials are independent of each other.

(4) The success probability p stays the same in each trial.

SITUATION: The random variable X that we are interested in is the *total* number of successes in the n trials. Such a random variable is said to have a Binomial distribution, write:

$$X \sim \text{Binomial}(n,p) \text{ or } X \sim \text{Bin}(n,p)$$

with parameters n and p, Its PMF is denoted by $b(x,n,p)$.

Examples: Identify n and p for the following Binomial situations:

- The number of heads in ten coin tosses.
- The number of girls in a family with ten children.
- The number of times you'll win the jackpot out of the 52 weeks you'll play the lottery next year.

Poll Question 5.6

PMF: The probability to have exactly x successes in n Bernoulli trials with success probability p is:

$$b(x, n, p) =$$

Poll Question 5.7

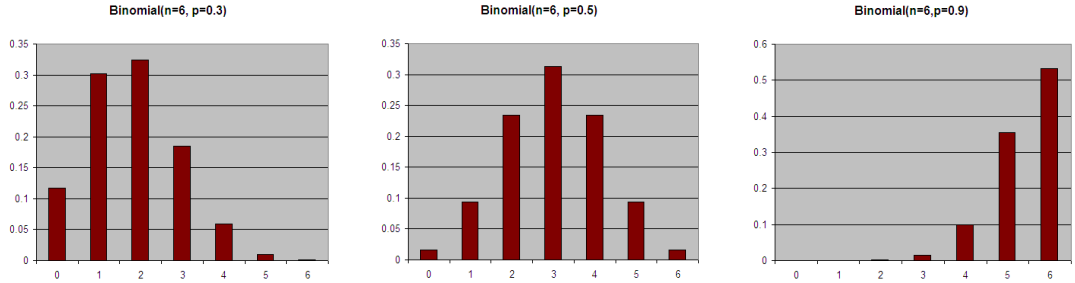
PROPOSITION: The CDF of a Binomial random variable X will be denoted by $B(x, n, p)$ and can be computed as

$$B(x, n, p) = P(X \leq x) = \sum_{y=0}^{\lfloor x \rfloor} b(x, n, p), \quad x \in \mathbf{R}$$

PROPOSITION: The expected value and variance of a Binomial random variable X are

$$E(X) = np, \qquad V(X) = np(1-p).$$

SHAPE:



Example: What is the probability to see three heads in five tosses of a biased coin that tosses heads with probability 0.6?

Define the random variable you are working with ($X = \ldots$) and state its distribution along with all relevant parameters.

The Hypergeometric Distribution

Binomial random variables can be used to model "drawing with replacement" experiments.

Example: You have a bag with ten red, eight green and two blue marbles. You randomly and *with replacement* draw three marbles. Let X be the number of green marbles you got. Then $X \sim \text{Binomial}(n = \quad, p = \quad)$.

If the same experiment were repeated *without* replacement, then X would not have a Binomial distribution. Why?

Poll Question 5.8

SITUATION: Suppose you have a population of size N, in which each individual can be classified as either a success (S) or a failure (F). The number of successes in the population is M. A sample of size n is selected at random and *without* replacement from the population.

DEFINITION: The random variable that we are interested in is X = number of successes in the sample. It has a Hypergeometric distribution with parameters n, M, and N.

NOTATION: $X \sim \text{Hypergeometric}(n, M, N)$, the PMF is denoted $h(x; n, M, N)$.

Examples: Identify the parameters n, M, N for the number of aces in a poker hand.

Poll Question 5.9

PMF: The possible values of a Hypergeometric random variable are $0, 1, \ldots, \min(M, n)$. Its PMF has the form

$$h(x; n, M, N) =$$

PROPOSITION: Expected value and variance of a Hypergeometric(n, M, N)

$$E(X) = n\frac{M}{N}, \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

Remark: If N is very large compared to n ($N \geq 20n$) it makes no significant difference if we draw with or without replacement. In these cases, we can approximate a Hypergeometric(n, M, N) random variable, by a Binomial(n, p) random variable with $p = \frac{M}{N}$.

The Geometric Distribution

Recall: For the Binomial and Hypergeometric distribution, the number of trials that are conducted is fixed at n and we count the number of successes.

SITUATION: Independent trials are conducted with success probability p. The Geometric distribution is used to count the number of trials until a success occurs.

NOTATION: $X \sim$ Geometric(p), the PMF is denoted $g(x;p)$.

Examples:

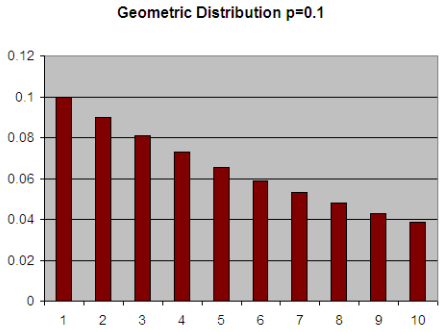
- The number of times you flip a fair coin until you get the first head.
- The number of times you have to play the lottery until you win the first jackpot.

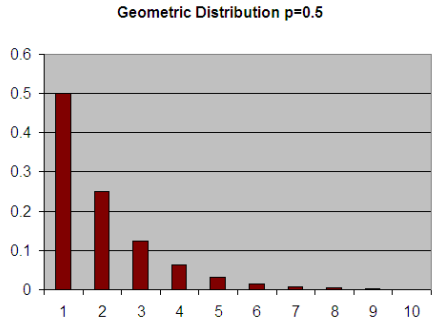
Poll Question 5.10

PMF: The possible values of a Geometric random variable X with success probability p are $1, 2, \ldots$ and the probability that the first success occurs on the x^{th} trial is

$$g(x;p) = P(X = x) = \qquad\qquad , x = 1, 2, \ldots$$

SHAPE:





PROPOSITION: The expected value and variance of a Geometric random variable are

$$E(X) = \frac{1}{p}, \qquad V(X) = \frac{1-p}{p^2}$$

REMARK: The Geometric distribution has the "lack-of-memory" property, meaning that it does not matter when we start observing. The number of trials until the *next* success always has a Geometric distribution.

DISCLAIMER: Your textbook does not define a Geometric random variable. Instead, it defines the more general Negative Binomial random variable, for which the Geometric situation is a special case.

Example: You roll a fair die until you see the first six. Define the random variable you are working with in words ($X = \ldots$) and state its distribution along with all relevant parameters.

(a) How many rolls do you expect to need?

Poll Question 5.11

(b) What is the probability that the first six occurs on the fifth roll?

(c) What is the probability that you will need more than ten rolls?

Example: Find a general formula for the cumulative distribution function of a Geometric random variable with parameter p.

Example: A roulette wheel has 38 slots, numbered $0, 00, 1, 2, \ldots, 36$. You have been watching the game for a while and you have observed that in the last ten spins the number five has not been hit. What is the probability that the first "five" comes up 15 spins from now?

The Negative Binomial Distribution

SITUATION: Independent Bernoulli(p) trials are performed until we see the r^{th} success. We are interested in the total number of trials performed.

NOTATION: $X \sim$ Negative Binomial(r, p), the PMF is denoted $nb(x; r, p)$.

Examples:

- The number of times you flip a fair coin until you get heads for the fourth time.
- The number of children you'll have if you keep having children until you have two girls.

PMF: The possible values of a Negative Binomial random variable are $x = r, r+1, \ldots$ and the probability that the r^{th} success occurs on the x^{th} trial is

$$nb(x; r, p) =$$

FACT: If $X_1, \ldots, X_r$ are independent Geometric(p) random variables, then

$$X = \sum_{i=1}^{r} X_i \sim \text{ Negative Binomial } (r, p)$$

PROPOSITION: The expected value and variance of a Negative Binomial random variable with parameters r and p are

$$E(X) = \frac{r}{p}, \qquad V(X) = r\frac{1-p}{p^2}$$

Example: You roll a fair six-sided die until you see the third six. What is the probability that this happens on the tenth roll?

Poll Question 5.12

Disclaimer: This definition of theNegative Binomial random variable differs slightly from the version found in your textbook (number of trials vs. number of failures). The definition above is the one much more commonly found in the literature. You are welcome to use either definition. But if you choose to work with the version from the book (rather than this one), you need to clearly state this in your work.

The Poisson Distribution

Recall: The Bernoulli, Binomial and Hypergeometric distribution can be used to model the number of successes in a experiment consisting of one or more trials.

The Poisson distribution is used to model the number of times a rare event occurs.

Examples:

- The number of hurricanes in the central U.S. in a month.
- The number of misprints on a page of some document.
- The number of macademia nuts in a macademia nut cookie.

PROPOSITION: A random variable X is said to have a Poisson distribution with parameter λ ($\lambda > 0$), if its PMF is

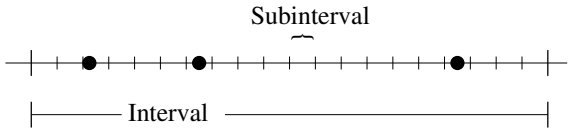
$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$$

Write $X \sim \text{Poisson}(\lambda)$. Here the parameter λ is the RATE at which the event occurs.

PROPOSITION: If X has a Poisson distribution with parameter λ, then its expected value and variance are:

$$E(X) = \lambda, \qquad V(X) = \lambda$$

Remark: Poisson variables can be derived from Binomial random variables.To count the number of rare events, imagine an interval (that can stand for time, or a page, or a volume of cookie dough) split up into n little subintervals.



It is always possible to make the subintervals small enough (by making n large), such that there is at most one event in a subinterval. Suppose that the probability that a subinterval has an event in it is p.

We are interested in the number of times X the event occurs in the interval. Strictly speaking, X has a Binomial(n, p) distribution but with a very large n and a small p (since the events are "rare"). What happens to the Binomial PMF, if $n \to \infty$? Let $\lambda = np$. Then

$$\begin{aligned} p_X(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \\ &\xrightarrow{n\to\infty} \frac{\lambda^x}{x!} e^{-\lambda} \end{aligned}$$

Example: Researchers from University College, London, UK, have studied the hurricane activity in the Atlantic ocean. The annual atlantic hurricane season lasts from June 1st through November 30th (6 months). On average, over the last few decades 5.9 hurricanes were observed per season of which 2.5 were classified as major hurricanes (categoy 3 or greater).

(a) During the 2020 hurricane season 13 named storms developed into hurricanes. What is the probability that this would have happened by chance?

(b) What is the probability that next hurricane season we will have at least two major hurricanes? (In 2020 there were six: Laura, Teddy, Delta, Epsilon, Eta and Iota).

(c) What is the probability that we will have 12 named storms during the next two hurricane seasons combined?

Poll Question 5.13

(d) What is the probability that we will have 6 named storms during *each* of the next two hurricane seasons?